1   **Phylodynamic analysis reveals disparate transmission dynamics of *Mycobacterium***

2   ***tuberculosis*-complex lineages in Botswana**

3

4   Qiao Wang,[1,2] Ivan Barilar,[3] Volodymyr M. Minin,[4] Chawangwa Modongo,[5] Patrick K. Moonan,[6]

5   Alyssa Finlay,[7] Rosanna Boyd,[7] John E. Oeltmann,[6] Tuduetso L. Molefi,[8] Nicola M. Zetola,[9]

6   Timothy F. Brewer,[1,10] Stefan Niemann[*,3,11] Sanghyuk S. Shin[*2]

7

8   1. Department of Epidemiology, Fielding School of Public Health, University of California Los Angeles,

9   Los Angeles, CA, USA; 2. Sue & Bill Gross School of Nursing, University of California Irvine, Irvine,

10   CA, USA; 3. Molecular Mycobacteriology, Forschungszentrum Borstel, Borstel, Germany; 4. Department

11   of Statistics, University of California Irvine, Irvine, CA, USA; 5. Victus Global Botswana Organisation,

12   Gaborone, Botswana; 6. Division of Global HIV and Tuberculosis, US Centers for Disease Control and

13   Prevention, Atlanta, GA, USA; 7. Division of Tuberculosis Elimination, US Centers for Disease Control

14   and Prevention, Atlanta, GA, USA; 8. National TB Program, Botswana Ministry of Health, Gaborone,

15   Botswana; 9. Augusta University School of Medicine, Augusta, GA, USA; 10. Division of Infectious

16   Diseases, University of California Los Angeles, Los Angeles, CA, USA; 11. German Center for Infection

17   Research (DZIF), Partner Site Hamburg-Lübeck-Borstel-Riems, Borstel, Germany.

18

19   [*]These authors have contributed equally to this work.

20

21

Abstract

Tuberculosis epidemics have traditionally been conceptualized as arising from a single uniform pathogen. However, *Mycobacterium tuberculosis*-complex (Mtbc), the pathogen causing tuberculosis in humans, encompasses multiple lineages exhibiting genetic and phenotypic diversity that may be responsible for heterogeneity in TB transmission. We analysed a population-based dataset of 1,354 Mtbc whole-genome sequences collected over four years in Botswana, a country with high HIV and tuberculosis burden. We identified Lineage 4 (L4) as the most prevalent (87.4%), followed by L1 (6.4%), L2 (5.3%), and L3 (0.9%). Within L4, multiple sublineages were identified, with L4.3.4 being the predominant sublineage. Phylodynamic analysis revealed L4.3.4 expanded steadily from late 1800s to early 2000s. Conversely, L1, L4.4, and L4.3.2 showed population trajectories closely aligned with the HIV epidemic. Meanwhile, L2 saw rapid expansion throughout most of the 20th century but declined sharply in early 1990s. Additionally, pairwise genome comparison of Mtbc highlighted differences in clustering proportions due to recent transmission at the sublineage level. These findings emphasize the diverse transmission dynamics of strains of different Mtbc lineages and highlight the potential for phylodynamic analysis of routine sequences to refine our understanding of lineage-specific behaviors.

Introduction

Despite being preventable and curable, tuberculosis (TB) claims an estimated 1.4 million lives each year worldwide, surpassing deaths caused by any other infectious disease besides severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2).[1] The SARS-CoV-2 pandemic highlighted the public health importance of monitoring the spread of different genomic variants globally and in local settings.[2] Indeed, genomic epidemiology has been instrumental in outbreak investigation and public health over the previous decade by tracing the spread of various infectious pathogens and informing key stakeholders. For instance, genomic surveillance of viral pathogens guided Ebola vaccine allocation in west Africa during the 2013–2016 outbreak,[3] identified likely source and routes of zika virus introductions into the United States during the 2016 outbreak,[4] and has continuously mapped transmission networks and detected high-risk groups in the ongoing HIV pandemic.[5, 6] These findings are facilitated through the use of phylodynamics, a statistical framework that integrates high-resolution genetic data with individual-level epidemiological data to shed light on the transmission dynamics of infectious pathogens, providing timely guidance for public health interventions.[7, 8]

For TB research, large-scale comparative genomic analysis have provided valuable insights into the origin and genetic diversity of the human-adapted TB bacteria, collectively known as the *Mycobacterium tuberculosis*–complex (Mtbc).[9, 10] There are currently a total of 9 major lineages of Mtbc and 64 sublineages that characterize their genetic diversity according to a 62-single nucleotide polymorphism (SNP) barcode.[11, 12, 13] Strain diversity in Mtbc can lead to important phenotypic variations in virulence, antibiotic resistance patterns, and transmissibility.[14, 15] For example, Lineages 2 and 4 are widespread globally and noted for their enhanced transmissibility, as evidenced by comparing Mtbc genomes (*e.g.,* genotypic clustering

62   proportions) collected in many settings.[16, 17] By contrast, Lineage 1 is less likely to cause disease

63   shortly after infection and may have a reduced ability to spread.[14, 16, 17] However, most

64   epidemiological studies and surveillance efforts still do not distinguish between Mtbc strains,

65   and the conventional approach of treating TB as a uniform epidemic fails to account for pathogen

66   heterogeneity and potential differences in the transmission dynamics of Mtbc strains, which

67   could vary significantly across different regions and populations.[14, 18, 19] Much like the insights

68   for viral transmission dynamics gained from genomic surveillance of SARS-CoV-2 variants,[2]

69   analysing Mtbc sequences to uncover long-term epidemiological trends of specific Mtbc strains

70   may offer a more granular understanding of the TB transmission dynamics at the population

71   level, which could improve our ability to predict the trajectory of the TB epidemic and design

72   more tailored interventions for a given setting.

73   Located in southern Africa, Botswana continues to face one of the most severe TB

74   endemics globally and remains one of the 30 high TB/HIV burden countries.[1] Despite the public

75   health implications, the evolutionary history and transmission dynamics of Mtbc lineages remain

76   poorly understood in this high burden country. In this study, we sought to expand our

77   understanding of the Mtbc population dynamics underlying the long-standing TB endemic in

78   Botswana by leveraging a countrywide dataset of whole genome sequences (WGS) from Mtbc

79   isolates collected over four years.[20] Our aim was to delineate the long-term spread of Mtbc

80   subpopulations in the context of sociohistorical events and the generalized HIV epidemic in the

81   region. In addition to the broader long-term trends revealed through phylodynamic analysis, we

82   also examined clustering of cases by Mtbc lineages based on pairwise genetic distances to track

83   recent transmission patterns. We hypothesized that TB in Botswana is composed of a collection

84   of heterogeneous epidemics driven by distinct transmission dynamics of Mtbc lineages.

85

Results

Distribution of Mtbc lineages in Botswana

During the study period, 1,426 Mtbc isolates underwent successful WGS and were assigned into known lineages based on lineage-specific SNPs. We excluded 72 isolates with evidence of mixed Mtbc strain infection for subsequent phylogenetic and phylodynamic analysis. Among the remaining 1,354 isolates, 1,184 belonged to lineage 4 (L4; 87.4%), 86 (6.4%) were classified as lineage 1 (L1), 72 (5.3%) were lineage 2 (L2), and 12 (0.9%) were lineage 3 (L3). L4 was found to encompass a number of sublineages, including L4.1 (n=2; 0.1%), L4.1.1 (n=141; 10.4%), L4.1.2 (n=149; 11.0%), L4.2.2 (n=1; 0.1%), L4.3.2 (n=117; 8.6%), L4.3.3 (n=27; 2.0%), L4.3.4 (n=400; 29.5%), L4.4 (n=189; 14.0%), L4.6 (n=2; 0.1%), L4.7 (n=14; 1.0%), L4.8 (n=92; 6.8%), and L4.9 (n=50; 3.7%). Within L1, 22 (1.6%) were L1.1 and 64 (4.7%) were L1.2. All of the L2 isolates were classified as L2.2 (Figure 1). Furthermore, Gaborone harbored more Mtbc strain diversity compared to Ghanzi. L4.3.4, the most prevalent sublineage, was found in higher proportion in Ghanzi than Gaborone (67% vs. 21%; Table 1). Host characteristics were similar among the major Mtbc lineages (Table 2).

Historical origins and population dynamics of Mtbc lineages

Demographic reconstruction was performed for L1, L2, L4.1.1, L4.1.2, L4.3.2, L4.3.4, L4.4, and L4.8. The time to the most recent common ancestor (MRCA) refers to the estimated age of a common ancestral population from which current populations or strains of Mtbc have descended. We inferred the MRCA to have emerged around year 1727 (95% highest posterior density interval [HPDI], 1607–1828) and 1900 (95% HPDI 1854–1938) for strains of L1 and L2, respectively. Time to MRCA varied among L4 sublineages, ranging from year 1695 for L4.1.2

108   (95% HPDI 1546–1817) to 1941 for L4.3.2 (95% HPDI 1911–1966; Table 3). Our analysis

109   revealed that strains of L4.3.4 experienced a steady expansion from late 1800s to early 2000s,

110   and remained the most prevalent sublineage in recent years. Similar population trajectory was

111   observed for L1 and L4.4 strains, with a substantial expansion observed from the late 1970s for

112   two decades, followed by a continuous decline from the late 1990s. For L4.3.2 strains, we

113   observed significant growth between 1980 to early 2000s, followed by a sharp reduction

114   afterward. L2 strains experienced a steady increase in population growth between 1940 and early

115   1990s, followed by a steep decline until 2000, after which the population plateaued followed by

116   a second decline around 2012. Other L4 sublineages experienced several waves of expansion and

117   contraction. Except for L4.1.1, all other Mtbc lineages underwent contraction after early 2000s

118   (Figure 2 and Supplementary Figure 1). For the inferred clock rate and maximum clade

119   credibility trees of these lineages, refer to Supplementary Table 1 and Supplementary Figure 2a-h.

120   Genomic clustering analysis

121   Using a 5-SNPs cutoff as an indicator of recent transmission, 48% of the isolates

122   (652/1354) were clustered. The cluster sizes varied, ranging from 2 isolates per cluster to 23

123   isolates in the largest cluster. The median cluster size for each lineage was as follows: L1 and L2:

124   2 (range: 2–7), L4.1.1: 3 (range: 2–23), L4.1.2: 3 (range: 2–12), L4.3.2 and L4.4: 2 (range 2–8),

125   L4.3.4: 2 (range: 2–19), and L4.8: 3.5 (range: 2–9). Compared to L1, we found that isolates

126   belonging to any other lineages were more likely to be part of a cluster (Table 4). Isolates

127   belonging to L2, L4.1.1, and L4.8 had the highest cluster proportion. Additionally, the cluster

128   proportion for the predominant L4.3.4 was significantly higher in Ghanzi than Gaborone (57%

129   vs. 40%; *p*-value <0.001; data not shown) These findings remained robust when we repeated the

130   analysis using a 12-SNPs cutoff (Supplementary Table 2).

131

132 **Discussion**

133   Few studies to date have examined lineage-specific, long-term transmission dynamics of

134 Mtbc strains in a high HIV/TB setting, largely due to the lack of population-level WGS datasets

135 over extended sampling periods.[16, 21] By reconstructing the evolutionary history and spread of

136 Mtbc strains of different lineages using WGS collected over a four-year period, we observed

137 varying transmission dynamics and demographic trajectories for extant lineages circulating in

138 Botswana. The expansion of many lineages coincided with the onset of the HIV/AIDS epidemic,

139 while their contraction aligned with the implementation of antituberculosis and antiretroviral

140 treatment programs.

141   With the exception of strains of L4.1.1, L4.1.2 and L4.8, strains of most lineages

142 experienced an expansion of various magnitude in the second half of the 20th century,

143 correlating closely with the region's high TB burden.[22] Notably, this period coincides with the

144 emergence and rapid escalation of the HIV epidemic. Botswana experienced one of the world's

145 most severe HIV epidemics, with initial cases emerging in the 1980s.[23] A phylodynamic study of

146 HIV dated the virus's emergence in Botswana around 1960, followed by rapid growth until

147 stabilization post-1990.[24] HIV increases the susceptibility of TB disease and is the strongest risk

148 factor for reactivating latent TB infections due to compromised immunity.[25] Further, the HIV

149 crisis was exacerbated by rapid urbanization driven by economic growth in the mining industry.

150 A network analysis of micro-census data from Botswana between 1981 and 2011 suggests a

151 highly mobile population with considerable movement between urban and rural areas.[26] Mining

152 towns, where HIV cases initially emerged, acted as major migration hubs for both inflow and

153 outflow of populations. In Selebi Phikwe, a mining town for copper and nickel in Botswana, HIV

154 prevalence had surged to 50% in 2000.[27] These towns also became hubs of TB transmission due

155 to the high-risk conditions associated with mining and the growing HIV prevalence.[28, 29] As the

156 HIV epidemic escalated, the number of individuals susceptible to Mtbc infection and reactivation

157 increased, driving the expansion of Mtbc populations. Thus, the intersection of the HIV epidemic,

158 mining-driven urbanization, and high population mobility likely played a critical role in the

159 significant expansion of strains from most Mtbc lineages in the latter part of the 20th century.

160 Nevertheless, this expansion was not observed across all lineages. It's possible that Mtbc strains

161 may exhibit varying capacities for TB transmission or reactivation influenced by their

162 interactions with the host in the context of HIV-related immunosuppression.[15, 17, 30] We observed

163 a marked decline in Mtbc populations at the turn of the 21st century, which aligned with a

164 consistent decrease in TB incidence from 2000 to 2016.[31] Several nationwide programs likely

165 contributed to the reduction. The most significant decrease in L1, L4.3.2, L4.3.4, L4.4, and L4.8

166 occurred beginning the early 2000s, coinciding with the launch of the Masa Antiretroviral

167 Therapy (ART) Programme in 2002.[23] This program, also known as the Masa Programme,

168 initially aimed to provide free ART to individuals with advanced immune suppression but

169 eligibility expanded over the years. "Masa" means "new dawn" in Setswana, symbolizing hope

170 and a fresh start for those affected by the epidemic. Shortly after, the President's Emergency Plan

171 for AIDS Relief (PEPFAR) was launched, focusing on combating HIV/AIDS globally, with a

172 significant emphasis on sub-Saharan Africa.[32] Botswana, with its high HIV prevalence rates, was

173 a key focus country for PEPFAR. Both initiatives aimed to support those affected by HIV/AIDS

174 through comprehensive treatment, prevention, and care programs, thereby reducing HIV

175 transmission.[23, 32, 33] As a result, Botswana had significantly increased number of people

176 receiving ART, reduced the rate of new HIV infections, and improved the overall healthcare

177   system's capacity to manage HIV/AIDS.[32, 33, 34] While PEPFAR and Masa Programme primarily

178   targeted the HIV/AIDS epidemic, their impact on reducing HIV transmission indirectly led to a

179   significant reduction in the burden of TB in Botswana.

180   Our study highlighted significant heterogeneity in transmission dynamics among strains

181   of different Mtbc lineages analysed. We found that L1 strains maintained a relatively stable

182   effective population size until experiencing a significant expansion following the emergence of

183   HIV. L1 is most prevalent in regions bordering the Indian Ocean, and several large-scale

184   phylogeographic studies have shown that South Asia and eastern Africa played critical roles in

185   dispersing this lineage via maritime trade.[35, 36] It is possible that MRCA of L1 strains was

186   introduced into Botswana through subsequent migrations between eastern and southern regions

187   of the continent. Post establishment, evidence suggests that L1 likely expanded due to changing

188   environmental conditions (*e.g.* increased population densities, malnutrition, immunosuppression),

189   rather than from large migration or transmission events.[35] Consistent with previous findings, our

190   study found that L1 had the lowest proportion of cases that belonged to clusters.[14, 16, 37] A

191   phylodynamic study from Vietnam, an L1-endemic region, concluded that compared to L2 and

192   L4, the disease burden due to L1 in the region was associated with reactivation of long-term

193   remote infections, evidenced by larger SNP differences and limited inferred migration events

194   among L1 isolates.[38] Similarly, in Botswana, reactivation of latent TB infections likely

195   contributes to new cases of L1 and may have driven this lineage's expansion as population

196   immunity waned due to HIV. L4.4 strains followed similar demographic trends but showed a

197   slower decline in population size after nationwide introduction of ART interventions, which may

198   be linked to its higher proportion of clustered cases due to recent transmission compared to L1

199   strains.

200   Notably, L2 strains underwent rapid expansion in effective population size shortly after

201   their estimated emergence in the early 20th century. Phylogenetic studies suggest that L2.2, often

202   associated with the Beijing genotype, likely originated in China.[21, 39] The 19th and 20th centuries

203   were characterized by intense interaction between Europe and China, driven by trade, imperial

204   ambitions, and cultural exchanges. Following the Opium Wars, numerous Chinese cities became

205   treaty ports where European powers had extraterritorial rights and conducted trade with relative

206   freedom. Prominent treaty ports included Shanghai, Guangzhou, and Tianjin, which became hubs

207   of international trade and cultural exchange. Additionally, during the late 19th and early 20th

208   centuries, Chinese laborers were often recruited for mining and railway construction projects

209   worldwide, including in Africa.[40] South Africa, which borders Botswana, had a significant

210   number of Chinese laborers working in its mines.[51] This movement could have facilitated the

211   introduction and spread of L2.2 (with putative origins in Beijing, China) to southern Africa.[50] L2

212   has been associated with hypervirulence (ability to cause disease shortly after infection) and high

213   transmissibility, supported by both clinical evidence and animal experiments.[41, 42] In our study,

214   we also observed that L2 had one of the highest proportions of genomic clustering due to recent

215   transmission. However, L2 strains experienced a decline in the early 1990s before the

216   implementation of country-wide ART programs. This decline might reflect a combination of

217   factors, including implementation of facility-based directly observed therapy, local pathogen-

218   host dynamics and its challenges in competing with other well-adapted indigenous lineages.[15]

219   The discrepancy between high genotypic clustering proportion and population dynamic of L2

220   strains in Botswana highlights that genotypic clustering analysis alone does not necessarily

221   capture the dynamics of an epidemic, nor does it always correlate with the population size of a

222   specific Mtbc lineage. Our finding also contrasts sharply with the phylodynamic study in

223 Vietnam, where L2 showed clear patterns of overtaking endemic L1 strains over time.[38] Overall,

224 these findings imply that the competitive dynamics between strains of different lineages can vary

225 significantly depending on the region and the specific lineages involved, and that studying

226 lineage features without accounting for geographic or host genetic background can obscure

227 important host-pathogen interactions.[15, 43]

228 Our analysis indicates that most TB cases in Botswana were caused by L4 strains,

229 specifically L4.3 strains. This finding aligns with previous reports of L4-belonging Latin

230 American and Mediterranean sublineage's dominant presence in southern Africa region.[44, 45, 46, 47]

231 Based on the sampled genetic diversity, we estimated that most L4 sublineages trace back to

232 MRCAs that emerged between the late 17th and 19th centuries. Recent genomic studies of Mtbc

233 suggests that the introduction and spread of L4 coincide with historical events involving

234 European colonization and subsequent socioeconomic changes.[44, 48] European colonization of

235 South Africa, Botswana's neighboring country, began in earnest with the establishment of a

236 Dutch settlement at the Cape of Good Hope in 1652.[49] Later, British colonization intensified in

237 early 19th century, and Botswana (then Bechuanaland) became a British protectorate in 1885.

238 The establishment of colonial administration and subsequent movements of European settlers

239 between Europe, South Africa and Botswana during this period may have facilitated the

240 establishment and spread of L4 strains still observed today.

241 The most common Mtbc lineage found in Botswana was L4.3.4, and it was

242 overwhelmingly more prevalent in Ghanzi than Gaborone (67% vs. 21%). Another study using a

243 different genotyping technique noted this, but was unable to differentiate whether it was due to

244 recent transmission or reactivation of latent infections.[50] The effective population size of L4.3.4

245 strains grew steadily since its introduction around 1880 and only began to decline slightly after

11

246   country-wide ART programs began. Unlike other lineages, L4.3.4's expansion during the HIV

247   epidemic was modest. This contrast may be explained by the higher prevalence of this lineage in

248   Ghanzi, which has lower HIV prevalence compared to Gaborone. Prior to the study,

249   approximately 36% of TB patients in Ghanzi were coinfected with HIV, compared to 70% in

250   Gaborone,[51] suggesting a less pronounced impact of the HIV epidemic on TB transmission

251   dynamics in Ghanzi. In our analysis, L4.3.4 displayed moderate clustering proportions due to

252   recent transmission, falling between the lowest (L1) and the highest groups (e.g., L2, L4.1.1,

253   L4.8). However, it exhibited significantly higher cluster proportions in Ghanzi compared to

254   Gaborone. Taken collectively, we propose that the success of L4.3.4 is driven primarily by

255   sustained ongoing transmission, rather than reactivation of latent infections.[50] This highlights the

256   need for public health efforts to prioritize activities that interrupt transmission, such as enhancing

257   contact tracing,[52] improving infection control measures,[53] and ensuring timely diagnosis and

258   treatment to reduce disease burden, especially in a rural setting like Ghanzi.[54]

259   L4.3.4 displayed clear competitive advantage as evidenced by its dominant presence in

260   both districts despite not being the earliest lineage to emerge in Botswana. It's also interesting to

261   observe that L4.3.2, a strain closely related to L4.3.4, emerged around the time of estimated HIV

262   introduction in Botswana.[24] However, unlike L4.3.4's steady growth, L4.3.2 experienced a rapid

263   expansion followed by a sharp decline, mirroring remarkably well with the timing of the HIV

264   epidemic. The striking difference in population dynamics between two closely related strains

265   warrant further investigation to determine whether strain-specific mutations contribute to distinct

266   phenotypes, and/or if potential host-pathogen interactions influence their respective trajectories

267   and the success of L4.3.4.

268   L4.1.1, L4.1.2, and L4.8 strains share similar demographic trends with large overlapping

269   CIs. Notably, L4.1.1 and L4.8 exhibited the highest cluster proportions due to recent

270   transmission, almost doubling the estimate compared to L4.3.4. This variation in genotypic

271   clustering within L4 sublineages was also observed in San Francisco, USA,[55] underscoring

272   sublineage-level heterogeneity and the necessity of investigating them individually. Furthermore,

273   L4.1.1 was the only lineage that expanded after the implementation of country-wide ART

274   programs. In addition to possible sublineage-specific characteristics (*e.g.* virulence), other

275   socioecological drivers, for example, overcrowding or malnutrition, may be responsible for the

276   expansion and contraction of this sublineage.[56] However, the relatively wide CIs associated with

277   the effective population size of these sublineages limit our ability to analyze local features of

278   their curves.[57]

279   Our study had limitations. First, the available genetic data may not fully represent the

280   diversity of Mtbc strains circulating in Botswana. However, the greater Gaborone and Ghanzi

281   district encompass a considerable segment (20%–25%) of the total population, and sampling in

282   Gaborone likely captured a substantial portion of Mtbc strain diversity due to its role as a central

283   urban hub for inflow and outflow of population. Second, Mtbc isolates were limited to sputa. It is

284   possible that some lineages may be under-represented because of the predication for pulmonary

285   clinical presentation.[58] Third, coalescent models (*e.g.* Skyride model applied in our analysis) use

286   pathogen genomic data to estimate the timescale of population dynamics.[59] The relatively low

287   mutation rate of Mtbc can result in broad uncertainty intervals for estimated effective population

288   sizes and divergence times, making it challenging to pinpoint precise timings for events such as

289   introductions or expansions of lineages. Inclusion of Mtbc sequences as they become available in

290   the future may improve the reconstruction of Mtbc population dynamics and reduce uncertainty

291 associated with population parameters. Lastly, we acknowledge the limitation of using a fixed-

292 SNP threshold to determine transmission clusters.[60] The chosen threshold is often arbitrary and

293 might not accurately reflect recent transmission. However, a SNP cutoff provides a standardized

294 approach, making it easier to compare results across different studies and useful for public health

295 investigations. The chosen 5-SNP cutoff for determining TB transmission clusters is common in

296 molecular epidemiology and has been shown to provide an informative overview of local

297 transmission patterns.[61, 62]

298 In summary, our study provides insight into the recent evolutionary origins and

299 population dynamics of strains of different Mtbc lineages in Botswana. We highlight the

300 heterogeneous transmission dynamics of key lineages circulating in Botswana and emphasize

301 that increasing awareness of these heterogeneities may be essential to further reducing TB

302 burden. Similar to viral pathogens, we recommend routine molecular surveillance and

303 phylodynamic analysis to monitor local TB transmission. This could enable public health

304 programs to effectively track the spread of different Mtbc strains, identify those that are

305 expanding over time, and tailor control strategies based on the epidemiological trends of

306 circulating Mtbc strains within a setting.

307

Methods

309 *Data source and study setting*

310 Data were obtained from a population-based TB study conducted between 2012 and 2016.

311 Study procedures have been described in detail elsewhere.[20] Briefly, the study enrolled

312 individuals of all ages with presumed TB in the greater Gaborone and Ghanzi districts of

313 Botswana. Gaborone, located in the southeast and bordering South Africa, is the capital city with

14

314   a high population density of 1,370 persons/km^2 in 2011. In contrast, Ghanzi is a rural district

315   located in western Botswana with the lowest population density in the country according to the

316   2011 census, at 0.37 persons/km^2.[63] Despite Ghanzi's low population density, it consistently has

317   the highest TB incidence in the country. Prior to the study, the TB incident rate per 100,000

318   people in Botswana, and Gaborone and Ghanzi districts were approximately 445, 440, and 722

319   cases, respectively.[64]

320   During the enrollment process, participants without a documented HIV status or with

321   negative test results over 12 months old were offered rapid HIV testing. Each participant was

322   required to provide at least one expectorated sputum sample. Sputum induction with nebulized

323   hypertonic saline solution was performed for those who had difficulty producing sputum

324   spontaneously. The Kopanyo Study was approved by the U.S. Centers for Disease Control and

325   Prevention, the Botswana Ministry of Health and Wellness, and the University of Pennsylvania

326   Institutional Review Boards. Additionally, approval of this secondary data analysis was obtained

327   from University of California at Irvine Institutional Review Board. Written informed consent

328   was obtained from all participants prior to enrollment.

329   *Culture and DNA extraction*

330   The sputum specimens underwent decontamination using the N-acetyl-L-cysteine and

331   sodium hydroxide method. After processing, the specimens were inoculated into Mycobacteria

332   Grow Indicator Tubes (MGIT) 960 system (Becton Dickinson Microbiology Systems, Sparks,

333   MD, USA) at the Botswana national reference laboratory in Gaborone. Weekly monitoring of

334   mycobacteria growth occurred, and if no growth was observed within 8 weeks, it was classified

335   as culture negative. For culture positive colonies, Mtbc DNA was extracted with GenoLyse kit

336   (Hain Lifescience, Germany) following the manufacturer's protocol. Genomic extracts were

15

337   stored at -80°C, and those containing a sufficient amount of DNA (at least 0.05ng/uL) were sent

338   to the Research Center Borstel for WGS.

339   *Whole genome sequencing and bioinformatics*

340   Genomic libraries were prepared with Illumina Nextera XT kit to generate 2x150bp

341   paired-end reads for sequencing using the Illumina NextSeq 500 platform. An automated

342   bioinformatics pipeline, MTBseq, was used for WGS data analysis.[65] MTBseq combined all

343   necessary steps needed for analysis of WGS data. Briefly, raw sequences were aligned to the

344   pan-susceptible, H37Rv reference genome (GenBank NC000962.2) with BWA-mem software

345   and Samtools. Next, base call recalibration and realignment of reads around insertions and

346   deletions was performed with GATKv3. Variant calling was performed using Samtools mpileup

347   and in-house scripts, and positions were considered reliable with 75% allele frequency, at least 4

348   calls with a phred score of 20 or more, and a minimum of 4 reads mapped in each direction

349   (forward and reverse orientation). Libraries were considered as high quality and further analyzed

350   when at least 50x coverage was achieved and 95% of the reference genome covered. Next,

351   genomes were annotated with known associations to antibiotic resistance. Finally,

352   phylogenetically informative SNPs based on existing literature were used for lineage

353   classification of the input samples,[13] excluding genes associated with antibiotic resistance,

354   insertions and deletions, repetitive regions (PPE and PE-PGRS gene families), and consecutive

355   variants in a 12bp window.

356   *Phylogenetic reconstruction*

357   We extracted variable sites from whole genome alignments of Mtbc isolates to create the

358   fasta files used for phylogenetic analysis. To illustrate the overall population structure of Mtbc

359   lineages in Botswana, we used IQ-TREE v1.6.12 to reconstruct maximum likelihood

16

360 phylogenies for each major Mtbc lineage, with isolates possibly containing mixed Mtbc strains

361 excluded from the dataset.[66] The Hasegawa, Kishino, and Yano (HKY) model of nucleotide

362 evolution was assumed,[67] and SNP ascertainment bias correction was specified. Additionally,

363 time-measured phylogenetic trees for specific lineages included in the coalescent-based analysis

364 were estimated using Bayesian method with details described below.

365 *Coalescent-based demographic reconstruction*

366 We reconstructed Mtbc population dynamics for L1, L2, and L4 sublineages that had at

367 least 70 samples using BEAST v1.10.4.[68] The HKY nucleotide substitution model with gamma-

368 distributed rate heterogeneity (discretized to 4 categories) was used.[67] We specified a strict

369 molecular clock model, with the substitution rate assumed to be distributed lognormal with a

370 mean of 1 and a sigma of 1.25. This prior translates to 95% of the probability mass between 0.09

371 and 4.2 SNPs per site per year (s/s/y). We also explored alternative priors more aligned with the

372 published substitution rate of Mtbc,[69] such as a lognormal distribution with a mean of 0.001 and

373 a sigma of 2 (i.e. 95% of probability mass between 2.4E-10 and 0.001 s/s/y), as well as a uniform

374 distribution with upper and lower bounds set to 5E-7 and 1E-8 s/s/y, respectively. However, the

375 posterior clock rate remained robust to these different prior specifications. To estimate the

376 mycobacterial population size change through time, we adopted the coalescent-based Gaussian

377 Markov random fields Bayesian Skyride tree prior.[59] This approach offers greater flexibility

378 compared to the Bayesian Skyline prior and does not require strong prior assumptions on the

379 number of population size change points. Furthermore, to account for ascertainment bias, we

380 made manual adjustments in the xml file to incorporate the number of invariant sites. This

381 correction was applied individually for each xml file corresponding to a specific lineage. The

382 Markov chain Monte Carlo (MCMC) chain for each lineage ran for 500 million iterations, with

383   parameters and trees sampled every 50,000th iteration. We verified adequate mixing and

384   posterior convergence using Tracer v1.7.2 after discarding the first 10% as burn-in. For a

385   parameter to be considered sufficiently sampled, we aimed for an effective sample size of at least

386   200. If any parameter fell short of this threshold, we ran a second chain of 500 million iterations

387   and combined the chains using LogCombiner v1.10.4. We note that L4.3.4 – the largest

388   sublineage in our dataset – was the only one that required a second MCMC chain of 500 million

389   iterations to achieve sufficient mixing and sampling. Finally, a consensus tree summarizing the

390   sampled posterior trees for each lineage was constructed using the maximum clade credibility

391   method.

392   *Genomic clustering analysis*

393   We explored differences in patterns of recent transmission between Mtbc lineages that

394   were included in the coalescent analysis. Genetically similar Mtbc strains under a pre-specified

395   SNP threshold (usually between 5 and 12-SNPs) are interpreted as a consequence of recent

396   transmission events[61]. We assigned cases into clusters, which was defined as two or more

397   individuals with Mtbc strains sharing ≤5-SNPs differences, as determined by pairwise SNP

398   comparison.[61, 62] Next, we regressed clustered or non-clustered case on Mtbc lineage in a logistic

399   model, adjusting for potential confounding by age, gender, enrollment district, and HIV status.

400   To evaluate the robustness of our findings, we also conducted logistic regression analysis using a

401   12-SNPs cutoff for identifying clustered cases. All statistical analysis and phylogenetic tree

402   visualizations were conducted in R version 4.2.0.[70] WGS data from this study has been submitted

403   to the European Nucleotide Archive (ENA) under study accession PRJEB:62480.

404

References

1. World Health Organization. Global Tuberculosis Report 2023. World Health Organization (2023).

2. Walensky, R. P., Walke H. T., Fauci A. S. SARS-CoV-2 variants of concern in the United States-challenges and opportunities. *JAMA* **325**, 1037-1038 (2021).

3. Kinganda-Lusamaki, E.*, et al.* Integration of genomic sequencing into the response to the Ebola virus outbreak in Nord Kivu, Democratic Republic of the Congo. *Nature Medicine* **27**, 710-716 (2021).

4. Grubaugh, N. D.*, et al.* Genomic epidemiology reveals multiple introductions of Zika virus into the United States. *Nature* **546**, 401-405 (2017).

5. Yuan, D.*, et al.* HIV-1 genetic transmission networks among people living with HIV/AIDS in Sichuan, China: a genomic and spatial epidemiological analysis. *Lancet Reg Health West Pac* **18**, 100318 (2022).

6. Skaathun, B.*, et al.* HIV-1 transmission dynamics among people who inject drugs on the US/Mexico border during the COVID-19 pandemic: a prosepective cohort study. *Lancet Reg Health Am* **33**, 100751 (2024).

7. Attwood, S. W., Hill S. C., Aanensen D. M., Connor T. R., Pybus O. G. Phylogenetic and phylodynamic approaches to understanding and combating the early SARS-CoV-2 pandemic. *Nat Rev Genet* **23**, 547-562 (2022).

8. Rife, B. D.*, et al.* Phylodynamic applications in 21(st) century global infectious disease research. *Glob Health Res Policy* **2**, 13 (2017).

19

427   9.   Bos, K. I.*, et al.* Pre-Columbian mycobacterial genomes reveal seals as a source of New

428   World human tuberculosis. *Nature* **514**, 494-497 (2014).

429   10.   Comas, I.*, et al.* Out-of-Africa migration and Neolithic coexpansion of Mycobacterium

430   tuberculosis with modern humans. *Nat Genet* **45**, 1176-1182 (2013).

431   11.   Coscolla, M.*, et al.* Phylogenomics of Mycobacterium africanum reveals a new lineage

432   and a complex evolutionary history. *Microb Genom* **7**, (2021).

433   12.   Ngabonziza, J. C. S.*, et al.* A sister lineage of the Mycobacterium tuberculosis complex

434   discovered in the African Great Lakes region. *Nat Commun* **11**, 2917 (2020).

435   13.   Coll, F.*, et al.* A robust SNP barcode for typing Mycobacterium tuberculosis complex

436   strains. *Nat Commun* **5**, 4812 (2014).

437   14.   Freschi, L.*, et al.* Population structure, biogeography and transmissibility of

438   Mycobacterium tuberculosis. *Nat Commun* **12**, 6099 (2021).

439   15.   Gagneux, S. Host-pathogen coevolution in human tuberculosis. *Philosophical*

440   *transactions of the Royal Society of London Series B, Biological sciences* **367**, 850-859

441   (2012).

442   16.   Guerra-Assuncao, J. A.*, et al.* Large-scale whole genome sequencing of M. tuberculosis

443   provides insights into transmission in a high prevalence area. *Elife* **4**, (2015).

444   17.   Coscolla, M., Gagneux S. Consequences of genomic diversity in Mycobacterium

445   tuberculosis. *Semin Immunol* **26**, 431-444 (2014).

446   18.   Mathema, B.*, et al.* Epidemiologic consequences of microvariation in Mycobacterium

447   tuberculosis. *J Infect Dis* **205**, 964-974 (2012).

448   19.   Gagneux, S., *et al.* Variable host-pathogen compatibility in Mycobacterium tuberculosis.

449   *Proc Natl Acad Sci U S A* **103**, 2869-2873 (2006).

450   20.   Zetola, N. M., *et al.* Protocol for a population-based molecular epidemiology study of

451   tuberculosis transmission in a high HIV-burden setting: the Botswana Kopanyo study.

452   *BMJ open* **6**, e010046 (2016).

453   21.   Liu, Q., *et al.* China's tuberculosis epidemic stems from historical expansion of four

454   strains of Mycobacterium tuberculosis. *Nat Ecol Evol* **2**, 1982-1992 (2018).

455   22.   Dodd, P. J., *et al.* Transmission modeling to infer tuberculosis incidence prevalence and

456   mortality in settings with generalized HIV epidemics. *Nat Commun* **14**, 1639 (2023).

457   23.   Ramogola-Masire, D., *et al.* Botswana's HIV response: Policies, context, and future

458   directions. *J Community Psychol* **48**, 1066-1070 (2020).

459   24.   Wilkinson, E., Engelbrecht S., de Oliveira T. History and origin of the HIV-1 subtype C

460   epidemic in South Africa and the greater southern African region. *Sci Rep* **5**, 16897

461   (2015).

462   25.   Corbett, E. L., *et al.* The growing burden of tuberculosis: global trends and interactions

463   with the HIV epidemic. *Archives of internal medicine* **163**, 1009-1021 (2003).

464   26.   Song, J., *et al.* Population mobility and the development of Botswana's generalized HIV

465   epidemic: a network analysis. *medRxiv*, (2023).

466   27.   UNAIDS. Botswana: Country Factsheets. UNAIDS (2004).

467   28.   Stuckler, D., Basu S., McKee M., Lurie M. Mining and risk of tuberculosis in sub-

468   Saharan Africa. *Am J Public Health* **101**, 524-530 (2011).

469   29.   Basu, S., Stuckler D., Gonsalves G., Lurie M. The production of consumption: addressing

470   the impact of mineral mining on tuberculosis in southern Africa. *Global Health* **5**, 11

471   (2009).

472   30.   Fenner, L.*, et al.* HIV infection disrupts the sympatric host-pathogen relationship in

473   human tuberculosis. *PLoS Genet* **9**, e1003318 (2013).

474   31.   The World Bank. Incidence of tuberculosis in Botswana. The World Bank Group (2022).

475   32.   Chin, R. J., Sangmanee D., Piergallini L. PEPFAR funding and reduction in HIV

476   infection rates in 12 focus sub-Saharan African countries: a quantitative analysis. *Int J*

477   *MCH AIDS* **3**, 150-158 (2015).

478   33.   Farahani, M.*, et al.* Outcomes of the Botswana national HIV/AIDS treatment programme

479   from 2002 to 2010: a longitudinal analysis. *Lancet Glob Health* **2**, e44-50 (2014).

480   34.   Mine, M.*, et al.* Progress towards the UNAIDS 95-95-95 targets in the Fifth Botswana

481   AIDS Impact Survey (BAIS V 2021): a nationally representative survey. *Lancet HIV* **11**,

482   e245-e254 (2024).

483   35.   O'Neill, M. B.*, et al.* Lineage specific histories of Mycobacterium tuberculosis dispersal

484   in Africa and Eurasia. *Mol Ecol* **28**, 3241-3256 (2019).

485   36.   Menardo, F.*, et al.* Local adaptation in populations of Mycobacterium tuberculosis

486   endemic to the Indian Ocean Rim. *F1000Res* **10**, 60 (2021).

487   37.   Couvin, D., Reynaud Y., Rastogi N. Two tales: Worldwide distribution of Central Asian

488   (CAS) versus ancestral East-African Indian (EAI) lineages of Mycobacterium

489 tuberculosis underlines a remarkable cleavage for phylogeographical, epidemiological

490 and demographical characteristics. *PLoS One* **14**, e0219706 (2019).

491 38. Holt, K. E.*, et al.* Frequent transmission of the Mycobacterium tuberculosis Beijing

492 lineage and positive selection for the EsxW Beijing variant in Vietnam. *Nat Genet* **50**,

493 849-856 (2018).

494 39. Merker, M.*, et al.* Evolutionary history and global spread of the Mycobacterium

495 tuberculosis Beijing lineage. *Nat Genet* **47**, 242-249 (2015).

496 40. Altan, S. *Chinese Workers of the World: Colonialism, Chinese Labor, and the Yunnan–*

497 *Indochina Railway*. Stanford University Press (2024).

498 41. Hanekom, M.*, et al.* Mycobacterium tuberculosis Beijing genotype: a template for

499 success. *Tuberculosis (Edinb)* **91**, 510-523 (2011).

500 42. Parwati, I., van Crevel R., van Soolingen D. Possible underlying mechanisms for

501 successful emergence of the Mycobacterium tuberculosis Beijing genotype strains. *The*

502 *Lancet Infectious diseases* **10**, 103-111 (2010).

503 43. Gagneux, S. Ecology and evolution of Mycobacterium tuberculosis. *Nat Rev Microbiol*

504 **16**, 202-213 (2018).

505 44. Stucki, D.*, et al.* Mycobacterium tuberculosis lineage 4 comprises globally distributed

506 and geographically restricted sublineages. *Nat Genet* **48**, 1535-1543 (2016).

507 45. Mogashoa, T.*, et al.* Genetic diversity of Mycobacterium tuberculosis strains circulating

508 in Botswana. *PLoS One* **14**, e0216306 (2019).

509   46.   Chihota, V. N.*, et al.* Geospatial distribution of Mycobacterium tuberculosis genotypes in

510   Africa. *PLoS One* **13**, e0200632 (2018).

511   47.   Viegas, S. O.*, et al.* Molecular diversity of Mycobacterium tuberculosis isolates from

512   patients with pulmonary tuberculosis in Mozambique. *BMC Microbiol* **10**, 195 (2010).

513   48.   Brynildsrud, O. B.*, et al.* Global expansion of Mycobacterium tuberculosis lineage 4

514   shaped by colonial migration and local adaptation. *Sci Adv* **4**, eaat5869 (2018).

515   49.   Oliver, E., Oliver W. H. The colonisation of South Africa: a unique case. *HTS*

516   *Theological Studies* **73**, 1-8 (2017).

517   50.   Click, E. S.*, et al.* Phylogenetic diversity of Mycobacterium tuberculosis in two

518   geographically distinct locations in Botswana - The Kopanyo Study. *Infect Genet Evol* **81**,

519   104232 (2020).

520   51.   Statistics Botswana. Botswana AIDS Impact Survey IV 2013 (BAIS IV): Report.

521   National AIDS & Health Promotion Agency (2013).

522   52.   Moonan, P. K.*, et al.* A neighbor-based approach to identify tuberculosis exposure, the

523   Kopanyo study. *Emerging infectious diseases* **26**, 1010-1013 (2020).

524   53.   Smith, J. P.*, et al.* High-resolution characterization of nosocomial Mycobacterium

525   tuberculosis transmission events in Botswana. *Am J Epidemiol* **192**, 503-506 (2023).

526   54.   Smith, J. P.*, et al.* Characterizing tuberculosis transmission dynamics in high-burden

527   urban and rural settings. *Sci Rep* **12**, 6780 (2022).

528   55.   Anderson, J.*, et al.* Sublineages of lineage 4 (Euro-American) Mycobacterium

529   tuberculosis differ in genotypic clustering. *Int J Tuberc Lung Dis* **17**, 885-891 (2013).

530 56. Lopez, M. G.*, et al.* Deciphering the tangible spatio-temporal spread of a 25-year

531   tuberculosis outbreak boosted by social determinants. *Microbiol Spectr* **11**, e0282622

532   (2023).

533 57. Du, D. H.*, et al.* The effect of M. tuberculosis lineage on clinical phenotype. *PLOS Glob*

534   *Public Health* **3**, e0001788 (2023).

535 58. Click, E. S., Moonan P. K., Winston C. A., Cowan L. S., Oeltmann J. E. Relationship

536   between Mycobacterium tuberculosis phylogenetic lineage and clinical site of

537   tuberculosis. *Clin Infect Dis* **54**, 211-219 (2012).

538 59. Minin, V. N., Bloomquist E. W., Suchard M. A. Smooth skyride through a rough skyline:

539   Bayesian coalescent-based inference of population dynamics. *Mol Biol Evol* **25**, 1459-

540   1471 (2008).

541 60. Hatherell, H. A.*, et al.* Interpreting whole genome sequencing for investigating

542   tuberculosis transmission: a systematic review. *BMC medicine* **14**, 21 (2016).

543 61. Walker, T. M.*, et al.* Whole-genome sequencing to delineate Mycobacterium tuberculosis

544   outbreaks: a retrospective observational study. *The Lancet Infectious diseases* **13**, 137-

545   146 (2013).

546 62. Zhang, X.*, et al.* Exploring programmatic indicators of tuberculosis control that

547   incorporate routine Mycobacterium tuberculosis sequencing in low incidence settings: a

548   comprehensive (2017-2021) patient cohort analysis. *Lancet Reg Health West Pac* **41**,

549   100910 (2023).

550   63.   Statistics Botswana. Botswana Population and Housing Census. Statistics Botswana

551   (2011).

552   64.   Zetola, N. M.*, et al.* Population-based geospatial and molecular epidemiologic study of

553   tuberculosis transmission dynamics, Botswana, 2012–2016. *Emerging infectious diseases*

554   **27**, 835-844 (2021).

555   65.   Kohl, T. A.*, et al.* MTBseq: a comprehensive pipeline for whole genome sequence

556   analysis of Mycobacterium tuberculosis complex isolates. *PeerJ* **6**, e5895 (2018).

557   66.   Nguyen, L. T., Schmidt H. A., von Haeseler A., Minh B. Q. IQ-TREE: a fast and effective

558   stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* **32**,

559   268-274 (2015).

560   67.   Hasegawa, M., Kishino H., Yano T. Dating of the human-ape splitting by a molecular

561   clock of mitochondrial DNA. *J Mol Evol* **22**, 160-174 (1985).

562   68.   Suchard, M. A.*, et al.* Bayesian phylogenetic and phylodynamic data integration using

563   BEAST 1.10. *Virus Evol* **4**, vey016 (2018).

564   69.   Menardo, F., Duchene S., Brites D., Gagneux S. The molecular clock of Mycobacterium

565   tuberculosis. *PLoS Pathog* **15**, e1008067 (2019).

566   70.   R Core Team. R: a language and environment for statistical computing. R Foundation for

567   Statistical Computing.

568

572 **Author contributions**

573 These authors contributed equally: Sanghyuk S. Shin and Stefan Neimann.

574 Q.W. contributed to project conception, performed the data analysis, and drafted the manuscript.

575 I.B. and S.N. processed the whole genome sequencing data and performed the bioinformatic

576 analysis. C.M., N.Z., P.K.M., A.F., R.B., J.E.O., and T.L.M. contributed to the Kopanyo Study

577 design, participant recruitment, data collection, and quality control. V.M.M., S.S.S., and T.F.B.

578 contributed to project conception, supervised data analysis and interpretation of results. All

579 authors were involved in revising the manuscript and approved the final version.

580

581 **Corresponding author**

582 Correspondence to Sanghyuk S. Shin.

583

584 **Competing interests**

585 The authors declare no competing interests.

586   **Table 1. Distribution of *Mycobacterium tuberculosis* complex lineages by district in**
587   **Botswana, 2012–2016**

	District		
	Gaborone n (%)	**Ghanzi n (%)**	**Total n (%)**
Lineage 1 Indo-Oceanic	83 (7.6)	3 (1.2)	86 (6.4)
L1.1	22 (2.0)	0	22 (1.6)
L1.2	61 (5.6)	3 (1.2)	64 (4.7)
Lineage 2 East Asian			
L2.2	65 (5.9)	7 (2.7)	72 (5.3)
Lineage 3 East African-Indian	12 (1.1)	0	12 (0.9)
Lineage 4 Euro-American	938 (85.4)	246 (96.1)	1184 (87.4)
L4.1	2 (0.2)	0	2 (0.1)
L4.1.1	136 (12.4)	5 (2.0)	141 (10.4)
L4.1.2	129 (11.7)	20 (7.8)	149 (11.0)
L4.2.2	1 (0.1)	0	1 (0.1)
L4.3.2	107 (9.7)	10 (3.9)	117 (8.6)
L4.3.3	24 (2.2)	3 (1.2)	27 (2.0)
L4.3.4	229 (20.9)	171 (66.8)	400 (29.5)
L4.4	164 (14.9)	25 (9.8)	189 (14.0)
L4.6	2 (0.2)	0	2 (0.1)
L4.7	13 (1.2)	1 (0.4)	14 (1.0)
L4.8	84 (7.7)	8 (3.1)	92 (6.8)
L4.9	47 (4.3)	3 (1.2)	50 (3.7)
Total	1098 (100)	256 (100)	1354 (100)

588

589

590 **Table 2. Participant characteristics by *Mycobacterium tuberculosis*–complex (Mtbc)**
591 **lineages in Botswana, 2012–2016**

	Mtbc lineage				
	L1 n (%)	L2 n (%)	L3 n (%)	L4 n (%)	Total n (%)
Sex					
Female	35 (40.7)	38 (52.8)	6 (50.0)	522 (44.1)	601 (44.4)
Male	51 (59.3)	34 (47.2)	6 (50.0)	662 (55.9)	753 (55.6)
Age					
<=15	2 (2.3)	3 (4.2)	1 (8.3)	21 (1.8)	27 (2.0)
16-24	10 (11.6)	9 (12.5)	1 (8.3)	226 (19.1)	246 (18.2)
25-40	47 (54.7)	46 (63.9)	7 (58.3)	632 (53.4)	732 (54.1)
41-64	24 (27.9)	14 (19.4)	3 (25.0)	280 (23.6)	321 (23.7)
>=65	3 (3.5)	0	0	25 (2.1)	28 (2.1)
Previous TB					
No	63 (73.3)	57 (79.2)	10 (83.3)	967 (81.7)	1097 (81.0)
Yes	23 (26.7)	15 (20.8)	2 (16.7)	217 (18.3)	257 (19.0)
Isoniazid preventive therapy					
No	83 (96.5)	64 (88.9)	12 (100.0)	1127 (95.2)	1286 (95.0)
Yes	2 (2.3)	5 (6.9)	0	42 (3.5)	49 (3.6)
Unknown	1 (1.2)	3 (4.2)	0	15 (1.3)	19 (1.4)
AFB smear results					
Negative	5 (5.8)	7 (9.7)	3 (25.0)	76 (6.4)	91 (6.7)
Positive	61 (70.9)	46 (63.9)	8 (66.7)	752 (63.5)	867 (64.0)
Not done	20 (23.3)	19 (26.4)	1 (8.3)	356 (30.1)	396 (29.2)
TB type					
Pulmonary	79 (91.9)	65 (90.3)	11 (91.7)	1075 (90.8)	1230 (90.8)
Extrapulmonary	7 (8.1)	6 (8.3)	0	105 (8.9)	118 (8.7)
Unknown	0	1 (1.4)	1 (8.3)	4 (0.3)	6 (0.4)
HIV status					
Negative	27 (31.4)	27 (37.5)	5 (41.7)	533 (45.0)	592 (43.7)
Positive	56 (65.1)	42 (58.3)	7 (58.3)	616 (52.0)	721 (53.2)
Unknown	3 (3.5)	3 (4.2)	0	35 (3.0)	41 (3.0)
ART status[1]					
Never taken ART	26 (46.4)	17 (40.5)	1 (14.3)	278 (45.1)	322 (44.7)
Taking ART	25 (44.6)	20 (47.6)	3 (42.9)	277 (45.0)	325 (45.1)
Took ART but stopped	1 (1.8)	1 (2.4)	2 (28.6)	11 (1.8)	15 (2.1)
Unknown	4 (7.1)	4 (9.5)	1 (14.3)	50 (8.1)	59 (8.2)

592 Abbreviations: AFB: acid-fast bacilli. ART: antiretroviral therapy.
593 [1]Among HIV-positive individuals only.
594

Table 3. Time to most recent common ancestor (tMRCA) of *Mycobacterium tuberculosis*–complex lineages in Botswana

Lineage	tMRCA (Posterior mean)	tMRCA (Posterior median)	95% HPD interval
L1	1727	1736	1607 – 1828
L2	1900	1903	1854 – 1938
L4.1.1	1729	1739	1587 – 1842
L4.1.2	1695	1705	1546 – 1817
L4.3.2	1941	1943	1911 – 1966
L4.3.4	1808	1814	1723 – 1881
L4.4	1739	1748	1621 – 1847
L4.8	1853	1858	1782 – 1912

Abbreviations: HPD: highest posterior density.

599 **Table 4. Genomic cluster proportions (based on a 5-SNPs cutoff) of the *Mycobacterium***
600 ***tuberculosis*–complex lineages**

Lineage	Cluster proportions	Crude OR (95% CI)	Adjusted[1] OR (95% CI)
L1	25/86	1	1
L2	42/72	3.42 (1.78, 6.69)	3.26 (1.68, 6.47)
L4.1.1	83/141	3.49 (1.99, 6.28)	3.44 (1.94, 6.25)
L4.1.2	74/149	2.41 (1.38, 4.29)	2.26 (1.28, 4.07)
L4.3.2	52/117	1.95 (1.09, 3.56)	1.91 (1.05, 3.52)
L4.3.4	190/400	2.21 (1.35, 3.71)	1.68 (1.00, 2.89)
L4.4	82/189	1.87 (1.09, 3.27)	1.78 (1.03, 3.16)
L4.8	55/92	3.63 (1.96, 6.86)	3.27 (1.74, 6.26)

601 Abbreviations: SNP: single nucleotide polymorphism. OR: odds ratio. CI: confidence interval.
602 [1]Adjusted for age, gender, HIV status, and district.
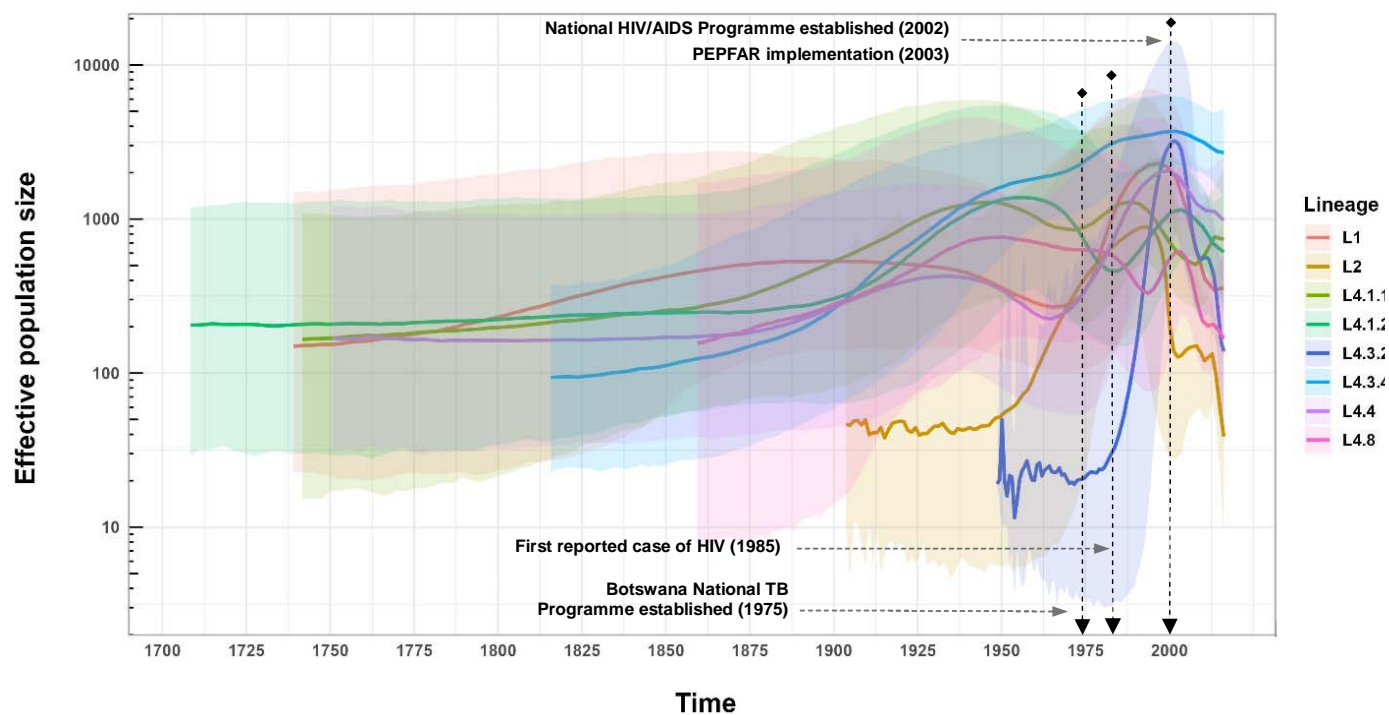603

604

Figure 1. Maximum likelihood phylogeny of *Mycobacterium tuberculosis*–complex (Mtbc) isolates collected in Botswana, 2012–2016 (n = 1,354).

A. Phylogeny of Lineage 1 Indo-Oceanic Mtbc isolates (n = 86). B. Phylogeny of Lineage 2 East Asian Mtbc isolates (n = 72). C. Phylogeny of Lineage 3 East African-Indian Mtbc isolates (n = 12). D. Phylogeny of Lineage 4 Euro-American Mtbc isolates (n = 1184)
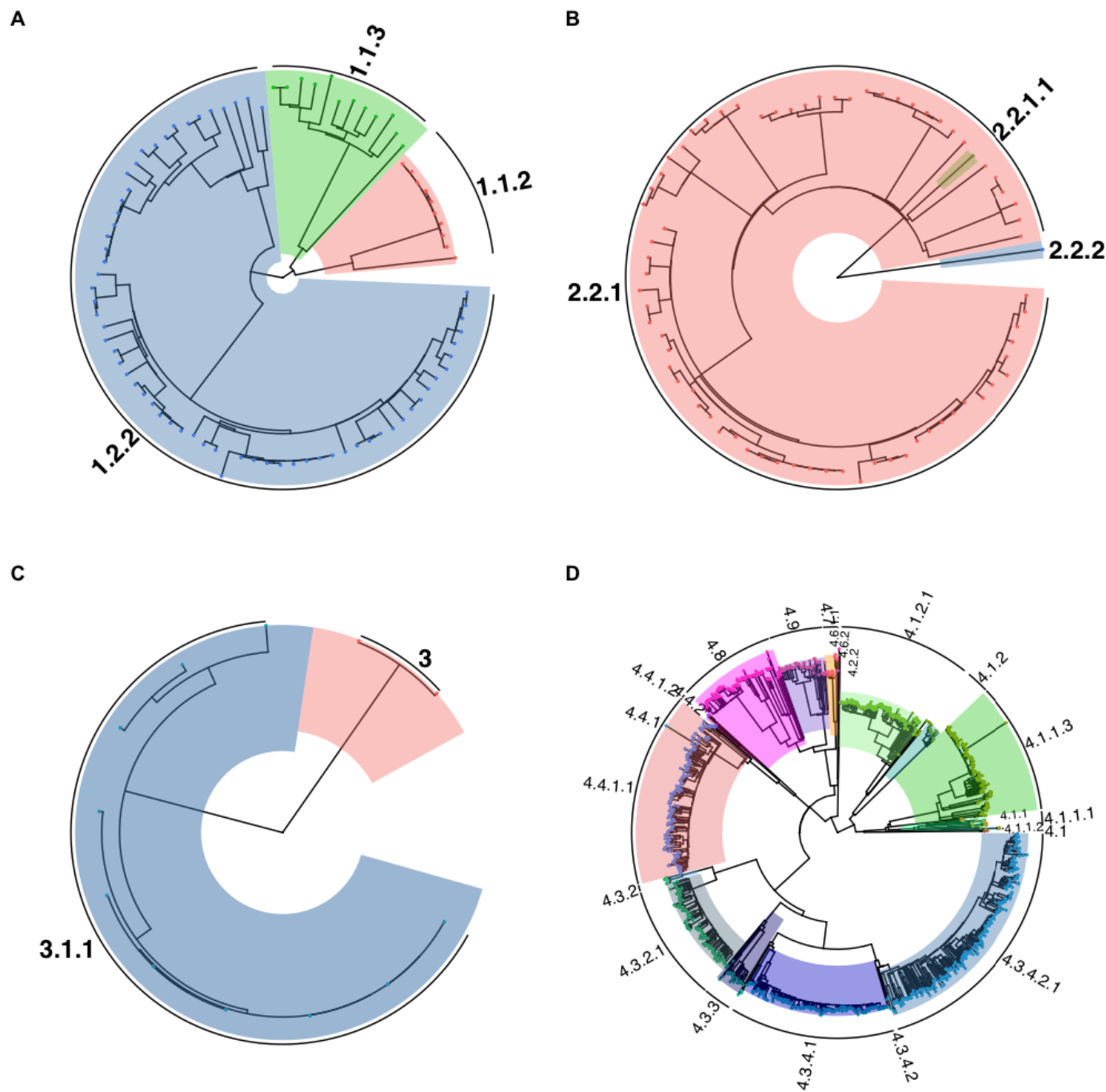
612   **Figure 2. Inferred effective population size over time for *Mycobacterium tuberculosis*–**
613   **complex lineages circulating in Botswana**

614

615

616

617