Title: Human genetic ancestry, *Mycobacterium tuberculosis* diversity and tuberculosis disease severity in Dar es Salaam, Tanzania

Short title: Human and bacterial genomic background and TB disease severity

Michaela Zwyer[1,2], Zhi Ming Xu[3,4], Amanda Ross[1,2], Jerry Hella[5], Mohamed Sasamalo[5], Maxime Rotival[6] , Hellen Hiza[1,2], Liliana K. Rutaihwa[7], Sonia Borrell[1,2], Klaus Reither[1,2], Jacques Fellay[3,4,8], Damien Portevin[1,2], Lluis Quintana-Murci[6,9], Sebastien Gagneux[1,2*] and Daniela Brites[1,2*]

*Corresponding authors:

Daniela Brites: d.brites@swisstph.ch (DB)

Sebastien Gagneux: sebastien.gagneux@swisstph.ch (SB)

[1] Swiss Tropical and Public Health Institute, Basel, Switzerland

[2] University of Basel, Basel, Switzerland

[3] Swiss Institute of Bioinformatics, Lausanne, Switzerland

[4] School of Life Sciences, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

[5] Department of Intervention and Clinical Trials, Ifakara Health Institute, Bagamoyo, Tanzania

[6] Human Evolutionary Genetics Unit, Institut Pasteur, Université Paris Cité, CNRS UMR2000, Paris, France

[7] FIND, Foundation for Innovative New Diagnostics

[8] Precision Medicine Unit, Lausanne University Hospital and University of Lausanne, Lausanne, Switzerland

[9] Chair of Human Genomics and Evolution, Collège de France, 75005 Paris, France

Author contributions:

MZ: Conceptualization, Data Curation, Formal Analysis, Project Administration, c, Visualization, Writing – Original Draft Preparation; ZMX: Resources, Writing – Review & Editing; AR: Formal Analysis, Methodology, Writing – Review & Editing; JH: Resources; MS: Resources; MR: Methodology, Writing – Review & Editing; HH: Resources, Writing – Review & Editing; LKR: Resources; MR: Methodology; SB: Resources; KR: Funding Acquisition, Writing – Review & Editing, JF: Funding Acquisition, Writing – Review & Editing; DP: Funding Acquisition, Writing – Review & Editing; LQM: Conceptualization, Writing – Review & Editing; SG: Funding Acquisition, Supervision, Project administration, Writing – Original Draft Preparation; DB: Supervision, Conceptualization, Funding Acquisition, Investigation, Methodology, Writing – Original Draft Preparation

Abstract

Infectious diseases have affected humanity for millennia and are among the strongest selective forces. Tuberculosis (TB) is an ancient disease, caused by the human-adapted members of the *Mycobacterium tuberculosis* complex (MTBC). The outcome of TB infection and disease is highly variable, and co-evolution between human populations and MTBC strains may account for some of this variability. Particular human genetic ancestries have been associated with higher susceptibility to TB, but socio-demographic aspects of the disease can confound such associations. Here, we studied 1,000 TB patients from Dar es Salaam, Tanzania, together with their respective MTBC isolates, by combining human and bacterial genomics with clinical data. We found that the genetic background of the TB patient population was strongly influenced by migrationsof Bantu-speaking populations from West Africa, which is in contrast to the corresponding MTBC genotypes that were mainly introduced from outside Africa. These findings suggest a recent evolutionary history of co-existence between the human and MTBC populations in Dar es Salaam. We detected no evidence of an effect of human genetic ancestry, or MTBC phylogenetic diversity alone, nor their interaction, on TB disease severity. There was also no evidence of an association between human variation genome-wide and TB disease severity.Treatment-seeking, social and environmental factors are likely to be the main determinants of disease severity at the point of care in this patient population.

Author Summary

Tuberculosis (TB) is an ancient infectious disease that continues to challenge global health efforts. Here, we explored the interplay between human and bacterial genetics on TB in Dar es Salaam, Tanzania. We found that neither the genetic ancestry of the patient, nor the bacterial genotype alone, nor their interaction, influenced the severity of TB. Our finding indicate that in this patient population, social and environmental factors may be the main determinants of TB disease severity.

Introduction

Africa harbors the largest human genetic diversity worldwide (1). This continent is also inhabited by numerous ethnic and linguistic groups (2). While the long evolutionary history of modern humans in Africa and their large effective population sizes (3) explain this high genetic diversity, more recent migration events within and from outside Africa during the last 5,000 years, as well as admixture between historically separated populations, have resulted in some degree of homogenization (4). Hence, African populations nowadays are often composed of a mixture of different genetic ancestries (5, 6). One human migration that had a major influence on the population structure of present-day Africans is the so-called 'Bantu expansion', where Bantu-speaking groups migrated from central Western Africa southwards and eastwards, spreading farming technologies across sub-Saharan Africa, and admixing with local groups of hunter-gatherers and pastoralists (4, 7). Recent human genetic studies have identified moderate population structuring among Bantu-speaking populations (8, 9), yet, admixture with local populations has impacted immune responses to infectious diseases (10). Since pathogens have been one of the strongest selective forces driving human evolution (11), disease susceptibility and clinical outcomes can differ markedly between human populations. Genome-wide association studies (GWAS) have identified mutations altering the susceptibility to various infectious diseases (12-15), including tuberculosis (TB), which remains the main cause of human death due to a single infectious agent (16).

The bacteria that cause TB belong to the *Mycobacterium tuberculosis* complex (MTBC) and can be classified into ten human-adapted phylogenetic lineages: Lineage 1 (L1) to L10, plus several lineages adapted to different wild and domestic animal host species (17, 18). While TB is a global problem, the human-adapted MTBC lineages differ in their geographical distribution. L2 and L4 occur worldwide and other lineages are restricted to specific regions. Specifically, L1 and L3 mainly occur around the Indian Ocean (19), L5 and L6 that are limited

to West Africa (20), and L7 only occurs in Ethiopia (21). This phylogeographic population structure of the human-adapted MTBC led to the hypothesis that certain MTBC genotypes are locally adapted to their sympatric human populations (19). This hypothesis is supported by findings from cosmopolitan settings, where sympatric associations between the geographical origin of TB patients and their infecting MTBC strains were observed (22-25). Moreover, in immune-compromised individuals with HIV co-infection, these sympatric associations were disrupted (26). The various MTBC genotypes also differ phenotypically (27), regarding disease progression (28), transmission (29, 30), disease presentation (31) and immune activation in vitro (32, 33) and ex vivo (34).

Human genetic diversity has also been linked to differences in TB susceptibility. While for example *TYK2* has been associated with TB disease worldwide (35), several human genetic loci were not consistently associated with TB in populations from different geographical regions but specific to certain populations (36-40). Particular human genetic ancestries have also been found to play a role in the context of TB. People with higher proportions of native Peruvian genetic ancestry showed a higher risk of progressing to active TB (41), and a higher proportion of San genetic ancestry was associated with an increased risk for TB among South African Coloured individuals (42). Differences in ethnicity have also been associated with different inflammatory profiles of TB patients at the time of presentation (34). In addition to the effects of human and bacterial genetic diversity on TB, many social and environmental factors, as well as co-morbidities are known drivers of TB. These include malnutrition (43), HIV infection, diabetes (44, 45), poverty (46), smoking (47), and alcohol consumption (48).While associations between TB and individual host, pathogen or environmental factors have been found (36-41, 43), studies considering all these components simultaneously, remain scarce (49-51).

Symptomatic TB has a wide spectrum of severity, and studies conducted before anti-TB drugs became available, suggest that mild disease is associated with higher odds of natural recovery (52, 53). If mortality and natural recovery from symptomatic TB vary depending on the spectrum of disease, conceivably human populations exhibit genetic variation underlying disease severity. Furthermore, given that humans and the MTBC have coexisted for millennia, some of this variation could be linked to human ancestry. Here, we characterized the genetic variation including the genetic ancestry of a cohort of TB patients from Dar es Salaam, Tanzania, the phylogenetic lineage of their correspondent MTBC isolate, and investigated the association of both with TB disease severity while accounting for demographic, social-economical and environmental variables.

Results

Genetic ancestry of Tanzanian TB patients

Genetic ancestries were estimated for 7,479 individuals from 249 populations (Supplementary Figure S2) including 1,444 Tanzanian TB patients, using the software Admixture (54) with 53,255 SNPs (Figure 1, Supplementary Figure S1).

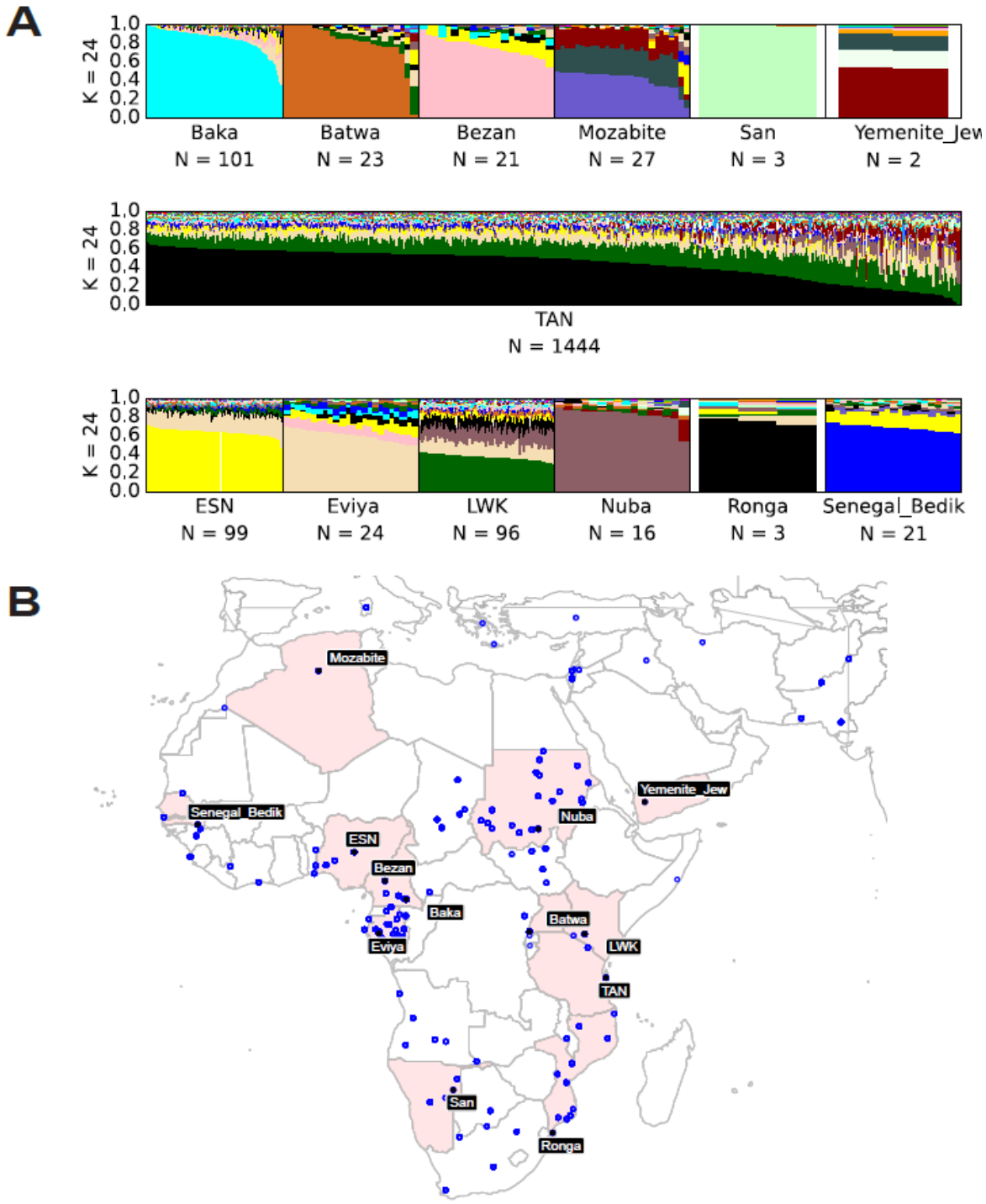


Figure 1 – Genetic ancestry analyses of Tanzanian TB patients.
A) Genetic ancestry proportions of the 1,444 Tanzanian TB patients and representative human populations who shared at least 1% of their most common genetic ancestry with the Tanzanians for K = 24. (ESN: Esan from Nigeria (1000G), LWK: Luhya from Kenya (1000G)). For all populations included in our study, see Supplementary Figure S2 for their geographic distribution and Figure S5 for the ancestry composition of all African populations included in this study. **B**) The geographical location of the representative populations shown in A are depicted with black circles, and the corresponding country is highlighted. The remaining African populations included in the analysis are represented by blue circles.

The optimal number of source populations to describe our dataset was 24, based on the lowest cross-validation error (55) (Supplementary Figure S3). For the purpose of this study, we named the genetic ancestries according to the geographical distribution and/or ethnicity of the reference populations that they are most prevalent in. The genetic ancestry with the highest contribution among Tanzanians with a mean of 44% (maximum 68%, minimum 0%) was also the most abundant in Bantu-speaking (BS) people from Southern and southeastern Africa (e.g. the Ronga population in Figure 1), and hence hereafter, we will refer to this ancestry as "Southeastern BS" (Figure 1). The second most common genetic ancestry with a mean of 22% (maximum 42%, minimum 6%) in the Tanzanian TB patients was most common among Kenyans (e.g. the Luhya population in Figure 1, 1000G and HGDP, see methods), and will be referred to as the "Eastern BS" genetic ancestry. Additionally, the Tanzanian TB patients had a mean of 9% (maximum 53%, minimum 0%) of a genetic ancestry that was most common among Bantu-speaking populations from western Central Africa, (e.g. the Eviya population in Figure 1); we will refer to it as the "Western BS" genetic ancestry. Furthermore, the Tanzanian TB patients contained on average 4% of a genetic ancestry most abundant among Nigerians represented by the Esan population ("Nigerian" genetic ancestry in Figure 1), and 4% of a genetic ancestry most abundant in people from Chad and Sudan (represented by the Nuba population in Figure 1). In addition, the genetic ancestry of the Tanzanian TB patients was composed of 3% of a genetic ancestry most prevalent in people from Western Africa (Gambia and Senegal represented by the Senegal Bedik population, in Figure 1), as well as 3% of a genetic ancestry most prevalent in individuals from western Africa rainforest hunter-gatherer populations (Bezan population in Figure 1). A mean of 2% belonged to a genetic ancestry most common among Bedouin individuals (represented by the Yemenite Jew population in Figure 1). The proportions of the remaining genetic ancestries were all smaller than 2%

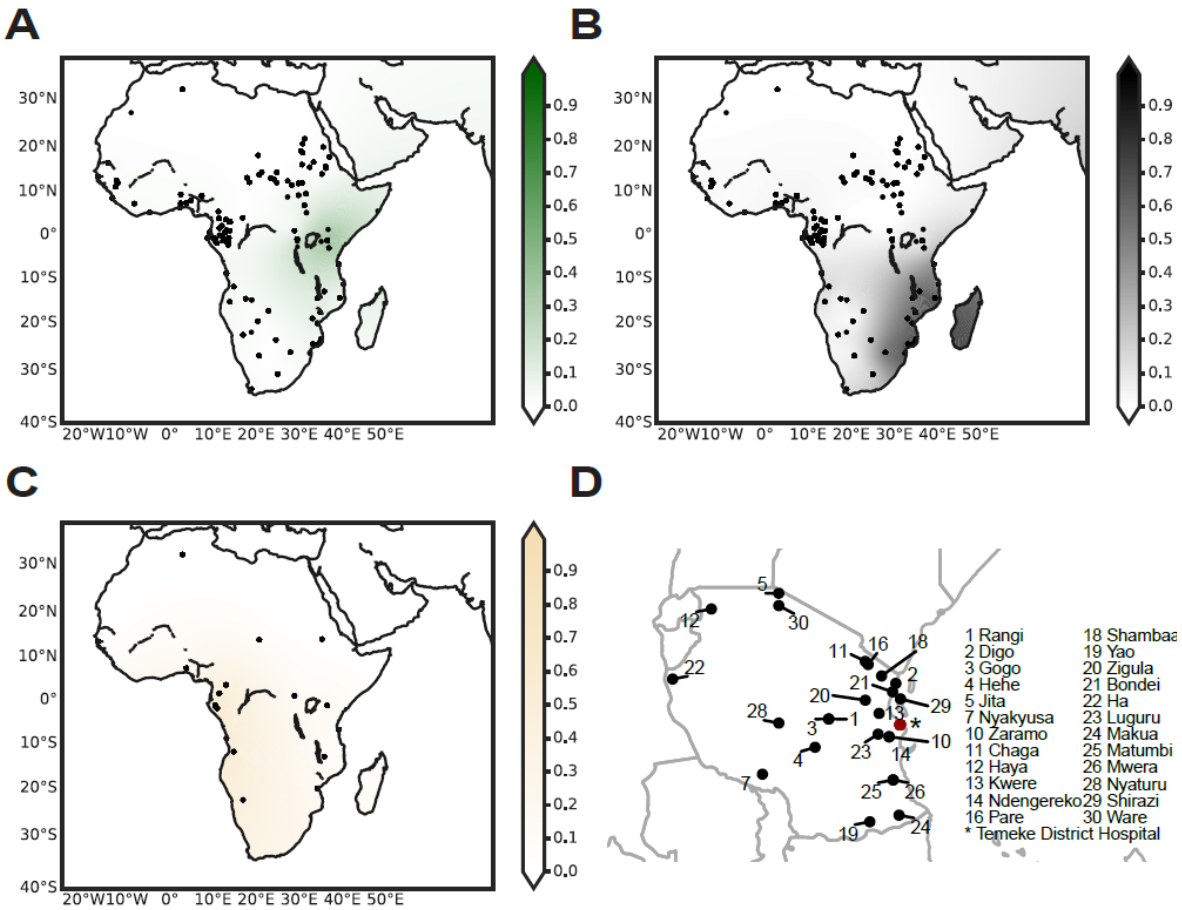
(Supplementary Figure S4, admixture plots for all African populations included can be found in Supplementary Figure S5). Finally, most Tanzanian TB patients had little non-African genetic ancestry (mean 5%, maximum 65%), with only 14 patients (~1%) showing more than 30% non-African genetic ancestry. In summary, the ancestry of Tanzanian TB patients was, for the most part, a mixture of three different Bantu components. Thus, for the remaining sections, we will focus on the ancestries termed Eastern BS (eBS) , Southeastern BS (seBS), and Western BS (wBS) .

Insights into the Bantu-speaking genetic ancestries in Africa and Tanzania

Compiling several datasets, including many different BS populations, allowed for a closer look at the distribution of BS ancestries across African populations. Like Tanzanians, the populations from the neighboring Kenya and Mozambique showed strong contributions of BS ancestries resulting from different admixture events (Figure 1). While the eBS genetic ancestry was highest in Kenya and Tanzania, decreasing from there to the south and to the west of the continent (Figure **2**A), the seBS genetic ancestry generally increased towards the south and decreased towards the west as observed by others (Figure **2**B) (8). The wBS genetic ancestry was mainly seen in BS populations from Gabon and Cameroon, as well as in South African populations (Figure **2**C). At a continental-wide level, the geographical distribution and the genetic distances of the human populations analyzed were significantly correlated (Mantel test: veridical correlation = 0.18, p-value < 0.001, Supplementary Figure S6).

Even though all our TB patients were recruited in Dar es Salaam, we found them to belong to a variety of self-defined ethnic groups linked to different geographical regions within Tanzania (Figure 2D). Even at this smaller scale, we found a significant correlation between the geographic distance of the self-defined ethnic groups of our TB patients and their genetic

distances (Mantel test: veridical correlation = 0.12, p-value < 0.001). The geographical structuring of human genetic diversity within Tanzania was further supported by our finding that the seBS genetic ancestry increased from West to East, and from North to South. The eBS genetic ancestry increased from South to North and decreased from West to East. The wBS genetic ancestry increased to the North and decreased to the East (Supplementary Figure S6A).



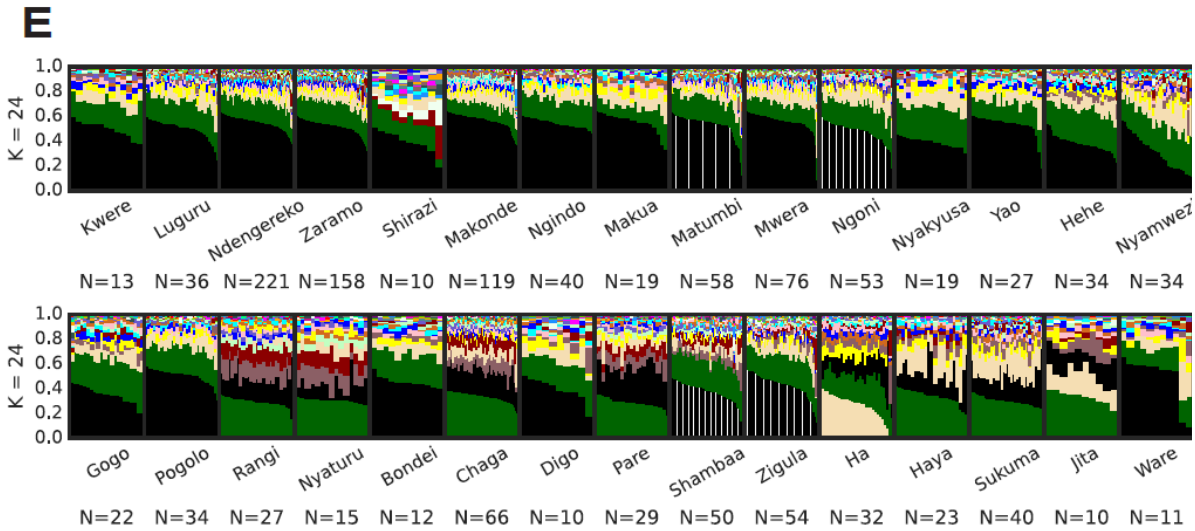


Figure 2 - Spatial visualizations of the BS genetic ancestries and the genetic ancestries of the different self-identified ethnic groups among the TB patients in Tanzania. The genetic ancestry was inferred by admixture with K = 24, and the interpolation of the ancestries was performed by using the pykrige module in Python (see methods). **A**) eBS genetic ancestry, **B**) seBS genetic ancestry, and **C**) wBS genetic ancestry. The populations included for spatial interpolations are marked with a black dot on the maps. The maps were created using the basemap module in Python. **D**) Geographical origin of the ethnic groups among our TB patient cohort. The Temeke Distric hospital in Dar es Salaam where the patients were recruited is marked with a red point. Note that for some ethnic groups, no geographical origin could be identified (Supplementary Table 1). **E**) Ancestry plots for the different ethnic groups with at least 10 patients from our TB patient cohort.

The MTBC genotypes circulating in Tanzania and their association with TB disease severity

In previous work (56), we investigated the MTBC genotypes causing TB in the patient cohort analyzed here. We found a high diversity of MTBC genotypes, with approximately half of the TB cases caused by four main MTBC genotypes estimated to have been introduced into Tanzania starting 320 years ago. After adding the genomic information of an additional 389 MTBC isolates to our dataset (new total N=1,471), the prevalence of the four dominant MTBC genotypes was very similar to our previous findings (50). The most frequent genotype within L3.1.1 (referred to as "Introduction 10") contributed 39% of the current TB cases, followed by a genotype within L1.1.2 ("Introduction 9") with 9%, a genotype within L4.3.4 ("Introduction 5") with 6%, and a genotype within L2.2.1 ("Introduction 1") with 5%. The remaining TB cases were caused by a variety of other genotypes within L1-L4, but occurred at frequencies of 1% or less (56). Despite a 36% increase in sample size compared to our previous analysis (56), we still found no evidence of an association between the different MTBC genotypes circulating in Dar es Salaam and TB disease severity using as proxies; X-ray scores (mild or severe), bacterial load (inferred from GeneXpert cycle threshold) and TB-score (mild or severe) (Table 1, Table 3) (see Methods).

Table 1- Human and bacterial genotypes by the severity measures

		TB-score				Lung damage (X-ray score)					Bacterial load (Ct-value)		
	levels	**Total N (%)**	**Missing N (%)**	**Severe N (%)**	**Mild N (%)**	**Total N**	**Missing N**	**severe**	**mild**	**not available**	**Total N**	**Missing N**	**Mean (SD)**
Total N (%)				624 (33)	1280 (67)			177 (9)	849 (45)	878 (46)			
MTBC Genotype	other	1471 (77.3)	433	207 (42)	406 (41)	764 (74.5)	262	51 (39)	269 (43)	293 (41)	863 (78.3)	239	19.4 (4.9)
	L2.2.1 - Intro 1			22 (4)	49 (5)			7 (5)	28 (4)	36 (5)			20.5 (5.1)
	L3.1.1 - Intro 10			184 (38)	388 (40)			59 (45)	239 (38)	274 (39)			19.1 (4.7)
	L4.3.4 - Intro 5			26 (5)	60 (6)			4 (3)	38 (6)	44 (6)			19.5 (4.1)
	L1.1.2 - Intro 9			51 (10)	78 (8)			11 (8)	58 (9)	60 (8)			19.1 (4.7)
seBS Bantu	Mean (SD)	1442 (75.7)	462	0.44 (0.2)	0.44 (0.2)	840 (81.9)	186	0.45 (0.1)	0.43 (0.2)	0.45 (0.2)	810 (73.5)	292	19.9 (5.2)
eBS	Mean (SD)	1442 (75.7)	462	0.23 (0.1)	0.22 (0.1)	840 (81.9)	186	0.22 (0.1)	0.23 (0.1)	0.22 (0.1)	810 (73.5)	292	19.9 (5.2)
wBS	Mean (SD)	1442 (75.7)	462	0.08 (0.1)	0.09 (0.1)	840 (81.9)	186	0.08 (0.1)	0.09 (0.1)	0.08 (0.1)	810 (73.5)	292	19.9 (5.2)
Other ancestry	Mean (SD)	1442 (75.7)	462	0.25 (0.1)	0.25 (0.1)	840 (81.9)	186	0.25 (0.1)	0.25 (0.1)	0.25 (0.1)	810 (73.5)	292	19.9 (5.2)

5 **Relationship between the human and MTBC population structures**

6 We previously found that the four dominant MTBC genotypes in Dar es Salaam differed in
7 their transmission rate and in the duration of the infectious period (56). Here, we assessed
8 whether there might be an additional host genetic contribution to these differences. We first
9 compared the genetic ancestry proportions between patients infected with the four dominant
10 genotypes, and then tested whether patients who were genetically more closely related were
11 infected with MTBC genotypes that were also more closely related as would be expected from
12 a co-evolutionary process (57). However, the human genetic ancestry proportions differed only
13 marginally between the TB patients infected by the four main MTBC genotypes (Table 2).
14 Moreover, there was no correlation between the human and bacterial genetic distances (Mantel
15 test: veridical correlation = -0.02, p-value = 0.85). Taken together, we found no statistically
16 significant relationship between the human and bacterial genetic population structure in Dar es
17 Salaam. These results also suggest that the genetic composition of this human population is
18 unlikely to have a measurable effect on the differences in bacterial transmission rate and
19 duration of the infectious period reported previously (56).

Table 2 - **Characteristics of MTBC genotypes for all patients with either human or bacterial genetic data available.**

	N (%)[1]	Missing N	levels	Other genotype	L2.2.1 - Intro 1	L3.1.1 - Intro 10	L4.3.4 - Intro 5	L1.1.2 - Intro 9	No bacterial data available
Total N (%)				613 (32)	71 (4)	572 (30)	86 (5)	129 (7)	433 (23)
Sex[2]	1471 (100.0)	0	male (%)	424 (69)	52 (73)	425 (74)	55 (64)	87 (67)	302 (70)
			female (%)	189 (31)	19 (27)	147 (26)	31 (36)	42 (33)	131 (30)
Age in years	1471 (100.0)	0	Median (IQR)	33.0 (28.0-40.0)	31.0 (24.5-38.0)	33.0 (26.0-41.0)	31.5 (27.0-38.8)	35.0 (26.0- 45.0)	35.0 (27.0- 43.0)
HIV Status[2]	1452 (98.7)	19	infected (%)	90 (15)	10 (14)	103 (18)	19 (22)	31 (25)	97 (23)
Smoker[2]	1443 (98.1)	28	yes (%)	127 (21)	18 (26)	149 (27)	11 (13)	26 (20)	97 (23)
Cough duration (weeks)	1454 (98.8)	17	Median (IQR)	4.0 (3.0- 4.0)	4.0 (3.0- 4.0)	4.0 (3.0-4.0)	3.5 (2.2-4.0)	4.0 (2.0-5.0)	3.0 (2.0-4.0)
Socioeconomic status	1452 (98.7)	19	Median (IQR)	80.0 (50.0-125.0)	75.0 (50.0-125.0)	87.5 (50.0-126.2)	85.4 (50.0-122.9)	75.0 (50.0-125.0)	83.3 (50.0-133.3)
Education[2]	1471 (100.0)	0	no education (%)	90 (14.7)	5 (7.0)	76 (13.3)	10 (11.6)	15 (11.6)	76 (17.6)
			primary (%)	391 (63.8)	54 (76.1)	390 (68.2)	61 (70.9)	92 (71.3)	273 (63.0)
			secondary (%)	105 (17.1)	12 (16.9)	90 (15.7)	13 (15.1)	18 (14.0)	64 (14.8)
			university (%)	27 (4.4)	0 (0.0)	16 (2.8)	2 (2.3)	4 (3.1)	20 (4.6)
Malnutrition[2]	1471 (100.0)	0	yes (%)	83 (13.5)	9 (12.7)	71 (12.4)	11 (12.8)	13 (10.1)	56 (12.9)
Relapse/Reinfection[2]	1447 (98.4)	24	yes (%)	9 (1.5)	4 (5.6)	18 (3.2)	2 (2.4)	3 (2.3)	12 (2.8)
Drug resistance status[2]	1471 (100.0)	0	susceptible (%)	574 (93.6)	70 (98.6)	556 (97.2)	80 (93.0)	119 (92.2)	0

			INH-Mono (%)	37 (6.0)	1 (1.4)	16 (2.8)	6 (7.0)	10 (7.8)	0
			MDR (%)		0 (0.0)	0 (0.0)	0 (0.0)	2 (0.3)	0
seBS	1009 (68.6)	462	Median (IQR)	0.5 (0.3-0.6)	0.5 (0.4-0.5)	0.5 (0.4-0.6)	0.5 (0.4-0.6)	0.5 (0.4-0.6)	0.5 (0.3-0.6)
eBS	1009 (68.6)	462	Median (IQR)	0.2 (0.2-0.3)	0.2 (0.2-0.3)	0.2 (0.2-0.3)	0.2 (0.2-0.3)	0.2 (0.2-0.3)	0.2 (0.2-0.3)
wBS	1009 (68.6)	462	Median (IQR)	0.1 (0.0-0.1)	0.1 (0.1-0.1)	0.1 (0.0-0.1)	0.1 (0.0-0.1)	0.1 (0.0-0.1)	0.1 (0.0-0.1)
Other ancestry	1009 (68.6)	462	Median (IQR)	0.2 (0.2-0.3)	0.2 (0.2-0.3)	0.2 (0.2-0.3)	0.2 (0.2-0.2)	0.2 (0.2-0.2)	0.2 (0.2-0.3)

[1] The column N (%) indicates the total number of patients with bacterial genetic data available that contained a value for the respective variable.

[2] The percentage of a MTBC genotype that have the respective characteristic (e.g percentage of males among patients infected with an Intro 1 MTBC genotype).

22

23

24

25 **Association of human genetic ancestry with TB disease severity**

26 In a previous publication on the same cohort, we found that the bacterial genetic background
27 could not explain the differences observed in TB disease severity (56). Since TB disease is
28 shaped by the bacterial and the human genetic background as well as environmental factors
29 (58), we next investigated whether human genetic ancestry could have contributed to the
30 differences in disease severity observed between our TB patients. We assessed the associations
31 between ancestry and the three proxies of TB severity in HIV-negative patients using logistic
32 regression models.. We included the three human genetic ancestries with the highest
33 proportions among the Tanzanian TB patients (seBS, eBS, wBS) as covariates, together with
34 age, sex, smoking, socioeconomic status, TB category (relapse or not), malnutrition, education
35 level, and drug resistance status to control for potential confounding. We found no evidence of
36 an association between human genetic ancestry and any of these three proxies of TB disease
37 severity (Table 1). We repeated the analysis by adding cough duration as covariate to possibly
38 account for the disease duration before recruitment but the results remained unchanged.
39 Additionally, we conducted a genome-wide association study to test for associations between
40 specific human genetic variants and the three proxies of TB disease severity. We included the
41 top three human genetic principal components (PCs), HIV status and the same covariates as
42 for the regression described above. No evidence of an association was found (Supplementary
43 Figure S7).

44 **The combined effect of human and MTBC genetic diversity on TB disease severity**

45 For a subset of 1,000 TB patients, we had both an MTBC genome and a human genome or
46 genotype available. Genetic interactions between the host and the pathogen have been shown
47 to affect TB severity in other settings (41, 59). To test for potential interactions between human

48 and bacterial diversity on TB severity, we added to the ancestry logistic regression models
49 described in the previous section the most common MTBC genotype as an additional
50 explanatory variable (L3.1.1-Introduction 10), as well as the interaction between human
51 ancestry and MTBC genotype. We only tested L3.1.1-Introduction 10, since the numbers of
52 patients with the other genotypes were too few for meaningful testing. However, we found no
53 evidence for any interaction between the MTBC genotype and human ancestry influencing TB
54 disease severity in this patient population (Table 3, Supplementary Figure S8).

Table 3 – **Estimated associations between disease severity, human genetic ancestry and bacterial genotype.** Three variables as proxies for disease severity were included: Lung damage (mild versus severe), TB-score (mild versus severe), Bacterial load (continuous, log10 transformed). Binomial logistic regressions were performed on the data of HIV-negative patients and adjusting was done for age, sex, smoking, socioeconomic status, level of education, malnutrition, TB Type (relapse or new infection), and drug resistance status by including these variables in the model. For the ancestries and the interactions, the p-values were retrieved by performing a likelihood ratio test comparing a model including the ancestries and interactions to a model without them. This table combines the results of two logistic regressions per disease severity measure, one including an interaction and one without. The ancestries were transformed and categorized (see Methods) with category 1 comprising the lowest amount of the respective ancestry and category 3 (4 in the case of wBS) the highest amount.

				Disease severity measure			
		Lung damage		**TB-score**		**Bacterial load (Ct-value)**	
		OR adjusted	p-value adjusted	OR adjusted	p-value adjusted	OR adjusted	p-value adjusted
MTBC genotype[1]	L3.1.1-Introduction 10	1.60 (0.95-2.67)	0.07	1.00 (0.72-1.40)	0.98	0.99 (0.97-1.00)	0.13
	other genotypes	1		1		1.00	
Human ancestry[2]	seBS category 3	2.29 (1.04-5.42)	0.19	1.06 (0.66-1.69)	0.17	1.00 (0.98-1.03)	0.13
	seBS category 2	2.40 (1.15-5.41)		0.62 (0.40-0.94)		1.02 (0.99-1.04)	
	seBS category 1	1		1		1.00	
	eBS category 3	1.30 (0.51-3.79)		1.09 (0.64-1.84)		1.02 (0.99-1.05)	
	eBS category 2	1.50 (0.61-4.28)		1.17 (0.69-1.94)		1.00 (0.97-1.03)	

	eBS category 1	1		1		1.00	
	wBS category 4	0.58 (0.23-1.53)		0.94 (0.49-1.76)		1.02 (0.98-1.06)	
	wBS category 3	0.54 (0.21-1.46)		1.09 (0.57-2.04)		1.03 (0.99-1.07)	
	wBS category 2	0.73 (0.26-2.15)		1.03 (0.50-2.09)		1.03 (0.99-1.08)	
	wBS category 1	1		1		1.00	
Interaction[2]	seBS category 3* L3.1.1-Intro 10	1.05 (0.20-5.43)	0.06 [(1)]	1.05 (0.40-2.71)	0.92	0.99 (0.94-1.04)	0.83
	seBS category 2* L3.1.1-Intro 10	0.39 (0.08-1.94)		1.29 (0.52-3.15)		0.98 (0.93-1.03)	
	seBS category 1* L3.1.1-Intro 10	1		1		1.00	
	eBS category 3* L3.1.1-Intro 10	8.32 (0.91-193.14)		1.45 (0.48-4.34)		1.01 (0.94-1.07)	
	eBS category 2* L3.1.1-Intro 10	6.38 (0.71-146.86)		1.25 (0.42-3.70)		1.02 (0.96-1.09)	
	eBS category 1* L3.1.1-Intro 10	1		1		1.00	

	wBS category 4* L3.1.1-Intro 10	0.17 (0.02-1.23)		1.79 (0.49-6.56)		0.97 (0.90-1.05)	
	wBS category 3* L3.1.1-Intro 10	0.33 (0.04-2.51)		1.48 (0.40-5.51)		0.99 (0.92-1.07)	
	wBS category 2* L3.1.1-Intro 10	0.87 (0.09-8.06)		2.40 (0.53-11.05)		0.98 (0.90-1.07)	
	wBS category 1* L3.1.1-Intro 10	1		1		1.00	

[1] The odds ratio for genotype represents the odds of severe disease for Introduction 10 compared to the odds for other genotypes.

[2] The odds ratios represent the estimated multiple in the odds of severe disease for a one-unit increase in the additive log-transformed ancestry variable.

60 **Discussion**

61 In this study, we analyzed the genetic ancestry of TB patients, the MTBC diversity underlying
62 their TB infection, and estimated the associations of both with disease severity in Dar es
63 Salaam, Tanzania. We found a strong component of BS genetic ancestries among the
64 Tanzanian TB patients, similar to those of neighboring populations from Mozambique and
65 Kenya, and little non-African genetic ancestry. Genetic ancestry proportions did not differ
66 between patients infected with different MTBC genotypes. There was no evidence that the
67 patient genetic ancestry or the MTBC genotype on their own, nor their interaction, had any
68 effect on TB disease severity.

69

70 Despite the fact that Tanzania is one of the few countries in sub-Saharan Africa where all four
71 main African linguistic groups co-exist (60), and that its largest city and economic capital, Dar
72 es Salaam, is strongly influenced by different human populations from within and outside
73 Africa, our cohort of TB patients mostly comprised BS ethnicities. Comparing this TB patient
74 cohort to a large number of modern human populations revealed major components of eBS and
75 seBS genetic ancestries. This genetic population structure probably resulted from several
76 admixture events estimated to have happened between 1500 to 150 years ago, between local
77 populations and BS populations who migrated from West Africa to the East and South of the
78 continent (4). The TB patients investigated here were recruited in one district hospital of Dar
79 es Salaam. Yet, we found the genetic distances between the patients to be correlated with the
80 original geographic range of their self-identified ethnicities, suggesting that the corresponding
81 human populations are not fully admixed. The population of Dar es Salaam has increased by
82 several millions in the last 40 years, mainly as a result of immigration from rural parts of
83 Tanzania (61). Thus, our findings suggest that our TB patient population mostly represents

84 recent migrants to Dar es Salaam from other regions of Tanzania. Moreover, we found little
85 evidence of Eurasian genomic influence in the TB patient population (on average 5% genetic
86 ancestry). This is in strong contrast to coastal Swahili populations, as recent findings comparing
87 modern and medieval Swahili people revealed large components of genetic ancestry derived
88 from exchanges between local East African BS populations and people from India, Persia, and
89 Arabia, starting as early as 1000 AD (62). We conclude that our TB patient population does
90 not represent the full spectrum of human genetic diversity in Tanzania.

91 In contrast to the genetic ancestry of the TB patients, we found that the MTBC genotypes
92 infecting these patients descend from multiple historical introductions, which mainly resulted
93 from the human exchanges that took place across the Indian Ocean during the last few centuries
94 (56). Some of these recently introduced MTBC genotypes became dominant, in particular the
95 MTBC genotype L3.1.1-Introduction 10, which caused TB in approximately 40% of our
96 patients. These strains descended from an introduction that occurred approximately 300 years
97 ago from South or Central Asia to East-Africa (56). Hence, while our TB patient population
98 reflects little historical gene flow from non-African populations, the underlying MTBC
99 diversity indicates that the MTBC genotypes introduced from outside successfully spread in
100 this newly encountered host population, eventually outcompeting native MTBC genotypes
101 (63).

102

103 We previously reported that in this TB patient cohort, some of the dominant MTBC genotypes
104 had a higher transmission rate than others, while some other MTBC genotypes induced patients
105 to remain infectious for longer (56). Based on the similar proportions of MTBC genotypes
106 among self-reported ethnic groups we observed at the time, we had already hypothesized that
107 human genetic heterogeneity of the host population is unlikely to be responsible for those

108 differences (56). Here we formally addressed this hypothesis, and found that there was no
109 evidence that the human genetic ancestry proportions differed between patients infected with
110 different MTBC genotypes in our cohort. This finding is consistent with the notion that the
111 differences in epidemiological parameters we reported previously are mainly determined by
112 the pathogen genotype.

113 Disease severity is one aspect of the clinical presentation of TB with a direct impact on patient
114 mortality and morbidity, as well as on pathogen transmission, as it influences patient
115 infectiousness. It is thus likely that both host and pathogen genetic characteristics can modulate
116 TB disease severity (57). Experimental infections in various animal models suggest that
117 different MTBC strains vary in virulence (64-66). However, *in clinico* evidence for differences
118 in disease severity caused by different MTBC genotypes are inconsistent (67, 68). We found
119 no evidence of differences in disease severity at the point of care caused by the different MTBC
120 genotypes in our study. Moreover, we did not observe any association between human genetic
121 ancestry and disease severity, which is in contrast to a recent study from Peru, where human
122 genetic ancestry was found to influence progression to active TB (41).

123 Our previous work found evidence of such genetic interactions when considering the complete
124 genomes of the TB patients and their infecting MTBC strain (50), and identified associations
125 between human and pathogen variants. Such associations reflect host-pathogen genetic
126 interactions that determine susceptibility to symptomatic TB or intra-host selection during
127 mycobacterial replication. Here, in the context of co-evolutionary history between humans and
128 MTBC, we specifically tested whether an interaction between the main human ancestry
129 components and being infected with the most dominant MTBC strain in Dar es Salaam could
130 explain the variability in TB disease severity. However, we did not find evidence of such an
131 effect. Others have reported an association between TB disease severity, a particular bacterial
132 genotype and a particular human SNP in Uganda, but did not explicitly link this human SNP

133 to a particular human genetic ancestry (59, 69). Several factors could account for the lack of
134 effect we observed. First, our patient population was relatively genetically homogeneous, given
135 that the different BS components represent populations with only moderate levels of genetic
136 differentiation (70). Second, there is likely to be selection bias in our cohort since only patients
137 presenting at the clinic were recruited. The disease severity measures included in this study
138 mainly reflect disease stages at which patients felt ill enough to go to the hospital; i.e. we did
139 not consider intermediate, more contrasting disease states that are known to occur between
140 infection and the development of symptoms (71). To at least partially account for that, we
141 included the number of weeks a patient was coughing as a covariate in our analyses. Third, the
142 lack of a measurable interaction between host genetic ancestry and MTBC genotype could
143 reflect the relatively recent presence of these MTBC genotypes in Tanzania, and the distinct
144 (i.e. allopatric) geographical origins of the host and pathogen populations. This indicates that
145 none of the ancestral human populations that compose modern Tanzanians has lived in
146 sympatry with the ancestors of the modern MTBC genotypes that circulate in Dar es Salaam
147 today.. With the exception of West Africa, where the geographically restricted West-African
148 MTBC lineages L5 and L6 remain an important cause of human TB (72), the situation in
149 Tanzania might be representative of the TB epidemics in many African countries, as evidence
150 suggests that many MTBC genotypes dominating the continent today have been introduced
151 from outside Africa recently in the history of human populations (63, 73-75).

152 In conclusion, our study shows that the TB patients from Dar es Salaam were mainly of BS
153 genetic ancestry reflecting limited Eurasian genetic influx. Neither the human genetic ancestry
154 or the MTBC genotype alone, nor their interaction, were associated with TB disease severity.
155 Our results highlight the dominant role of social and environmental factors in human TB in
156 Tanzania.

157

158 **Methods**

159 **Study population**

160 This study is based on a previously described dataset (50, 56). Briefly, adult active TB patients
161 with pulmonary disease (sputum smear-positive and GeneXpert-positive) were recruited
162 between November 2013 and June 2022 at the Temeke District Hospital in Dar es Salaam,
163 Tanzania when they first presented for care. Sputum and blood samples were collected from
164 each patient to extract DNA for sequencing of the MTBC strain, and genotyping or whole-
165 genome sequencing (WGS) of the patient. Additionally, clinical and sociodemographic
166 information was obtained from every patient. In total, there was either human or bacterial data
167 available for 1,904 patients. Of those 1,444 patients had human genetic data and 1,471 had
168 bacterial genetic data available, respectively. A total of 1,000 patients had both types of data
169 available after quality-based filtering. The geographical locations of the self-indicated ethnic
170 group of each patient were retrieved by searching for the original region of the respective ethnic
171 group, and if they originated from a single region, the geographic coordinates according to
172 Wikipedia were taken. If two neighboring regions were among the origins, then a random
173 location between the two regions was taken as surrogate (Supplementary Table S1).

174 **Bacterial and human sample processing**

175 The MTBC bacteria were cultured on solid Löwenstein-Jensen media at the TB laboratory of
176 the Ifakara Health Institute in Bagamoyo, Tanzania. Before 2018, the MTBC isolates were
177 transferred to Switzerland for DNA extraction and later, DNA extraction was carried out in
178 Bagamoyo. Bacterial WGS was done using the Illumina short-read technology at the
179 Department of Biosystems Science and Engineering of ETH Zurich in Basel (DBSSE). Human
180 WGS was done at the Health 2030 Genome Center in Geneva, Switzerland using an Illumina
181 NovaSeq 6000 sequencer. The human genotyping was done at the iGE3 Genomics platform at

182 the University of Geneva in Switzerland using the Illumina Infinium H3Africa genotyping
183 microarray (Version 2; https://chipinfo.h3abionet.org) plus custom Tanzanian-specific SNP
184 add-ons (76).

185 **Human genetic data**

186 The processing of the human genetic data has been described in detail by Xu et al (50). Briefly,
187 we used the GRCh38 as a human reference genome to map the WGS reads of 118 patients
188 using BWA aligner (v0.7.17) (77). Duplicate reads were then marked with the markduplicates
189 module of Picard (v2.8.14) (78). Variant calling was first done for each sample individually
190 following GATK best practices for germline short variant discovery. Samples with a coverage
191 below 5 were excluded followed by a joint calling of the variants. A Quality Score
192 Recalibration (VQSR) based filter was applied (real sensitivity of 99.7, excess heterozygosity
193 of 54.69) and samples with more than 50% missing genotype calls were removed.

194 For the genotyping data, we used the Illumina GenomeStudio software (v2.0.5,
195 https://support.illumina.com/array/array_software/genomestudio/downloads.html) to analyze
196 the raw microarray data. Samples with a low quality, that were badly clustered, or that had a
197 call rate below 0.97, were excluded. The PLINK format was then converted to VCF format
198 using PLINK (v1.9) (79). The first round of imputation was performed with the African
199 Genome Resources (AFGR, https://www.apcdr.org/) reference panel on the sanger imputation
200 server with EAGLE (80) for phasing and Positional Burrows-Wheeler transform (PBWT) (81)
201 for imputing. The second round was performed with a reference panel created in house that
202 was based on 118 patients with available whole-genome sequences (76) with SHAPEIT4 for
203 phasing and Minimac3 for imputing. For each SNP, the reference panel with the highest
204 imputation quality score was used to determine the final genotype call. SNPs with an INFO
205 score below 0.8 were discarded.

206 Bcftools (v1.15) was used to merge the WGS and genotyping samples after identifying SNPs
207 shared between the two methods that were missing in fewer than 10 samples.

208 **Whole-genome sequence analysis of the MTBC bacteria**

209 We analyzed all FASTQ files using the WGS analysis pipeline described previously (82). In
210 summary, Trimmomatic (83) v. 0.33 (SLIDINGWINDOW:5:20) was used to remove the
211 Illumina adaptors and to trim low quality reads. Only reads with a length of at least 20 bp were
212 kept for further analysis. Overlapping paired-end reads were mergend using SeqPrep v. 1.2
213 (84) (overlap size = 15). We then mapped the resulting reads to a reconstructed ancestral
214 sequence of the MTBC (85) with BWA v. 0.7.13 (mem algorithm) (77). Picard v. 2.9.1 (86)
215 was then applied to mark and exclude duplicated reads. Furthermore, the
216 RealignerTargetCreator and IndelRealigner modules of GATK v. 3.4.0 (87) were used to
217 perform local realignment of reads around INDELs. Reads having an alignment score lower
218 than $(0.93 \times read\ length) - (read\ length\ \times 4\ \times 0.07)$, corresponding to more than 7
219 miss-matches per 100bp, were excluded using Pysam v. 0.9.0 (88). SNP calling was then
220 performed with SAMtools v. 1.2 mpileup (89) and VarScan v. 2.4.1 (90) with the following
221 thresholds: a minimum mapping quality of 20, a minimum base quality at a position of 20,
222 minimum read depth at a position of 7 and no strand bias. Positions in repetitive regions such
223 as PE, PPE, and PGRS genes or phages were excluded, as described in (91). A whole-genome
224 Fasta file was created from the resulting VCF file. We applied some additional filters; genomes
225 were excluded from downstream analysis if they had a sequencing coverage of lower than 15
226 or if they contained SNPs suggestive of different MTBC lineages (i.e. mixed infections). We
227 identified lineages and sublineages using the SNP-based classification by Steiner et al. (92),
228 and Coll et al. (93), respectively. In addition, we identified drug resistance mutations published
229 in the WHO catalogue (94) and determined the respective drug resistance profile. The majority

230 of strains were susceptible (95%), while 70 (5%) were isoniazid mono-resistant and two
231 contained isoniazid and rifampicin resistance and were considered as multi-drug resistant.

232 **Identification of bacterial SNPs diagnostic for the successful MTBC introductions**

233 We previously identified several successful MTBC introductions into Dar es Salaam (49). For
234 these, we aimed to obtain a set of diagnostic SNPs that would allow assigning MTBC strains
235 not included in our previous study to these genotypes. For that, we merged the VCF files from
236 the 1,082 MTBC genomes included in that previous dataset by using BCFtools (v1.9). We then
237 used VCFtools (v0.1.16) to remove Indels and positions that were variable in less than 12
238 genomes (12 was the minimal threshold selected when identifying the successful introductions
239 in our previous publication (56)). By using the R package VariantAnnotation (95) and a
240 customized Python script, SNPs specific to one of the most successful introductions were
241 extracted. To ensure the SNPs identified as markers for the introductions were specific, we also
242 identified phylogenetic SNPs on a bigger and global dataset representing the human-adapted
243 MTBC diversity (82), and tested whether any of the phylogenetic SNPs identified for any of
244 the successful Introductions was present in any of the MTBC lineages or sublineages. We
245 compiled a subset of 25 SNPs (Supplementary Table S2) that we used as phylogenetic markers
246 for the different MTBC introductions and identified strains belonging to one of the four most
247 successful MTBC introductions (Introduction 1, Introduction 5, Introduction 9, and
248 Introduction 10) in the expanded MTBC dataset based on these SNPs by using a customized
249 script.

250 **Measures of TB disease severity**

251 We used three different proxies for TB disease severity. The first one was the TB-score, which
252 is a clinical score adapted from Wejse et al. (96) that consists of several signs and symptoms
253 including the presence of fever and the Body Mass Index (BMI). A point was given for each

254 of the following symptoms or clinical measures if present: BMI below 18, BMI below 16, mid
255 upper arm circumference (MUAC) below 220, MUAC below 200, body temperature higher
256 than 37°C, cough, hemoptysis, dyspnea, chest pain, night sweat, abnormal auscultation,
257 anaemia. A maximum of 12 points could be achieved, and a TB-score below 6 was considered
258 as mild and everything above as severe. As a second proxy, we assessed the amount of lung
259 damage. Two independent radiologists assessed chest X-ray pictures of the patients and gave
260 a Ralph score (97). X-ray scores above 71 were considered as severe and everything below as
261 mild. The Ralph score has been validated to grade chest X-ray severity in adult pulmonary TB
262 patients and 71 was the optimal cut-off point for predicting unfavorable outcome (89).
263 Furthermore Ralph scores higher than 71 have been associated with a longer duration of
264 symptoms, a lower BMI, and higher clinical scores (99). As a third proxy we used the bacterial
265 load in the sputum represented by the difference between the first (early cycle) and the last
266 (late cycle) threshold during quantitative PCR (Ct-value) as determined by GeneXpert
267 MTB/RIF assays. For each sputum sample, we took five probes, run a quantitative PCR each
268 and reported the lowest Ct-value.

269 **Genetic ancestry analysis of TB patients**

270 To investigate the genetic ancestry of the TB patients, we combined our dataset with the data
271 from ten other projects: The Gambian Genome Variation Project (GGVP) (100), the 1000
272 Genomes Project (1000G) (101), the Human Genome Diversity Project (HGDP) (102), Simons
273 Genome Diversity Project (103), as well as data generated by Patin et al. 2014 and 2017 (4,
274 104), Hollfelder et al. 2017 (105), Semo et al. 2019 (8), Schlebusch et al. 2012 (106), and
275 Fortes-Lima et al. 2022 (107). We used the GRCh37 version of all datasets. The dataset of
276 HGDP was in GRCh38 version and we thus did a lift over to GRCh37 using the picard
277 (v2.26.10) (78) tool LiftoverVcf. For all the datasets including only populations from one
278 single continent , we excluded variants with a missingness of more than 10% and only included

279 variants that did not deviate from Hardy-Weinberg equilibrium ($p < 1e-5$) using PLINK
280 (version 1.9b, www.cog-genomics.org/plink/1.9/) (79). For the 1000G, SGDP, and HGDP
281 data, we first identified variants that deviated from the Hardy-Weinberg equilibrium ($p < 1e-$
282 5) in each superpopulation using PLINK (version 2.0a, www.cog-genomics.org/plink/2.0/) and
283 removed them from the whole dataset. We additionally removed variants with a high
284 missingness (> 10%) from the full datasets using PLINK (version 1.9b). After extraction of
285 103,262 nucleotide positions common to all datasets, we merged the datasets using PLINK
286 (version 1.9b). From the merged dataset, we removed second-degree relatives using PLINK
287 (version 2.0a, king cutoff of 0.088) (N=369) and patients from our cohort where the sex
288 according to the genetic data did not correspond with the sex indicated in the clinical data,
289 patients who were genetic outliers based on principal component analysis (PCA) or who did
290 not cluster with any other African samples (N=83). In addition, we removed regions of high
291 linkage disequilibrium
292 (https://genome.sph.umich.edu/wiki/Regions_of_high_linkage_disequilibrium_(LD)) and
293 applied additional filters to the merged dataset (missingness > 10%, minimum allele frequency
294 of 5%, removal of sex chromosomes, variant pruning with –indep-pairwise 50 10 0.1, only
295 biallelic positions) ending up with 53,255 positions and 7,479 individuals from 249
296 populations.

297 To infer the ancestry proportions of the Tanzanian TB patients, we used ADMIXTURE
298 (version 1.3.0) (54). We estimated the number of ancestral populations (K) by running
299 ADMIXTURE 15 times for each value for K from K=2 until K=29 with the option --cv. The -
300 -cv option performs 5-fold cross-validation and allows to identify the value for K resulting in
301 the lowest cross-validation error (Supplementary Figure S3). The cross-validation error was
302 lowest for K = 24. From the 15 runs performed with K = 24, we selected a representative of an
303 output that was supported by most of the 15 runs (6/15) to extract the ancestry proportions of

304 each individual. A PCA was performed using PLINK (version 1.9b) on all African populations
305 included.

306 **Spatial interpolation of human genetic ancestry proportions**

307 To visualize the distributions of the different patient ancestries, we performed spatial
308 interpolation using the OrdinaryKriging function from the pykrige module in Python
309 (variogram_model = "linear", grid space of 500). For the eBS and seBS ancestries, we included
310 all African populations for the interpolation. For the wBS, the interpolation failed when using
311 all African populations with an insufficient slope, suggesting little spatial variability. Since
312 many populations were sampled in the region with the highest proportions of wBS genetic
313 ancestry and among them many hunter-gatherer populations containing little to no wBS genetic
314 ancestry, we repeated the interpolation with only a subset of the non-hunter-gatherer
315 populations.

316 **Correlation between distance matrices**

317 To assess, whether MTBC genotypes that are more closely related tend to infect people that
318 are also more closely related genetically which would be compatible with co-evolution, we
319 investigated the correlation between the human and bacterial distance matrices. To calculate
320 the pairwise bacterial genetic distances, alignments of variable positions where data was
321 missing in less than 10% of the genomes were generated and used to create SNP distance
322 matrices according to the Hamming distance (https://git.scicore.unibas.ch/TBRU/tacos).
323 Insertions and deletions were considered as missing data. To get the human pairwise genetic
324 distances, we calculated the Euclidean distance based on the first two principal components.
325 When only looking at the Tanzanian TB patients, we calculated the principal components for
326 the Tanzanians only while for the continental dataset we included all available African
327 populations.

328 To investigate, whether human populations that are geographically more distant are also
329 genetically more distant, we calculated the correlation between the geographical and the
330 genetic distance on an African level as well as on the level of Tanzania. For the geographical
331 distance matrix, we calculated the Euclidean distance based on the latitude and the longitude.
332 At the level of Tanzania, the broad geographic location of the original area of the ethnic group
333 was considered (Supplementary Table S1) and for the continental level, the coordinates of the
334 hospital in Temeke were taken for the Tanzanian TB patients considering that for the other
335 studies only the sampling locations were known. The human and bacterial genetic distance
336 matrices as well as the human genetic and geographic distance matrices were tested for
337 correlations using the mantel.test() function from the mantel module in Python (options: perms
338 = 10000, method = "pearson", tail = "upper").

339 **Statistical analyses**

340 Clinical and sociodemographic characteristics, MTBC genotype as well as human genetic
341 ancestry were summarized by the different disease severity measures using proportions or
342 means and standard deviations. Similarly, the human genetic ancestries were summarized by
343 MTBC genotype.

344 To estimate the associations between disease severity and the explanatory variables bacterial
345 genotype (binary, belonging to Introduction 10 or not) and human ancestries (seBS , eBS,
346 wBS), we used a logistic regression model. We included three variables as proxies for disease
347 severity: Lung damage based on X-ray score (mild versus severe), TB-score (mild versus
348 severe), bacterial load (continuous, log10 transformed).Categorization or transformation of the
349 disease severity measures were performed because the distributions of the residues were not
350 normal, violating the assumption of linear regression. We tested only for Introduction 10
351 because there were few observations of the other MTBC genotypes. To account for the

352 compositional nature of the human ancestries (i.e. that they sum up to 1), we used the additive
353 log transformation from the R package 'compositions' (108). The ancestry proportions were
354 transformed and categorized with category 1 comprising the lowest amount of the respective
355 ancestry and category 3 (4 in the case of wBS) the highest. The categories were chosen to have
356 roughly equal numbers of patients in each. We used categories to allow a non-linear
357 relationship without specifying polynomials, and avoid having difficulty in interpretation but
358 we recognize that a small amount of information is lost. Similar results were also obtained with
359 other parameterizations. We assessed whether there was an interaction between ancestry and
360 genotype on TB severity using the likelihood ratio test. For that, we compared a model
361 including the interaction between the MTBC genotype and human genetic ancestry to a model
362 without the interaction using the 'lmtest' package (109). The estimates were adjusted for: age,
363 sex, smoking, the number of weeks with cough, the socioeconomic status (ratio between
364 household income and number of household members), education (no, primary, secondary,
365 university), drug resistance profile, TB category (new infection or relapse/reinfection), and
366 malnutrition. Only HIV-negative patients were included in the analysis. All statistical analyses
367 were carried out in R (version 4.1.2). Code for the statistical analysis can be found
368 (https://github.com/mzwyer/TB-Dar_Mtb).

369 To conduct the genome-wide association study (GWAS), we used logistic regression (TB
370 Score, X-ray Score) and a linear model (Ct-value) implemented by PLINK (version 1.9). A
371 total of 6,665,541 common (MAF > 0.05) human genetic variants were included. Genetic
372 principal components were calculated using GCTA (version 1.93.2). As covariates, we
373 included the top three human genetic principal components (PCs) to correct for population
374 stratification, along with sex, HIV status, age, cough duration, smoking, socioeconomic status,
375 TB category, malnutrition, education, and drug resistance profile. Genomic inflation factors
376 were below 1 of all tested outcomes (Supplementary Figure S9).

Acknowledgements

Calculations were performed at the sciCORE (http://scicore.unibas.ch/) scientific computing core facility at University of Basel. We thank all TB-DAR staff and study participants.

Data availability

The bacterial WGS data can be found under the bioprojects PRJEB49562 and PRJNA670836 on ENA and the human WGS and genotyping data under EGAS00001005850 and EGAS00001007216, respectively. Clinical data are available on request.

Funding Statement

Swiss National Science Foundation (Grant No: CRSII5_177163, CRSII5_213514, 320030L-231163 and 320030-227432) and the European Research Council (Grant: 883582). The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Tishkoff SA, Verrelli BC. Patterns of human genetic diversity: Implications for human evolutionary history and disease. Annual Review of Genomics and Human Genetics. 2003;4:293-340.
2. Reed FA, Tishkoff SA. African human diversity, origins and migrations. Current Opinion in Genetics & Development. 2006;16(6):597-605.
3. Garrigan D, Kingan SB, Pilkington MM, Wilder JA, Cox MP, Soodyall H, et al. Inferring human population sizes, divergence times and rates of gene flow from mitochondrial, X and Y chromosome resequencing data. Genetics. 2007;177(4):2195-207.
4. Patin E, Lopez M, Grollemund R, Verdu P, Harmant C, Quach H, et al. Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America. Science. 2017;356(6337):543-6.
5. Bird N, Ormond L, Awah P, Caldwell EF, Connell B, Elamin M, et al. Dense sampling of ethnic groups within African countries reveals fine-scale genetic structure and extensive historical admixture. Science Advances. 2023;9(13).
6. Pfennig A, Petersen LN, Kachambwa P, Lachance J. Evolutionary Genetics and Admixture in African Populations. Genome Biol Evol. 2023;15(4).
7. Wang K, Goldstein S, Bleasdale M, Clist B, Bostoen K, Bakwa-Lufu P, et al. Ancient genomes reveal complex patterns of population movement, interaction, and replacement in sub-Saharan Africa. Science Advances. 2020;6(24).
8. Semo A, Gaya-Vidal M, Fortes-Lima C, Alard B, Oliveira S, Almeida J, et al. Along the Indian Ocean Coast: Genomic Variation in Mozambique Provides New Insights into the Bantu Expansion. Molecular Biology and Evolution. 2020;37(2):406-16.
9. Sengupta D, Choudhury A, Fortes-Lima C, Aron S, Whitelaw G, Bostoen K, et al. Genetic substructure and complex demographic history of South African Bantu speakers. Nature Communications. 2021;12(1).

414 10. Uren C, Moller M, van Helden PD, Henn BM, Hoal EG. Population structure and infectious
415 disease risk in southern Africa. Molecular Genetics and Genomics. 2017;292(3):499-509.
416 11. Fumagalli M, Sironi M, Pozzoli U, Ferrer-Admetlla A, Pattini L, Nielsen R. Signatures of
417 environmental genetic adaptation pinpoint pathogens as the main selective pressure through human
418 evolution. PLoS Genet. 2011;7(11):e1002355.
419 12. Chapman SJ, Hill AV. Human genetic susceptibility to infectious disease. Nat Rev Genet.
420 2012;13(3):175-88.
421 13. Newport MJ, Finan C. Genome-wide association studies and susceptibility to infectious
422 diseases. Brief Funct Genomics. 2011;10(2):98-107.
423 14. Lee JC, Espeli M, Anderson CA, Linterman MA, Pocock JM, Williams NJ, et al. Human SNP
424 links differential outcomes in inflammatory and infectious disease to a FOXO3-regulated pathway.
425 Cell. 2013;155(1):57-69.
426 15. Akcay IM, Katrinli S, Ozdil K, Doganay GD, Doganay L. Host genetic factors affecting hepatitis
427 B infection outcomes: Insights from genome-wide association studies. World J Gastroenterol.
428 2018;24(30):3347-60.
429 16. WHO. Global Tuberculosis Report 2023. 2023.
430 17. Zwyer M, Cavusoglu C, Ghielmetti G, Pacciarini ML, Scaltriti E, Van Soolingen D, et al. A new
431 nomenclature for the livestock-associated Mycobacterium tuberculosis complex based on
432 phylogenomics. Open Res Eur. 2021;1:100.
433 18. Guyeux C, Senelle G, Le Meur A, Supply P, Gaudin C, Phelan JE, et al. Newly Identified
434 Mycobacterium africanum Lineage 10, Central Africa. Emerg Infect Dis. 2024;30(3):560-3.
435 19. Gagneux S. Ecology and evolution of Mycobacterium tuberculosis. Nat Rev Microbiol.
436 2018;16(4):202-13.
437 20. de Jong BC, Antonio M, Gagneux S. Mycobacterium africanum--review of an important cause
438 of human tuberculosis in West Africa. PLoS Negl Trop Dis. 2010;4(9):e744.
439 21. Wiens KE, Woyczynski LP, Ledesma JR, Ross JM, Zenteno-Cuevas R, Goodridge A, et al.
440 Global variation in bacterial strains that cause tuberculosis disease: a systematic review and meta-
441 analysis. Bmc Medicine. 2018;16.
442 22. Reed MB, Pichler VK, McIntosh F, Mattia A, Fallow A, Masala S, et al. Major Mycobacterium
443 tuberculosis lineages associate with patient country of origin. J Clin Microbiol. 2009;47(4):1119-28.
444 23. Gagneux S, DeRiemer K, Van T, Kato-Maeda M, de Jong BC, Narayanan S, et al. Variable host-
445 pathogen compatibility in Mycobacterium tuberculosis. Proc Natl Acad Sci U S A. 2006;103(8):2869-
446 73.
447 24. Baker L, Brown T, Maiden MC, Drobniewski F. Silent nucleotide polymorphisms and a
448 phylogeny for Mycobacterium tuberculosis. Emerg Infect Dis. 2004;10(9):1568-77.
449 25. Groschel MI, Perez-Llanos FJ, Diel R, Vargas R, Jr., Escuyer V, Musser K, et al. Differential
450 rates of Mycobacterium tuberculosis transmission associate with host-pathogen sympatry. Nat
451 Microbiol. 2024;9(8):2113-27.
452 26. Fenner L, Egger M, Bodmer T, Furrer H, Ballif M, Battegay M, et al. HIV infection disrupts the
453 sympatric host-pathogen relationship in human tuberculosis. PLoS Genet. 2013;9(3):e1003318.
454 27. Du DH, Geskus RB, Zhao Y, Codecasa LR, Cirillo DM, van Crevel R, et al. The effect of M.
455 tuberculosis lineage on clinical phenotype. medRxiv. 2023.
456 28. de Jong BC, Hill PC, Aiken A, Awine T, Antonio M, Adetifa IM, et al. Progression to active
457 tuberculosis, but not transmission, varies by Mycobacterium tuberculosis lineage in The Gambia. J
458 Infect Dis. 2008;198(7):1037-43.
459 29. Asare P, Asante-Poku A, Prah DA, Borrell S, Osei-Wusu S, Otchere ID, et al. Reduced
460 transmission of Mycobacterium africanum compared to Mycobacterium tuberculosis in urban West
461 Africa. International Journal of Infectious Diseases. 2018;73:30-42.
462 30. Holt KE, McAdam P, Thai PVK, Thuong NTT, Ha DTM, Lan NN, et al. Frequent transmission of
463 the Mycobacterium tuberculosis Beijing lineage and positive selection for the EsxW Beijing variant in
464 Vietnam. Nat Genet. 2018;50(6):849-56.

465 31. Click ES, Moonan PK, Winston CA, Cowan LS, Oeltmann JE. Relationship Between
466 Mycobacterium tuberculosis Phylogenetic Lineage and Clinical Site of Tuberculosis. Clinical Infectious
467 Diseases. 2012;54(2):211-9.
468 32. Manca C, Reed MB, Freeman S, Mathema B, Kreiswirth B, Barry CE, 3rd, et al. Differential
469 monocyte activation underlies strain-specific Mycobacterium tuberculosis pathogenesis. Infect
470 Immun. 2004;72(9):5511-4.
471 33. Arbues A, Schmidiger S, Reinhard M, Borrell S, Gagneux S, Portevin D. Soluble immune
472 mediators orchestrate protective in vitro granulomatous responses across Mycobacterium
473 tuberculosis complex lineages. Elife. 2025;13.
474 34. Coussens AK, Wilkinson RJ, Nikolayevskyy V, Elkington PT, Hanifa Y, Islam K, et al. Ethnic
475 variation in inflammatory profile in tuberculosis. PLoS Pathog. 2013;9(7):e1003468.
476 35. Boisson-Dupuis S, Ramirez-Alejo N, Li Z, Patin E, Rao G, Kerner G, et al. Tuberculosis and
477 impaired IL-23-dependent IFN-gamma immunity in humans homozygous for a common TYK2
478 missense variant. Sci Immunol. 2018;3(30).
479 36. Stein CM, Sausville L, Wejse C, Sobota RS, Zetola NM, Hill PC, et al. Genomics of human
480 pulmonary tuberculosis: from genes to pathways. Curr Genet Med Rep. 2017;5(4):149-66.
481 37. Phelan J, Gomez-Gonzalez PJ, Andreu N, Omae Y, Toyo-Oka L, Yanai H, et al. Genome-wide
482 host-pathogen analyses reveal genetic interaction points in tuberculosis disease. Nat Commun.
483 2023;14(1):549.
484 38. Bai H, Song M, Lei S, Jiao L, Hu X, Wu T, et al. Genome-wide association study of tuberculosis
485 in the western Chinese Han and Tibetan population. MedComm (2020). 2023;4(2):e250.
486 39. Png E, Alisjahbana B, Sahiratmadja E, Marzuki S, Nelwan R, Balabanova Y, et al. A genome
487 wide association study of pulmonary tuberculosis susceptibility in Indonesians. BMC Med Genet.
488 2012;13:5.
489 40. Zheng R, Li Z, He F, Liu H, Chen J, Chen J, et al. Genome-wide association study identifies two
490 risk loci for tuberculosis in Han Chinese. Nat Commun. 2018;9(1):4072.
491 41. Asgari S, Luo Y, Huang CC, Zhang Z, Calderon R, Jimenez J, et al. Higher native Peruvian
492 genetic ancestry proportion is associated with tuberculosis progression risk. Cell Genom. 2022;2(7).
493 42. Daya M, van der Merwe L, van Helden PD, Moller M, Hoal EG. The role of ancestry in TB
494 susceptibility of an admixed South African population. Tuberculosis. 2014;94(4):413-20.
495 43. Macallan DC. Malnutrition in tuberculosis. Diagnostic Microbiology and Infectious Disease.
496 1999;34(2):153-7.
497 44. Selwyn PA, Hartel D, Lewis VA, Schoenbaum EE, Vermund SH, Klein RS, et al. A Prospective-
498 Study of the Risk of Tuberculosis among Intravenous Drug-Users with Human Immunodeficiency
499 Virus-Infection. New England Journal of Medicine. 1989;320(9):545-50.
500 45. Dooley KE, Chaisson RE. Tuberculosis and diabetes mellitus: convergence of two epidemics.
501 Lancet Infectious Diseases. 2009;9(12):737-46.
502 46. Spence DPS, Hotchkiss J, Williams CSD, Davies PDO. Tuberculosis and Poverty. British
503 Medical Journal. 1993;307(6907):759-61.
504 47. Bates MN, Khalakdina A, Pai M, Chang L, Lessa F, Smith KR. Risk of tuberculosis from
505 exposure to tobacco smoke - A systematic review and meta-analysis. Archives of Internal Medicine.
506 2007;167(4):335-42.
507 48. Imtiaz S, Shield KD, Roerecke M, Samokhvalov AV, Lonnroth K, Rehm J. Alcohol consumption
508 as a risk factor for tuberculosis: meta-analyses and burden of disease. Eur Respir J. 2017;50(1).
509 49. McHenry ML, Wampande EM, Joloba ML, Malone LL, Mayanja-Kizza H, Bush WS, et al.
510 Interaction between M. tuberculosis Lineage and Human Genetic Variants Reveals Novel Pathway
511 Associations with Severity of TB. Pathogens. 2021;10(11).
512 50. Xu ZM, Zwyer M, Brites D, Hiza H, Sasamalo M, Reinhard M, et al. Genome-to-genome
513 analysis reveals associations between human and mycobacterial genetic variation in tuberculosis
514 patients from Tanzania. medRxiv. 2023:2023.05. 11.23289848.

515 51. Ogarkov O, Mokrousov I, Sinkov V, Zhdanova S, Antipina S, Savilov E. 'Lethal' combination of
516 Mycobacterium tuberculosis Beijing genotype and human CD209-336G allele in Russian male
517 population. Infection Genetics and Evolution. 2012;12(4):732-6.
518 52. Leavitt SV, Rodriguez CA, Bouton TC, Horsburgh CR, Abel Zur Wiesch P, Nichols BE, et al.
519 Outcomes for people with TB by disease severity at presentation. Int J Tuberc Lung Dis.
520 2024;28(3):142-7.
521 53. Alling DW, Bosworth EB. The after-history of pulmonary tuberculosis. VI. The first fifteen
522 years following diagnosis. Am Rev Respir Dis. 1960;81:839-49.
523 54. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated
524 individuals. Genome Res. 2009;19(9):1655-64.
525 55. Alexander DH, Lange K. Enhancements to the ADMIXTURE algorithm for individual ancestry
526 estimation. BMC Bioinformatics. 2011;12:246.
527 56. Zwyer M, Rutaihwa LK, Windels E, Hella J, Menardo F, Sasamalo M, et al. Back-to-Africa
528 introductions of Mycobacterium tuberculosis as the main cause of tuberculosis in Dar es Salaam,
529 Tanzania. Plos Pathogens. 2023;19(4).
530 57. Brites D, Gagneux S. Co-evolution of Mycobacterium tuberculosis and Homo sapiens.
531 Immunol Rev. 2015;264(1):6-24.
532 58. Comas I, Gagneux S. The past and future of tuberculosis research. PLoS Pathog.
533 2009;5(10):e1000600.
534 59. McHenry ML, Bartlett J, Igo RP, Wampande EM, Benchek P, Mayanja-Kizza H, et al.
535 Interaction between host genes and Mycobacterium tuberculosis lineage can affect tuberculosis
536 severity: evidence for coevolution? Plos Genetics. 2020;16(4).
537 60. Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, Froment A, et al. The Genetic
538 Structure and History of Africans and African Americans. Science. 2009;324(5930):1035-44.
539 61. LLC M. [Available from: https://www.macrotrends.net/global-metrics/cities/22894/dar-es-
540 salaam/population.
541 62. Brielle ES, Fleisher J, Wynne-Jones S, Sirak K, Broomandkhoshbacht N, Callan K, et al.
542 Entwined African and Asian genetic roots of medieval peoples of the Swahili coast. Nature.
543 2023;615(7954):866-73.
544 63. Comas I, Hailu E, Kiros T, Bekele S, Mekonnen W, Gumi B, et al. Population Genomics of
545 Mycobacterium tuberculosis in Ethiopia Contradicts the Virgin Soil Hypothesis for Human
546 Tuberculosis in Sub-Saharan Africa. Curr Biol. 2015;25(24):3260-6.
547 64. Lopez B, Aguilar D, Orozco H, Burger M, Espitia C, Ritacco V, et al. A marked difference in
548 pathogenesis and immune response induced by different Mycobacterium tuberculosis genotypes.
549 Clin Exp Immunol. 2003;133(1):30-7.
550 65. Aguilar D, Hanekom M, Mata D, Gey van Pittius NC, van Helden PD, Warren RM, et al.
551 Mycobacterium tuberculosis strains with the Beijing genotype demonstrate variability in virulence
552 associated with transmission. Tuberculosis (Edinb). 2010;90(5):319-25.
553 66. Dormans J, Burger M, Aguilar D, Hernandez-Pando R, Kremer K, Roholl P, et al. Correlation of
554 virulence, lung pathology, bacterial load and delayed type hypersensitivity responses after infection
555 with different Mycobacterium tuberculosis genotypes in a BALB/c mouse model. Clin Exp Immunol.
556 2004;137(3):460-8.
557 67. Coscolla M, Gagneux S. Does M. tuberculosis genomic diversity explain disease diversity?
558 Drug Discov Today Dis Mech. 2010;7(1):e43-e59.
559 68. Stanley S, Spaulding CN, Liu Q, Chase MR, Ha DTM, Thai PVK, et al. Identification of bacterial
560 determinants of tuberculosis infection and treatment outcomes: a phenogenomic analysis of clinical
561 strains. Lancet Microbe. 2024.
562 69. McHenry ML, Williams SM, Stein CM. Genetics and evolution of tuberculosis pathogenesis:
563 New perspectives and approaches. Infect Genet Evol. 2020;81:104204.
564 70. Cavalli-Sforza LL, Menozzi P, Piazza A. The history and geography of human genes: Princeton
565 university press; 1994.

566 71. Frascella B, Richards AS, Sossen B, Emery JC, Odone A, Law I, et al. Subclinical Tuberculosis
567 Disease-A Review and Analysis of Prevalence Surveys to Inform Definitions, Burden, Associations,
568 and Screening Methodology. Clin Infect Dis. 2021;73(3):e830-e41.
569 72. Silva ML, Ca B, Osorio NS, Rodrigues PNS, Maceiras AR, Saraiva M. Tuberculosis caused by
570 Mycobacterium africanum: Knowns and unknowns. PLoS Pathog. 2022;18(5):e1010490.
571 73. Rutaihwa LK, Menardo F, Stucki D, Gygli SM, Ley SD, Malla B, et al. Multiple Introductions of
572 Mycobacterium tuberculosis Lineage 2-Beijing Into Africa Over Centuries. Frontiers in Ecology and
573 Evolution. 2019;7.
574 74. Menardo F, Rutaihwa LK, Zwyer M, Borrell S, Comas I, Conceicao EC, et al. Local adaptation
575 in populations of Mycobacterium tuberculosis endemic to the Indian Ocean Rim. F1000Res.
576 2021;10:60.
577 75. Chihota VN, Niehaus A, Streicher EM, Wang X, Sampson SL, Mason P, et al. Geospatial
578 distribution of Mycobacterium tuberculosis genotypes in Africa. PLoS One. 2018;13(8):e0200632.
579 76. Xu ZM, Rueger S, Zwyer M, Brites D, Hiza H, Reinhard M, et al. Using population-specific add-
580 on polymorphisms to improve genotype imputation in underrepresented populations. PLoS Comput
581 Biol. 2022;18(1):e1009628.
582 77. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform.
583 Bioinformatics. 2009;25(14):1754-60.
584 78. Broad Institute. Picard Toolkit. Broad Institute, GitHub Repository. 2019.
585 79. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising
586 to the challenge of larger and richer datasets. Gigascience. 2015;4:7.
587 80. Loh PR, Danecek P, Palamara PF, Fuchsberger C, Y AR, H KF, et al. Reference-based phasing
588 using the Haplotype Reference Consortium panel. Nat Genet. 2016;48(11):1443-8.
589 81. Durbin R. Efficient haplotype matching and storage using the positional Burrows-Wheeler
590 transform (PBWT). Bioinformatics. 2014;30(9):1266-72.
591 82. Menardo F, Loiseau C, Brites D, Coscolla M, Gygli SM, Rutaihwa LK, et al. Treemmer: a tool to
592 reduce large phylogenetic datasets with minimal loss of diversity. BMC Bioinformatics.
593 2018;19(1):164.
594 83. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data.
595 Bioinformatics. 2014;30(15):2114-20.
596 84. SeqPrep. https://github.com/jstjohn/SeqPrep [
597 85. Comas I, Chakravartti J, Small PM, Galagan J, Niemann S, Kremer K, et al. Human T cell
598 epitopes of Mycobacterium tuberculosis are evolutionarily hyperconserved. Nat Genet.
599 2010;42(6):498-503.
600 86. Picard. https://github.com/broadinstitute/picard [
601 87. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome
602 Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.
603 Genome Res. 2010;20(9):1297-303.
604 88. Pysam. https://github.com/pysam-developers/pysam [
605 89. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and
606 population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27(21):2987-
607 93.
608 90. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: somatic
609 mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res.
610 2012;22(3):568-76.
611 91. Stucki D, Brites D, Jeljeli L, Coscolla M, Liu Q, Trauner A, et al. Mycobacterium tuberculosis
612 lineage 4 comprises globally distributed and geographically restricted sublineages. Nat Genet.
613 2016;48(12):1535-43.
614 92. Steiner A, Stucki D, Coscolla M, Borrell S, Gagneux S. KvarQ: targeted and direct variant
615 calling from fastq reads of bacterial genomes. BMC Genomics. 2014;15:881.

616 93. Coll F, McNerney R, Guerra-Assuncao JA, Glynn JR, Perdigao J, Viveiros M, et al. A robust SNP
617 barcode for typing Mycobacterium tuberculosis complex strains. Nat Commun. 2014;5:4812.
618 94. Organization WH. Catalogue of mutations in *Mycobacterium tuberculosis* complex and their
619 association with drug resistance. 2021.
620 95. Obenchain V, Lawrence M, Carey V, Gogarten S, Shannon P, Morgan M. VariantAnnotation:
621 a Bioconductor package for exploration and annotation of genetic variants. Bioinformatics.
622 2014;30(14):2076-8.
623 96. Wejse C, Gustafson P, Nielsen J, Gomes VF, Aaby P, Andersen PL, et al. TBscore: Signs and
624 symptoms from tuberculosis patients in a low-resource setting have predictive value and may be
625 used to assess clinical course. Scand J Infect Dis. 2008;40(2):111-20.
626 97. Ralph AP, Ardian M, Wiguna A, Maguire GP, Becker NG, Drogumuller G, et al. A simple, valid,
627 numerical score for grading chest x-ray severity in adult smear-positive pulmonary tuberculosis.
628 Thorax. 2010;65(10):863-9.
629 98. Michel AL, Bengis RG, Keet DF, Hofmeyr M, Klerk LM, Cross PC, et al. Wildlife tuberculosis in
630 South African conservation areas: implications and challenges. Vet Microbiol. 2006;112(2-4):91-100.
631 99. Chakraborthy A, Shivananjaiah AJ, Ramaswamy S, Chikkavenkatappa N. Chest X ray score
632 (Timika score): an useful adjunct to predict treatment outcome in tuberculosis. Adv Respir Med.
633 2018;86(5):205-10.
634 100. Band G, Le QS, Clarke GM, Kivinen K, Hubbart C, Jeffreys AE, et al. Insights into malaria
635 susceptibility using genome-wide data on 17,000 individuals from Africa, Asia and Oceania. Nature
636 Communications. 2019;10.
637 101. Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, et al. A global
638 reference for human genetic variation. Nature. 2015;526(7571):68-+.
639 102. Bergstrom A, McCarthy SA, Hui RY, Almarri MA, Ayub Q, Danecek P, et al. Insights into
640 human genetic variation and population history from 929 diverse genomes. Science.
641 2020;367(6484):1339-+.
642 103. Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, et al. The Simons Genome
643 Diversity Project: 300 genomes from 142 diverse populations. Nature. 2016;538(7624):201-+.
644 104. Patin E, Siddle KJ, Laval G, Quach H, Harmant C, Becker N, et al. The impact of agricultural
645 emergence on the genetic history of African rainforest hunter-gatherers and agriculturalists. Nat
646 Commun. 2014;5:3163.
647 105. Hollfelder N, Schlebusch CM, Gunther T, Babiker H, Hassan HY, Jakobsson M. Northeast
648 African genomic variation shaped by the continuity of indigenous groups and Eurasian migrations.
649 Plos Genetics. 2017;13(8).
650 106. Schlebusch CM, Skoglund P, Sjodin P, Gattepaille LM, Hernandez D, Jay F, et al. Genomic
651 variation in seven Khoe-San groups reveals adaptation and complex African history. Science.
652 2012;338(6105):374-9.
653 107. Fortes-Lima C, Triska P, Cizkova M, Podgorna E, Diallo MY, Schlebusch CM, et al.
654 Demographic and Selection Histories of Populations Across the Sahel/Savannah Belt. Mol Biol Evol.
655 2022;39(10).
656 108. van den Boogaart KG, Tolosana-Delgado R, Bren M. compositions: Compositional Data
657 Analysis 2022 [Available from: https://CRAN.R-project.org/package=compositions.
658 109. A. Z, T. H. Diagnostic Checking in Regression Relationships. R News. 2002;2(3):7-10.

659

660

661 **Supporting information captions**

662 Supplementary Figure S1 - The first two PCA components for all African populations
663 included in this study (n=116).

664 Supplementary Figure S2 - Populations included in the Admixture analysis. Populations are
665 coloured according to the study from which the data was obtained (4, 8, 44, 91-93, 95-98).

666 Supplementary Figure S3 - Boxplots of cross-validation errors for values of K between 2 and
667 29 resulting from 15 runs. The lowest cross-validation error was obtained for K = 24.

668 Supplementary Figure S4 - Boxplots of the proportions of the 24 genetic ancestries among the
669 Tanzanian TB patients. The ancestries were named after the population(s) where they were
670 most abundant.

671 Supplementary Figure S5 - Ancestry plots for all African populations included (K = 24).

672 Supplementary Figure S6 - Heatmap showing the correlations between the genetic ancestries
673 and geographical location. A) Correlations at the country level of Tanzania between the latitude
674 and longitude of the ethnic group the patients belong to. B) Correlations on a continental level
675 including all African populations. For the latitude, a negative correlation indicates higher
676 genetic ancestry proportions towards the South and for the longitude a negative correlation
677 indicates lower genetic ancestry proportions towards the East.

678 Supplementary Figure S7 - Manhattan plot for GWAS conducted using A, B, C) TB-score, X-
679 ray score, and Ct-value. Red line indicates GWAS significance threshold of 5e-8.

680 Supplementary Figure S8 – Association between the Bantu genetic ancestries and TB-score
681 and X-ray score for each of the most successful introductions. A, C, E) Southeastern, Eastern,
682 and Western Bantu on mild or severe TB-score, B, D, E) Southeastern, Eastern, and Western
683 Bantu genetic ancestry lung damage assessed by lung X-ray.

684 Supplementary Figure S9 - QQ Plot and genomic inflation factor for GWAS conducted using
685 A, B, C) TB-score, X-ray score, and Ct-value.

686 Supplementary Table S1 - The different ethnic groups with at least 10 members in our cohort
687 and the region and broad geographic location of the original area of the ethnic group. The
688 latitude and longitude are given in decimal degrees. In the case of two associated regions, a
689 location close to the border of the two regions was selected.

690 Supplementary Table S2 - Phylogenetic markers selected to identify the Introductions. The
691 position is based on the reconstructed reference of the ancestor (79) and the derived base
692 indicates the base present in the respective Introduction. Intro1 refers to Introduction 1 within
693 L2.2.1, Intro5 to Introduction 5 within L4.3.4, Intro9 to Introduction 9 within L1.1.2, and
694 Intro10 to Introduction 10 within L3.1.1.

695

696