Host-pathogen co-adaptation shapes susceptibility to infection with *Mycobacterium tuberculosis*

Authors

Matthias I Gröschel*, Francy J. Pérez-Llanos*, Roland Diel, Roger Vargas Jr, Vincent Escuyer, Kimberlee Musser, Lisa Trieu, Jeanne Sullivan Meissner, Jillian Knorr, Don Klinkenberg, Peter Kouw, Susanne Homolka, Wojciech Samek, Barun Mathema, Dick van Soolingen, Stefan Niemann[#], Shama Ahuja, Maha R Farhat[#].

Affiliations

*[#] Equal contribution

MIG, MRF: Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA

PK: Department of Infectious Diseases, Public Health Service, Amsterdam, The Netherlands

VE, KM: Wadsworth Center, New York State Department of Health, Albany, New York, USA

LT, JSM, JK, SA: New York City Department of Health and Mental Hygiene, Bureau of Tuberculosis Control, Queens, New York

SN: German Center for Infection Research, Partner Site Hamburg-Lübeck-Borstel-Riems, Borstel, Germany

RD: Institute for Epidemiology, University Medical Hospital Schleswig-Holstein, Kiel, Germany

RD: Lungenclinic Grosshansdorf, Airway Research Center North (ARCN), Member of the German Center for Lung Research (DZL), Grosshansdorf, Germany

FJPL, SH, SN: Molecular and Experimental Mycobacteriology, Research Center Borstel, Borstel, Germany

DK, DvS: National Tuberculosis Reference Laboratory, Center for Infectious Disease Control, National Institute for Public Health and the Environment (RIVM), Bilthoven, The Netherlands

MRF: Division of Pulmonary and Critical Care Medicine, Massachusetts General Hospital, Boston, MA, USA

WS: Department of Electrical Engineering and Computer Science, Technical University Berlin and Department of Artificial Intelligence, Fraunhofer Heinrich Hertz Institute, Berlin, Germany

BM: Mailman School of Public Health, Columbia University, New York City, NY, USA

Keywords

Tuberculosis, Transmission, Co-adaptation, Contact Tracing, Whole Genome Sequencing, Host-pathogen interaction

Word count

Correspondence: {matthias_groeschel, maha_farhat}@hms.harvard.edu

SUMMARY

The study of host-pathogen co-evolution is fundamental to understanding the emergence and spread of infectious diseases. The obligate human pathogen of the *Mycobacterium* tuberculosis complex (*Mtbc*) separates genetically into nine lineages with distinct patterns of geographical distribution that in some cases parallel that of human subpopulations. Based on these observations, geographically restricted *Mtbc* lineages have been hypothesized to be niche *specialists* that preferentially infect particular human subpopulations, but this is yet to be confirmed while controlling for social networks and risk of disease among exposed hosts. To address this question, we used a multi-site cohort of tuberculosis index cases with pathogen sequence data and linked contacts. Our data show that strains of specialist (spec) *Mtbc* lineages L1,$L2_{spec}$,L3,$L4_{spec}$,L5,L6 are intrinsically less transmissible than generalist *Mtbc* lineages ($L2_{gen}$,$L4_{gen}$) across Western European and North American cosmopolitan populations. Comparing transmissibility between sympatric and allopatric host-pathogen pairs, we found the first controlled evidence for co-adaptation between *Mtbc* strains and their human hosts; allopatric host-pathogen exposures had a 32% decrease in the odds of infection among contacts compared with sympatric exposures. We measured 10-fold decreased phagocytosis and growth rates of L6 specialist strains compared to $L4_{gen}$ in *in vitro* allopatric macrophage infections. In conclusion, long term co-adaptation between *Mtbc* strains and humans has resulted in differential transmissibility between allopatric and sympatric hosts for the specialist lineages. Understanding the specific genetic and immunological underpinnings of this co-adaptation may inform rational vaccine design and TB control.

INTRODUCTION

Humans and their ancestors have continuously inhabited an earth replete with a wide diversity of microbial species. While many of these species are harmless or even beneficial to humans, some have the ability to infect human tissues and inflict harm, including pathogenic species that evolved the concurrent ability to transmit between humans (e.g., via contact, droplets or airborne transmission). The study of pathogen evolution and host-pathogen relationships at long-time scales can shed novel insights on the biology of infection that guide disease control efforts.

Until recently, the study of host-pathogen interactions was methodologically limited to model organisms in model environments, as direct investigation of naturally occurring co-adaptation and long-term generational effects was technically and logistically difficult (*1*). High-throughput whole genome sequencing (WGS) has since enabled the direct study of hundreds of thousands of pathogens sampled directly from their human host, facilitating exploration of host-pathogen interactions such as associations between regions of the *P. falciparum* genome and sickle cell trait that increase the risk of severe malaria and inform clinical and public health practice (*2*).

The obligate human pathogen *Mycobacterium tuberculosis* complex (*Mtbc*) is hypothesized to have co-evolved with its human host (*3*, *4*). Based on studies describing the phylo-geographical distribution of *Mtbc*, strains within geographically-restricted lineages of *Mtbc* are hypothesized to be niche *specialists* that preferentially infect local human subpopulations, while strains of other lineages occurring worldwide have been classified as *generalists*, potentially based on a wider host spectrum (*4*–*9*). Stable host ancestry (i.e. pathogen lineage associations among secondary tuberculosis (TB) cases) has been observed in two cosmopolitan cities (*10*, *11*). However, as *Mtbc* is an airborne pathogen, observations suggesting such co-adaptation (also known as *sympatry*) have been alternatively attributed to differential exposure to *Mtbc* lineages as a result of social mixing

patterns, genetic drift, or founder effects rather than true co-adaptation (*4*). These observations have also been dismissed as unlikely based on the relatively minute amount of diversity (<0.1%) separating *Mtbc's* main genetic lineages (*5*).

Here, we leverage detailed public health contact tracing data and whole genome sequencing from three low TB incidence cosmopolitan cities to study transmissibility and co-adaptation of *Mtbc* strains while controlling for founder effects, host factors, and extent of exposure. We follow up the epidemiological analysis with *in vitro* experiments demonstrating strong evidence for co-adaptation between *Mtbc* lineages and humans of different ancestry.

RESULTS

A multi-site cosmopolitan cohort of tuberculosis cases and close contacts

A major barrier to investigating *Mtbc* co-adaptation is the limitation in size and depth of contact tracing data directly measuring transmissibility in cosmopolitan societies with mixing of human and pathogen populations. To address this, we built a large cohort of TB cases (n = 5,256) and their contacts (n=28,889) from New York City (NYC), Amsterdam, and Hamburg. The three settings routinely and systematically surveyed for TB cases and contacts as part of TB control activities and performed whole genome sequencing (WGS) for strains from nearly all active culture-positive TB cases in this cohort (>99%, 5,219/5,256). WGS allowed *Mtbc* lineage classification and informed transmission assessment.

After excluding extrapulmonary TB as potential source cases and removing contacts without significant exposure, the final analysis cohort comprised 2,279 patients with pulmonary TB linked to 12,749 close contacts (**Methods, Table 1, Figure 1a**). The median number of close contacts investigated per index TB case was respectively 4 contacts (IQR 2 – 6), 4 (IQR 2 – 8), and 3 (IQR 1 – 7) in NYC, Amsterdam, and Hamburg (**Figure 1b**).

Patient characteristics differed across study sites

The three source cities differed in key drivers of TB case burden and transmission, including birth in a country with high TB incidence, proportion of TB in the adolescent to young adult age group, sputum *Mtbc* smear positivity, alcohol or substance use disorder, and homelessness (*12*) (**Table 1**). On average across study sites, 69% of TB cases and 48% of close contacts were born in a country with a high TB incidence country, defined as >30/100,000 TB incidence based WHO's 2016 data (*13*); TB patients and close contacts from NYC were more likely to be born in a high-incidence country (Chi squared, $p<0.001$ for both comparisons corrected for the false discovery rate). TB patients in NYC were older by 9 and 14 years on median compared with patients in Hamburg or Amsterdam, respectively (Wilcoxon rank sum, $p<0.001$). TB patients from Hamburg were less often sputum smear

positive (Chi squared, p=0.002 for Amsterdam and p<0.001 NYC), but more often used alcohol (p < 0.001 for both comparisons), or were homeless (p = 0.036 Amsterdam, p < 0.001 NYC). NYC TB cases more often used illicit substances (p = 0.009 Hamburg, p = 0.018 Amsterdam). Factors recognized to increase the contact risk of progression to active TB after infection, including HIV and other immunosuppression, did not differ by city. Diabetes was more common among TB cases in NYC (p < 0.001 for both comparisons). TB patients in NYC were more often isoniazid mono-resistant than in Hamburg (p = 0.005); there was no difference in multidrug resistance (MDR, resistance to isoniazid and rifampicin) across the cities.

Specialist lineages cause a measurable proportion of TB disease burden

We next used the WGS data to characterize the circulating *Mtbc* lineages in the three cities. Among the 5,087 isolate sequences from patients infected with *Mtbc sensu stricto* or *M. africanum*, we identified 297,061 single nucleotide substitutions (SNSs) that allowed the calculation of a maximum likelihood phylogeny (**Figure 1c**). Using canonical signature SNSs, the *Mtbc* strains analyzed were classified into six of the nine known human-adapted *Mtbc* complex lineages (*5*): the most common being strains of the Euro-American Lineage (L4, n = 2,976, 59%), followed in order by strains of the East Asia Lineage (L2, n = 802, 16%), East Africa Central Asia (L3, n = 709, 14%), and East African, The Philippines, rim of the Indian Ocean lineages (L1, n = 551, 11%). A minority of strains belonged to *M. africanum* 1 (L5, n = 17) or 2 (L6, n = 32). We categorized *Mtbc* strains as *specialists* if they belonged to lineages found to be phylo-geographically restricted in previous literature (*3*–*7*) (L1, L3, L5, L6, ancestral branches of L2 (2.1 'Proto-Beijing', 2.2.2 'Asia Ancestral 1', or 2.1.2.2.1 'Asia Ancestral 2'), or L4 sub-lineages: 4.11, 4.2.1.1, 4.3.i2, 4.5, 4.6.2.2 'Cameroon', 4.6.1.1.1 'Uganda') (**Suppl. Table S1**) and compared them to the remaining *generalist* sublineages of L2 and L4 (*4*, *5*, *8*).

Of 2,279 pulmonary TB patients, 28% (n = 643) were infected with strains of a specialist lineage. The proportion of specialist strain infections remained stable over the study period, and there was no association with patient age or sex (Chi squared, p>0.17, **Figure 1d-e**). Strains of specialist lineages were equally likely to harbor isoniazid mono-resistance (p = 0.8) and slightly less likely to harbor MDR compared with generalist strains (OR 0.98, CI_{95} 0.97 to 1, Wald p = 0.021), but the overall rate of MDR was low (n = 60, 2.6%, **Figure 1c**).

Strains of specialist *Mtbc* lineages differ in phylogenetic metrics of transmissibility

We hypothesized that strains of specialist lineages were less transmissible than generalist strains in the cosmopolitan source communities. To investigate this, we compared genetic diversity metrics between strains of specialist and generalist lineages using the total sample phylogeny. Terminal branch lengths, which represent an upper bound of evolutionary time since transmission for each *Mtbc* case, were significantly longer for strains of specialist lineages (mean 73 SNSs, CI_{95} 70 to 76, vs mean 47 SNSs CI_{95} 45 to 48 for generalists, $p < 2.2 \times 10^{-16}$ using Kolmogorov-Smirnov test, **Figure 2a**). This difference was driven by a combination of the long terminal branch lengths of strains of L1 (mean 96 SNSs, CI_{95} 90 to 102), $L2_{spec}$ (mean 87 SNSs, CI_{95} 75 to 99) and $L4_{spec}$ (mean 69 SNSs, CI_{95} 62 to 77) compared to shorter branch lengths for $L2_{gen}$, L3, and $L4_{gen}$ (means 54, 52, 45 SNSs, respectively) (**Suppl. Figure S1a**). Within main lineages, strains of $L2_{spec}$ and $L4_{spec}$ had longer terminal branch lengths than strains of $L2_{gen}$ and $L4_{gen.}$ respectively ($p < 1.8 \times 10^{-7}$ Kolmogorov Smirnov test for both comparisons). Possible-transmission clusters, defined as strains within 12-SNSs distance of each other, were more common for generalist lineage strains than specialist lineage strains (822/3,778 of generalist lineage strains clustered in 239/293 total clusters vs 242/1,309 of specialist lineage strains clustered in 54/293 clusters, p = 0.035, Chi Squared, **Figure 2b**). The specialist lineages L5/L6 comprised one L6 12-SNS cluster with two strains. The average size of clusters was slightly smaller for generalists than specialists lineage strains (3.4 patients vs 4.5 patients per cluster, p = 0.07, two-sample t-test).

Strains of specialist *Mtbc* lineages demonstrate intrinsically lower transmissibility

We isolated the effect of the infecting *Mtbc* lineage on the rate of infection among close contacts from other exposure characteristics and contact susceptibility as follows: we focused on TB infection as measured by a positive QuantiFERON®-TB Gold In-Tube test (IGRA ≥0.35 IU/ml) among close contacts (**Methods**), and (1) modeled the count of new TB infections in each index cases' group of close contacts as a quasi-Poisson distributed random variable with the mean estimated as a linear function of *Mtbc* lineage and index case characteristics and a subset of contact susceptibility characteristics that was available across the three source sites, and (2) modelled the probability of TB infection of each individual close contact as a function of index case infectiousness and a subset of contact susceptibility variables using generalized estimating equation (GEE) with a logistic link function (**Methods**). The latter approach allowed for more fine-grain control of contact characteristics but is expected to have lower power than the quasi-Poisson model. Exposure to a specialist lineage strain was associated with a lower rate of TB infections among close contacts (adjusted incidence rate ratio (IRR_{adj}) 0.86, CI_{95} 0.78 to 0.95, Wald P-value = 0.004) (**Table 2**, **Figure 2c**). We adjusted for a higher proportion of close contacts born in a high TB-incidence country as a confounder and found its effect on IGRA positivity to be significant (IRR_{adj} 1.98, CI_{95} 1.75 to 2.25, Wald P-value < 0.001). Host predictors of TB infection among close contacts that were significantly associated in the model included younger age and sputum smear positivity of the index case (**Table 2, Figure 2c**). In the logistic GEE, exposure to a specialist lineage was associated with a lower odds of TB infection among close contacts (OR_{adj} 0.82, CI_{95} 0.72 to 0.94, Wald P-value = 0.004), in agreement with the models above (**Suppl. Table S2**).

We similarly modeled the interplay of *Mtbc* lineage and host characteristics on the rate of secondary active TB among close contacts using a quasi-Poisson model. Secondary pulmonary and extrapulmonary cases were linked to the source case based on a Bayesian

posterior probability of transmission ≥0.5 estimated using Phybreak that uses genetic distance, sampling interval, and priors on incubation period and within-host pathogen dynamics **(methods)** (*14*). Phybreak estimated that 364 of the 2,279 pulmonary TB index cases had one close contact progress to active disease, 53 had two close contacts progress, and 10 had three progress (**Figure 2d**). Most of these transmitting TB index cases (77%, 322/427) were infected with a generalist *Mtbc* lineage strain. Exposure to a specialist lineage was associated with significantly lower risk of secondary active TB cases among close contacts controlling for index characteristics (IRR_{adj} 0.55, CI_{95} 0.32 to 0.90, Wald P-value = 0.023) (**Table 2, Figure 2c**). In the logistic GEE, the exposure to a specialist lineage had a similar trend towards lower risk of secondary active TB but was not statistically significant (OR_{adj} 0.77, CI_{95} 0.57 to 1.04, Wald P-value = 0.09) (**Suppl. Table S2**). We used the effect estimates of all variables of the quasi-Poisson model of disease progression to calculate a host infectiousness score (*12*). The infectiousness score did not differ between index cases infected with strains of specialist or generalist *Mtbc* lineages (Wilcoxon signed-rank test, p = 0.072), supporting the conclusion that host factors alone do not explain the observed differences in transmissibility (**Figure 2f**).

The above models only partially account for the effect of contact susceptibility on TB transmission and disease progression, as data on contact characteristics was limited for contacts in Hamburg and Amsterdam. We hence performed a secondary logistic GEE analysis using only NYC data to confirm the association between specialist *Mtbc* lineages and intrinsic transmissibility controlling for the additional characteristics of diabetes and HIV (n = 8,481 close contacts). The model confirmed young age, male sex, and diabetes as independent predictors of secondary active TB among close contacts (**Suppl. Table S3**). HIV status among contacts was not associated with secondary active TB possibly because the majority of contacts had well controlled HIV (78% of all estimated people with HIV in New York City were virally suppressed based on data from 2020) (*15*). Only four contacts reported using illicit drugs and hence this variable was not included in the model. Exposure

to a specialist *Mtbc* lineage was associated with lower odds of secondary active TB than estimated using the logistic GEE on the larger cohort with control for a subset of contact characteristics (OR_{adj} 0.54, CI_{95} 0.37 to 0.78, Wald P-value = 0.001 *vs.* OR_{adj} 0.77, CI_{95} 0.57 to 1.04, Wald P-value = 0.09). The effect estimate did not reach statistical significance for the model predicting TB infection (**Suppl. Table S3**).

We next evaluated the specific contributions of *Mtbc* sub-lineages on the observed differential transmissibility between specialist and generalist strains across the three cosmopolitan cities using the quasi-Poisson model. L1 strains were associated with a lower rate of TB infections and secondary active cases among close contacts compared to $L4_{gen}$ strains (IRR_{adj} 0.71, CI_{95} 0.60 to 0.85, P-value < 0.001, and 0.23, CI_{95} 0.05 to 0.69, P-value = 0.027 respective) (**Figure 2e, Suppl. Table S4**). Although L3 strains demonstrated high proportions of clustering based on genetic distance in the Netherlands (**Suppl. Figure S1b**), the adjusted incidence rate ratio of L3 strains was 0.41 (CI_{95} 0.16 to 0.88) indicating a lower rate of secondary active cases among close contacts compared with $L4_{gen}$ strains (P-value < 0.032, **Figure 2e, Suppl. Table S4**).

Allopatric host-pathogen relationships decrease susceptibility of TB infection

The observation of lower specialist transmissibility in Western cosmopolitan cities may relate to their poor ability to transmit among allopatric human populations, i.e., where there has not been long-term host-pathogen co-adaptation. To investigate this, we first mapped countries linked to specialist *Mtbc* lineages and assessed previously reported stable host ancestry-*Mtbc* lineage combinations among the TB index cases of the present analysis (**Figure 3a-b**) (*10*, *11*). Seventy-five percent of index cases were infected with a sympatric *Mtbc* specialist strain, significantly more than expected by chance (**Figure 3c-d**). To add validity to our definition of sympatry, we evaluated the geographic origin and lineage of 25,243 *Mtbc* strains with publicly available WGS (**Suppl. Figure S2a).** To test this co-adaptation hypothesis, we measured the transmissibility of specialists *Mtbc* lineages among exposed

sympatric *vs.* allopatric close contacts using contact place of birth as a proxy for host ancestry (**Methods**).

We used a logistic GEE to study the effect of contact sympatry on specialist *Mtbc* strain infection controlling for index and other contact characteristics. We excluded 108/12,749 close contacts with mixed ancestry, and 667/12,749 close contacts without information on country of birth. Contact allopatry was associated with a lower odds of *Mtbc* strain infection among close contacts of index cases infected with a specialist lineages strain (OR_{adj} for pooled L1,L3,$L2_{spec}$, $L4_{spec}$,L5,L6: 0.68, CI_{95} 0.5 – 0.92, n = 2,988, p = 0.014, **Figure 3e, Suppl. Figure S2b, Suppl. Table S5**) with similar additional host factors associated with *Mtbc* strain infection as measured above in models comparing specialists across all contacts. The effect of contact allopatry was extinguished when assuming the modern L2 sub lineages to be sympatric to East Asian ancestry, in line with evidence of modern L2's generalist phenotype and global spread (*16*). We found the association of allopatry to *Mtbc* strain infection to be robust to ancestry misclassification based on using place of birth in a sensitivity analysis (OR_{adj} 0.8, CI_{95} 0.65 – 0.99) (**Methods, Figure 3f**) including among the subset of data from NYC with additional contact susceptibility variables, diabetes, and HIV (**Suppl. Table S6**).

Human macrophages differ in susceptibility to infection with *Mtbc* L4 and *M. africanum* strains by ancestry

We looked for evidence of co-adaptation using *in vitro* macrophage infection models using strains of different *Mtbc* lineages. Human blood monocyte derived macrophages (MDMs) were obtained from donors of self-identified European Caucasian (Germany) or African ethnicities (Nigeria, Cameroon, Ghana, Sudan) (see **Suppl. Figure S3** for the approach). We compared the *Mtbc* phagocytosis/uptake and intracellular growth of three generalist L4 strains with that of three specialist *Mtbc* L6 strains, in each group we included the references strains: ATCC H37Rv (L4) and *M. africanum* ATCC reference strain (L6). For MDMs from

African donors, uptake did not differ between L4 and L6 (mean 34.2% vs 31.6%, respectively, $p = 0.66$, unpaired T-test) (**Figure 4a-b, i-j**), while European MDMs phagocytosed L6 strains at significantly lower proportions, i.e., on average 33.9% of L4 strains vs 2.9% of L6 strains ($p < 0.001$, unpaired T-test) (**Figure 4c-d, i-j**). Macrophages of either ancestry phagocytosed L4 strains at similar rates (**Figure 4a,c, i-j**).

We further investigated whether L4 or L6 strains differed in their intracellular growth kinetics after uptake into MDMs of either ancestry over time (4h, 24h, 96h, and 168h post infection). We ensured equal infection doses by comparing colony forming units (CFUs) of the inoculum (**Figure 4 f and h**). In MDM's from African donors, there was no significant difference in growth between L4 and L6 at two of the three time points (**Figure 4e-f**). However, growth in MDMs from African donors was higher for L6 compared with L4 strains at 24h pi ($p = 0.003$) (**Figure 4 e-f**). Conversely, in the MDM's from European Caucasian donors, L4 strains grew better than L6 at all the time points tested ($p < 0.001$ at 24h and 168h post infection, $p < 0.05$ at 96h, unpaired T-test and ANOVA with Bonferroni correction) (**Figure 4g-h**).

Overall, these results provide evidence that differential transmissibility between allopatric and sympatric pairs may relate to the observed weaker uptake and growth of specialist lineages in allopatric macrophages than increased growth and uptake of these lineages in sympatric macrophages. This suggests that co-adaptation predominantly diminishes infectivity of specialist *Mtbc* lineages in allopatric hosts.

DISCUSSION

In this study, we provide the first evidence for co-adaptation between humans and obligate respiratory pathogens of the *M. tuberculosis* complex in a controlled study. Our results demonstrate intrinsically reduced transmissibility of strains of specialist *Mtbc* lineages compared with strains of generalists *Mtbc* lineages in three cosmopolitan cities. While strains of specialist lineages are able to infect sympatric hosts, they display lower transmissibility in allopatric host populations. We validate our epidemiological observations using a macrophage infection model, finding evidence that the reduced transmissibility of specialist strains may relate to lower uptake and growth of *Mtbc* in allopatric macrophages after initial exposure. Our central discovery is that the reduced infectivity of strains of specialist *Mtbc* lineages in allopatric populations is likely responsible for their limited success in driving the global spread of TB compared to strains of modern generalist L2/L4 lineages.

Our results support the theory that interactions between the human immune system and pathogens have genetically shaped both populations over time (*1*, *17*). The recent wide adoption of high throughput sequencing of human and pathogen populations has provided a new approach to disentangling natural host-pathogen interactions. Specifically, this approach has provided molecular evidence of co-adaptation between humans and several pathogen species, e.g. *P. falciparum*, *H. pylori*, and *Y. pestis* with differing effects on the directionality of sympatry towards more or less pathogenicity (*2*, *18*, *19*). In *P. falciparum*, sympatric human-pathogen co-adaptation results in apparent loss of the protective effect of the HbS allele on disease severity, i.e., evolution towards higher pathogen virulence (*2*). While colonization with allopatric *H. pylori* strains was correlated with severe gastric cancer, sympatric *H. pylori* infections carry the best prognosis (*18*). In contrast, in *Mtbc*, we discover that co-adaptation decreases susceptibility to infection among allopatric populations without affecting the rate of transmission in sympatric populations.

Recent work in human paleogenomics characterized the bulk of European genetic adaptation to have occurred in response to infectious diseases after the start of the Bronze Age less than 4,500 years ago (*17*). The date of emergence of modern *Mtbc* lineages is estimated to have occurred at a similar time scale (4449 years before present) (*4*, *20*, *21*). Along with other infectious diseases with high mortality, *Mtbc* infection likely contributed to shaping human genetic evolution over this time period; and given the obligate human niche, *Mtbc* was reciprocally shaped in a process of host-pathogen genetic co-adaptation.

Elucidating molecular mechanisms of resistance to *Mtbc* infection will take time, but our findings have immediate implications for public health practice. Although several patient risk factors (e.g., age, sex, HIV, and diabetes) are currently recognized to influence individual risk of TB infection and disease after exposure, they are rarely used quantitatively in practice to support decisions in contact tracing or the initiation of chemoprophylaxis after exposure. Here, we demonstrate that allopatric infections constitute about 52% of total contact exposures to strains of specialist *Mtbc* lineages in low prevalence settings and influence the odds of infection to a similar extent as recognized risk factors such as index case smear grade. The incorporation of *Mtbc* sympatry for specialists *Mtbc* lineages as well as host factors into the baseline assessment of risk using statistical models may help refine the use of chemoprevention and guide contact tracing activities, e.g., triage patient exposures to lower risk allopatric exposures vs higher risk sympatric exposures in public health practice.

The presence of measurable co-adaptation between humans and *Mtbc* has implications for infection prevention and vaccine design. Characterizing the molecular mechanisms of co-adaptation within both the host and the pathogen could help elucidate targets of protection from infection. This can fill a needed knowledge gap. Despite early work linking TB susceptibility to human genetic diversity there have not been consistent signals of association between human gene variants and TB infection or progression (*22*, *23*). A recent large study implicated Major Histocompatibility Complex Class II variants to be associated

with TB susceptibility (*24*), a genetic region that plays a key role in immune responses and host pathogen interactions. Other variants that increase the genetic risk of TB infection were identified in genes involved in dendric cell migration (*25*) and tissue regeneration (*26*). These studies have rarely controlled for pathogen genotype and, along with the lack of a coherent signal of genetic risk for TB, highlight the need for multinational studies that also represent pathogen diversity.

The investigation of co-adaptation between humans and *Mtbc* strains requires a diverse cohort that ensures both infections with sympatric and allopatric strains are adequately represented in the sample. This study was enabled by the availability of robust TB case data, pathogen WGS, and contact tracing data and illustrates that these data can successfully be repurposed and pooled across several public health systems. Yet, this study was not without limitations. The TB index case and contact information relied on social contact tracing that may be subject to recall bias. Information on contacts of TB cases is not routinely collected in Hamburg and Amsterdam and as a result some key risk factors of TB susceptibility (contact HIV, contact diabetes) were not available. The analysis of secondary active TB relied on modeling rather than epidemiological linkage which may include false positive transmission links, and some contacts of TB index cases may not be visible to the public health departments due to contact relocation between cities or states. We used country of birth, or in case of minors, the parent's countries of birth, as a surrogate for ancestry. We performed several sensitivity analyses to address these limitations. We randomly assigned the sympatry label for 14% of our sample (the proportion of Dutch residents who are foreign-born, Statistics Netherlands) to account for ancestry misclassification. Estimated transmission links were only accepted based on sample size and support value thresholds that were shown to provide robust effect estimates in our study and other work (*27*). We built models of transmission at contact level in a subset of data where HIV and diabetes information was available, as these are important effect modifiers of TB susceptibility.

In summary, we demonstrate co-adaptation between strains of specialist *Mtbc* lineages and host ancestry that manifests in lower infectivity in allopatric host-pathogen combinations. This suggests that strains of generalist lineages are likely to increase in prevalence relative to strains of specialist lineages in a progressively globalized world. Future work should aim to understand the molecular underpinnings of allopatric versus sympatric infections.

METHODS

Study design, participants, and associated metadata

We harmonized TB surveillance data from public health departments tracing TB in three large cities Amsterdam, Hamburg, and New York City, where TB is reportable to the respective public health agencies**.** The data consists of clinical characteristics and pathogen whole genome sequence data of TB index cases, as well as information on their close contacts including exposure time (close contact versus non-close contact), demographics, relevant comorbidities, and Interferon Gamma Release Assay (IGRA) results. We detail each TB surveillance sample by site below:

Amsterdam and the Netherlands: The National Institute for Public Health and the Environment (RIVM) in Bilthoven, The Netherlands, contains the Mycobacterial reference laboratory for the diagnosis of all TB cases in the Netherlands. Since retrospective data in this study was de-identified by name, WGS data from the RIVM and clinical and demographic data for patients were linked based on sex, date of birth, year of diagnosis and postal code. All notified *M. tuberculosis* culture-positive cases from the city of Amsterdam between December 2015 and December 2019 and their contacts were included. In addition, we obtained information on all active cases diagnosed elsewhere in The Netherlands in this period. In the case of patients with multiple isolates, only the isolate with the earliest date of diagnosis was included. Close contact was defined as daily contact in a room of in total ≥48 hours or ≥6 hours in a car.

Hamburg: We obtained data from the Public Health Office Hamburg, Germany for TB index cases from 2008 through 2017. Sequencing of Hamburg cases was performed at the National Reference Center for Mycobacteria at the Research Center Borstel. Close contact was defined as a contact time of ≥8 hours.

New York City: The New York City Department of Health and Mental Hygiene notifies, and traces contacts of all infectious TB index cases. We obtained clinical and demographic data of TB index cases diagnosed between 2016 and 2020 and their contacts. Pathogen WGS is performed at the New York State Wadsworth Center in Albany, NY, for culture-positive cases. In the case of patients with multiple isolates, only the isolate with the earliest date of diagnosis was included. Close contact was defined as a contact time of ≥8 hours per week.

We compared characteristics of index cases and their contact group networks to explore comparability of contact tracing activity across the three cities' public health departments. Across the three cities, index case age positively correlated with mean contact group age (pooled Pearson's correlation coefficient 0.4 [CI_{95} 0.39 - 42], and the count of TB infected close contacts correlated moderately with the contact group size (Pearson's 0.5 [CI_{95} 0.53 – 0.56]) (**Figure S4a-b**). We excluded extrapulmonary TB cases from the source case analysis due to their recognized significantly lower infectiousness (n = 903/5,256,17%). We excluded strains not typed as *Mtb* sensu stricto or *M. africanum* (i.e., *M. canettii* or *M. bovis*, n = 132/5,219, 0.03%). Among contacts, we excluded those without blood test result for *Mtbc* infection (Interferon Gamma Release Assay, or IGRA, 12%, 3,483/28,889), excluded contacts not considered close (42%, 12,228/28,889) and contacts not linkable to an index case (17%, 4,891/28,889).

Index case and contact data linkage

Demographics, comorbidity information, and data on social risk factors for TB were extracted from the three public health databases. The investigated contacts were considered to have TB infection if their IGRA values were above 0.35 IU/ml based on existing test specifications. Contacts were linked to TB index cases by epidemiological investigation. This linkage of contacts and active cases is coded as a unique identifier that was subsequently used to compute aggregated contact group information (count and percent of high-incidence country birthplaces) per index case. Conversely, index case information was linked to the contact

data (age, sex, sputum microscopy, co-morbidities) as these were important covariates to control for in the modeling.

Ethical approval

The study is approved by the Institutional Review Boards at Harvard Medical School (IRB19-1910), the University of Lübeck (AZ-19-071), the New York City Department of Hygiene and Mental Health (20-078), and the Registration Committee of the Netherlands Tuberculosis Register NTR (11-2019).

Phenotypic drug resistance

At RIVM (The Netherlands), the Research Center Borstel (Germany) and the New York State Wadsworth Center phenotypic drug susceptibility testing was performed with the MGIT 960 method using the World Health Organization recommended critical concentrations of 1mg/L for rifampicin and 0.1 mg/L for isoniazid.

Whole Genome Sequencing and analysis

At the time of writing, whole genome sequencing of all culture positive TB cases is routinely performed in the Netherlands and New York State since 2016.

Isolates from the Netherlands were sequenced at the RIVM Bilthoven as follows: genomic DNA was extracted from positive MGIT culture tubes (Becton Dickinson, NJ, USA) using the QIAamp DNA mini kit (QIAGEN GmbH, Hilden, Germany) and sequencing was performed on Illumina HiSeq2500 with 2 x 125bp reads following Nextera XT library preparation.

At the Wadsworth Center of the New York State Department of Health, genomic DNA was extracted from MGIT culture tubes (Becton Dickinson, NJ, USA) using the InstaGene/FastPrep method as previously described (*28*). Paired-end 250-bp DNA sequencing was carried out using the Illumina MiSeq platform (Illumina, San Diego, CA, USA) after Nextera XT library preparation.

Patient isolates from Hamburg were sequenced for a previous research project at the Research Center Borstel as follows (*29*): Genomic DNA was extracted from positive MGIT culture tubes (Becton Dickinson, NJ, USA) using the QIAamp DNA mini kit method (QIAGEN GmbH, Hilden, Germany). Sequencing libraries were constructed from genomic DNA using a modified Illumina Nextera protocol for 2 x 151 bp reads on the Illumina NextSeq 500 (Illumina, San Diego, CA, USA).

Raw sequencing reads across the three sites were pooled and analyzed on Harvard Medical Schools High Performance Cluster O2. Analysis and quality control was performed using the same bioinformatics pipeline. Reads were first trimmed to remove low quality bases and adapter sequences using *fastp* v0.20.1 (*30*). The trimmed reads were aligned to the reference genome H37Rv (accession NC_000962.3) using *BWA mem* v0.7.17 (parameters: default) (*31*). Duplicate reads were removed with *Picard* v2.25.1 (http://broadinstitute.github.io/picard/). We called single nucleotide substitutions (SNSs) and indels with *Pilon* v1.24 (parameters: default) requiring confidence base calls (*pass* filter) (*32*).

Variable position alignment and phylogenetic analysis

To analyze the genetic variation in our sample we created a genotypes matrix based on the variant call format files. We detected SNS sites at 347,471 H37Rv reference positions (of which 9,641 SNSs were not biallelic) among our global sample of 5,219 strains. We constructed a 347,471x5,219 genotypes matrix (coded as 0:A, 1:C, 2:G, 3:T, 9:Missing) and filled in the matrix for the allele supported at each SNS site (row) for each isolate. We excluded 7,406 SNS sites that had an Empirical Base-level Recall (EBR) score of <0.9 (*33*), 3,555 SNS sites in mobile genetic elements (e.g., transposases, integrases, phages, or insertion sequences), and 933 SNS sites located in overlapping coding sequences. We next dropped 10,169 SNS sites that are missing in >10% of our strains. An additional 133 SNS sites were dropped because the minor allele is present in <1 isolate. If a base call at a

specific reference position for an isolate did not meet the filter criteria that allele was coded as missing. Our filtered genotypes matrix had dimensions of 325,275x5,219 SNSs representing 325,275 SNS sites across 5,219 strains. The phylogenetic lineage was determined using a refined lineage-calling scheme based on 96-SNSs with *fast-lineage-caller* (*5*).

To infer the population structure of our strains we merged the variant call format files using *bcftools* v1.12 (*34*) and removed regions with repetitive sequence content with low EBR scores (*33*), mobile genetic elements, coding for genes involved in antibiotic resistance as well as all invariant sites. A multiple sequence alignment of 297,061characters found in 5,087 *Mtbc* L1-L6 strains was created using *vcf2phylip* (v1.5, https://doi.org/10.5281/zenodo.1257057) and fed to *iqtree* v2.1.2 for phylogenetic inference (*35*). The final maximum likelihood tree was inferred using a general time reversible model of nucleotide substitution, a Gamma model of rate heterogeneity, empirical base frequencies (GTR+F+R2) and 1000 ultrafast bootstrap replicates. The tree was annotated using iTol (*36*).

Definition of clusters based on genetic distance

Mtbc strains from patients linked by transmission using epidemiological data rarely differ by more than 5 SNSs (*29*, *37*). To define clustered strains in our dataset, we first excluded the small number of non-biallelic SNSs to simplify distance calculation. Of the 325,275 SNS sites, 316,274 were biallelic, i.e., had only a single alternate allele among our strains. We calculated the pairwise distance as the Hamming distance of SNSs across the total 5,219 strains and subset this matrix for to obtain one pairwise distance matrix for each *Mtbc* lineage and city.

We next performed graph-based clustering using the *gengraph* function implemented in the adegenet R package v2.1.5 (*38*) for each *Mtbc* lineage and city specific pairwise distance

matrix as we were only interested in transmission within each city. We used a 12 SNS cut-off as possible transmission with single linkage clustering, i.e., strains in a cluster differed at most 12 SNSs from at least one other isolate in the cluster.

Index case inference and infector/infectee relationships

We used the R package phybreak v0.5.2 to identify the infector case for each active TB case in a cluster (*14*). Cases that were not in clusters were assumed to not have secondary active TB cases. (If contacts of TB cases moved to a different city and developed secondary active TB this would not have been detected using this approach.) Phybreak identifies infector-infectee relationships by combining models for transmission, case observation, within-host pathogen dynamics, and mutation assuming that all cases of a transmission cluster have been sequenced. The most likely transmission tree is sampled from the posterior distribution through Bayesian inference and Monte Carlo Markov Chain sampling. The priors for the mean and shape of the sampling interval (time between infection and sampling of a case) and generation interval (i.e., time between infection of a primary case and a secondary case by this primary case) were assumed to follow a gamma distribution as in previous work (*27*, *39*). We ran five independent MCM chains with 10,000 cycles of burn-in and 50,000 sampling cycles for each cluster. We dropped each estimated infector-infectee relationship where the effective sample size was <200 which can occur if sampling dates occur within a short period of time. We excluded active cases that did not have an infector with a posterior probability (PP) of at least 0.5 from further analysis after a sensitivity analysis evaluating different PP values in the transmission models (**Suppl. Fig S4**).

Statistical analysis and models of transmission

The python libraries *pandas* v1.3.5 (*40*) and *numpy* v1.21.5 (*41*) were used to manipulate epidemiological data. All statistical analyses and visualizations were performed in R v4.1.2 (*42*) using libraries included in the *tidyverse* v.1.3.1 (*43*). Effect estimates of regression analyses were visualized using the *sJPlot* package 2.8.10. Missing values for the covariates

HIV coinfection and alcohol were imputed using a bootstrapped logistic regression model implemented in the package *mice* v3.14.0 (*44*). We coded age into brackets (0 to 14, 15 to 24, 25 to 34, 35 to 44, 45 to 54, 55 to 64, 65 to 74, and 75 and above).

We used a multivariate Quasi-Poisson rate regression model (Eq1) to identify predictors of the number of secondary active TB cases or *Mtbc* infections among close contacts. The group size of close contacts per index case was used as the offset to denote the exposed group in the Poisson regression. Adjusted incidence rate ratios are presented with 95% confidence intervals.

Eq1:

$$\begin{aligned}\boldsymbol{log}(Count\ TB\ infections_{index}) \\ &= \beta_0 + \beta_1 age_{index} + \beta_2 sex_{index} + \beta_3 sputum\ smear_{index} + \beta_4\ alcohol_{index} \\ &+ \beta_5 drug\ use_{index} + \beta_6 high\ incidence\ country_{contact} \\ &+ \beta_7 one\ child\ contact_{index} + \beta_8 Mtb\ lineage_{index} \\ &+ log(offset(contact\ group\ size))\end{aligned}$$

Generalized estimating equations (GEE) were used to validate the Quasi-Poisson models and to quantify the effect of host-pathogen sympatry on TB infection for each close contact individually (Eq2). The *gee* package v4.13-20 was used for these analyses with the option ‘exchangeable’ to imply that any two observations per contact group are equally correlated without correlation between observations from different contact groups (*45*). This allowed to adjust the effect estimates for correlations in the data. Adjusted odds ratios and 95% CI are presented.

Eq2:

$$\boldsymbol{log}\left[\frac{P(ContactTBinfection)}{1-P(ContactTBinfection)}\right]$$

$$= \beta_0 + \beta_1 age_{contact} + \beta_2 sex_{contact} + \beta_3 diabetes_{contact} + \beta_4\, HIV_{contact}$$

$$+ \beta_5 high\ incidence\ country_{contact} + \beta_6 sputum\ smear_{index} + \beta_7\, alcohol_{index}$$

$$+ \beta_8 drug\ use_{index} + \beta_9 sympatric\ exposure_{contact}$$

Frequencies or proportions of binary variables were compared using Chi Squared, binary to continuous variables were compared using the Wilcoxon Rank Sum test, and continuous data was compared using ANOVA or t-tests. All statistical tests were two-sided and a *P* value cut-off of 0.05 corrected for multiple testing with the false discovery rate.

Host infectiousness scores

The adjusted estimates from the multivariate Quasi-Poisson regression model were each multiplied as weights to calculate each index case's infectiousness score, similar to previous work on the propensity to propagate (*12*).

Human macrophage donors

Blood donation was approved by the ethics committee of the Research Center Borstel (RCB) (Borstel, Germany) and the University of Luebeck (Luebeck, Germany) (AZ-19-071). A total of ~ 300 ml of whole blood (heparin- anti-coagulated) was obtained via venepuncture by the blood service unit at the RCB. Human healthy donors were from the European Caucasian and African ethnicities with a medical check-up no longer than two years ago.

***Mtbc* strains and culture conditions**

A panel of six *Mtbc* strains representing the modern and ancient human-adapted clades 1 and 2 (L4 and L6) was selected from a reference collection at the RBC. Strains were recovered from clinical samples or provided by the American Type Culture Collection

(ATCC). Strains were characterized by drug susceptibility testing and sequencing as previously described (*46*, *47*). *Mtbc* strains were grown on Loewenstein-Jensen (LJ) medium and further passaged in 10 ml of mycobacteria liquid media (MLM): Middlebrook 7H9 broth (Becton Dickinson) supplemented with 10% (w/v) oleic acid albumin dextrose catalase (OADC), 0.2% (w/v) glycerol, 0.05% (w/v) Tween 80. *Mtbc* bacilli were grown shaking 100 revolutions per minute (rpm) until mid-log phase at an Optical Density $(OD)_{600nm}$ of 0.5 - 0.6 at 37 °C, harvested, and stored as stocks in MLM at -80 °C. Precultures were performed in 30 ml square media bottles (Nalgene) by inoculating 1000 µl of unfrozen *Mtbc* stock in 9 ml of MLM and were grown to shake (100 rpm) until mid-log phase (OD_{600nm} of 0.5 - 0.6) at 37 °C, harvested, and stored in MLM with 20% (w/v) of glycerol at -80 °C. After three weeks of storage, six aliquots per isolate were selected randomly and titrated in triplicates by Colony-Forming Unit (CFU) assays onto Middlebrook 7H10 agar (Becton Dickinson) plates supplemented with 10% (w/v) heat-inactivated bovine serum and cycloheximide (10 mg/ml), followed by incubation at 37 °C for 3-4 weeks. CFU (cells /ml) was calculated as colonies count*dilution factor/volume plated (µl).

In vitro generation of macrophages

Blood-derived human European Caucasian and African monocytes were obtained by density gradient centrifugation and positive immunomagnetic separation with anti-human CD14 antibodies (Milteny Biotec, Germany) following manufacturer instructions. Isolated monocytes had a purity and viability greater than 85-90% which was assessed by flow cytometry and electronic current exclusion technology with CASY. Monocytes were seeded in 24 well plates at a density of $0.5x10^6$ cells in 500 µl of Macrophage Culture Media (MCM): RPMI 1640 (Biochrom), penicillin/streptomycin 1% (Biochrom), L-glutamine 1% (Biochrom) with AB + heat-inactivated human serum 1% and further differentiated into monocyte-derived macrophages (MDM) after seven days of culture in the presence of 10 ng/ml of human recombinant macrophage colony stimulating factor (M-CSF). Cells were incubated at 37°C /5% CO_2, and MCM was changed every three days until cells achieved ~ 80% confluence.

Cell infection assays

Mtbc stocks previously grown at mid-log phase (OD_{600} 0.5-0.6) and CFU titrated were unfrozen for 10 min in a heat block at 37 °C and centrifugated at 4500 rpm (Eppendorf 5427 R FA-45-24-11 max. 16220 rpm) for 10 min at 4 °C to remove medium traces. Next, supernatants were removed, and the respective *Mtbc* bacilli pellet was resuspended in 1 ml of Macrophage Infection Media (MIM) by passing through a 1 ml syringe with a 25-gauge needle ten times. MIM compromised 1% L- Glutamine (Biochrom), 10% FSC (Biochrom), in RPMI- 1640 (Biochrom). Infection bacterial suspension (*Mtbc* infectious dose) was prepared per strain in MIM at a multiplicity of infection (MOI) ~1:1 based on the CFU titration of the frozen stocks and the number of assay wells per plate. Each experimental setup comprised four plates for different time points (4 h, 24 h, 96 h, and 168 h post-infection (hpi)). 0.5×10^6 *Mtbc* bacilli per 500 µl MIM were used to infect 0.5×10^6 MDM per well. MIM only was replaced for the negative infection control. Plates wereincubated at 37 °C in 5% CO_2 for 4 h. CFU assays were performed to titrate the infectious dose. The percentage of *Mtbc* uptake was calculated as colonies count at 4hpi*100%/colonies count of infectious dose.

Before bacteria harvesting, the morphological characteristics of live macrophages were visually inspected with an inverted light microscope after completing 4 hpi and over the infection period of 7 days. Each time point included three replicates per *Mtbc* strain and three macrophages no infected wells. In all *in vitro* cell infection experiments, the number of intracellular *Mtbc* bacilli was determined by CFU assays immediately after phagocytosis (4hpi) and at different time points (24 hpi, 96 hpi, and 168 hpi) by lysing the monolayer with sodium dodecyl sulfate (SDS) 0.05% (w/v) in double-distilled water (ddH_2O). Serial 10-fold dilutions of the infectious macrophage lysates were then plated onto Middlebrook 7H10 agar plates in triplicates supplemented with 10% (w/v) heat-inactivated bovine serum and cycloheximide (10 mg/mL) for CFU counting. and

Three independent European macrophage infection experiments were performed of which one comprised one single donor. Due to insufficient cell yield the other two experiments

compromised two donors each. Of these, the first donor was infected with L4 and the second donor with L6 including H37Rv as a reference in all experiments. These data are represented together in one figure as no significant differences in H37Rv uptake were found among donors (**Suppl. Figure S6)** nine technical replicates of infections, and twenty-seven CFU technical replicates per strain). Four independent African macrophage infection experiments were conducted of which each compromised one single donor infected with L4 and L6 (Twelve technical replicates of infection, and thirty-six CFU technical replicates per strain)

Data availability

Bioprojects and accessions for each fastq file have been requested and will be available in Supplementary table X.

Acknowledgements

We thank Pascal Lapierre, Han de Neeling, and Thomas Kohl for transferring sequence data, Henrieke Schimmel and Erika Slump for help with data linkage of the Dutch dataset, the Dutch TB Registry (NTR commissie) as well as all study participants. We thank the Wadsworth Center Applied Genomics Technology Cluster for whole genome sequencing

Author contributions

MRF, and MIG conceived the idea for the epidemiological analysis. SN, SH, FJPL conceived the idea for the *in vitro* experiments. MRF supervised the project. MIG performed data curation and data analysis. MIG and MRF wrote the first draft. FJPL performed data curation and data analysis. RVJr and DK carried out data analysis. LT, PK, RD carried out data acquisition. VE, KM, JSM, SH, DvS, SA, SN supervised data acquisition and curation. WS and BM critically reviewed the drafts. All authors reviewed the draft and assisted in the manuscript preparation.

Funding

This work was funded by National Institutes of Health / National Institute of Allergy and Infectious Diseases R21 AI154089 to MRF. MIG is supported by the German Research Foundation (GR5643/1-1). FJPL, SN, and SH were funded by the Leibniz Science Campus EvoLUNG (Evolutionary Medicine of the Lung) http://evolung.fz-471borstel.de/, grant number W47/2019. The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

REFERENCES

1. M. E. J. Woolhouse, J. P. Webster, E. Domingo, B. Charlesworth, B. R. Levin, Biological and biomedical implications of the co-evolution of pathogens and their hosts. *Nat. Genet.* **32**, 569–577 (2002).

2. G. Band, E. M. Leffler, M. Jallow, F. Sisay-Joof, C. M. Ndila, A. W. Macharia, C. Hubbart, A. E. Jeffreys, K. Rowlands, T. Nguyen, S. Gonçalves, C. V. Ariani, J. Stalker, R. D. Pearson, R. Amato, E. Drury, G. Sirugo, U. d'Alessandro, K. A. Bojang, K. Marsh, N. Peshu, J. W. Saelens, M. Diakité, S. M. Taylor, D. J. Conway, T. N. Williams, K. A. Rockett, D. P. Kwiatkowski, Malaria protection due to sickle haemoglobin depends on parasite genotype. *Nature* (2021), doi:10.1038/s41586-021-04288-3.

3. I. Comas, M. Coscolla, T. Luo, S. Borrell, K. E. Holt, M. Kato-Maeda, J. Parkhill, B. Malla, S. Berg, G. Thwaites, D. Yeboah-Manu, G. Bothamley, J. Mei, L. Wei, S. Bentley, S. R. Harris, S. Niemann, R. Diel, A. Aseffa, Q. Gao, D. Young, S. Gagneux, Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans. *Nat. Genet.* **45**, 1176–1182 (2013).

4. D. Brites, S. Gagneux, Co-evolution of Mycobacterium tuberculosis and Homo sapiens. *Immunol. Rev.* **264**, 6–24 (2015).

5. L. Freschi, R. Vargas Jr, A. Husain, S. M. M. Kamal, A. Skrahina, S. Tahseen, N. Ismail, A. Barbova, S. Niemann, D. M. Cirillo, A. S. Dean, M. Zignol, M. R. Farhat, Population structure, biogeography and transmissibility of Mycobacterium tuberculosis. *Nat. Commun.* **12** (2021), doi:10.1038/s41467-021-26248-1.

6. L. Baker, T. Brown, M. C. Maiden, F. Drobniewski, Silent nucleotide polymorphisms and a phylogeny for Mycobacterium tuberculosis. *Emerg. Infect. Dis.* **10**, 1568–1577 (2004).

7. A. Asante-Poku, D. Yeboah-Manu, I. D. Otchere, S. Y. Aboagye, D. Stucki, J. Hattendorf, S. Borrell, J. Feldmann, E. Danso, S. Gagneux, Mycobacterium africanum is associated with patient ethnicity in Ghana. *PLoS Negl. Trop. Dis.* **9**, e3370 (2015).

8. D. Stucki, D. Brites, L. Jeljeli, M. Coscolla, Q. Liu, A. Trauner, L. Fenner, L. Rutaihwa, S. Borrell, T. Luo, Q. Gao, M. Kato-Maeda, M. Ballif, M. Egger, R. Macedo, H. Mardassi, M. Moreno, G. Tudo Vilanova, J. Fyfe, M. Globan, J. Thomas, F. Jamieson, J. L. Guthrie, A. Asante-Poku, D. Yeboah-Manu, E. Wampande, W. Ssengooba, M. Joloba, W. Henry Boom, I. Basu, J. Bower, M. Saraiva, S. E. G. Vaconcellos, P. Suffys, A. Koch, R. Wilkinson, L. Gail-Bekker, B. Malla, S. D. Ley, H.-P. Beck, B. C. de Jong, K. Toit, E. Sanchez-Padilla, M. Bonnet, A. Gil-Brusola, M. Frank, V. N. Penlap Beng, K. Eisenach, I. Alani, P. Wangui Ndung'u, G. Revathi, F. Gehre, S. Akter, F. Ntoumi, L. Stewart-Isherwood, N. E. Ntinginya, A. Rachow, M. Hoelscher, D. M. Cirillo, G. Skenders, S. Hoffner, D. Bakonyte, P. Stakenas, R. Diel, V. Crudu, O. Moldovan, S. Al-Hajoj,

L. Otero, F. Barletta, E. Jane Carter, L. Diero, P. Supply, I. Comas, S. Niemann, S. Gagneux, Mycobacterium tuberculosis lineage 4 comprises globally distributed and geographically restricted sublineages. *Nat. Genet.* **48**, 1535–1543 (2016).

9. M. Merker, C. Blin, S. Mona, N. Duforet-Frebourg, S. Lecher, E. Willery, M. G. B. Blum, S. Rüsch-Gerdes, I. Mokrousov, E. Aleksic, C. Allix-Béguec, A. Antierens, E. Augustynowicz-Kopeć, M. Ballif, F. Barletta, H. P. Beck, C. E. Barry 3rd, M. Bonnet, E. Borroni, I. Campos-Herrero, D. Cirillo, H. Cox, S. Crowe, V. Crudu, R. Diel, F. Drobniewski, M. Fauville-Dufaux, S. Gagneux, S. Ghebremichael, M. Hanekom, S. Hoffner, W.-W. Jiao, S. Kalon, T. A. Kohl, I. Kontsevaya, T. Lillebæk, S. Maeda, V. Nikolayevskyy, M. Rasmussen, N. Rastogi, S. Samper, E. Sanchez-Padilla, B. Savic, I. C. Shamputa, A. Shen, L.-H. Sng, P. Stakenas, K. Toit, F. Varaine, D. Vukovic, C. Wahl, R. Warren, P. Supply, S. Niemann, T. Wirth, Evolutionary history and global spread of the Mycobacterium tuberculosis Beijing lineage. *Nat. Genet.* **47**, 242–249 (2015).

10. S. Gagneux, K. DeRiemer, T. Van, M. Kato-Maeda, B. C. de Jong, S. Narayanan, M. Nicol, S. Niemann, K. Kremer, M. Cristina Gutierrez, M. Hilty, P. C. Hopewell, P. M. Small, Variable host–pathogen compatibility in Mycobacterium tuberculosis. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 2869–2873 (2006).

11. M. B. Reed, V. K. Pichler, F. McIntosh, A. Mattia, A. Fallow, S. Masala, P. Domenech, A. Zwerling, L. Thibert, D. Menzies, K. Schwartzman, M. A. Behr, Major Mycobacterium tuberculosis lineages associate with patient country of origin. *J. Clin. Microbiol.* **47**, 1119–1128 (2009).

12. H. Nebenzahl-Guimaraes, M. W. Borgdorff, M. B. Murray, D. van Soolingen, A novel approach - the propensity to propagate (PTP) method for controlling for host factors in studying the transmission of Mycobacterium tuberculosis. *PLoS One*. **9**, e97816 (2014).

13. World Health Organization, “Global Tuberculosis Report 2017” (World Health Organization, 2017).

14. D. Klinkenberg, J. A. Backer, X. Didelot, C. Colijn, J. Wallinga, Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. *PLoS Comput. Biol.* **13**, e1005495 (2017).

15. “HIV surveillance annual report 2020” (New York City Department of Health and Mental Hygiene, 2020), (available at chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www1.nyc.gov/assets/doh/downloads/pdf/dires/hiv-surveillance-annualreport-2020.pdf).

16. K. E. Holt, P. McAdam, P. V. K. Thai, N. T. T. Thuong, D. T. M. Ha, N. N. Lan, N. H. Lan, N. T. Q. Nhu, H. T. Hai, V. T. N. Ha, G. Thwaites, D. J. Edwards, A. P. Nath, K. Pham, D. B. Ascher, J. Farrar, C. C. Khor, Y. Y.

Teo, M. Inouye, M. Caws, S. J. Dunstan, Frequent transmission of the Mycobacterium tuberculosis Beijing lineage and positive selection for the EsxW Beijing variant in Vietnam. *Nat. Genet.* **50**, 849–856 (2018).

17. G. Kerner, A.-L. Neehus, L. Abel, J.-L. Casanova, E. Patin, G. Laval, L. Quintana-Murci, Genetic adaptation to pathogens and increased risk of inflammatory disorders in post-Neolithic Europe. *bioRxiv* (2022), , doi:10.1101/2022.07.02.498543.

18. N. Kodaman, A. Pazos, B. G. Schneider, M. B. Piazuelo, R. Mera, R. S. Sobota, L. A. Sicinschi, C. L. Shaffer, J. Romero-Gallo, T. de Sablet, R. H. Harder, L. E. Bravo, R. M. Peek Jr, K. T. Wilson, T. L. Cover, S. M. Williams, P. Correa, Human and Helicobacter pylori coevolution shapes the risk of gastric disease. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 1455–1460 (2014).

19. A. Immel, F. M. Key, A. Szolek, R. Barquera, M. K. Robinson, G. F. Harrison, W. H. Palmer, M. A. Spyrou, J. Susat, B. Krause-Kyora, K. I. Bos, S. Forrest, D. I. Hernández-Zaragoza, J. Sauter, U. Solloch, A. H. Schmidt, V. J. Schuenemann, E. Reiter, M. S. Kairies, R. Weiß, S. Arnold, J. Wahl, J. A. Hollenbach, O. Kohlbacher, A. Herbig, P. J. Norman, J. Krause, Analysis of genomic DNA from medieval plague victims suggests long-term effect of Yersinia pestis on human immunity genes. *Mol. Biol. Evol.* **38**, 4059–4076 (2021).

20. K. I. Bos, K. M. Harkins, A. Herbig, M. Coscolla, N. Weber, I. Comas, S. A. Forrest, J. M. Bryant, S. R. Harris, V. J. Schuenemann, T. J. Campbell, K. Majander, A. K. Wilbur, R. A. Guichon, D. L. Wolfe Steadman, D. C. Cook, S. Niemann, M. A. Behr, M. Zumarraga, R. Bastida, D. Huson, K. Nieselt, D. Young, J. Parkhill, J. E. Buikstra, S. Gagneux, A. C. Stone, J. Krause, Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. *Nature*. **514**, 494–497 (2014).

21. C. S. Pepperell, A. M. Casto, A. Kitchen, J. M. Granka, O. E. Cornejo, E. C. Holmes, B. Birren, J. Galagan, M. W. Feldman, Correction: The role of selection in shaping diversity of natural M. tuberculosis populations. *PLoS Pathog.* **9** (2013), doi:10.1371/annotation/cff22061-44d5-4301-b853-41702d160203.

22. E. A. van der Eijk, E. van de Vosse, J. P. Vandenbroucke, J. T. van Dissel, Heredity versus environment in tuberculosis in twins: the 1950s United Kingdom Prophit Survey Simonds and Comstock revisited. *Am. J. Respir. Crit. Care Med.* **176**, 1281–1288 (2007).

23. G. W. Comstock, Tuberculosis in twins: a re-analysis of the Prophit survey. *Am. Rev. Respir. Dis.* **117**, 621–624 (1978).

24. G. Sveinbjornsson, D. F. Gudbjartsson, B. V. Halldorsson, K. G. Kristinsson, M. Gottfredsson, J. C. Barrett, L. J. Gudmundsson, K. Blondal, A. Gylfason, S. A. Gudjonsson, H. T. Helgadottir, A. Jonasdottir, A.

Jonasdottir, A. Karason, L. B. Kardum, J. Knežević, H. Kristjansson, M. Kristjansson, A. Love, Y. Luo, O. T. Magnusson, P. Sulem, A. Kong, G. Masson, U. Thorsteinsdottir, Z. Dembic, S. Nejentsev, T. Blondal, I. Jonsdottir, K. Stefansson, HLA class II sequence variants influence tuberculosis risk in populations of European ancestry. *Nat. Genet.* **48**, 318–322 (2016).

25. J. Curtis, Y. Luo, H. L. Zenner, D. Cuchet-Lourenço, C. Wu, K. Lo, M. Maes, A. Alisaac, E. Stebbings, J. Z. Liu, L. Kopanitsa, O. Ignatyeva, Y. Balabanova, V. Nikolayevskyy, I. Baessmann, T. Thye, C. G. Meyer, P. Nürnberg, R. D. Horstmann, F. Drobniewski, V. Plagnol, J. C. Barrett, S. Nejentsev, Susceptibility to tuberculosis is associated with variants in the ASAP1 gene encoding a regulator of dendritic cell migration. *Nat. Genet.* **47**, 523–527 (2015).

26. R. Zheng, Z. Li, F. He, H. Liu, J. Chen, J. Chen, X. Xie, J. Zhou, H. Chen, X. Wu, J. Wu, B. Chen, Y. Liu, H. Cui, L. Fan, W. Sha, Y. Liu, J. Wang, X. Huang, L. Zhang, F. Xu, J. Wang, Y. Feng, L. Qin, H. Yang, Z. Liu, Z. Cui, F. Liu, X. Chen, S. Gao, S. Sun, Y. Shi, B. Ge, Genome-wide association study identifies two risk loci for tuberculosis in Han Chinese. *Nat. Commun.* **9**, 4072 (2018).

27. S. M. Gygli, C. Loiseau, L. Jugheli, N. Adamia, A. Trauner, M. Reinhard, A. Ross, S. Borrell, R. Aspindzelashvili, N. Maghradze, K. Reither, C. Beisel, N. Tukvadze, Z. Avaliani, S. Gagneux, Prisons as ecological drivers of fitness-compensated multidrug-resistant Mycobacterium tuberculosis. *Nat. Med.* **27**, 1171–1177 (2021).

28. J. Shea, T. A. Halse, P. Lapierre, M. Shudt, D. Kohlerschmidt, P. Van Roey, R. Limberger, J. Taylor, V. Escuyer, K. A. Musser, Comprehensive whole-genome sequencing and reporting of drug resistance profiles on clinical cases of Mycobacterium tuberculosis in New York State. *J. Clin. Microbiol.* **55**, 1871–1882 (2017).

29. R. Diel, T. A. Kohl, F. P. Maurer, M. Merker, K. Meywald Walter, J. Hannemann, A. Nienhaus, P. Supply, S. Niemann, Accuracy of whole genome sequencing to determine recent tuberculosis transmission: an 11-year population-based study in Hamburg, Germany. *Eur. Respir. J.* (2019), doi:10.1183/13993003.01154-2019.

30. S. Chen, Y. Zhou, Y. Chen, J. Gu, fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*. **34**, i884–i890 (2018).

31. H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio.GN]* (2013), (available at http://arxiv.org/abs/1303.3997).

32. B. J. Walker, T. Abeel, T. Shea, M. Priest, A. Abouelliel, S. Sakthikumar, C. A. Cuomo, Q. Zeng, J. Wortman, S. K. Young, A. M. Earl, Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*. **9**, e112963 (2014).

33. M. Marin, R. Vargas Jr, M. Harris, B. Jeffrey, L. E. Epperson, D. Durbin, M. Strong, M. Salfinger, Z. Iqbal, I. Akhundova, S. Vashakidze, V. Crudu, A. Rosenthal, M. R. Farhat, Genomic sequence characteristics and the empiric accuracy of short-read sequencing. *bioRxiv* (2021), , doi:10.1101/2021.04.08.438862.

34. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. **25**, 2078–2079 (2009).

35. L.-T. Nguyen, H. A. Schmidt, A. von Haeseler, B. Q. Minh, IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).

36. I. Letunic, P. Bork, Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* **44**, W242-5 (2016).

37. T. M. Walker, C. L. C. Ip, R. H. Harrell, J. T. Evans, G. Kapatai, M. J. Dedicoat, D. W. Eyre, D. J. Wilson, P. M. Hawkey, D. W. Crook, J. Parkhill, D. Harris, A. S. Walker, R. Bowden, P. Monk, E. G. Smith, T. E. A. Peto, Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. *Lancet Infect. Dis.* **13**, 137–146 (2013).

38. T. Jombart, adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics*. **24**, 1403–1405 (2008).

39. M. A. Behr, P. H. Edelstein, L. Ramakrishnan, Revisiting the timetable of tuberculosis. *BMJ*. **362**, k2738 (2018).

40. J. Reback, jbrockmendel, W. McKinney, J. Van den Bossche, T. Augspurger, P. Cloud, S. Hawkins, M. Roeschke, gfyoung, Sinhrks, A. Klein, T. Petersen, P. Hoefler, J. Tratner, C. She, W. Ayd, S. Naveh, M. Garcia, J. H. M. Darbyshire, J. Schendel, A. Hayden, R. Shadrach, D. Saxton, M. E. Gorelli, F. Li, M. Zeitlin, V. Jancauskas, A. McMaster, P. Battiston, S. Seabold, *pandas-dev/pandas: Pandas 1.3.5* (Zenodo, 2021; http://dx.doi.org/10.5281/ZENODO.3509134).

41. C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. Del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, T. E. Oliphant, Array programming with NumPy. *Nature*. **585**, 357–362 (2020).

42. R Core Team, R: A language and environment for statistical computing. *R Foundation for Statistical Computing* (2013) (available at http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.470.5851&rep=rep1&type=pdf).

43. H. Wickham, M. Averick, J. Bryan, W. Chang, L. McGowan, R. François, G. Grolemund, A. Hayes, L. Henry, J. Hester, M. Kuhn, T. Pedersen, E. Miller, S. Bache, K. Müller, J. Ooms, D. Robinson, D. Seidel, V. Spinu, K. Takahashi, D. Vaughan, C. Wilke, K. Woo, H. Yutani, Welcome to the Tidyverse. *JOSS*. **4**, 1686 (2019).

44. W. H. Finch, M. E. H. Finch, M. Singh, Data imputation algorithms for mixed variable types in large scale educational assessment: a comparison of random forest, multivariate imputation using chained equations, and MICE with recursive partitioning. *Int. J. Quant. Res. Educ.* **3**, 129 (2016).

45. J. Yan, J. Fine, Estimating equations for association structures. *Stat. Med.* **23**, 859–74; discussion 875-7,879-80 (2004).

46. S. Homolka, S. Niemann, D. G. Russell, K. H. Rohde, Functional genetic diversity among Mycobacterium tuberculosis complex clinical isolates: delineation of conserved core and lineage-specific transcriptomes during intracellular survival. *PLoS Pathog.* **6**, e1000988 (2010).

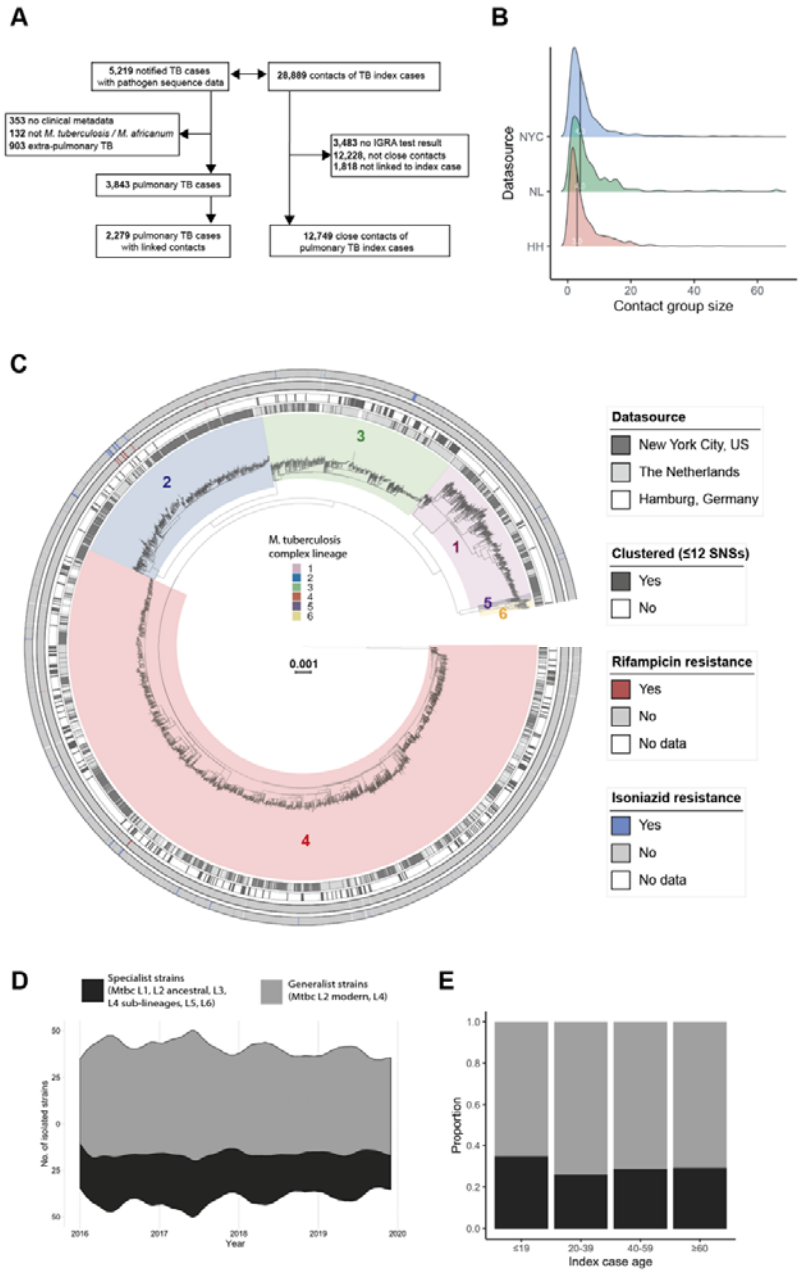
47. N. Reiling, S. Homolka, K. Walter, J. Brandenburg, L. Niwinski, M. Ernst, C. Herzmann, C. Lange, R. Diel, S. Ehlers, S. Niemann, Clade-specific virulence patterns of Mycobacterium tuberculosis complex strains in human primary macrophages and aerogenically infected mice. *MBio*. **4** (2013), doi:10.1128/mBio.00250-13.

Host-pathogen co-adaptation shapes susceptibility to infection with *M. tuberculosis*

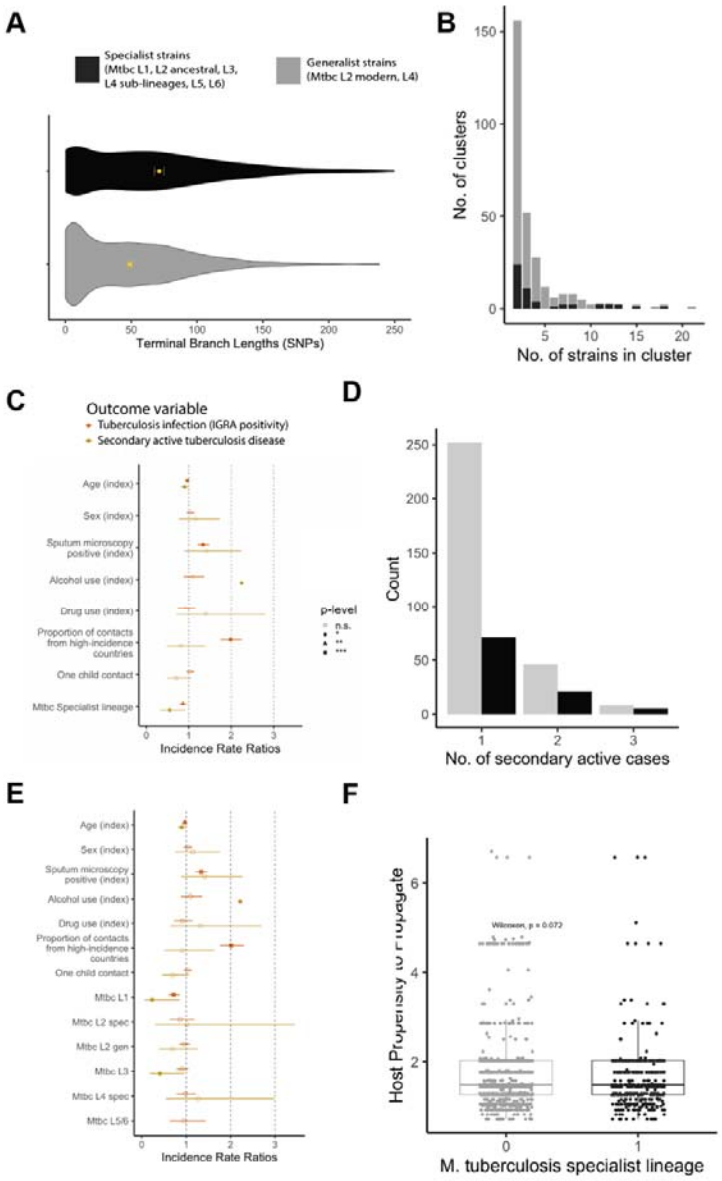
FIGURE 1-4

TABLE 1-2

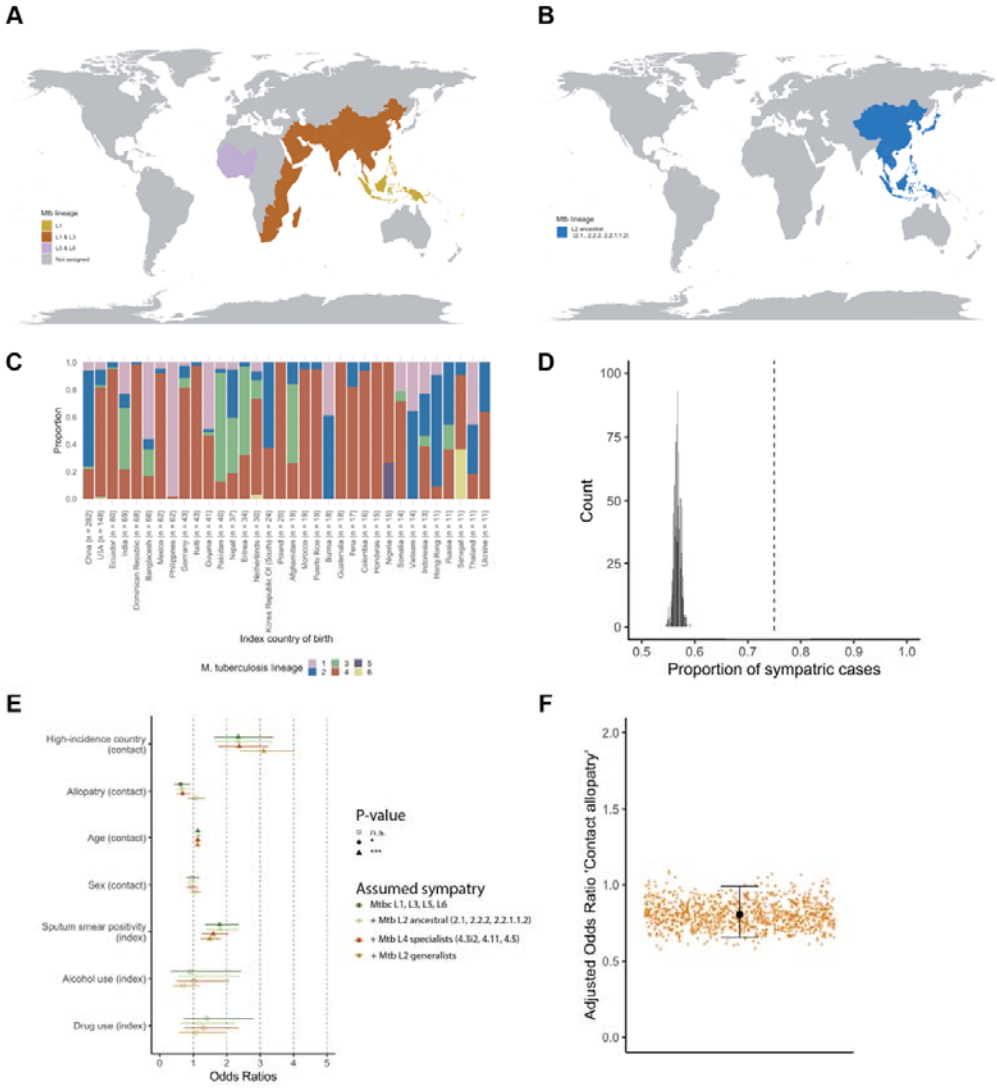
Supplementary figures and tables



Legend Figure 1: Properties of circulating *M. tuberculosis* complex specialist and generalist strains. a) Flowchart of included tuberculosis index cases and linked contacts; b) Density plot of contact group sizes per city; c) Maximum likelihood phylogenetic tree of included strains, clades are coloured by genetic lineage, the rings denote (from inner to outer) the city, clustering based on a 12 Single Nucleotide Substitution cut-off, phenotypic drug resistance to Rifampicin and Isoniazid; d) Proportion of isolated strains by specialist or generalist *Mtbc* lineages over the study period; e) Bar plot showing the proportions of isolated specialist or generalist *Mtbc* lineages by age groups. TB = tuberculosis, IGRA = Interferon Gamma Release Assay, SNS = Single Nucleotide Substitution, Mtbc = *M. tuberculosis* complex.



Legend Figure 2: ***M. tuberculosis* complex strain and tuberculosis index case characteristics associated with transmissibility and clustering**. a) Violin plot of the terminal branch lengths of *Mtbc* strains; b) Numbers of genetic clusters and number of clustered strains by specialist or generalist *Mtbc* lineage (legend as in a); c) Forest plot of a multivariable Poisson rate model quantifying the effect of several predictor variables (y-axis) on the outcome 'Count of tuberculosis infected contacts' (in orange) or 'Count of secondary active tuberculosis cases among contacts' (in yellow); d) Bar plot of secondary active tuberculosis among close contacts inferred by Phybreak stratified by specialist or generalist *Mtbc* lineage (legend as in a)); e) Forest plot of a multivariable Poisson rate model quantifying the effect of Mtbc lineage on tuberculosis infection (orange) or secondary active tuberculosis (yellow) among close contacts (same legend as in c); f) Host Propensity to Propagate by specialist or generalist Mtbc lineage; f) Host propensity to propagate stratified by *M. tuberculosis* complex lineage. Mtbc = *M. tuberculosis* complex



Legend for Figure 3: ***M. tuberculosis* complex sympatry and its effect on transmissibility.** a - b) Global maps displaying the countries that were considered endemic (or sympatric) to the specialist *Mtbc* lineages detailed in the figure legends. c) Bar plot showing the proportions of *Mtbc* lineages isolated from the 2,279 pulmonary tuberculosis index cases included in this study; d) Permutation analysis (n = 1000) of the expected proportion of sympatric host ancestry – *Mtbc* lineage proportions. Vertical dashed line represents the proportion (0.75) among the index cases included in the study; e) Forest plot of the effects estimated using a Generalized Estimating Equation model to quantify the variable 'sympatry'. The outcome is tuberculosis infection per contact, and sympatry was defined using the maps in panel a) and b). f) Permutation test assuming 14% mixed ancestry in the sample (flipping the sympatry variable from 1 to 0 and vice versa in 14% of the sample) while rerunning the model in panel e).

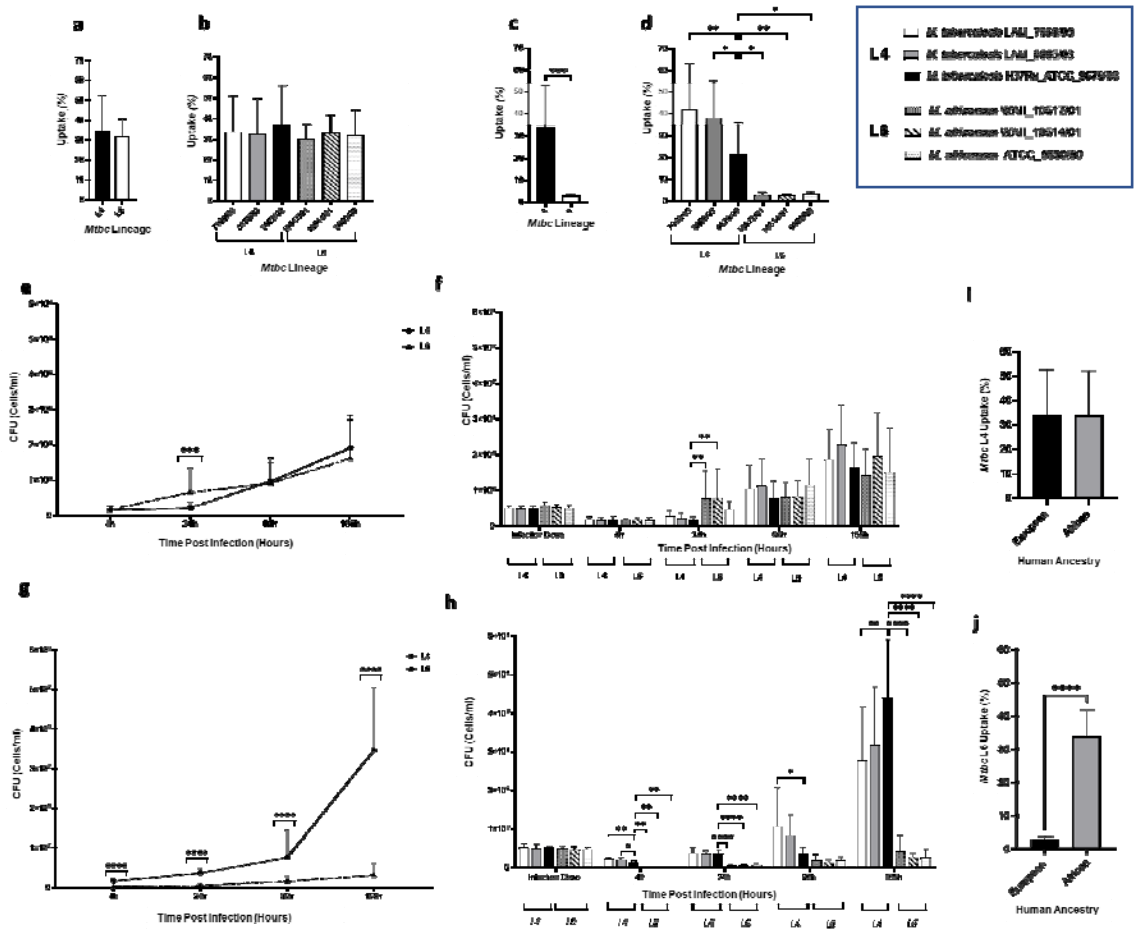


Figure 4. Comparative analysis of phagocytosis and intracellular growth of the Euro American (L4) and West African II (L6) *M. tuberculosis* complex strains in human macrophages of different ancestries. *Mtbc* uptake by lineage and strain in African (a-b) and European Caucasian (c-d) blood monocyte derived macrophages. *Mtbc* intracellular growth by lineage in African (e-f) and European (g-h) blood-derived macrophages. *Mtbc* uptake by macrophages (CFU at 4hpi x 100%/ CFU count of infection dose) was determined immediately at 4hpi and intracellular growth at 24h, 96h, and 168h post-infection. i) *Mtbc* L4 and j) *Mtbc* L6 uptake by human macrophage donor ancestry. Figures represent mean and standard deviations of European Caucasian macrophage infections compromising three independent experiments of which two compromised two donors (See supplementary Figure S3). Nine technical replicates of infection were performed and 27 CFU technical replicates per strain. African macrophages infections compromised 4 independent experiments, 12 technical replicates of infection, and 36 CFU replicates per strain. Macrophages were infected at MOI ~1:1 (0.5 x10^6 cells: 0.5 x10^6 *Mtbc* bacilli). Statistical results represent T-test among *Mtbc* lineages (average of strains belonging to a specific lineage) and One-way ANOVA with Bonferroni post hoc test considering CFU values between the clinical strains and the reference laboratory strain H37Rv (9679-00). Mean, standard deviation, and statistical significances (*P <0.05 **P < 0.01; ***P < 0.001 and ****P < 0.0001) are depicted in the figures.

Table 1: Pulmonary tuberculosis patients characteristics.

	N	Overall N = 2,279[1]	Hamburg N = 498[1]	Netherlands N = 266[1]	New York City N = 1,515[1]
Sex	2,279				
Female		849 (37%)	191 (38%)	98 (37%)	560 (37%)
Male		1,430 (63%)	307 (62%)	168 (63%)	955 (63%)
Age	2,279	48 (31, 66)	43 (29, 61)	38 (26, 55)	52 (33, 68)
Country of birth	2,279				
High TB incidence		1,574 (69%)	224 (45%)	158 (59%)	1,192 (79%)
Medium/low TB incidence		705 (31%)	274 (55%)	108 (41%)	323 (21%)
Sputum microscopy	2,279				
Negative		809 (35%)	230 (46%)	90 (34%)	489 (32%)
Positive		1,470 (65%)	268 (54%)	176 (66%)	1,026 (68%)
Symptomatic	2,254	1,985 (88%)	448 (92%)	203 (81%)	1,334 (88%)
Homeless	2,275	72 (3.2%)	32 (6.5%)	7 (2.6%)	33 (2.2%)
Drug use	2,279	108 (4.7%)	13 (2.6%)	6 (2.3%)	89 (5.9%)
Alcohol use	2,272	137 (6.0%)	77 (16%)	5 (1.9%)	55 (3.6%)
HIV	1,975	77 (3.9%)	21 (4.7%)	8 (3.6%)	48 (3.7%)
Diabetes	2,279	418 (18%)	24 (4.8%)	22 (8.3%)	372 (25%)
Immunosuppression	1,829	44 (2.4%)	6 (3.4%)	6 (4.3%)	32 (2.1%)
Isoniazid monoresistance	2,278	242 (11%)	35 (7.0%)	25 (9.4%)	182 (12%)
Multidrug resistance	2,278	60 (2.6%)	13 (2.6%)	3 (1.1%)	44 (2.9%)
***M. tuberculosis* complex lineage**	2,279				
Generalist		1,636 (72%)	396 (80%)	183 (69%)	1,057 (70%)
Specialist		643 (28%)	102 (20%)	83 (31%)	458 (30%)

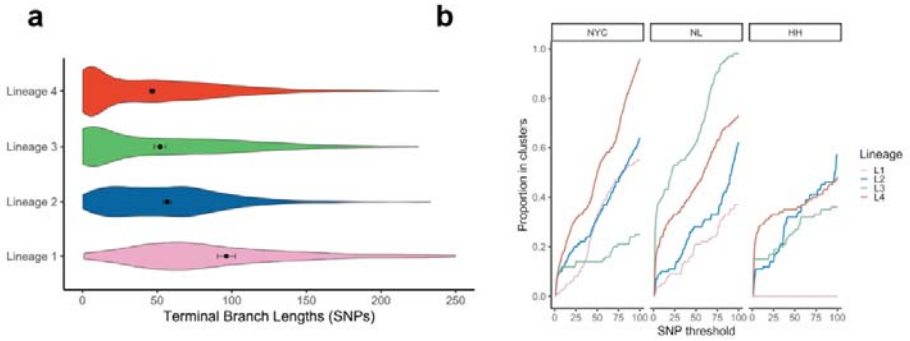
[1]n (%); Median (IQR); High TB incidence defined as >30/100.000 incidence based on 2016 WHO data

Table 2: Associations of participant characteristics and bacterial factors on transmission (quasi-Poisson model with overdispersion φ = 1.2 and φ = 1.4, respectively, N = 2,279)

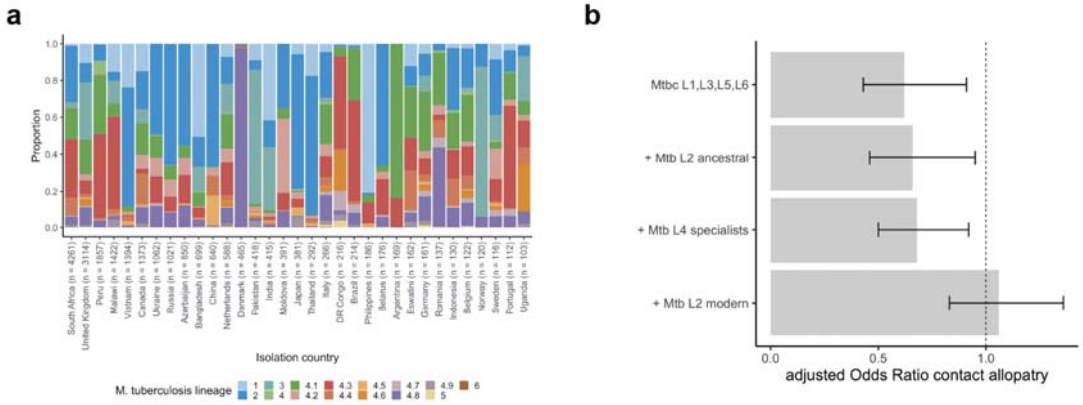
	Tuberculosis Infection			Secondary Active Tuberculosis		
Characteristic	IRR[1]	95% CI[1]	p-value	IRR[1]	95% CI[1]	p-value
Age (index)	0.96	0.94, 0.99	**<0.001**	0.90	0.82, 1.00	**0.042**
Sex (index)	1.04	0.95, 1.14	0.4	1.15	0.77, 1.75	0.5
Smear positive (index)	1.34	1.21, 1.48	**<0.001**	1.42	0.92, 2.28	0.13
Alcohol use (index)	1.10	0.87, 1.36	0.4	2.25	1.16, 4.01	**0.010**
Drug use (index)	0.93	0.74, 1.15	0.5	1.40	0.66, 2.65	0.3
High incidence contact group (%)	1.98	1.75, 2.25	**<0.001**	0.81	0.47, 1.39	0.5
One child contact (index)	1.03	0.94, 1.13	0.5	0.71	0.47, 1.05	0.091
Mtbc specialist lineage (index)	0.86	0.78, 0.95	**0.004**	0.55	0.32, 0.90	**0.023**

[1]IRR = Incidence Rate Ratio, CI = Confidence Interval, *Mtbc* = M. tuberculosis complex, P-values from Wald test

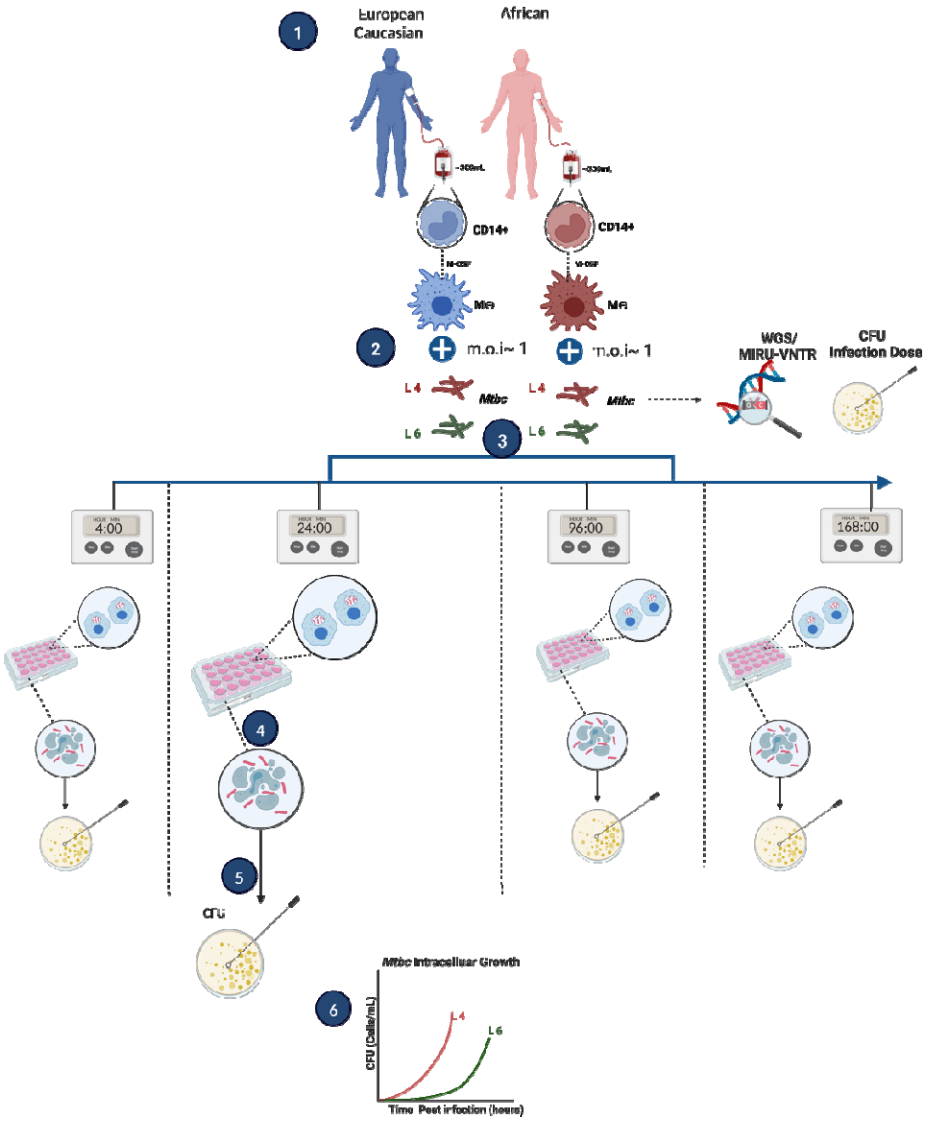
SUPPLEMENTARY MATERIAL



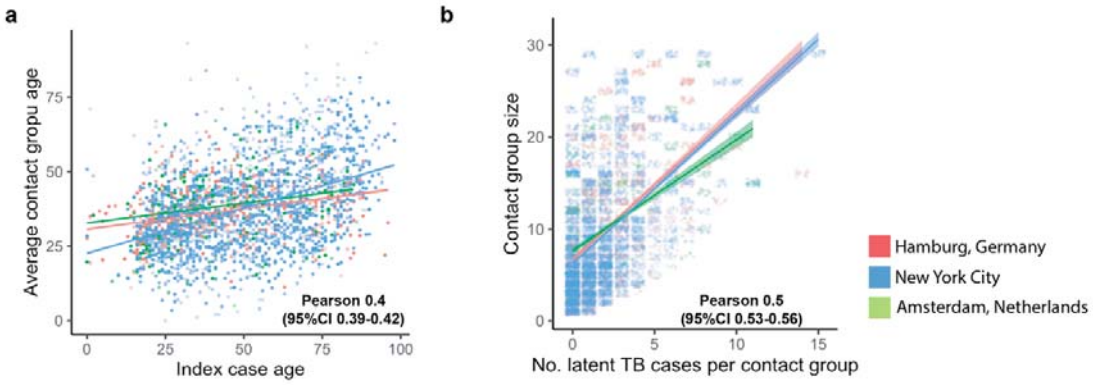
Legend for Suppl. Figure S1: Genetic characteristics of *M. tuberculosis* complex strain. a) Violin plot of the terminal branch lengths of the included *Mtbc* genetic lineages. b) Proportions of strains in clusters based on several different Single Nucleotide Substitution thresholds by genetic lineage and city.



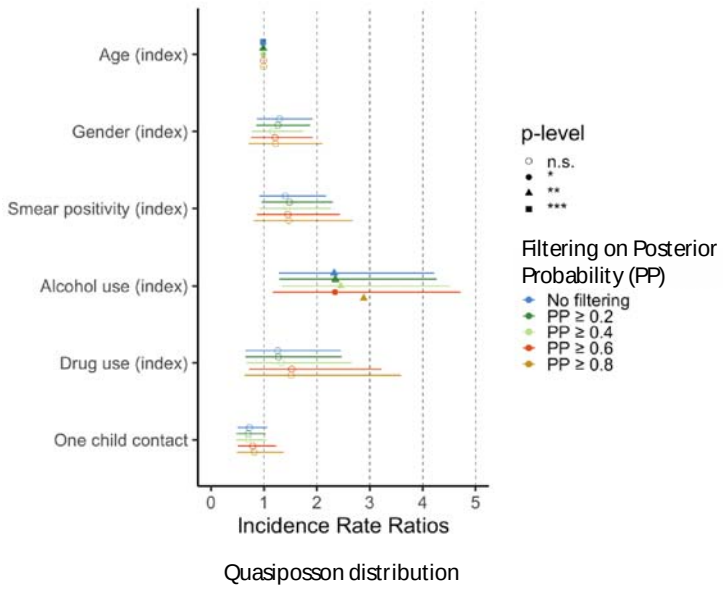
Legend for Suppl. Figure S2: Relationship of index case ancestry and *M. tuberculosis* complex lineage. a) Bar plot detailing the proportions of isolation country and *Mtbc* lineage in a global sample of 25,243 strains. b) Adjusted odds ratios for the variable contact allopatry using different sympatry assumptions from multivariate Generalized Estimation Equation models (see Figure 3e in main text).



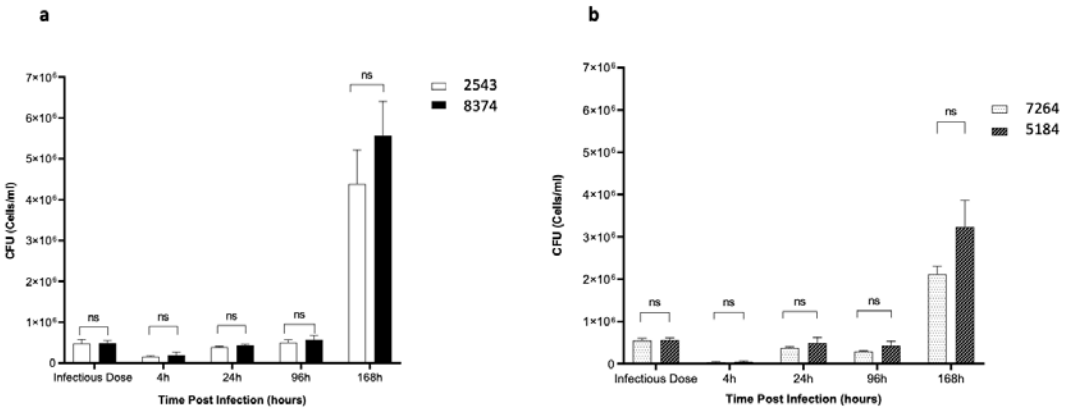
Legend for Suppl. Figure S3: Experimental Design of Host (Macrophage)-Pathogen Interaction. 1) ~300 mL of total blood was obtained from healthy donors. CD14+ monocytes were obtained by magnetic separation with beads conjugated to monoclonal antihuman CD14 antibodies and further differentiated to macrophages in the presence of M-CSF for 7 days. 2) Macrophages were infected with *M. tuberculosis* complex bacilli previously genotyped by whole genome sequencing (WGS) and mycobacterial interspersed repetitive unit – variable number of tandem repeats typing (MIRU) and colony forming unit (CFU) titrated. 3) Macrophages were incubated with *M. tuberculosis* complex bacilli for 4h, 24 h, 96h, and 168h. 4) Infected macrophages were lysed, and intracellular growth was assessed by CFU assay (6). The figure was designed with Bio-Render.



Legend for Suppl. Figure S4: Comparison of index case and contact group characteristics. a) Dot plot of the index case age (x-axis) versus mean contact group age (y-axis) for each of the included cities; b) Dot plot of the No. of tuberculosis infections per contact group (x-axis) and the size of the contact group (y-axis).



Legend for Suppl. Figure S5: Impact of different support value thresholds on effect estimates of transmission. Forest plot of effect estimates from a quasi-Poisson count regression model predicting the count of secondary active tuberculosis cases for each index among his close contacts. Several different posterior probability values were evaluated after inference of transmission relationships using Phybreak (1).



Legend for Suppl. Figure S6: Comparison of intracellular survival of *M. tuberculosis* H37Rv (9679-00) laboratory reference strain among two distinct donors compromising an independent experiment of the European Caucasian macrophage infection experiments. a) Experiment one (donor 2543 + donor 8374). b) Experiment three (donor 7264+ donor 5184). Blood monocyte derived macrophages were infected at MOI~1:1 ($0.5x10^6$ cells: $0.5x10^6$ bacteria) and the numbers of intracellular bacteria (CFU) were determined immediately after uptake at 4h post-infection. Colony-forming units (CFU) per ml (x-axis) and time post-infection in hours (y-axis). Statistic results of experiments one and three were determined by two-way ANOVA with Bonferroni´s multiple comparison test. No statistical differences among donors within experiments are depicted in the figure (ns >0.05). Error bars represent the standard error of the mean of nine CFU technical replicates (three replicates of infections) for each time point and three CFU technical replicates for infection dose.

Suppl. Table S1: ***M. tuberculosis*** **complex lineages and sub-lineages considered geographically restricted specialists.**

Lineage	Simpsons diversity index
1	-
2.1 (‘Proto-Beijing’)	0.19
2.2.2 (‘Asia Ancestral 1’)	0.67
2.2.1.1.2 (‘Asia Ancestral 2’)	0.28
3	-
4.3.i2	0.25
4.11	0.11
4.5	-
4.6.2.2 (‘Cameroon’)	0.28
4.6.1.1.1 (‘Uganda’)	0.28
4.2.1.1	0.09
5 (M. africanum West African 1)	-
6 (M. africanum West African 2)	-

Simpsons diversity index based on Freschi et al. (2).

Suppl. Table S2: **Tuberculosis index case and contact characteristics and their association with transmission using Generalized Estimation Equations** (N = 12,749).

	Tuberculosis Infection			Secondary Active Tuberculosis		
Characteristic	OR[1]	95% CI[1]	p-value	OR[1]	95% CI[1]	p-value
Age (contact)	1.14	1.11, 1.17	**<0.001**	0.96	0.90, 1.01	0.13
Sex (contact)	1.01	0.92, 1.10	0.9	1.71	1.44, 2.04	**<0.001**
High-inc. country born (contact)	2.18	1.96, 2.42	**<0.001**	0.61	0.49, 0.76	**<0.001**
Age (index)	0.91	0.88, 0.93	**<0.001**	0.77	0.72, 0.83	**<0.001**
Smear positive (index)	1.48	1.30, 1.68	**<0.001**	0.94	0.70, 1.26	0.7
Alcohol use (index)	1.01	0.77, 1.32	>0.9	3.06	1.73, 5.42	**<0.001**
Drug use (index)	0.95	0.67, 1.34	0.8	0.85	0.47, 1.54	0.6
Specialist lineage (index)	0.82	0.72, 0.94	**0.004**	0.77	0.57, 1.04	0.088

[1]OR = Odds Ratio, CI = Confidence Interval, P-values from Wald test

Suppl. Table S3: **New York City tuberculosis index case and contact characteristics and their association with transmission using Generalized Estimation Equations** (N = 8,481).

	Tuberculosis Infection			Secondary Active Disease		
Characteristic	OR[1]	95% CI[1]	p-value	OR[1]	95% CI[1]	p-value
Age (contact)	1.12	1.09, 1.16	**<0.001**	0.84	0.78, 0.92	**<0.001**
Sex (contact)	0.95	0.85, 1.05	0.3	1.67	1.29, 2.16	**<0.001**
High-inc. country-born (contact)	2.24	1.97, 2.55	**<0.001**	1.93	1.39, 2.68	**<0.001**
HIV (contact)	0.99	0.57, 1.70	>0.9	1.28	0.41, 3.98	0.7
Diabetes (contact)	1.56	1.15, 2.12	**0.004**	4.23	2.36, 7.59	**<0.001**
Age (index)	0.91	0.88, 0.95	**<0.001**	0.95	0.87, 1.03	0.2
Smear positive (index)	1.62	1.38, 1.90	**<0.001**	1.14	0.79, 1.66	0.5
Alcohol use (index)	0.73	0.48, 1.09	0.13	2.24	1.05, 4.78	**0.037**
Drug use (index)	0.98	0.67, 1.44	>0.9	2.04	1.21, 3.44	**0.007**
Specialist lineage (index)	0.92	0.78, 1.08	0.3	0.54	0.37, 0.78	**0.001**

[1]OR = Odds Ratio, CI = Confidence Interval, P-values from Wald test

Suppl. Table S4: Associations of study members' characteristics and bacterial lineage on transmission (N = 2,279).

	Tuberculosis Infection			Secondary Active Tuberculosis		
Characteristic	IRR[1]	95% CI[1]	p-value	IRR[1]	95% CI[1]	p-value
Age (index)	0.97	0.95, 0.99	**0.003**	0.90	0.81, 1.00	**0.043**
Sex (index)	1.04	0.94, 1.14	0.5	1.15	0.76, 1.78	0.5
Smear positive (index)	1.34	1.21, 1.48	**<0.001**	1.41	0.90, 2.31	0.15
Alcohol use (index)	1.09	0.87, 1.35	0.4	2.21	1.11, 4.01	**0.015**
Drug use (index)	0.91	0.72, 1.14	0.4	1.32	0.60, 2.56	0.5
High incidence contact group (%)	2.01	1.76, 2.30	**<0.001**	0.91	0.50, 1.62	0.8
One child contact (index)	1.03	0.94, 1.13	0.5	0.70	0.46, 1.05	0.092
Mtbc lineage (index)						
$L4_{gen}$	—	—		—	—	
L1	0.71	0.60, 0.85	**<0.001**	0.23	0.05, 0.69	**0.027**
$L2_{spec}$	0.87	0.62, 1.17	0.4	1.02	0.23, 2.89	>0.9
$L2_{gen}$	0.96	0.85, 1.08	0.5	0.70	0.38, 1.22	0.2
L3	0.91	0.78, 1.05	0.2	0.41	0.16, 0.88	**0.038**
$L4_{spec}$	0.98	0.77, 1.22	0.8	1.27	0.49, 2.75	0.6
L5/L6	0.95	0.61, 1.40	0.8			

[1]IRR = Incidence Rate Ratio, Mtbc = M. tuberculosis complex, CI = Confidence Interval, P-values from Wald test

Suppl. Table S5: Effect of contact allopatry on infection assuming different country-lineage sympatry relationships.

	Sympatry for *Mtbc* L1,L3,L5,L6				plus *Mtb* L2 ancestral (2.1, 2.2.2, 2.2.1.1.2 in East Asia)				plus *Mtb* L4 specialists (4.3.i2, 4.11, 4.5)				plus *Mtb* L2 modern (in East Asia)			
Characteristic	N	OR[1]	95% CI[1]	p-value	N	OR[1]	95% CI[1]	p-value	N	OR[1]	95% CI[1]	p-value	N	OR[1]	95% CI[1]	p-value
High incidence birth (contact)	2,556	2.34	1.62, 3.39	**<0.001**	2,788	2.35	1.63, 3.38	**<0.001**	2,988	2.37	1.74, 3.24	**<0.001**	4,830	3.10	2.39, 4.02	**<0.001**
Allopatry (contact)	2,556	0.62	0.43, 0.91	**0.013**	2,788	0.66	0.46, 0.95	**0.025**	2,988	0.68	0.50, 0.92	**0.014**	4,830	1.06	0.83, 1.36	0.6
Age (contact)	2,556	1.14	1.07, 1.21	**<0.001**	2,788	1.15	1.08, 1.22	**<0.001**	2,988	1.14	1.07, 1.20	**<0.001**	4,830	1.13	1.08, 1.18	**<0.001**
Sex (contact)	2,556	0.97	0.79, 1.18	0.8	2,788	0.95	0.79, 1.15	0.6	2,988	0.96	0.80, 1.15	0.6	4,830	1.07	0.93, 1.22	0.4
Smear positivity (index)	2,556	1.78	1.35, 2.36	**<0.001**	2,788	1.80	1.37, 2.37	**<0.001**	2,988	1.60	1.24, 2.05	**<0.001**	4,830	1.49	1.22, 1.82	**<0.001**
Alcohol (index)	2,556	0.90	0.34, 2.43	0.8	2,788	1.13	0.53, 2.41	0.7	2,988	1.02	0.50, 2.07	>0.9	4,830	0.69	0.40, 1.19	0.2
Drug use (index)	2,556	1.41	0.71, 2.80	0.3	2,788	1.19	0.63, 2.25	0.6	2,988	1.31	0.72, 2.35	0.4	4,830	1.07	0.58, 1.96	0.8

[1]OR = Odds Ratio, CI = Confidence Interval, *Mtbc* = *M. tuberculosis* complex

Suppl. Table S6: Generalized estimation equation model to quantify the effect of contact sympatry on *M. tuberculosis* complex infection among New York City contacts (N = 2,012).

Characteristic	OR[1]	95% CI[1]	p-value
Age (contact)	1.11	1.04, 1.19	**0.002**
Sex (contact)	0.92	0.73, 1.16	0.5
High-inc. country-born (contact)	2.27	1.56 3.22	**<0.001**
Allopatry (contact)	0.068	0.48, 0.96	**0.03**
HIV (contact)	0.00	0.00, 0.00	**<0.001**
Diabetes (contact)	1.01	0.54, 1.89	0.9
Smear positive (index)	1.75	1.31, 2.34	**<0.001**
Alcohol (index)	0.83	0.37, 1.86	0.7
Drug use (index)	1.82	0.98, 3.38	0.059

[1]OR = Odds Ratio, CI = Confidence Interval, P-values from Wald test

References Supplementary material

1. D. Klinkenberg, J. A. Backer, X. Didelot, C. Colijn, J. Wallinga, Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. PLoS Comput. Biol. **13**, e1005495 (2017).

2. L. Freschi, R. Vargas Jr, A. Husain, S. M. M. Kamal, A. Skrahina, S. Tahseen, N. Ismail, A. Barbova, S. Niemann, D. M. Cirillo, A. S. Dean, M. Zignol, M. R. Farhat, Population structure, biogeography and transmissibility of Mycobacterium tuberculosis. Nat. Commun. **12** (2021), doi:10.1038/s41467-021-26248-1.