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], 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; 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 their 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 the Bantu migrations 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. 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 (22-24). Moreover, in immune-compromised individuals with HIV co-infection, these sympatric associations were disrupted (25). The various MTBC genotypes also differ phenotypically (26), regarding disease progression (27), transmission (28, 29), and disease presentation (30).

Human genetic diversity has also been linked to differences in TB susceptibility. While for example *TYK2* has been associated with TB disease worldwide (31), several human genetic loci were not consistently associated with TB in populations from different geographical regions but specific to certain populations (32-36). 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 (37), and a higher proportion of San genetic ancestry was associated with an increased risk for TB among South African Coloured individuals (38). 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 (39), HIV infection, diabetes (40, 41), poverty (42), smoking (43), and alcohol consumption (44).While associations between TB and individual host, pathogen or environmental factors have been found (32-37, 39), studies considering all these components simultaneously, remain scarce (45-47).

Here, we characterized 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.

117

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 (48) with 53,255 SNPs (Figure 1, Supplementary Figure S1).

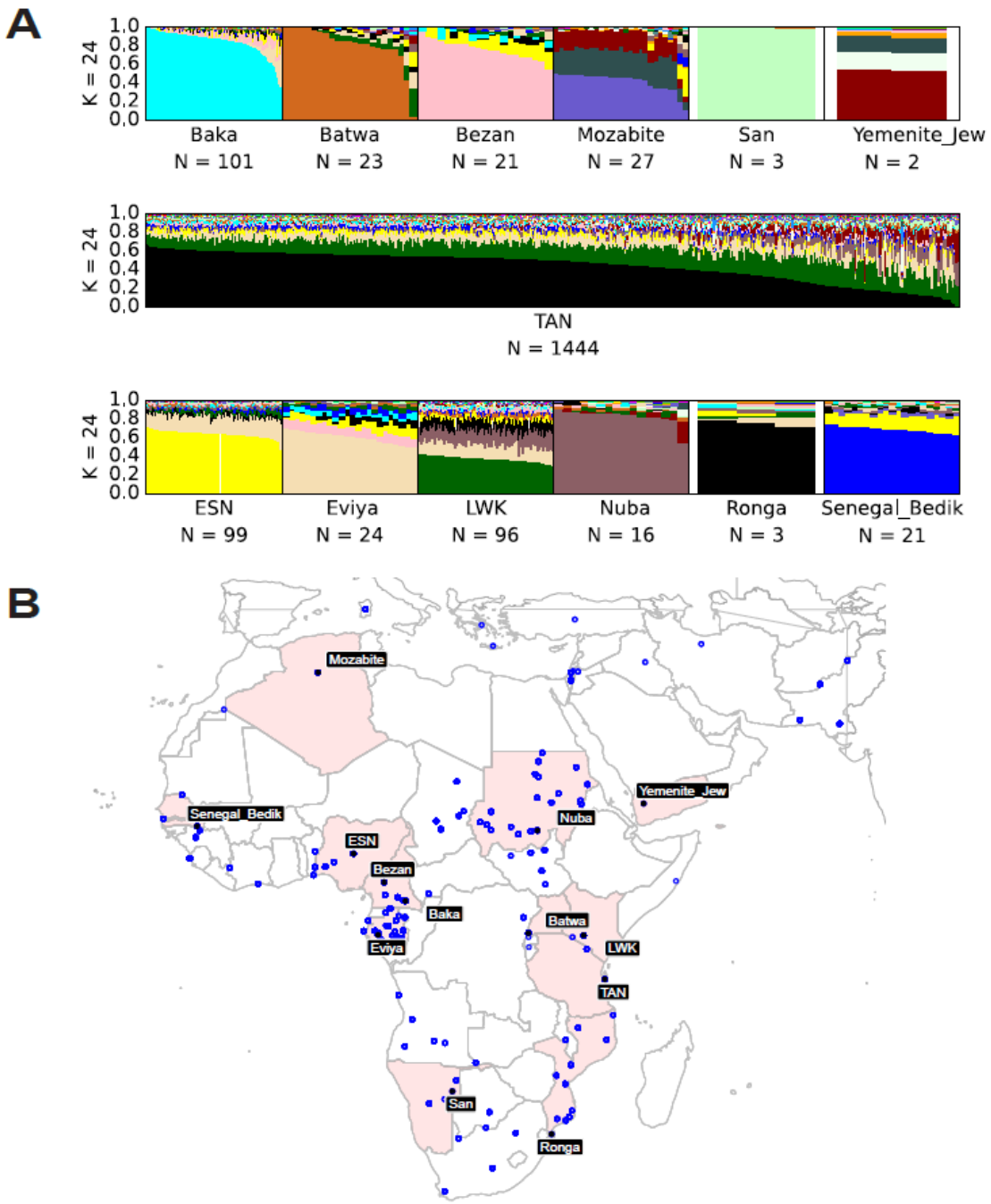


Figure 1 – Genetic ancestry analyses of Tanzanian TB patients.
A) Genetic ancestry proportions of the 1,444 Tanzanian TB patients and representative human populations who shared at least 1% of their most common genetic ancestry with the Tanzanians for K = 24. (ESN: Esan from Nigeria (1000G), LWK: Luhya from Kenya (1000G)). For all populations included in our study, see Supplementary Figure S2 for their geographic distribution and 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 (49) (**Error! Reference source not found.**). 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 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 Bantu" (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 Bantu" 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 Bantu" 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% (**Error! Reference source not found.**, admixture plots for all African populations included can be found in **Error! Reference source not found.**). 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 Bantu, Southeastern Bantu, and Western Bantu.

Insights into the Bantu genetic ancestries in Africa and Tanzania

Compiling several datasets, including many different Bantu populations, allowed for a closer look at the distribution of Bantu ancestries across African populations. Like Tanzanians, the populations from the neighboring Kenya and Mozambique showed strong contributions of Bantu ancestries resulting from different admixture events (Figure 1). While the Eastern Bantu genetic ancestry was highest in Kenya and Tanzania, decreasing from there to the south and to the west of the continent (

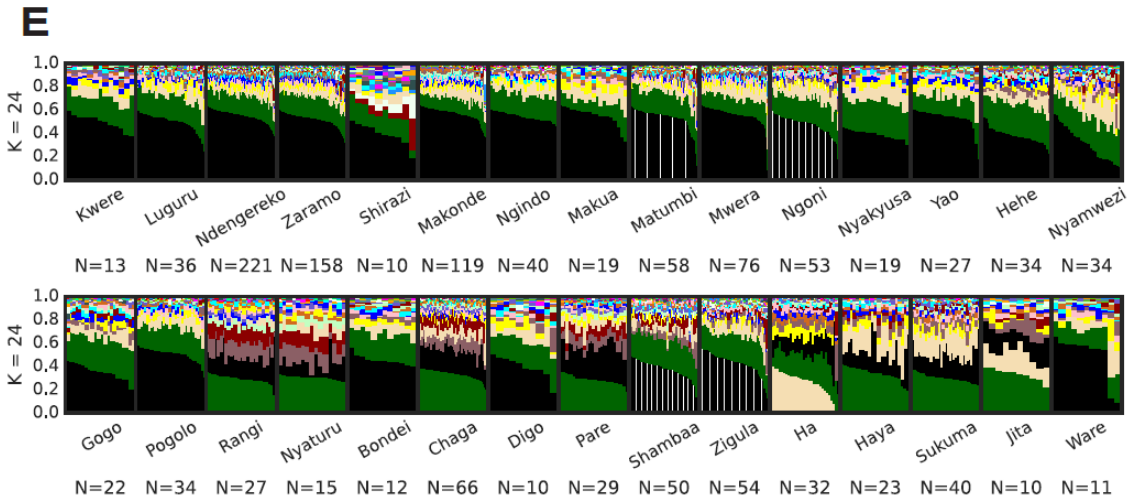


Figure 2A), the Southeastern Bantu genetic ancestry generally increased towards the south

and decreased towards the west as observed by others (

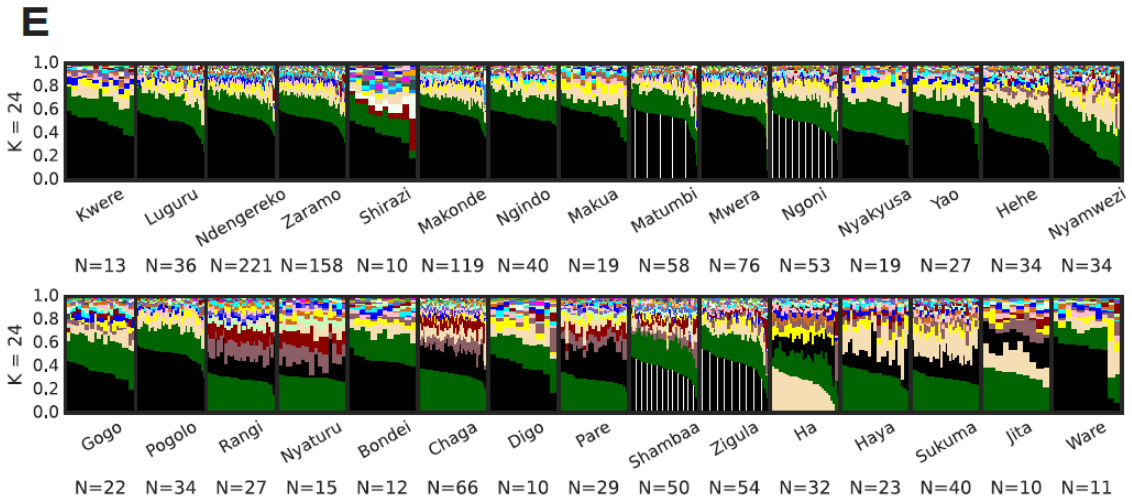


Figure 2B) (8). The Western Bantu genetic ancestry was mainly seen in Bantu-speaking populations from Gabon and Cameroon, as well as in South African populations (

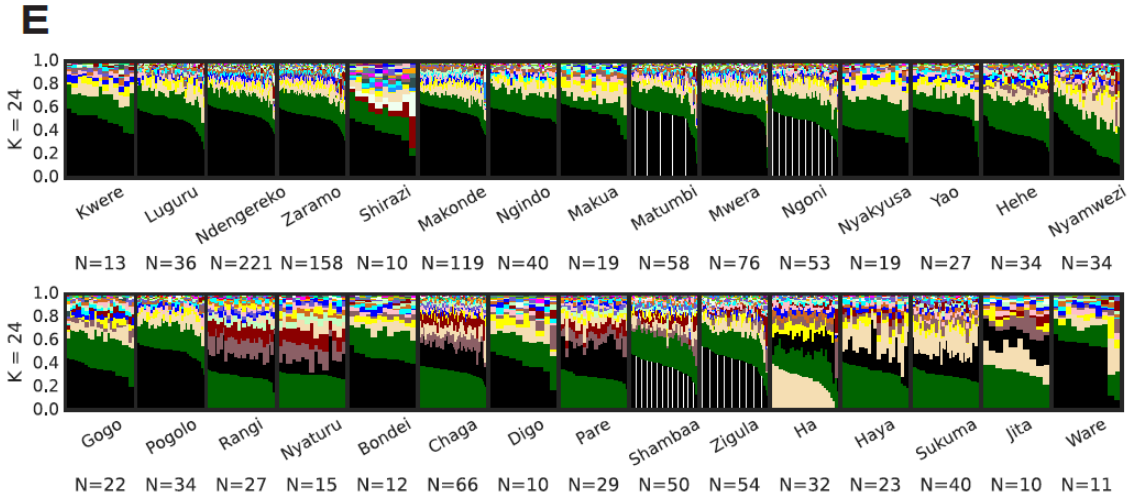
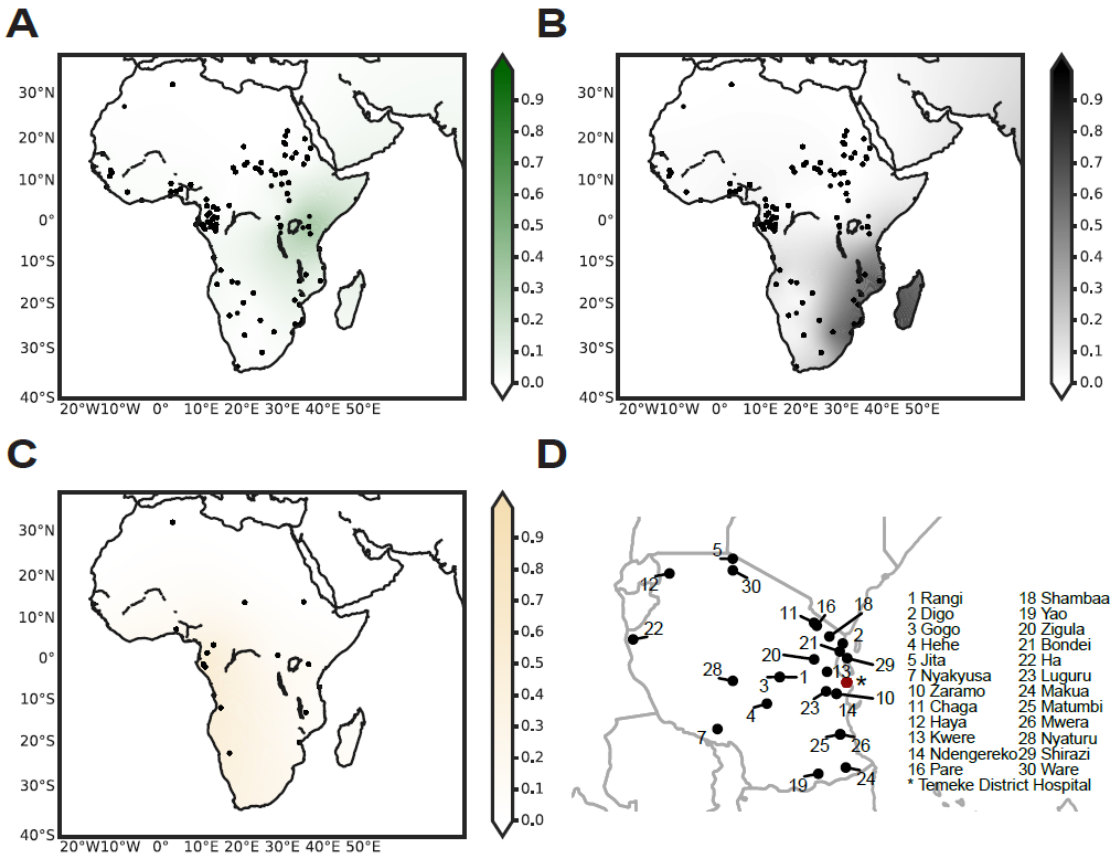


Figure 2C). At a continental-wide level, the geographical distribution and the genetic distances of the human populations analyzed were significantly correlated (Mantel test: veridical correlation = 0.18, p-value < 0.001, Supplementary Figure S6).

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 Southeastern Bantu genetic ancestry increased from West to East, and from North to South. The Eastern Bantu genetic ancestry increased from South to North and decreased from West to East. The Western Bantu genetic ancestry increased to the North and decreased to the East (**Error! Reference source not found.**A).



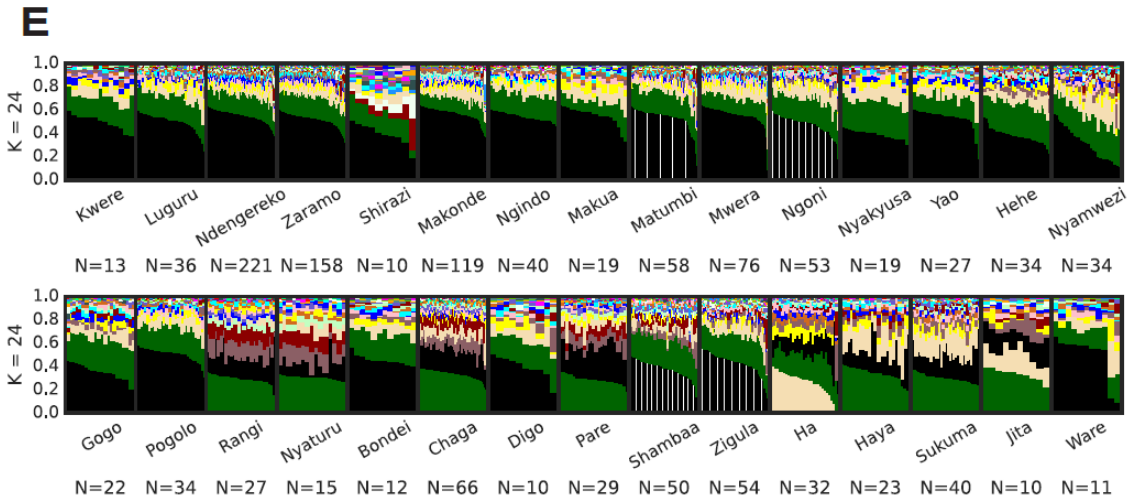


Figure 2 - Spatial visualizations of the Bantu genetic ancestries and the genetic ancestries of the different self-identified ethnic groups among the TB patients in Tanzania. The genetic ancestry was inferred by admixture with K = 24, and the interpolation of the ancestries was performed by using the pykrige module in Python (see methods). **A**) Eastern Bantu genetic ancestry, **B**) Southeastern Bantu genetic ancestry, and **C**) Western Bantu 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 (50), 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 (total N=1,471), the prevalence of the four dominant MTBC genotypes was very similar to our previous findings (Ref). The most successful 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 (50). Despite a 36% increase in sample size compared to our previous analysis (50), we still find 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)
Southeastern 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)
Eastern Bantu	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)
Western Bantu	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)

Relationship between the human and MTBC population structures

We previously found that the four dominant MTBC genotypes in Dar es Salaam differed in their transmission rate and in the duration of the infectious period (50). Here, we assessed whether there might be an additional host genetic contribution to these differences. We first compared the genetic ancestry proportions between patients infected with the four dominant genotypes, and then tested whether patients who were genetically more closely related were infected with MTBC genotypes that were also more closely related as would be expected from a co-evolutionary process (51). However, the human genetic ancestry proportions differed only marginally between the TB patients infected by the four main MTBC genotypes (Table 2). Moreover, there was no correlation between the human and bacterial genetic distances (Mantel test: veridical correlation = -0.02, p-value = 0.85). Taken together, we found no evidence for any statistically significant relationship between the human and bacterial genetic population structure in Dar es Salaam. These results also suggest that the genetic composition of this human population is unlikely to have a measurable effect on the differences in bacterial transmission rate and duration of the infectious period reported previously (50).

Table 2- Summary of ancestry proportions, sociodemographic and clinical data stratified by MTBC genotype.

	Total N (%)	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	Missing
Total N (%)				613 (32)	71 (4)	572 (30)	86 (5)	129 (7)	433 (23)
Sex	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	1471 (100.0)	0	Mean (SD)	34.64 (10.4)	32.03 (8.4)	34.59 (10.5)	33.85 (11.0)	36.01 (12.4)	35.94 (11.4)
HIV Status	1452 (98.7)	19	negative	515 (85)	60 (86)	462 (82)	67 (78)	95 (75)	330 (77)
			infected	90 (15)	10 (14)	103 (18)	19 (22)	31 (25)	97 (23)
Smoker	1443 (98.1)	28	no	477 (79)	52 (74)	408 (73)	73 (87)	102 (80)	334 (77)
			yes	127 (21)	18 (26)	149 (27)	11 (13)	26 (20)	97 (23)
Cough duration (weeks)	1454 (98.8)	17	Mean (SD)	3.69 (1.9)	3.74 (2.0)	3.76 (1.8)	3.77 (2.0)	3.68 (1.7)	3.61 (1.9)
Southeastern Bantu	1009 (68.6)	462	Mean (SD)	0.43 (0.2)	0.45 (0.1)	0.45 (0.2)	0.46 (0.2)	0.46 (0.1)	0.44 (0.2)
Eastern Bantu	1009 (68.6)	462	Mean (SD)	0.23 (0.1)	0.22 (0.1)	0.22 (0.1)	0.23 (0.1)	0.22 (0.1)	0.22 (0.1)
Western Bantu	1009 (68.6)	462	Mean (SD)	0.09 (0.1)	0.09 (0.1)	0.08 (0.1)	0.08 (0.1)	0.08 (0.1)	0.09 (0.1)
Other ancestry	1009 (68.6)	462	Mean (SD)	0.26 (0.1)	0.24 (0.1)	0.24 (0.1)	0.23 (0.1)	0.23 (0.1)	0.25 (0.1)

Association of human genetic ancestry with TB disease severity

We next investigated whether human genetic ancestry could have contributed to the differences in disease severity observed between our TB patients. We fitted logistic regression models using the three proxies of TB disease severity mentioned previously as outcome variables. We included the three human genetic ancestries with the highest proportions among the Tanzanian TB patients (Southeastern Bantu, Eastern Bantu, Western Bantu) as covariates, together with age, sex, HIV status, age, smoking and cough duration to control for potential confounding. We found no evidence of an association between human genetic ancestry and any of these three proxies of TB disease severity (Table 1).

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

For a subset of 1,000 TB patients, we had both an MTBC genome and a human genome or genotype available. Genetic interactions between the host and the pathogen have been show to affect TB severity in other settings (37, 52). To test for potential interactions between human and bacterial diversity on TB severity, we added to the logistic regression models described in the previous section the most common MTBC genotype as an additional explanatory variable (L3.1.1-Introduction 10), as well as the interaction between human ancestry and MTBC genotype. We only tested L3.1.1-Introduction 10, since the numbers of the other genotypes were too few for meaningful testing. However, there we found no evidence for any interaction between the MTBC genotype and human ancestry influencing TB disease severity in this patient population (Table 3).

Table 3 - Associations between the different proxies for disease severity, human genetic ancestry, bacterial genotype and the interaction between human genetic ancestry and bacterial genotype. Binomial logistic regressions were performed on the data of HIV-negative patients, adjusting for age, sex, smoking and cough duration.

		Disease severity measures					
		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	1.59 (0.95-2.66)	0.08	1.05 (0.77-1.44)	0.76	0.99 (0.97-1.01)	0.24
	other genotypes	1		1		1.00	
Human ancestry[2]	Southeastern Bantu category 3	2.26 (1.06-5.21)	0.17	1.09 (0.71-1.68)	0.29	1.00 (0.97-1.02)	0.22
	Southeastern Bantu category 2	2.27 (1.10-5.10)		0.69 (0.46-1.02)		1.01 (0.99-1.05)	
	Southeastern Bantu category 1	1		1		1.00	
	Eastern Bantu category 3	1.36 (0.55-3.90)		0.93 (0.56-1.52)		1.02 (0.99-1.05)	
	Eastern Bantu category 2	1.55 (0.64-4.37)		1.09 (0.67-1.76)		1.00 (0.97-1.03)	
	Eastern Bantu category 1	1		1		1.00	
	Western Bantu category 4	0.54 (0.22-1.42)		1.16 (0.64-2.06)		1.01 (0.98-1.05)	
	Western Bantu category 3	0.51 (0.20-1.39)		1.30 (0.72-2.32)		1.03 (0.99-1.06)	

	Western Bantu category 2	0.70 (0.25-2.05)		1.13 (0.58-2.17)		1.03 (0.99-1.06)	
	Western Bantu category 1	1		1		1.00	
Interaction between MTBC genotype and human ancestry	Southeastern Bantu category 3 & L3.1.1-Intro 10	1.10 (0.21-5.65)	0.06	1.03 (0.42-2.53)	0.90	0.99 (0.94-1.04)	0.91
	Southeastern Bantu category 2 & L3.1.1-Intro 10	0.39 (0.08-1.91)		1.28 (0.55-2.98)		0.98 (0.93-1.03)	
	Southeastern Bantu category 1 & L3.1.1-Intro 10	1		1		1.00	
	Eastern Bantu category 3 & L3.1.1-Intro 10	7.79 (0.87-177.55)		1.69 (0.59-4.76)		1.01 (0.94-1.07)	
	Eastern Bantu category 2 & L3.1.1-Intro 10	6.55 (0.74-147.98)		1.38 (0.50-3.83)		1.02 (0.96-1.09)	
	Eastern Bantu category 1 & L3.1.1-Intro 10	1		1		1.00	
	Western Bantu category 4 & L3.1.1-Intro 10	0.21 (0.03-1.46)		1.44 (0.44-4.75)		0.97 (0.90-1.05)	
	Western Bantu category 3 & L3.1.1-Intro 10	0.45 (0.05-3.45)		1.06 (0.32-3.55)		0.99 (0.92-1.07)	
	Western Bantu category 2 & L3.1.1-Intro 10	1.19 (0.12-11.04)		1.62 (0.41-6.54)		0.98 (0.91-1.07)	
	Western Bantu 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.

Discussion

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

Despite the fact that Tanzania is one of the few countries in sub-Saharan Africa where all four main African linguistic groups co-exist (53), and that its largest city and economic capital, Dar es Salaam, is strongly influenced by different human populations from within and outside Africa, our cohort of TB patients mostly comprised Bantu-speaking ethnicities. Comparing this TB patient cohort to a large number of modern human populations revealed major components of Eastern and Southeastern Bantu genetic ancestries. This genetic population structure probably resulted from several admixture events estimated to have happened between 1500 to 150 years ago, between local populations and Bantu-speaking populations who migrated from West Africa to the East and South of the continent (4). The TB patients investigated here were recruited in one district hospital of Dar es Salaam. Yet, we found the genetic distances between the patients to be correlated with the original geographic range of their self-identified ethnicities, suggesting that the corresponding human populations are not fully admixed. The population of Dar es Salaam has increased by several millions in the last 40 years, mainly as a result of immigration from rural parts of Tanzania (54). Thus,

our findings suggest that our TB patient population mostly represents recent migrants to Dar es Salaam from other regions of Tanzania. Moreover, we found little evidence of Eurasian genomic influence in the TB patient population (on average 5% genetic ancestry). This is in strong contrast to coastal Swahili populations, as recent findings comparing modern and medieval Swahili people revealed large components of genetic ancestry derived from exchanges between local East African Bantu populations and people from India, Persia, and Arabia, starting as early as 1000 AD (55). We conclude that our TB patient population does not represent the full spectrum of human genetic diversity in Tanzania.

In contrast to the genetic ancestry of the TB patients, we found that the MTBC genotypes infecting these patients descend from multiple historical introductions, which mainly resulted from the human exchanges that took place across the Indian Ocean during the last few centuries (50). Some of these recently introduced MTBC genotypes became dominant, in particular the MTBC genotype L3.1.1-Introduction 10, which caused TB in approximately 40% of our patients. These strains descended from an introduction that occurred approximately 300 years ago from South or Central Asia to East-Africa (50). Hence, while our TB patient population reflects little historical gene flow from non-African populations, the underlying MTBC diversity indicates that the MTBC genotypes introduced from outside successfully spread in this newly encountered host population, eventually outcompeting native MTBC genotypes (56).

We previously reported that in this TB patient cohort, some of the dominant MTBC genotypes had a higher transmission rate than others, while some other MTBC genotypes induced patients to remain infectious for longer (50). Based on the similar proportions of MTBC genotypes among self-reported ethnic groups we observed at the time, we had already

hypothesized that human genetic heterogeneity of the host population is unlikely to be responsible for those differences (50). Here we formally addressed this hypothesis, and found that there was no evidence that the human genetic ancestry proportions differed between patients infected with different MTBC genotypes in our cohort. This finding further supports the notion that the differences in epidemiological parameters we reported previously are mainly determined by the pathogen genotype.

Disease severity is one aspect of the clinical presentation of TB with a direct impact on patient mortality and morbidity, as well as on pathogen transmission, as it influences patient infectiousness. It is thus likely that both host and pathogen genetic characteristics can modulate TB disease severity (51). Experimental infections in various animal models suggest that different MTBC strains vary in virulence (57-59). However, *in clinico* evidence for differences in disease severity caused by different MTBC genotypes are inconsistent (60, 61). We found no evidence of differences in disease severity at the point of care caused by the different MTBC genotypes in our study. Moreover, we did not observe any association between human genetic ancestry and disease severity, which is in contrast to a recent study from Peru, where human genetic ancestry was found to influence progression to active TB (37).

Our previous work found evidence of such genetic interactions when considering the complete genomes of the TB patients and their infecting MTBC strain (46) has identified associations between human and pathogen variants. Such associations reflect host-pathogen genetic interactions that determine susceptibility to symptomatic TB or intra-host selection during mycobacterial replication. Here, in the context of co-evolutionary history between humans and MTBC, we specifically tested whether an interaction between the main human

ancestry components and being infected with the most dominant MTBC strain in Dar es Salaam could explain the variability in TB disease severity. However, we did not find any evidence of such an effect. Others have reported an association between TB disease severity, a particular bacterial genotype and a particular human SNP in Uganda, but did not explicitly link this human SNP to particular human genetic ancestry (52, 62). Several factors could account for the lack of effect we observed. First, our patient population was relatively genetically homogeneous, given that the different Bantu components represent populations with only moderate levels of genetic differentiation (63). Second, there is likely to be selection bias in our cohort since only patients presenting at the clinic were recruited. The disease severity measures included in this study mainly reflect disease stages, at which patients felt ill enough to go to the hospital; i.e. we did not consider intermediate, more contrasting disease states that are known to occur between infection and the development of symptoms (64). To at least partially account for that, we included the number of weeks a patient was coughing as a covariate in our analyses. Third, the lack of a measurable interaction between host genetic ancestry and MTBC genotype could reflect the relatively recent presence of these MTBC genotypes in Tanzania, and the distinct (i.e. allopatric) geographical origins of the host and pathogen populations. This indicates that none of the ancestral human populations that compose the modern Tanzanians has lived in sympatry with the ancestors of the modern MTBC genotypes that circulate in Dar es Salaam today, thus rendering human susceptibility alleles linked to genetic ancestry unlikely. With the exception of West Africa, where the geographically restricted West-African MTBC lineages L5 and L6 remain an important cause of human TB (65), the situation in Tanzania might be representative of the TB epidemics in many African countries, as evidence suggests that many MTBC genotypes dominating the continent today have been introduced from outside Africa during recent centuries (56, 66-68).

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

Methods

Study population

This study is based on a previously described dataset (46, 50). Briefly, adult active TB patients (sputum smear-positive and GeneXpert-positive) were recruited between November 2013 and June 2022 at the Temeke District Hospital in Dar es Salaam, Tanzania when they first presented for care. Sputum and blood samples were collected from each patient to extract DNA for sequencing of the MTBC strain, and genotyping or whole-genome sequencing (WGS) the patient. Additionally, clinical and sociodemographic information was obtained from every patient. In total, there was either human or bacterial data available for 1,906 patients. Of those 1,444 patients had human genetic data and 1,471 had bacterial genetic data available, respectively. A total of 1,000 patients had both types of data available after quality-based filtering. The geographical locations of the self-indicated ethnic group of each patient were retrieved by searching for the original region of the respective ethnic group, and if they originated from a single region, the geographic coordinates according to Wikipedia were taken. If two neighboring regions were among the origins, then a random location between the two regions was taken as surrogate (Supplementary Table S1).

Bacterial and human sample processing

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

platform at the University of Geneva in Switzerland using the Illumina Infinium H3Africa genotyping microarray (Version 2; https://chipinfo.h3abionet.org) plus custom Tanzanian-specific SNP add-ons (69).

Human genetic data

The processing of the human genetic data has been described in detail by Xu et al (46). Briefly, we used the GRCh38 as a human reference genome to map the WGS reads of 118 patients using BWA aligner (v0.7.17) (70). Duplicate reads were then marked with the markduplicates module of Picard (v2.8.14) (71). Variant calling was first done for each sample individually following GATK best practices for germline short variant discovery. Samples with a coverage below 5 were excluded followed by a joint calling of the variants. A Quality Score Recalibration (VQSR) based filter was applied (real sensitivity of 99.7, excess heterozygosity of 54.69) and samples with more than 50% missing genotype calls were removed.

For the genotyping data, we used the Illumina GenomeStudio software (v2.0.5, https://support.illumina.com/array/array_software/genomestudio/downloads.html) to analyze the raw microarray data. Samples with a low quality, that were badly clustered, or that had a call rate below 0.97, were excluded. The PLINK format was then converted to VCF format using PLINK (v1.9) (72). The first round of imputation was performed with the African Genome Resources (AFGR, https://www.apcdr.org/) reference panel on the sanger imputation server with EAGLE (73) for phasing and Positional Burrows-Wheeler transform (PBWT) (74) for imputing. The second round was performed with a reference panel created in house that was based on 118 patients with available whole-genome sequences (69) with SHAPEIT4 for phasing and Minimac3 for imputing. For each SNP, the reference panel with

the highest imputation quality score was used to determine the final genotype call. SNPs with an INFO score below 0.8 were discarded.

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

Whole-genome sequence analysis of the MTBC bacteria

We analyzed all FASTQ files using the WGS analysis pipeline described previously (75). In summary, Trimmomatic (76) v. 0.33 (SLIDINGWINDOW:5:20) was used to remove the Illumina adaptors and to trim low quality reads. Only reads with a length of at least 20 bp were kept for further analysis. Overlapping paired-end reads were mergend using SeqPrep v. 1.2 (77) (overlap size = 15). We then mapped the resulting reads to a reconstructed ancestral sequence of the MTBC (78) with BWA v. 0.7.13 (mem algorithm) (70). Picard v. 2.9.1 (79) was then applied to mark and exclude duplicated reads. Furthermore, the RealignerTargetCreator and IndelRealigner modules of GATK v. 3.4.0 (80) were used to perform local realignment of reads around INDELs. Reads having an alignment score lower than $(0.93 \times read\ length) - (read\ length\ \times 4\ \times 0.07)$, corresponding to more than 7 miss-matches per 100bp, were excluded using Pysam v. 0.9.0 (81). SNP calling was then performed with SAMtools v. 1.2 mpileup (82) and VarScan v. 2.4.1 (83) with the following thresholds: a minimum mapping quality of 20, a minimum base quality at a position of 20, minimum read depth at a position of 7 and no strand bias. Positions in repetitive regions such as PE, PPE, and PGRS genes or phages were excluded, as described in (84). A whole-genome Fasta file was created from the resulting VCF file. We applied some additional filters; genomes were excluded from downstream analysis if they had a sequencing coverage of lower than 15 or if they contained SNPs suggestive of different MTBC lineages (i.e. mixed

infections). We identified lineages and sublineages using the SNP-based classification by Steiner et al. (85), and Coll et al. (86), respectively.

Identification of bacterial SNPs diagnostic for the successful MTBC introductions

We previously identified several successful MTBC introductions into Dar es Salaam (49). For these, we aimed to obtain a set of diagnostic SNPs that would allow assigning MTBC strains not included in our previous study to these genotypes. For that, we merged the VCF files from the 1,082 MTBC genomes included in that previous dataset by using BCFtools (v1.9). We then used VCFtools (v0.1.16) to remove Indels and positions that were variable in less than 12 genomes (12 was the minimal threshold selected when identifying the successful introductions in our previous publication (50)). By using the R package VariantAnnotation (87) and a customized Python script, SNPs specific to one of the most successful introductions were extracted. To ensure the SNPs identified as markers for the introductions were specific, we also identified phylogenetic SNPs on a bigger and global dataset representing the human-adapted MTBC diversity (75), and tested whether any of the phylogenetic SNPs identified for any of the successful Introductions was present in any of the MTBC lineages or sublineages. We compiled a subset of 25 SNPs (**Error! Reference source not found.**) that we used as phylogenetic markers for the different MTBC introductions and identified strains belonging to one of the four most successful MTBC introductions (Introduction 1, Introduction 5, Introduction 9, and Introduction 10) in the expanded MTBC dataset based on these SNPs by using a customized script.

Measures of TB disease severity

We used three different proxies for TB disease severity. The first one was the TB-score, which is a clinical score adapted from Wejse et al. (88) that consists of several signs and symptoms including the presence of fever and the Body Mass Index (BMI). A point was

given for each of the following symptoms or clinical measures if present: BMI below 18, BMI below 16, mid upper arm circumference (MUAC) below 220, MUAC below 200, body temperature higher than 37°C, cough, hemoptysis, dyspnea, chest pain, night sweat, abnormal auscultation, anaemia. A maximum of 12 points could be achieved, and a TB-score below 6 was considered as mild and everything above as severe. As a second proxy, we assessed the amount of lung damage. Two independent radiologists assessed chest x-ray pictures of the patients and gave a Ralph score (89). X-ray scores above 71 were considered as severe and everything below as mild. As a third proxy we used the bacterial load in the sputum represented by the difference between the first (early cycle) and the last (late cycle) threshold during quantitative PCR (Ct-value). For each sputum sample, we took five probes, run a quantitative PCR each and reported the lowest Ct-value.

Genetic ancestry analysis of TB patients

To investigate the genetic ancestry of the TB patients, we combined our dataset with the data from ten other projects: The Gambian Genome Variation Project (GGVP) (90), the 1000 Genomes Project (1000G) (91), the Human Genome Diversity Project (HGDP) (92), Simons Genome Diversity Project (93), as well as data generated by Patin et al. 2014 and 2017 (4, 94), Hollfelder et al. 2017 (95), Semo et al. 2019 (8), Schlebusch et al. 2012 (96), and Fortes-Lima et al. 2022 (97). We used the GRCh37 version of all datasets. The dataset of HGDP was in GRCh38 version and we thus did a lift over to GRCh37 using the picard (v2.26.10) (71) tool LiftoverVcf. For all the datasets including only populations from one single continent , we excluded variants with a missingness of more than 10% and only included variants that did not deviate from Hardy-Weinberg equilibrium ($p < 1e-5$) using PLINK (version 1.9b, www.cog-genomics.org/plink/1.9/) (72). For the 1000G, SGDP, and HGDP data, we first identified variants that deviated from the Hardy-Weinberg equilibrium ($p < 1e-5$) in each superpopulation using PLINK (version 2.0a, www.cog-genomics.org/plink/2.0/) and removed

them from the whole dataset. We additionally removed variants with a high missingness (> 10%) from the full datasets using PLINK (version 1.9b). After extraction of 103,262 nucleotide positions common to all datasets, we merged the datasets using PLINK (version 1.9b). From the merged dataset, we removed second-degree relatives using PLINK (version 2.0a, king cutoff of 0.088) (N=369) and patients from our cohort where the sex according to the genetic data did not correspond with the sex indicated in the clinical data, patients who were genetic outliers based on principal component analysis (PCA) or who did not cluster with any other African samples (N=83). In addition, we removed regions of high linkage disequilibrium (https://genome.sph.umich.edu/wiki/Regions_of_high_linkage_disequilibrium_(LD)) and applied additional filters to the merged dataset (missingness > 10%, minimum allele frequency of 5%, removal of sex chromosomes, variant pruning with –indep-pairwise 50 10 0.1, only biallelic positions) ending up with 53,255 positions and 7,479 individuals from 249 populations.

To infer the ancestry proportions of the Tanzanian TB patients, we used ADMIXTURE (version 1.3.0) (48). We estimated the number of ancestral populations (K) by running ADMIXTURE 15 times for each value for K from K=2 until K=29 with the option --cv. The --cv option performs 5-fold cross-validation and allows to identify the value for K resulting in the lowest cross-validation error (**Error! Reference source not found.**). The cross-validation error was lowest for K = 24. From the 15 runs performed with K = 24, we selected a representative of an output that was supported by most of the 15 runs (6/15) to extract the ancestry proportions of each individual. A PCA was performed using PLINK (version 1.9b) on all African populations included.

Spatial interpolation of human genetic ancestry proportions

To visualize the distributions of the different patient ancestries, we performed spatial interpolation using the OrdinaryKriging function from the pykrige module in Python (variogram_model = “linear”, grid space of 500). For the Eastern and Southern Bantu ancestries, we included all African populations for the interpolation. For the Western Bantu, the interpolation failed when using all African populations with an insufficient slope, suggesting little spatial variability. Since many populations were sampled in the region with the highest proportions of Western Bantu genetic ancestry and among them many hunter-gatherer populations containing little to no Western Bantu genetic ancestry, we repeated the interpolation with only a subset of the non-hunter-gatherer populations.

Correlation between distance matrices

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

To investigate, whether human populations that are geographically more distant are also genetically more distant, we calculated the correlation between the geographical and the

genetic distance on an African level as well as on the level of Tanzania. For the geographical distance matrix, we calculated the Euclidean distance based on the latitude and the longitude. At the level of Tanzania, the broad geographic location of the original area of the ethnic group was considered (Supplementary Table S1) and for the continental level, the coordinates of the hospital in Temeke were taken for the Tanzanian TB patients considering that for the other studies only the sampling locations were known. The human and bacterial genetic distance matrices as well as the human genetic and geographic distance matrices were tested for correlations using the mantel.test() function from the mantel module in Python (options: perms = 10000, method = “pearson”, tail = “upper”).

Statistical analyses

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

To estimate the associations between disease severity and the explanatory variables bacterial genotype (binary, belonging to Introduction 10 or not) and human ancestries (Southeastern Bantu, Eastern Bantu, Western Bantu), we used a logistic regression model. We included three variables as proxies for disease severity: Lung damage based on X-ray score (mild versus severe), TB-score (mild versus severe), bacterial load (continuous, log10 transformed). We tested only for Introduction 10 because there were few observations of the other MTBC genotypes. To account for the compositional nature of the human ancestries (i.e. that they sum up to 1), we used the additive log transformation from the R package ‘compositions’ (98). The ancestry proportions were transformed and categorized with category 1 comprising the lowest amount of the respective ancestry and category 3 (4 in the case of Western Bantu)

the highest. The categories were chosen to have roughly equal numbers of patients in each. We used categories to allow a non-linear relationship without specifying polynomials, needing a large sample size or difficulty in interpretation but we recognize that a small amount of information is lost. Similar results were also obtained with other parameterizations. We assessed whether there was an interaction between ancestry and genotype on TB severity using the likelihood ratio test. For that, we compared a model including the interaction between the MTBC genotype and human genetic ancestry to a model without the interaction using the 'lmtest' package (99). The estimates were adjusted for: age, sex, smoking, and the number of weeks with cough. Only HIV-negative patients were included in the analysis. All statistical analyses were carried out in R (version 4.1.2). Code for the statistical analysis can be found (https://github.com/mzwyer/TB-Dar_Mtb).

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 available under request.

Funding Statement

Swiss National Science Foundation (Grant No: CRSII5_177163 , CRSII5_213514 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).

10. Uren C, Moller M, van Helden PD, Henn BM, Hoal EG. Population structure and infectious disease risk in southern Africa. Molecular Genetics and Genomics. 2017;292(3):499-509.
11. Fumagalli M, Sironi M, Pozzoli U, Ferrer-Admetlla A, Pattini L, Nielsen R. Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. PLoS Genet. 2011;7(11):e1002355.
12. Chapman SJ, Hill AV. Human genetic susceptibility to infectious disease. Nat Rev Genet. 2012;13(3):175-88.
13. Newport MJ, Finan C. Genome-wide association studies and susceptibility to infectious diseases. Brief Funct Genomics. 2011;10(2):98-107.
14. Lee JC, Espeli M, Anderson CA, Linterman MA, Pocock JM, Williams NJ, et al. Human SNP links differential outcomes in inflammatory and infectious disease to a FOXO3-regulated pathway. Cell. 2013;155(1):57-69.
15. Akcay IM, Katrinli S, Ozdil K, Doganay GD, Doganay L. Host genetic factors affecting hepatitis B infection outcomes: Insights from genome-wide association studies. World J Gastroenterol. 2018;24(30):3347-60.
16. WHO. Global Tuberculosis Report 2023. 2023.
17. Zwyer M, Cavusoglu C, Ghielmetti G, Pacciarini ML, Scaltriti E, Van Soolingen D, et al. A new nomenclature for the livestock-associated Mycobacterium tuberculosis complex based on phylogenomics. Open Res Eur. 2021;1:100.
18. Guyeux C, Senelle G, Le Meur A, Supply P, Gaudin C, Phelan JE, et al. Newly Identified Mycobacterium africanum Lineage 10, Central Africa. Emerg Infect Dis. 2024;30(3):560-3.
19. Gagneux S. Ecology and evolution of Mycobacterium tuberculosis. Nat Rev Microbiol. 2018;16(4):202-13.
20. de Jong BC, Antonio M, Gagneux S. Mycobacterium africanum--review of an important cause of human tuberculosis in West Africa. PLoS Negl Trop Dis. 2010;4(9):e744.
21. Wiens KE, Woyczynski LP, Ledesma JR, Ross JM, Zenteno-Cuevas R, Goodridge A, et al. Global variation in bacterial strains that cause tuberculosis disease: a systematic review and meta-analysis. Bmc Medicine. 2018;16.
22. Reed MB, Pichler VK, McIntosh F, Mattia A, Fallow A, Masala S, et al. Major Mycobacterium tuberculosis lineages associate with patient country of origin. J Clin Microbiol. 2009;47(4):1119-28.
23. Gagneux S, DeRiemer K, Van T, Kato-Maeda M, de Jong BC, Narayanan S, et al. Variable host-pathogen compatibility in Mycobacterium tuberculosis. Proc Natl Acad Sci U S A. 2006;103(8):2869-73.
24. Baker L, Brown T, Maiden MC, Drobniewski F. Silent nucleotide polymorphisms and a phylogeny for Mycobacterium tuberculosis. Emerg Infect Dis. 2004;10(9):1568-77.
25. Fenner L, Egger M, Bodmer T, Furrer H, Ballif M, Battegay M, et al. HIV infection disrupts the sympatric host-pathogen relationship in human tuberculosis. PLoS Genet. 2013;9(3):e1003318.
26. Du DH, Geskus RB, Zhao Y, Codecasa LR, Cirillo DM, van Crevel R, et al. The effect of M. tuberculosis lineage on clinical phenotype. medRxiv. 2023.
27. de Jong BC, Hill PC, Aiken A, Awine T, Antonio M, Adetifa IM, et al. Progression to active tuberculosis, but not transmission, varies by Mycobacterium tuberculosis lineage in The Gambia. J Infect Dis. 2008;198(7):1037-43.
28. Asare P, Asante-Poku A, Prah DA, Borrell S, Osei-Wusu S, Otchere ID, et al. Reduced transmission of Mycobacterium africanum compared to Mycobacterium tuberculosis in urban West Africa. International Journal of Infectious Diseases. 2018;73:30-42.
29. Holt KE, McAdam P, Thai PVK, Thuong NTT, Ha DTM, Lan NN, et al. Frequent transmission of the Mycobacterium tuberculosis Beijing lineage and positive selection for the EsxW Beijing variant in Vietnam. Nat Genet. 2018;50(6):849-56.
30. Click ES, Moonan PK, Winston CA, Cowan LS, Oeltmann JE. Relationship Between Mycobacterium tuberculosis Phylogenetic Lineage and Clinical Site of Tuberculosis. Clinical Infectious Diseases. 2012;54(2):211-9.

31. Boisson-Dupuis S, Ramirez-Alejo N, Li Z, Patin E, Rao G, Kerner G, et al. Tuberculosis and impaired IL-23-dependent IFN-gamma immunity in humans homozygous for a common TYK2 missense variant. Sci Immunol. 2018;3(30).
32. Stein CM, Sausville L, Wejse C, Sobota RS, Zetola NM, Hill PC, et al. Genomics of human pulmonary tuberculosis: from genes to pathways. Curr Genet Med Rep. 2017;5(4):149-66.
33. Phelan J, Gomez-Gonzalez PJ, Andreu N, Omae Y, Toyo-Oka L, Yanai H, et al. Genome-wide host-pathogen analyses reveal genetic interaction points in tuberculosis disease. Nat Commun. 2023;14(1):549.
34. Bai H, Song M, Lei S, Jiao L, Hu X, Wu T, et al. Genome-wide association study of tuberculosis in the western Chinese Han and Tibetan population. MedComm (2020). 2023;4(2):e250.
35. Png E, Alisjahbana B, Sahiratmadja E, Marzuki S, Nelwan R, Balabanova Y, et al. A genome wide association study of pulmonary tuberculosis susceptibility in Indonesians. BMC Med Genet. 2012;13:5.
36. Zheng R, Li Z, He F, Liu H, Chen J, Chen J, et al. Genome-wide association study identifies two risk loci for tuberculosis in Han Chinese. Nat Commun. 2018;9(1):4072.
37. Asgari S, Luo Y, Huang CC, Zhang Z, Calderon R, Jimenez J, et al. Higher native Peruvian genetic ancestry proportion is associated with tuberculosis progression risk. Cell Genom. 2022;2(7).
38. Daya M, van der Merwe L, van Helden PD, Moller M, Hoal EG. The role of ancestry in TB susceptibility of an admixed South African population. Tuberculosis. 2014;94(4):413-20.
39. Macallan DC. Malnutrition in tuberculosis. Diagnostic Microbiology and Infectious Disease. 1999;34(2):153-7.
40. Selwyn PA, Hartel D, Lewis VA, Schoenbaum EE, Vermund SH, Klein RS, et al. A Prospective-Study of the Risk of Tuberculosis among Intravenous Drug-Users with Human Immunodeficiency Virus-Infection. New England Journal of Medicine. 1989;320(9):545-50.
41. Dooley KE, Chaisson RE. Tuberculosis and diabetes mellitus: convergence of two epidemics. Lancet Infectious Diseases. 2009;9(12):737-46.
42. Spence DPS, Hotchkiss J, Williams CSD, Davies PDO. Tuberculosis and Poverty. British Medical Journal. 1993;307(6907):759-61.
43. Bates MN, Khalakdina A, Pai M, Chang L, Lessa F, Smith KR. Risk of tuberculosis from exposure to tobacco smoke - A systematic review and meta-analysis. Archives of Internal Medicine. 2007;167(4):335-42.
44. Imtiaz S, Shield KD, Roerecke M, Samokhvalov AV, Lonnroth K, Rehm J. Alcohol consumption as a risk factor for tuberculosis: meta-analyses and burden of disease. Eur Respir J. 2017;50(1).
45. McHenry ML, Wampande EM, Joloba ML, Malone LL, Mayanja-Kizza H, Bush WS, et al. Interaction between M. tuberculosis Lineage and Human Genetic Variants Reveals Novel Pathway Associations with Severity of TB. Pathogens. 2021;10(11).
46. Xu ZM, Zwyer M, Brites D, Hiza H, Sasamalo M, Reinhard M, et al. Genome-to-genome analysis reveals associations between human and mycobacterial genetic variation in tuberculosis patients from Tanzania. medRxiv. 2023:2023.05. 11.23289848.
47. Ogarkov O, Mokrousov I, Sinkov V, Zhdanova S, Antipina S, Savilov E. 'Lethal' combination of Mycobacterium tuberculosis Beijing genotype and human CD209-336G allele in Russian male population. Infection Genetics and Evolution. 2012;12(4):732-6.
48. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19(9):1655-64.
49. Alexander DH, Lange K. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. BMC Bioinformatics. 2011;12:246.
50. Zwyer M, Rutaihwa LK, Windels E, Hella J, Menardo F, Sasamalo M, et al. Back-to-Africa introductions of Mycobacterium tuberculosis as the main cause of tuberculosis in Dar es Salaam, Tanzania. Plos Pathogens. 2023;19(4).
51. Brites D, Gagneux S. Co-evolution of Mycobacterium tuberculosis and Homo sapiens. Immunol Rev. 2015;264(1):6-24.

52. McHenry ML, Bartlett J, Igo RP, Wampande EM, Benchek P, Mayanja-Kizza H, et al. Interaction between host genes and Mycobacterium tuberculosis lineage can affect tuberculosis severity: evidence for coevolution? Plos Genetics. 2020;16(4).
53. Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, Froment A, et al. The Genetic Structure and History of Africans and African Americans. Science. 2009;324(5930):1035-44.
54. LLC M. [Available from: https://www.macrotrends.net/global-metrics/cities/22894/dar-es-salaam/population.
55. Brielle ES, Fleisher J, Wynne-Jones S, Sirak K, Broomandkhoshbacht N, Callan K, et al. Entwined African and Asian genetic roots of medieval peoples of the Swahili coast. Nature. 2023;615(7954):866-73.
56. Comas I, Hailu E, Kiros T, Bekele S, Mekonnen W, Gumi B, et al. Population Genomics of Mycobacterium tuberculosis in Ethiopia Contradicts the Virgin Soil Hypothesis for Human Tuberculosis in Sub-Saharan Africa. Curr Biol. 2015;25(24):3260-6.
57. Lopez B, Aguilar D, Orozco H, Burger M, Espitia C, Ritacco V, et al. A marked difference in pathogenesis and immune response induced by different Mycobacterium tuberculosis genotypes. Clin Exp Immunol. 2003;133(1):30-7.
58. Aguilar D, Hanekom M, Mata D, Gey van Pittius NC, van Helden PD, Warren RM, Hernandez-Pando R. Mycobacterium tuberculosis strains with the Beijing genotype demonstrate variability in virulence associated with transmission. Tuberculosis (Edinb). 2010;90(5):319-25.
59. Dormans J, Burger M, Aguilar D, Hernandez-Pando R, Kremer K, Roholl P, et al. Correlation of virulence, lung pathology, bacterial load and delayed type hypersensitivity responses after infection with different Mycobacterium tuberculosis genotypes in a BALB/c mouse model. Clin Exp Immunol. 2004;137(3):460-8.
60. Coscolla M, Gagneux S. Does M. tuberculosis genomic diversity explain disease diversity? Drug Discov Today Dis Mech. 2010;7(1):e43-e59.
61. Stanley S, Spaulding CN, Liu Q, Chase MR, Ha DTM, Thai PVK, et al. Identification of bacterial determinants of tuberculosis infection and treatment outcomes: a phenogenomic analysis of clinical strains. Lancet Microbe. 2024.
62. McHenry ML, Williams SM, Stein CM. Genetics and evolution of tuberculosis pathogenesis: New perspectives and approaches. Infect Genet Evol. 2020;81:104204.
63. Cavalli-Sforza LL, Menozzi P, Piazza A. The history and geography of human genes: Princeton university press; 1994.
64. Frascella B, Richards AS, Sossen B, Emery JC, Odone A, Law I, et al. Subclinical Tuberculosis Disease-A Review and Analysis of Prevalence Surveys to Inform Definitions, Burden, Associations, and Screening Methodology. Clin Infect Dis. 2021;73(3):e830-e41.
65. Silva ML, Ca B, Osorio NS, Rodrigues PNS, Maceiras AR, Saraiva M. Tuberculosis caused by Mycobacterium africanum: Knowns and unknowns. PLoS Pathog. 2022;18(5):e1010490.
66. Rutaihwa LK, Menardo F, Stucki D, Gygli SM, Ley SD, Malla B, et al. Multiple Introductions of Mycobacterium tuberculosis Lineage 2-Beijing Into Africa Over Centuries. Frontiers in Ecology and Evolution. 2019;7.
67. Menardo F, Rutaihwa LK, Zwyer M, Borrell S, Comas I, Conceicao EC, et al. Local adaptation in populations of Mycobacterium tuberculosis endemic to the Indian Ocean Rim. F1000Res. 2021;10:60.
68. Chihota VN, Niehaus A, Streicher EM, Wang X, Sampson SL, Mason P, et al. Geospatial distribution of Mycobacterium tuberculosis genotypes in Africa. PLoS One. 2018;13(8):e0200632.
69. Xu ZM, Rueger S, Zwyer M, Brites D, Hiza H, Reinhard M, et al. Using population-specific add-on polymorphisms to improve genotype imputation in underrepresented populations. PLoS Comput Biol. 2022;18(1):e1009628.
70. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754-60.
71. Broad Institute. Picard Toolkit. Broad Institute, GitHub Repository. 2019.

749 72. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising
750 to the challenge of larger and richer datasets. Gigascience. 2015;4:7.
751 73. Loh PR, Danecek P, Palamara PF, Fuchsberger C, Y AR, H KF, et al. Reference-based phasing
752 using the Haplotype Reference Consortium panel. Nat Genet. 2016;48(11):1443-8.
753 74. Durbin R. Efficient haplotype matching and storage using the positional Burrows-Wheeler
754 transform (PBWT). Bioinformatics. 2014;30(9):1266-72.
755 75. Menardo F, Loiseau C, Brites D, Coscolla M, Gygli SM, Rutaihwa LK, et al. Treemmer: a tool to
756 reduce large phylogenetic datasets with minimal loss of diversity. BMC Bioinformatics.
757 2018;19(1):164.
758 76. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data.
759 Bioinformatics. 2014;30(15):2114-20.
760 77. SeqPrep. https://github.com/jstjohn/SeqPrep [
761 78. Comas I, Chakravartti J, Small PM, Galagan J, Niemann S, Kremer K, et al. Human T cell
762 epitopes of Mycobacterium tuberculosis are evolutionarily hyperconserved. Nat Genet.
763 2010;42(6):498-503.
764 79. Picard. https://github.com/broadinstitute/picard [
765 80. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome
766 Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.
767 Genome Res. 2010;20(9):1297-303.
768 81. Pysam. https://github.com/pysam-developers/pysam [
769 82. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and
770 population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27(21):2987-
771 93.
772 83. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: somatic
773 mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res.
774 2012;22(3):568-76.
775 84. Stucki D, Brites D, Jeljeli L, Coscolla M, Liu Q, Trauner A, et al. Mycobacterium tuberculosis
776 lineage 4 comprises globally distributed and geographically restricted sublineages. Nat Genet.
777 2016;48(12):1535-43.
778 85. Steiner A, Stucki D, Coscolla M, Borrell S, Gagneux S. KvarQ: targeted and direct variant
779 calling from fastq reads of bacterial genomes. BMC Genomics. 2014;15:881.
780 86. Coll F, McNerney R, Guerra-Assuncao JA, Glynn JR, Perdigao J, Viveiros M, et al. A robust SNP
781 barcode for typing Mycobacterium tuberculosis complex strains. Nat Commun. 2014;5:4812.
782 87. Obenchain V, Lawrence M, Carey V, Gogarten S, Shannon P, Morgan M. VariantAnnotation: a
783 Bioconductor package for exploration and annotation of genetic variants. Bioinformatics.
784 2014;30(14):2076-8.
785 88. Wejse C, Gustafson P, Nielsen J, Gomes VF, Aaby P, Andersen PL, Sodemann M. TBscore:
786 Signs and symptoms from tuberculosis patients in a low-resource setting have predictive value and
787 may be used to assess clinical course. Scand J Infect Dis. 2008;40(2):111-20.
788 89. Ralph AP, Ardian M, Wiguna A, Maguire GP, Becker NG, Drogumuller G, et al. A simple, valid,
789 numerical score for grading chest x-ray severity in adult smear-positive pulmonary tuberculosis.
790 Thorax. 2010;65(10):863-9.
791 90. Band G, Le QS, Clarke GM, Kivinen K, Hubbart C, Jeffreys AE, et al. Insights into malaria
792 susceptibility using genome-wide data on 17,000 individuals from Africa, Asia and Oceania. Nature
793 Communications. 2019;10.
794 91. Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, et al. A global
795 reference for human genetic variation. Nature. 2015;526(7571):68-+.
796 92. Bergstrom A, McCarthy SA, Hui RY, Almarri MA, Ayub Q, Danecek P, et al. Insights into
797 human genetic variation and population history from 929 diverse genomes. Science.
798 2020;367(6484):1339-+.

93. Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature. 2016;538(7624):201-+.
94. Patin E, Siddle KJ, Laval G, Quach H, Harmant C, Becker N, et al. The impact of agricultural emergence on the genetic history of African rainforest hunter-gatherers and agriculturalists. Nat Commun. 2014;5:3163.
95. Hollfelder N, Schlebusch CM, Gunther T, Babiker H, Hassan HY, Jakobsson M. Northeast African genomic variation shaped by the continuity of indigenous groups and Eurasian migrations. Plos Genetics. 2017;13(8).
96. Schlebusch CM, Skoglund P, Sjodin P, Gattepaille LM, Hernandez D, Jay F, et al. Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. Science. 2012;338(6105):374-9.
97. Fortes-Lima C, Triska P, Cizkova M, Podgorna E, Diallo MY, Schlebusch CM, Cerny V. Demographic and Selection Histories of Populations Across the Sahel/Savannah Belt. Mol Biol Evol. 2022;39(10).
98. van den Boogaart KG, Tolosana-Delgado R, Bren M. compositions: Compositional Data Analysis 2022 [Available from: https://CRAN.R-project.org/package=compositions.]
99. A. Z, T. H. Diagnostic Checking in Regression Relationships. R News. 2002;2(3):7-10.

Supporting information captions

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

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

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

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

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

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

Supplementary Figure S7 – Association between the Bantu genetic ancestries and TB-score and X-ray score for each of the most successful introductions. A, C, E) Southeastern, Eastern, and Western Bantu on mild or severe TB-score, B, D, E) Southeastern, Eastern, and Western Bantu genetic ancestry lung damage assessed by lung X-ray.

Supplementary Table S1 - The different ethnic groups with at least 10 members in our cohort and the region and broad geographic location of the original area of the ethnic group. The

latitude and longitude are given in decimal degrees. In the case of two associated regions, a location close to the border of the two regions was selected.

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