Evolutionary history and molecular epidemiology of *Mycobacterium tuberculosis* in Tanzania and across Africa

INAUGURALDISSERTATION
zur
Erlangung der Würde eines Doktors der Philosophie

vorgelegt der
Philosophisch-Naturwissenschaftlichen Fakultät
der Universität Basel

von
Liliana Kokusanilwa Rutaihwa
aus Nsisha, Tansania

Basel, 2020

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät auf Antrag von Herrn Prof. Dr. Sébastien Gagneux und Herrn Prof. Dr. Thierry Wirth.

Basel, den 11. Dezember 2018

Prof. Dr. Martin Spiess
Dekan

Summary

Humans have been affected by tuberculosis (TB) for millennia. Today, TB remains a global health problem and the leading cause of mortality due to a single infectious agent. TB in humans is primarily caused by seven human-adapted phylogenetic lineages of *Mycobacterium tuberculosis* (Mtb) complex. Mtb lineages differ in their geographical distribution, partly reflecting human demographic histories. Importantly, variation in Mtb is known to impact TB infection and clinical disease.

In recent years, advances in sequence-based molecular markers i.e. single nucleotide polymorphisms (SNPs) and whole genome sequencing (WGS) technologies have enabled robust classification of Mtb strains which ultimately have allowed researchers to address important questions regarding Mtb phenotypes, transmission patterns and the evolutionary history of TB. Remarkably, such investigations remain underexplored in high-endemic TB settings of sub-Saharan Africa.

By applying phylogenetically robust methods such as SNP-based typing complemented with WGS we can gradually disentangle the role of Mtb variation on TB epidemic in high burden clinical settings. On the other hand, with recent large-scale WGS, it is becoming clear that Mtb strains are heterogeneous at the lineage level. Several studies have explored the phylogenetic substructure of Lineage 2 and Lineage 4; the two most geographically widespread and more successful Mtb lineages. However, Lineage 1 and 3 are still important drivers of TB epidemics along the Indian Ocean rim, which includes parts of Africa. Yet to date, the phylogeographies of these two lineages have not been fully explored. By contrast, Lineage 2–Beijing seems to have emerged only recently in Africa. Among the seven Mtb lineages, Lineage 2–Beijing is highly virulent and associated with antibiotic resistance; thus, this calls for investigation of its origin on the African continent.

In this thesis, we aimed to gain countrywide insights into the genetic diversity of Mtb in Tanzania based on SNP-typing. Secondly, using a combination of SNP-typing and WGS techniques we describe the local diversity of Mtb and assessed for clinical phenotypes in urban and rural settings of Tanzania. We then studied the global phylogeographies of

Mtb Lineage 1 and 3 to infer their evolutionary histories and global spread. Finally, we analyzed the origin of Mtb Lineage 2–Beijing in Africa using WGS.

This thesis contains 7 chapters. The first two chapters provide the background on TB, Mtb lineages, and the objectives of the thesis. The remaining four chapters cover the conducted research performed during this PhD thesis. In the final chapter, we summarize the key findings, limitations and discuss the general implications of our work.

In *Chapter 1*, we highlight the global burden and control of TB, the outcome of TB infection and disease, the overview on the Mtb genetic diversity, different molecular markers and genotyping techniques, and the consequences of Mtb diversity.

In *Chapter 2* we state the objectives of the thesis.

In *Chapter 3*, we studied a countrywide population structure of Mtb in Tanzania based on SNP-typing and assessed relationships between Mtb lineages with patients' clinical and sociodemographic characteristics.

In *Chapter 4*, we zoomed into the local urban and rural settings of Temeke, Dar es Salaam and Ifakara, Morogoro in Tanzania, to identify clinically relevant Mtb phenotypes. In addition, we describe the local diversity and performed an exploratory analysis on transmission patterns in the urban setting.

In *Chapter 5*, we studied the phylogeography and the spread of Lineage 1 and 3 using global representative genomes from places where strains of the two lineages are frequent.

In *Chapter 6*, we used whole genome sequences of Mtb Lineage 2–Beijing to investigate the evolutionary history of this lineage in Africa. We reveal multiple introductions of Mtb Lineage 2–Beijing into Africa originating from Asia. We further show that these introductions occurred over the last 300 years, with most pre-dating the antibiotic era.

In *Chapter 7*, we summarize the key findings from this PhD thesis, discuss the implications and highlight future directions.

Contents

Acknowledgements

To the **Ifakara Health Institute-Bagamoyo**, where my career in TB research began and through which I got to learn about Swiss TPH. Special thanks to Dr. Klaus Reither, Dr. Levan Jugheli and the TB unit for the invaluable experiences that inspired me to develop a career in research.

To **Christine Mensch**, before I even set foot in Basel and once I did you have been nothing but helpful and supportive throughout. For that and more, vielen Dank!

To **Prof. Marcel Tanner**, your enthusiasm inspired my day to day life and made me a proud member of Swiss TPH as we all sailed on the same boat. Asante!

To the **lecturers and trainers** at the institute and outside, thank you for the sharing of knowledge and expertise.

To my fellow **PhDs and other colleagues**, thank you for the excursions, coffee breaks, etc., that contributed to a lively and friendly learning/working environment. To the **old office** (Natalie, Liza, Oli, Anton and Isidoros) thank you for being part of my PhD journey.

To all the **collaborators**, thank you for your efforts, time and dedication to contributing strains and sharing of data that made this thesis a success. Special thanks to Emilyn Costa, Janet Fyfe, Niaina Rakotosamimanana, Horng-Yunn Dou, Inaki Comas, Christophe Sola, Iñaki Comas and Darío García-de-Viedma.

To the **Central Tuberculosis Reference Laboratory – Tanzania**, thank you for your great contribution to the Tanzanian component of this thesis. I am particularly grateful to Dr. Basra Doulla for all the support, Amri Kingalu, Bryceson Malewo and Ally Kingazi for preparing the isolate collection and organizing logistics for the CTRL project.

To the **TBDAR cohort team**, thank you for your tireless efforts and dedication in making this cohort a great platform for research including this thesis. My special appreciation to Jerry Hella, Mohamed Sasamalo, Hellen Hiza, Lujeko Kamwela, Francis Mhimbira, George Sikalengo and Emilio Letang.

To the members of **Tuberculosis Research Unit-Swiss TPH**, thank you for the wonderful six years of your constant support, encouragement and wonderful social interactions! It was a great pleasure to work with you and to get to know you all. Special thanks to Julia and Miriam particularly for the support in the BSL3 laboratory and the sequencing; Daniela and Fabrizio; for the coaching and assisting in the genomics analyses and manuscript writing; Sonia for the BSL3 laboratory training; Mohamed and Aladino, for the "mentoring" experience and your invaluable contribution to the thesis; Anna for finalizing lab work in the BSL1 for the TBDAR and CTRL projects; Michaela, for assisting with uploading of the Lineage 2–Beijing sequences; Andrej, Ainhoa, Khadija, Peter and to all others. I extend my sincere gratitude to the **"PT" office**; you are terrific people, "believe me!" Thank you for your friendship and the memorable moments during the crazy PhD life. Rhastin, for figuring things out together from the early days in the group and for the good laughs; Monica, for your time and patience in coding-related issues and for the enjoyable moments and experiences outside the institute and Basel; Sebastian, for the "beer-o'clock" reminders and sessions accompanied with interesting conversations and good laughs, and for being a "writing buddy", which made the writing process manageable; Chloe, for the amazing pipeline that works like a charm, for spotting typos in scripts that saved me a lot time and frustration, for your constant support during the writing process, making sure I have my feet on the ground and most of all for simply being there! Kukusanilwa appreciates you all a lot. Lastly to David Stucki who inspired and motivated me to continue the sub-, subsub- and subsubsublineage work, indeed it made this book!

To **my family**: Faustine and Faustina, thank you for raising me to become the person I am today and for building me a solid foundation in life, Mastidia, Josephine, Evelyn and Richard, for being wonderful siblings and friends throughout the years, Florian, for always checking up on "auntie Lily", Charles, Hiza, nieces and nephews .Thank you all for the love, support and prayers, I am forever grateful.

To **my dearest friends** Nacky, Nura and Upendo, thank you for being my "tribe", Serej and Manu, Nicole, Tobi and Andreas for your wonderful friendship and for always having my back. Thank you all for everything!

To **Davide**, by default you became a PhD candidate. Thank you for being there through thick and thin. You have been my rock!

To **Prof. Thierry Wirth**, thank you for being part of my PhD committee. Your efforts to evaluate the thesis marked the finish line to it. I would like to extend my appreciation to Prof. **Christian Lengeler** for kindly agreeing to chair my PhD defense.

To **Dr. Lukas Fenner**, thank you for the opportunity to continue the work I started during my Masters, which began with your willingness to let me become part of the TBDAR team and for assisting me in setting up collaborations with the CTRL and IHI-Ifakara. Asante sana!

To **Prof. Sébastien Gagneux**, there are not enough words to express my gratitude. For six years you have mentored me, presented me with opportunities to learn and to grow. Thank for your constant professional and moral support, encouragement and mostly for making this thesis what it has become. Nashukuru sana sana!

For the beautiful gift of life and countless opportunities and experiences that have come along with it, thank you Almighty.

Abbreviations

BMI	Body Mass Index
BRTC	Bagamoyo Research and Training Center
CAS	Central Asian
DR	Drug resistance
DST	Drug Susceptibility Testing
EAI	East African Indian
EMB	Ethambutol
ETH	Ethionamide
FQ	Fluoroquinolone
HIV	Human Immunodeficiency Virus
IHI	Ifakara Health Institute
INH	Isoniazid
IQR	Inter Quartile Range
IS	Insertion Sequence
LAM	Latin America Mediterranean
LPA	Line Probe Assay
MDR	Multi-drug resistance
MIRU	Mycobacterial Interspered Repetetive Units
ML	Maximum Likelihood
MLST	Multi Locus Sequence Typing

MRCA	Most Recent Common Ancestor
Mtb	*Mycobacterium tuberculosis*
MTBC	*Mycobacterium tuberculosis* complex
PCR	Polymerase Chain Reaction
PZA	Pyrazinamide
SM	Streptomycin
SNP	Single Nucleotide Polymorphism
Spoligotyping	Spacer Oligonucleotide Typing
TB	Tuberculosis
VNTR	Variable Number of Tandem Repeats
WGS	Whole Genome Sequencing
WHO	World Health Organization

1. Introduction

1.1. Global burden and control of tuberculosis

For millennia, human beings have suffered from tuberculosis (TB), with a billion human deaths caused by TB during the last 200 years (Daniel, 2006). In 2017, TB claimed an estimated 1.6 million lives (including in 300'000 immunodeficiency virus [HIV] infected individuals), and an estimated 10 million people newly contracted the disease (WHO, 2018). Most of the global TB burden lies in the 30 high-burden countries, many of which are in sub-Saharan African (Figure 1.1). Africa, the second largest home to the world's population, carries one-quarter of the global burden of TB cases and has the highest TB related deaths (WHO, 2018). The most alarming challenges facing the control of TB are the HIV pandemic and the emergence of drug resistant (DR)–TB. The two aggravate the TB epidemics, which prior to these events had begun to decrease, at least in the developed world (Dye *et al.*, 2010). Figure 1.2 highlights the three high-burden lists for TB, TB/HIV and multidrug resistance (MDR)–TB with each category accounting for 90% of the global burden (WHO, 2018).

Individuals infected with HIV are immune-compromised and therefore unable to mount efficient immune responses against TB (Kwan *et al.*, 2011), putting them at up to a 37-fold risk of developing active TB disease compared to HIV-negative individuals (Getahun *et al.*, 2010). In addition, TB is the leading cause of mortality among people living with HIV. The WHO African region bears the largest TB/HIV burden where approximately 72% of HIV–associated TB in 2017 occurred in the region (WHO, 2018). Moreover, MDR and extensively-drug resistant (XDR) TB are two almost incurable forms of TB disease. Over 450,000 people were estimated to develop MDR–TB in 2017, of which 8.5% had XDR–TB (WHO, 2018). However, these figures could be underestimated as high-burden TB settings still lack appropriate tools to diagnose DR–TB. By contrast, the highest proportions of MDR–TB are in the former Soviet Union countries (Figure 1.2). Even though treatment for MDR and XDR is possible, the regimen is costly, lengthy, linked to many toxic side-effects, and associated with high rates of treatment failure and mortality

(Nathanson *et al.*, 2006). Worse yet, DR–strains resistant to all first-line and second-line drugs have emerged and they cause extremely drug-resistant (XXDR) or totally-drug resistant (TDR) forms of TB (Migliori *et al.*, 2007).

Despite the collective efforts, TB control still relies on out-dated diagnostic tools of poor sensitivity (smear microscopy), old drugs (isoniazid, rifampicin, ethambutol, pyrazinamide, streptomycin) and vaccine (Bacille Calmette-Guérin [BCG]) of questionable efficacy (Young *et al.*, 2008). Although BCG vaccination has an invaluable contribution in protection against paediatric disseminated forms of TB (Colditz *et al.*, 1995; Trunz *et al.*, 2006), its efficacy against adult pulmonary disease varies from absolute no efficacy to as high as 80% protection (Fine, 1995). These tools remain in use in parts of the world where TB is one of the most important public health concerns. Recently, new rapid molecular diagnostics such GeneXpert, line probe assays (LPA), nucleic acid amplification test (NAAT) have come into play. The endorsement of GeneXpert and other rapid molecular tests has enabled improved detection among TB/HIV co-infected individuals (who often perform poorly with smear microscopy) and in parallel genotypic drug resistance testing. However, the scale-up and cost-effectiveness of these tools under routine conditions remain a challenge, especially for low-income settings (Weyer *et al.*, 2011).

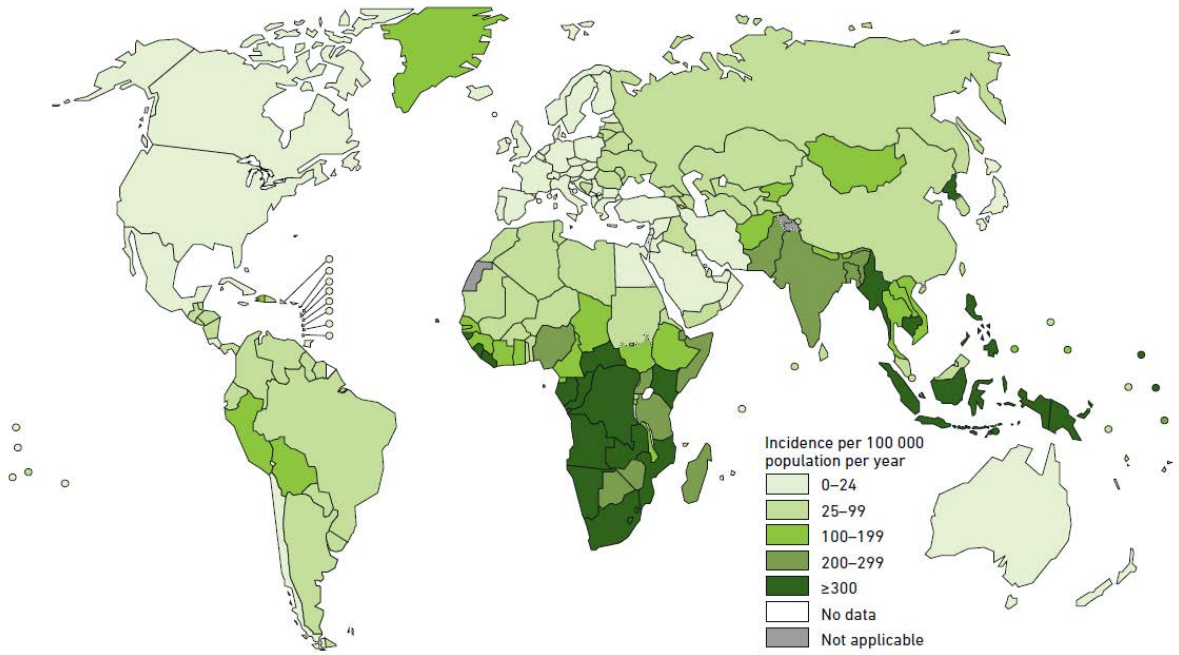


Figure 1.1.: **Estimated TB incidence rates in 2017.** Global Tuberculosis Report 2018, World Health Organization (WHO) (WHO, 2018).

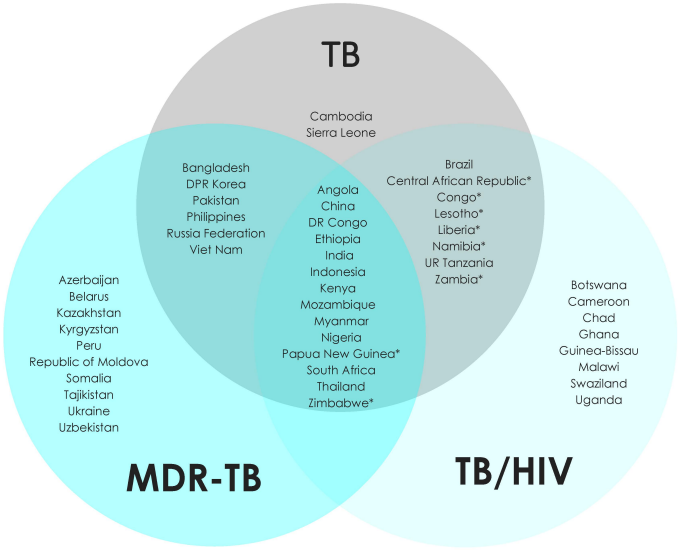


Figure 1.2.: **Countries in the three high burden lists for TB, TB/HIV and MDR-TB.** Modified from the Global Tuberculosis Report 2018 (WHO, 2018).

1.2. TB infection and disease outcome

TB is transmitted via air droplets from an infected individual with a pulmonary form of disease to a health individual, primarily through coughing. Upon exposure to the infectious droplets, a person could either clear the infection as a result of an efficient innate immune response, progress to active disease or contain the infection in latent form. One-quarter of the world's population is latently infected with TB and a potential reservoir for future active disease (Houben *et al.*, 2016). About 5-10% of the latently infected individuals have a life-time risk to progress to active TB disease where the risk is up to 50% in HIV co-infected individuals (Koul *et al.*, 2011). Moreover, the features of active TB disease can vary and include pulmonary and extra-pulmonary presentations.

The interplay of host, environmental factors, and the pathogen together influence both TB infection and disease dynamics. Human factors include immunity on the one hand, thus immune suppression i.e., due to HIV infection is a strong factor for TB progression to active disease (Kwan *et al.*, 2011), and genetic factors which influence TB susceptibility on the other (Casanova *et al.*, 2002). Environmental factors such as good ventilation play a role by reducing the risk of exposure to infectious droplets. Pathogen diversity is increasingly appreciated to have important phenotypic consequences (Coscolla *et al.*, 2014), further discussed in the next sections.

1.3. Diversity in *Mycobacterium tuberculosis* complex

Mycobacterium tuberculosis (Mtb) complex comprises bacterial species and sub-species that cause TB in a wide range of mammalian hosts; both human and animals (Figure 1.3). The animal-adapted Mtb include those infecting wild animals: *M. microti* (voles), *M. orygis* (oryx, antelopes, gazelles, waterbucks and deers), chimpanzee bacillus (chimpanzee), *M. pinnipedii* (sea lions and seals), *M. mungi* (mongoose), *M. suricattae* (meerkats), the Dassie bacillus (hyrax) and those infecting domestic animals: *M. caprae* (goats and sheep) and *M. bovis* (cattle) (Alexander *et al.*, 2010; Brosch *et al.*, 2002; Coscolla *et al.*, 2013; Cousins *et al.*, 1994; Ingen *et al.*, 2012; Parsons *et al.*, 2013). However, *M. bovis* can also cause bovine TB in humans. On the other hand, Mtb sensu stricto and *M. africanum* are the typical human-adapted species. *M. canettii* is a peculiar and distantly related member of the complex which belongs to the group of "smooth tubercle bacilli" (STB) (Gutierrez *et al.*, 2005). The STB are suggested to contain the putative ancestor of Mtb (Gutierrez *et al.*, 2005). So far, *M. canettii* has been isolated from the Horn of Africa and differs from its counterparts Mtb and *M. africanum* in colony morphology, higher genetic diversity and evidence of horizontal gene exchange (Supply *et al.*, 2013). In addition, *M. canettii* isolates show no evidence of human-to-human transmission, and epidemiological data suggest that they are environmental organisms (Soolingen *et al.*, 1997).

In general, strains of Mtb are genetically monomorphic compared to other bacteria (Achtman, 2008), and therefore had been considered "identical" in the past. It was not until three decades ago that strains of Mtb were known to exhibit differences at the levels of DNA sequence. Small insertions and deletions (indels), large duplications and insertion sequences (Embden *et al.*, 1993), large genomic deletions (Gagneux *et al.*, 2006) and single nucleotide polymorphisms (SNPs) (Stucki *et al.*, 2012) for instance are important sources of genetic diversity in Mtb and have been used as molecular markers to explore the Mtb diversity.

Earlier phylogenetic analyses defined Mtb into evolutionary "ancient" and "modern" groups based on the presence and absence, respectively, of a specific deletion, TbD1 (Brosch *et al.*, 2002). Presently, the human-adapted Mtb is classified into seven main phylogenetic lineages (Comas *et al.*, 2013). These include: Lineage 1 (Indo-Oceanic), Lineage 2 (East-Asian), Lineage 3 (East-African-Indian), Lineage 4 (Euro-American), Lineage 7; and the *M. africanum* lineages: Lineage 5 (West African 1), 6 (West African 2). Further, Lineage 1, 5 and 6 are considered evolutionary "ancient" lineages whilst Lineage 2, 3, and 4 are the

evolutionary "modern" lineages. The phylogenetic lineages show a defined phylogeographical distribution consisting of broadly distributed lineages, Lineage 2 and 4 or "generalist" and locally restricted lineages, Lineages 5, 6 and 7 or "specialist", exclusively restricted in Africa. Lineage 1 and 3 show an intermediate geographical distribution. It is proposed that the diversity and distribution of Mtb lineages has been partly a consequence of adaption to human populations and their demographic history and migration.

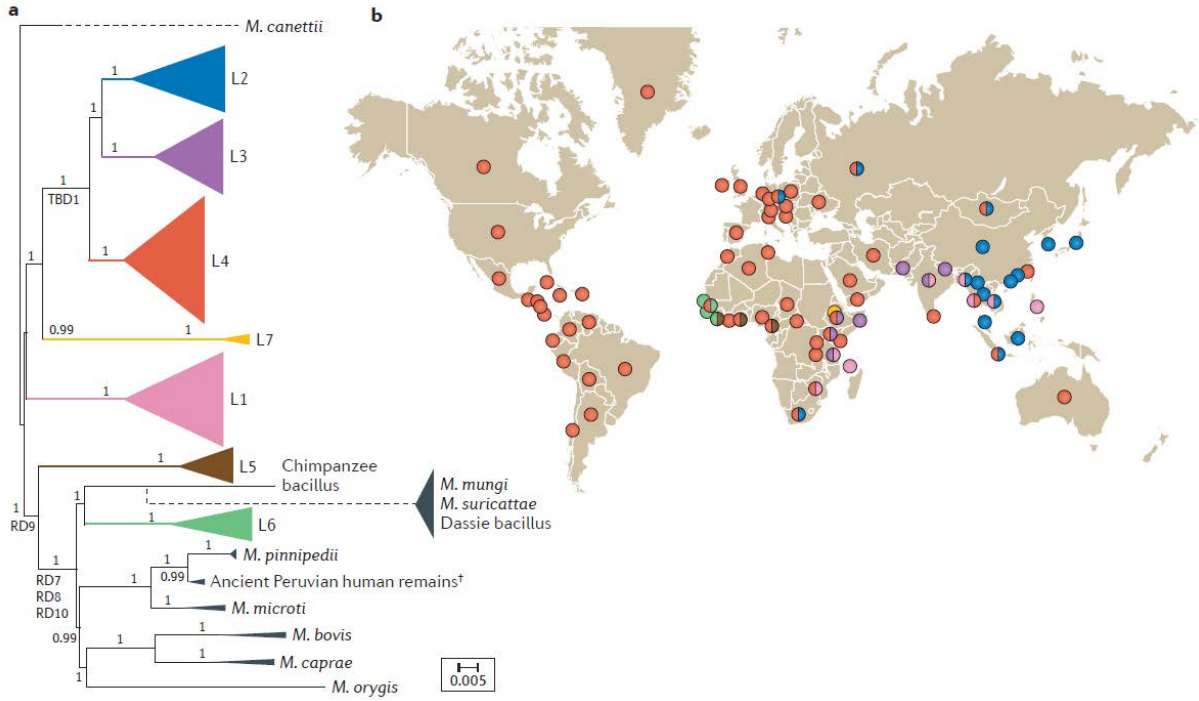


Figure 1.3.: **Global phylogeny and geographical distribution of the human-adapted Mtb complex.** a.) Whole genome-based phylogeny of the Mtb complex rooted on *M. canettii*. b.) Global distribution of the seven human-adapted Mtb lineages (Gagneux, 2018)

1.4. Molecular markers and typing of Mtb diversity

Since Mtb harbors low DNA sequence diversity, the standard sequence-based methods like multilocus sequence typing (MLST) only provide minimal phylogenetic resolution (Sreevatsan *et al.*, 1997). The early genotyping tools in Mtb were therefore based on mobile and repetitive genetic elements. Insertion Sequence (IS) 6110–restriction fragment length polymorphism (RFLP) (McEvoy *et al.*, 2007), was the first fingerprinting method used to classify Mtb based on the analysis of variable copy number of a mobile element IS6110 (Embden *et al.*, 1993). Polymerase chain reaction (PCR) based methods were devel-

oped afterwards and included spacer oligonucleotide typing (spoligotyping) (Kamerbeek *et al.*, 1997) and mycobacterial interspersed repetitive units–variable number of tandem repeats (MIRU–VNTR) (Supply *et al.*, 2000). The two former techniques classify Mtb based on repetitive elements; the clustered regularly interspaced short palindromic repeats (CRISPR) found at the direct repeat (DR) locus and the VNTRs, found at different loci in the Mtb genome.

The mobile and repetitive genetic elements are highly discriminatory markers given their rapid change and thus have been applied in the molecular epidemiology field (Kato-Maeda *et al.*, 2011). However, this property of rapid change of the markers leads to convergent evolution resulting into different Mtb strains independently acquiring similar fingerprint patterns (Comas *et al.*, 2009; Fenner *et al.*, 2011). This phenomenon is referred to as homoplasy and is a limitation for robust phylogenetic inferences (Comas *et al.*, 2009).

Ideally, phylogenetic markers should be stable and irreversible. In Mtb, large sequence polymorphisms (LSPs) or genomic deletions and SNPs have been described as such (Comas *et al.*, 2009). In addition, the two markers possess low rates of homoplasy. LSPs are robust phylogenetic markers because Mtb lack horizontal gene transfer (Gagneux *et al.*, 2006). However, genetic distances cannot be inferred from LSPs, and this hinders the possibility to explore evolutionary relationships among Mtb strains. The exponential increase in whole genome sequences of Mtb strains in recent years has allowed for the discovery of SNPs from which SNP-typing assays have been developed (Stucki *et al.*, 2012) both for lineage and within lineage i.e., sublineage typing (Stucki *et al.*, 2016). SNP-typing is a cost-effective approach to screen large collections of Mtb strains although WGS remains the most ideal phylogenetic tool as it gives the most detailed insight into the Mtb genome (Comas *et al.*, 2009).

1.5. Consequences of Mtb diversity

In addition to host and environment, Mtb variation forms a triad in determining the extent of TB infection and disease. Genetic variation in Mtb has been shown to translate to relevant biological and epidemiological phenotypes (Coscolla, 2017). Indeed, Mtb strains show different molecular and immunological phenotypes demonstrated in vitro or in vivo, and "in clinico" phenotypes demonstrated in epidemiological settings.

For example, recent advances in Mtb "-omics" revealed lineage specific transcriptional pro-

files and differences in mycolic acid biosynthesis. A well-established example of transcriptional phenotype is one involving a dormancy regulator (DosR). Lineage 2–Beijing strains have been shown to over express DosR, resulting to the accumulation of triglycerides, which may be a source of energy upon nutrient scarcity for example during transmission (Reed *et al.*, 2007; Rose *et al.*, 2013).

Lineage-specific patterns have also been observed in the biosynthesis of mycolic acids (MAs). Portevin *et al.*, 2014 reported significant variations in MAs profiles among Mtb phylogenetic lineages. Specific variations in MAs have been shown to play role in inflammatory responses elicited by different Mtb clinical strains. Moreover, production of other lipid classes such as a polyketide synthase-derived phenolic glycolipid, by strains of Lineage 2–Beijing was shown to inhibit / reduce the pro-inflammatory responses (Reed *et al.*, 2004).

The ability of Mtb strains to induce reduced or delayed inflammatory responses is often associated with higher virulence in infection models. Phylogenetically modern Mtb lineages have been demonstrated to elicit reduced or delayed immune responses in human monocyte-derived macrophages compared to ancient lineages: a property that could improve their survival and proliferation during the early stages of infection in order to achieve increased virulence at later stages (Portevin *et al.*, 2011).

The reduced pro-inflammatory immune responses are also linked to more rapid progression of TB. In an epidemiological setting in the Gambia, contacts exposed to TB patients infected with Lineage 2–Beijing were five times more likely to progress to active disease than those exposed to *M. africanum* Lineage 6 (Jong *et al.*, 2008). Studies conducted in other epidemiological settings show that Mtb strains also vary in severity of disease and their ability to transmit (Coscolla, 2017). Stavrum *et al.*, 2014 showed that TB patients in Tanzania infected with modern Lineage 4 had a lower body mass index and higher induced levels of acute phase reactants than those infected with ancient Lineage 1. These findings suggest enhanced virulence in modern strains. Reduced virulence and lower progression risk in ancient Lineage 6 for instance has been pinpointed to the presence of a non-synonymous SNP in PhoP, which is a virulence regulator (Gonzalo-Asensio *et al.*, 2014). Further evidence from epidemiological settings advocate modern lineages to show increased transmission compared to other lineages. Transmission potential is estimated from parameters such as tuberculin skin test conversion of household contacts (Jong *et al.*, 2008), genetic clustering of Mtb strains (Spuy *et al.*, 2003) and association with younger age of TB patient (Borgdorff *et al.*, 2013). Based on the afore mentioned transmission parameters, several studies have shown Lineage 2–Beijing strains to be highly transmissible

(Buu *et al.*, 2012; Holt *et al.*, 2018). Lineage 2–Beijing strains have also been associated with antibiotic resistance (Borrell *et al.*, 2009).

Taken together, evidence suggests Mtb modern lineages to be more virulent, thus considered more successful over their ancient counterpart, a suggestion that is also reflected by their wider geographical range. Linking Mtb genotype to phenotype is challenging as variation in Mtb strains occurs in various forms, including repetitive and mobile elements, deletions, duplications and SNPs, all of which have been shown to influence phenotypes (Coscolla *et al.*, 2014). Although for a long time, host genetics and the environment were known to be crucial determinants of TB infection and disease, pathogen factors cannot be excluded. Owing to their long-standing interactions, both the pathogen and host should be explored for better understanding of infection and disease; ideally, with the integration of socio-economic factors in a systems epidemiology manner (Comas *et al.*, 2009).

2. Aims and Objectives

2.1. Aims of the thesis

The overall aims of this thesis were i) to characterize the diversity of Mtb isolates in Tanzania at local and national level, ii) to study the epidemiological consequences of Mtb diversity, iii) to infer the origin and evolutionary history of Mtb Lineage 1 and 3, and African Lineage 2–Beijing.

2.2. Specific objectives

The following were the specific objectives of this PhD thesis:

Objective 1. To describe the countrywide diversity of Mtb in Tanzania (*Chapter 3*)

Objective 2. To study the molecular epidemiology of Mtb in urban and rural Tanzania (*Chapter 4*)

Objective 3. To study the evolutionary history of the Mtb Lineage 1 and Lineage 3 along the rim of the Indian Ocean (*Chapter 5*)

Objective 4. To investigate the origins of the Lineage 2–Beijing in Africa (*Chapter 6*)

2.3. Outline

In the next chapter (*Chapter 3*), we describe the genetic diversity of Mtb isolates from new and retreatment TB cases in Tanzania and describe associations of the lineages identified

with patients' clinical and demographic characteristics.

To get deeper insights into Mtb epidemiological phenotypes and understand transmission patterns, we focused on the urban and rural settings of Temeke district in Dar es Salaam and Ifakara in Morogoro Tanzania, respectively. This molecular epidemiological study is described in *Chapter 4*.

In *Chapter 5*, we studied the evolutionary history and global spread of Lineage 1 and 3. To achieve this, we used whole-genome sequencing data from global representative clinical strains to define their global phylogeography and infer their global dissemination.

In *Chapter 6*, we describe introductions of Lineage 2-Beijing into Africa followed by onward spread on the African continent.

In the last chapter (*Chapter 7*), we discuss the key findings, general implications and recommendations.

3. Insights into the genetic diversity of *Mycobacterium tuberculosis* in Tanzania

Liliana K. Rutaihwa[1,2,3*], Mohamed Sasamalo[1,2,3], Aladino Jaleco[1,2], Jerry Hella[1,2,3], Ally Kingazi[3], Lujeko Kamwela[1,2,3], Amri Kingalu[4,5], Bryceson Malewo[4,5], Raymond Shirima[4,5], Anna Doetsch[1,2], Julia Feldmann[1,2], Miriam Reinhard[1,2], Sonia Borrell[1,2], Daniela Brites[1,2], Klaus Reither[1,2], Basra Doulla[4,5], Lukas Fenner[1,2,6#], Sebastien Gagneux[1,2,#*]

[1] Swiss Tropical and Public Health Institute, Basel, Switzerland
[2] University of Basel, Basel, Switzerland
[3] Ifakara Health Institute, Bagamoyo, Tanzania
[4] Central Tuberculosis Reference Laboratory, Dar es Salaam, Tanzania
[5] National Tuberculosis and Leprosy Programme, Dar es Salaam, Tanzania
[6] Institute of Social and Preventive Medicine, University of Bern, Bern, Switzerland

[*] Corresponding authors
Email: liliana.rutaihwa@gmail.com (LKR) and sebastien.gagneux@swisstph.ch (SG)

[#] Equal contribution

This paper has been published in *PLoS ONE* 2019, 14(4):e0206334.

3.1. Abstract

Background

Human tuberculosis (TB) is caused by seven phylogenetic lineages of the *Mycobacterium tuberculosis* complex (MTBC), Lineage 1–7. Recent advances in rapid genotyping of MTBC based on single nucleotide polymorphisms (SNP), allow for phylogenetically robust strain classification, paving the way for defining genotype-phenotype relationships in clinical settings. Such studies have revealed that, in addition to host and environmental factors, strain variation in the MTBC influences the outcome of TB infection and disease. In Tanzania, such molecular epidemiological studies of TB however are scarce in spite of a high TB burden.

Methods and findings

Here we used SNP-typing to characterize a nationwide collection of 2,039 MTBC clinical isolates representative of 1.6% of all new and retreatment TB cases notified in Tanzania during 2012 and 2013. Four lineages, namely Lineage 1–4 were identified within the study population. The distribution and frequency of these lineages varied across regions but overall, Lineage 4 was the most frequent (n = 866, 42.5%), followed by Lineage 3 (n = 681, 33.4%) and 1 (n = 336, 16.5%), with Lineage 2 being the least frequent (n = 92, 4.5%). We found Lineage 2 to be independently associated with female sex (adjusted odds ratio [aOR] 2.14; 95% confidence interval [95% CI] 1.31 – 3.50, p = 0.002) and retreatment cases (aOR 1.67; 95% CI 0.95 – 2.84, p = 0.065) in the study population. We found no associations between MTBC lineage and patient age or HIV status. Our sublineage typing based on spacer oligotyping on a subset of Lineage 1, 3 and 4 strains revealed the presence of mainly EAI, CAS and LAM families. Finally, we detected low levels of multidrug resistant isolates among a subset of 144 retreatment cases.

Conclusions

This study provides novel insights into the MTBC lineages and the possible influence of pathogen–related factors on the TB epidemic in Tanzania.

3.2. Introduction

Tuberculosis (TB) is the leading cause of mortality due to an infectious disease (WHO, 2017). In 2017, an estimated 10 million people developed TB globally, with 1.6 million dying of the disease. Tanzania is among the 30 high burden countries, with a national average TB notification rate of 129 cases per 100,000; however, some regions show higher notification rates (NTLP, 2016). Like in most sub-Saharan African countries, the HIV epidemic contributes substantially to the high TB incidence in Tanzania, where a-third of the TB patients are co-infected with HIV (NTLP, 2016). Contrarily, drug-resistant TB is still low in this setting (Nagu *et al.*, 2015). Other risk factors such as poverty also influence the epidemiology of TB in Tanzania (MoHSW, 2013).

Transmission of TB occurs via infectious aerosols, where upon exposure individuals can either clear the infection, develop active disease or remain latently infected (Rieder, 1999). The complex dynamics of TB infection and disease are determined by the environment, the host and the pathogen (Comas *et al.*, 2009). Seven main phylogenetic lineages of the *Mycobacterium tuberculosis* complex (MTBC) (Lineage 1–7) cause TB in humans (Gagneux, 2018). These lineages are phylogeographically distributed, partially reflecting human migration histories (Gagneux *et al.*, 2006; Fenner *et al.*, 2013; Comas *et al.*, 2013). Genomic differences among the MTBC strains translate into relevant biological and epidemiological phenotypes (Coscolla, 2017). Epidemiologically speaking, these phenotypes are demonstrated by indicators such as transmission potential, disease severity and progression rates from infection to disease (Stavrum *et al.*, 2014; Holt *et al.*, 2018; Hanekom *et al.*, 2007; Cowley *et al.*, 2008). In general, strains of the widely distributed lineages, Lineage 2 and 4 or "generalists", appear to be more virulent than those of the geographically restricted lineages, Lineage 5 and 6 or "specialists" (Coscolla, 2017; Gagneux, 2018).

Studying genotype-phenotype relationships requires understanding the genetic diversity of MTBC clinical strains in a given clinical setting. In Tanzania, few studies have described the genetic diversity of the MTBC (Mfinanga *et al.*, 2014; Eldholm *et al.*, 2006; Kibiki *et al.*, 2007; Mbugi *et al.*, 2015). These previous works used conventional genotyping tools such as the spacer oligonucleotide typing (spoligotyping) technique and revealed the presence of mainly the East African Indian (EAI), Central Asian (CAS) and Latin American Mediterranean (LAM) spoligo families, and the Beijing family reported only at the lowest frequencies. Based on phylogenetically robust techniques, which include single nucleotide polymorphisms (SNPs), the spoligo families correspond to Lineage 1, 3, 4 and 2, respectively. These previous studies from Tanzania are limited as they only focused

on few specific geographical locations on the country and only one study profiled MTBC on a countrywide scale albeit with low sampling coverage (Mfinanga *et al.*, 2014). Moreover, despite the invaluable contribution of techniques like spoligotyping in the molecular epidemiology field, such techniques are suboptimal for phylogenetically robust strain classification due to high rates of convergent evolution (Comas *et al.*, 2009; Fenner *et al.*, 2011).

In this study, we applied for the first time a robust SNP typing method to classify the largest so far nationwide representative collection of clinical isolates to gain insights into unknown patterns of MTBC diversity in different regions of Tanzania. Given that only few studies have assessed and identified lineage-specific differences in clinical settings, we then looked for potential associations between the MTBC lineages and the available clinical and epidemiological characteristics of the patients in the study population.

3.3. Materials and Methods

3.3.1. Ethics statement

The study was approved by the National Tuberculosis and Leprosy Programme and the ethical clearance was provided by the National Institute for Medical Research of Tanzania (Dar es Salaam, Tanzania). The data in this study were analyzed anonymously.

3.3.2. The National Tuberculosis and Leprosy Program routine drug surveillance system

Our study was based on a nationwide convenience sample of sputum smear positive new and retreatment TB cases diagnosed between 2012 and 2013 in Tanzania. The collection was obtained via a platform established for routine TB drug resistance surveillance by the National Tuberculosis Leprosy Program (NTLP) of Tanzania, covering health facilities in all geographical regions of the country. Briefly, smear positive sputa specimens from approximately 25% of new TB cases and from all retreatment cases were obtained for the drug resistance surveillance. To obtain 25% sputa from new cases each region was allocated four months per annum, where the respective health facilities in the region submitted sputa samples to zonal reference laboratories for culture. The zonal laboratories include the Central Tuberculosis Reference Laboratory (CTRL) in Dar es Salaam, Bugando Medical Center (BMC) in Mwanza and Kilimanjaro Christian Medical Center

(KCMC) in Kilimanjaro, which serve the Coastal and Southern zone, the Lake zone, and the Northern zone, respectively. Isolates from the two zonal laboratories, BMC and KCMC were then sent to the CTRL for drug susceptibility testing (DST). For this study we included all the isolates we could retrieve from the culture archives at the CTRL.

3.3.3. Study population and data collection

We included a total of 2,039 unique (single patient) culture-confirmed TB cases, each of whom we could retrieve the respective culture isolate from the CTRL culture archives. This study population represents 1.6% of all the estimated TB notified cases in the country between 2012 and 2013 (Figure A.1). We also obtained corresponding socio-demographic and clinical information collected during patients' consultation at the respective health facilities. The demographic data collected included age, sex and geographical location of the patients, whereas clinical data included HIV status and disease category (i.e., new case and retreatment case).

3.3.4. Processing of culture isolates

The smear positive sputa samples were cultured on Loewenstein Jensen (LJ) growth medium according to laboratory protocols. For this study, we included MTBC clinical isolates retrieved from archived LJ media. We then prepared heat inactivated samples for the retrieved clinical isolates by suspending MTBC colonies into 1ml sterile water and heat inactivate at 95°C for one hour.

3.3.5. Molecular genotyping

We then classified the MTBC clinical isolates into main phylogenetic lineages by TaqMan realtime PCR according to standard protocols (Applied Biosystems, Carlsbad, USA) and as previously described (Stucki *et al.*, 2012). Briefly, the TaqMan PCR uses fluorescently labeled allele-specific probes for singleplex SNP-typing that are specific for each MTBC lineage. For comparisons, we also performed 43-spacer spoligotyping on a membrane for a subset of representative MTBC clinical strains following standard protocols (Kamerbeek *et al.*, 1997), since spoligotyping is still widely used as a gold standard for genotyping in similar settings. We randomly selected 107 samples out of the 2,039 representative of three lineages, Lineage 1, 3 and 4. We excluded Lineage 2 strains for this analysis given that such strains almost exclusively belong to the Beijing family. The clinical strains were

assigned to spoligotype families using the online database SITVITWEB (Demay *et al.*, 2012).

3.3.6. Drug resistance genotyping

We selected a subset of 144 clinical isolates from the 321 retreatment cases to perform molecular drug resistance testing. We used a previously described multiplex polymerase chain reaction (PCR) to target the rifampicin resistance determining region of *rpoB* gene (Malla *et al.*, 2012). The PCR assay targets both the tuberculous and non-tuberculous Mycobacteria (MTBC and NTMs, respectively) *rpoB* gene, so we could also rule out the presence of non-tuberculous isolates in our study sample using the assay. The amplified *rpoB* gene product was confirmed by electrophoresis on a 2% agarose gel and sent for Sanger sequencing. We analyzed the resulting sequences by Staden software package (Staden, 1996) and using MTBC H37Rv *rpoB* gene as reference sequence.

3.3.7. Statistical analysis

For statistical analysis, we applied descriptive statistics to delineate patients' characteristics. We used Chi-square or Fisher's exact tests for assessment of differences between groups in categorical variables, whenever applicable. We used univariate and multivariate logistic regression models to assess for the association between MTBC lineages and patients' clinical and demographic characteristics. The associations were assessed for Lineage 2 compared to all other lineages (Lineages 1, 3 and 4), adjusting for age, sex, disease category and HIV status. All statistical analyses were performed in R 3.5.0 (R Core Team, 2018).

3.4. Results

3.4.1. Patients' demographic and clinical characteristics

The patients' demographic and clinical information in our study included age, sex, geographical location, HIV and disease category (new or retreatment case). Table 3.1 describes patients' characteristics of the study population. The proportions of the observed and missing data for the study population are summarized in Figure A.2.

Our study population consisted of TB patients ranging between the age of 2 and 89 years with a median age of 28 years (interquartile range [IQR] 27–44). To further probe the age

distribution in the study population, we categorized the TB patients into five different age groups (Table 3.1). We detected approximately three-quarters of the TB cases to occur among the "young age" and "early adult" age groups. Further, our findings show that about 1.0% of the TB cases were pediatric cases (< 15 years).

Similar to other settings (WHO, 2017), we identified a higher proportion of male TB cases compared to female TB cases. However, the male-to-female ratio observed in our study population was higher than the national estimates for the two years of sampling (2.2:1 vs., 1.4:1). The striking gender imbalance among TB cases seems to peak at adolescence onwards and is less pronounced among pediatric TB cases (Table A.1). Additionally, a-third (32.2%, 517/1604) of the TB cases with available HIV status were HIV co-infected. In contrast, TB/HIV co-infected cases were more likely to be female (44.5%, CI 38.3–50.7% vs., 25.8%, 95% CI 20.6–31.0%) which is consistent with HIV being generally more prevalent in females than males in Tanzania (Hegdahl *et al.*, 2016). We found that our study population comprised 16.1% (321/2000) of TB retreatment cases, which was four-fold higher than the overall countrywide notifications (NTLP, 2013).

Finally, more than half (51.6%, 1029/1996) of the TB patients in our study population were diagnosed in the Coastal zone of Tanzania and about 40% were either diagnosed in the Lake and Northern zones. In addition to higher TB notification rates, the three former mentioned geographical zones contain the country's zonal TB reference laboratories. The remaining 10% of the patients were diagnosed in any of the remaining four geographical zones of Tanzania.

3.4.2. Main MTBC lineages in Tanzania

Using SNP-typing, we detected four of the seven known MTBC lineages (Figure 3.1), albeit at varying proportions. In our study population, Lineage 4 and Lineage 3 were the most frequent (866, 42.5% and 681, 33.4%, respectively) followed by Lineage 1 (336, 16.5%). Lineage 2 was the least frequent (92, 4.5%). The remaining 64 clinical isolates (3.1%) could not be assigned to any of the MTBC lineages possibly because there was insufficient amount of DNA in the samples (below the detection limit). Of the seven geographical zones, four (Coastal, Northern, Lake and Southern Highlands) were highly represented with more than 100 clinical strains each (Table 3.2). The distribution of the MTBC lineages varied within the geographical zones (Figure 3.1 and Figure A.3). Our findings reveal that Lineage 1 strains were more frequent in the Lake zone compared to the overall average frequency (20.9% vs. 16.8%), whereas the frequency of Lineage 3 in this

Table 3.1.: Clinical and demographic characteristics of the TB cases.

Characteristics	**Valid%**	**Total n = 2,039 (%)**
Age, median (IQR)		
35 (27-44)		
Age groups (years)		
Child age (<15)	9.9	20 (1.0)
Young age (15 - 24)	29.7	312 (15.3)
Early adult (25 - 44)	48.0	1,170 (57.4)
Late adult (45 - 64)	10.0	379 (18.6)
Old age (>65)	2.5	73 (3.6)
Not available		85 (4.2)
	valid n = 1,954	
Sex		
Female	32.4	645 (31.6)
Male	67.6	1,346 (66.0)
Not available		48 (2.4)
	valid n = 1,991	
HIV status		
Negative	67.7	1,086 (53.3)
Positive	32.2	517 (25.4)
Indeterminate	0.06	1 (0.1)
Not available		435 (21.3)
	valid n = 1,604	
Case		
New case	84.0	1,679 (82.3)
Retreatment	16.1	321 (15.7)
Not available		39 (1.9)
	valid n = 2,000	
Zonal region		
Central	1.1	22 (1.1)
Coastal	51.6	1,029 (50.5)
Lake	17.9	358 (17.6)
Northern	20.2	403 (19.8)
S. Highlands	8.1	162 (8.0)
Western	0.5	10 (0.5)
Zanzibar	0.6	12 (0.6)
Not available		43 (2.1)
	valid n = 1,996	

IQR, interquartile range; valid proportion, proportion excluding missing values; Total n, all values including NA (not available).

zone was lower (27.6% vs. 34.3%) compared to other geographical zones. By contrast, Lineage 4 was the most predominant in all geographical zones and showed relatively similar frequencies across the zones.

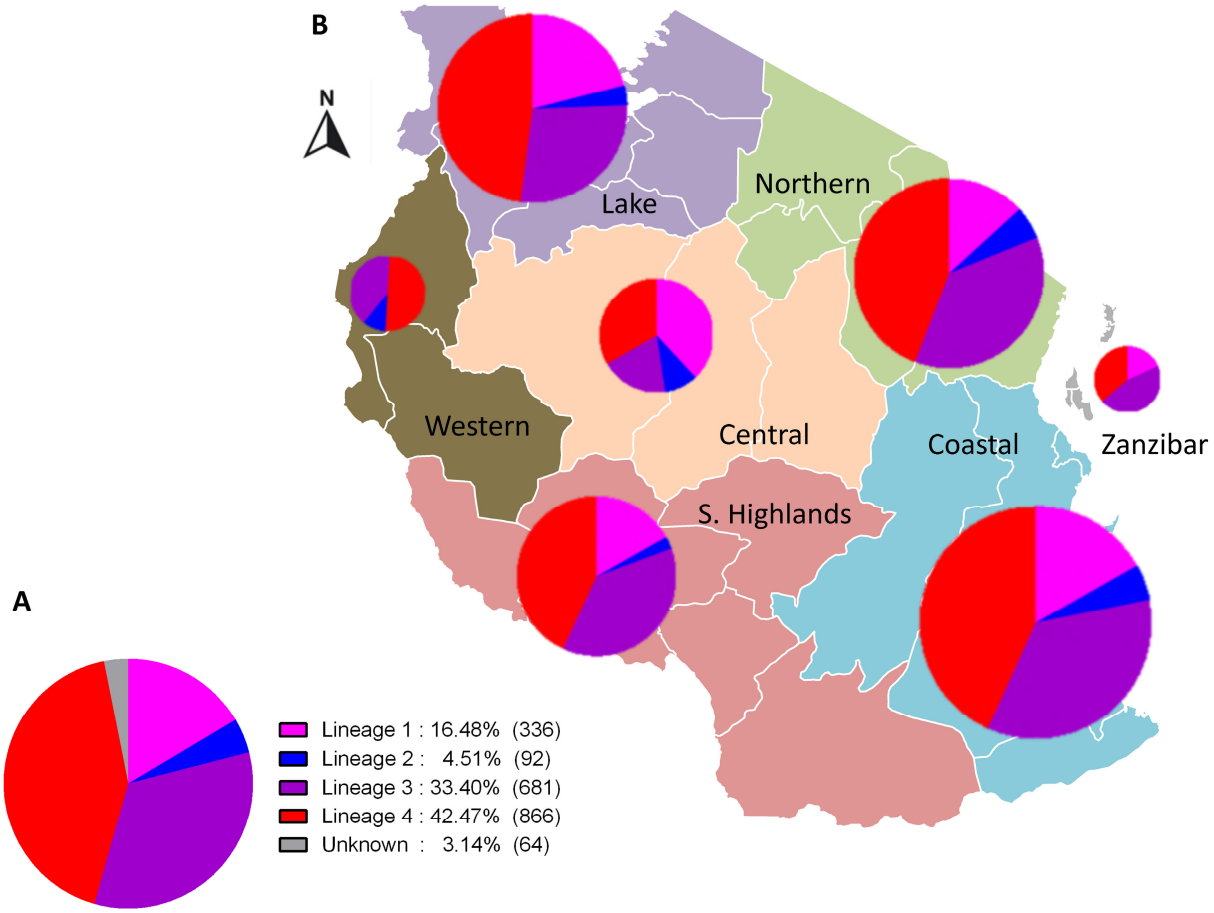


Figure 3.1.: **MTBC lineages in Tanzania.** A. MTBC lineage classification of 2,039 nationwide clinical strains. B. MTBC lineage frequencies and geographical distribution in Tanzania.

Table 3.2.: MTBC lineage distribution across geographical regions in Tanzania.

Geographical Zone	**Lineage**				**Total**
	Lineage 1 (%)	Lineage 2 (%)	Lineage 3 (%)	Lineage 4 (%)	
Central	8 (38.1)	2 (9.5)	4 (19)	7 (33.3)	21
Coastal	168 (16.8)	50 (5)	350 (35)	432 (43.2)	1,000
Lake	72 (20.9)	12 (3.5)	95 (27.6)	165 (48)	344
Northern	52 (13.3)	22 (5.6)	145 (37)	173 (44.1)	392
S. Highlands	27 (16.9)	4 (2.5)	60 (37.5)	69 (43.1)	160
Western	0 (0)	1 (10)	4 (40)	5 (50)	10
Zanzibar	2 (18.2)	0 (0)	5 (45.5)	4 (36.4)	11
Total	329 (17)	91 (4.7)	663 (34.2)	855 (44.1)	1,938

3.4.3. Sublineage classification

After we detected four main MTBC lineages, we next explored the respective subfamilies on a subset of Lineage 1, 3 and 4 strains using spoligotyping. Lineage 2 strains were excluded from this analysis since the global strains almost exclusively belong to one spoligotype family, Beijing with almost identical fingerprint pattern. We identified 24 spoligotypes (SITs; Spoligotype International Type) among the 107 clinical strains analyzed (Figure A.6). Twenty six (24.3%) of the spoligo patterns had not been previously reported in the SITVITWEB database and therefore we assigned them as orphan spoligotypes. Several spoligotypes were identified within each of the three lineages. EAI5 was the common spoligotype among the Lineage 1 and CAS1_Kili spoligotype among the Lineage 3 strains. Within Lineage 4 strains, LAM, T, and H families were detected and the LAM subfamily, particularly LAM_ZWE was the most frequent.

3.4.4. Association between lineage and patients' characteristics

Having described the circulating main lineages of the MTBC we then assessed the relationship between these lineages and patients' characteristics in the study population (Table 3.3). We detected a higher proportion of female sex among TB patients infected with Lineage 2 (52.1%) compared to those infected with the other three lineages (range from 31% to 34.5%, p = 0.009). Moreover, we observed that retreatment cases were frequently infected with Lineage 2 strains (26.8%), which was twofold higher compared to Lineage 1 and 4 strains ($p < 0.001$). We found no evidence for association between lineages and patients' characteristics such as age and HIV status (Table 3.3).

Lineage 2 has previously been associated with retreatment cases, drug resistance and lately also with female sex (Holt *et al.*, 2018; Malla *et al.*, 2012). We therefore investigated if similar associations exist in our study population using a subset of TB cases with complete clinical and demographic information (n = 1,535). To assess these associations we performed logistic regression analyses comparing Lineage 2 to all other lineages pooled together (Table 3.4). Our analyses revealed Lineage 2 to be independently associated with female sex (adjusted odds ratio [aOR] 2.14; 95% confidence interval [95% CI] 1.31 – 3.50, p = 0.002) and retreatment cases (aOR 1.67; 95% CI 0.95 – 2.84, p = 0.065). We did not detect any association between the lineages and patients' age and the HIV status.

Table 3.3.: Frequency distribution of MTBC main lineages across patients' characteristic groups.

Patient characteristics	**Lineage**			
	Lineage 1, n (%)	**Lineage 2, n (%)**	**Lineage 3, n (%)**	**Lineage 4, n (%)**
Age group				
Child age (<15)	2 (0.7)	0 (0)	7 (1.3)	8 (1.2)
Young age (15–24)	44 (16.2)	10 (13.9)	79 (14.9)	112 (16.9)
Early adult (25–44)	156 (57.4)	44 (61.1)	336 (63.5)	378 (57.1)
Late adult (45–64)	57 (21.0)	15 (20.8)	85 (16.1)	140 (21.1)
Old age (>65)	13 (4.8)	3 (4.2)	22 (4.2)	24 (3.6)
Sex				
Female	85 (31.3)	37 (51.4)	184 (34.8)	223 (33.7)
Male	187 (68.8)	35 (48.6)	345 (65.2)	439 (66.3)
HIV status				
Negative	184 (67.6)	45 (62.5)	349 (66.0)	459 (69.3)
Positive	88 (32.4)	27 (37.5)	180 (34.0)	203 (30.7)
Patient category				
New case	235 (86.4)	53 (76.6)	405 (76.6)	560 (84.6)
Retreatment	37 (13.6)	19 (26.4)	124 (23.4)	102 (15.4)
Total	272 (17.7)	72 (4.7)	529 (34.4)	662 (43.2)

Table 3.4.: Associations of patients' clinical and demographic characteristics with MTBC Lineage 2 (n = 72) compared to all other lineages (n = 1,463).

Patient characteristics	**Lineage 2**	**Unadjusted**		**Adjusted**	
	n (%)	**OR (95% CI)**	**p value**	**OR (95% CI)**	**p value**
Age, median (IQR)	35 (28–44)			0.99 (0.97 – 1.01)	0.329
Female sex	37 (51.4)	2.09 (1.30–3.36)	0.002	2.14 (1.31 – 3.50)	0.002
Retreatment case	19 (26.4)	1.64 (0.93–2.76)	0.075	1.67 (0.95 – 2.84)	0.065
HIV positive	27 (37.5)	0.79 (0.49–1.31)	0.349	0.90 (0.55–1.51)	0.91
Observations		1,535		1,535	

3.4.5. Mutations within *rpoB* gene in retreatment cases

To investigate whether drug resistance was linked to a particular lineage, we included in total 144 out of 321 retreatment cases for drug resistance genotyping of the *rpoB* gene that confers resistance to rifampicin. Out of these, 112 (77.8%) had no mutations compared to the H37Rv reference gene and 15 (10.4%) contained at least one mutation, either synonymous (3/15) or non-synonymous (12/15) (Figure A.4 and Table A.2). We could not determine mutation status in the *rpoB* gene of 17 (11.8%) retreatment cases due to PCR and sequencing failure. Among the 12 strains detected with non-synonymous *rpoB* mutations, five belonged to Lineage 2, four to Lineage 4, and three to Lineage 3 (Table A.3). Table 3.5 summarizes the non-synonymous *rpoB* mutations detected.

Table 3.5.: Non-synonymous mutations detected on the *rpoB* gene among retreatment cases.

Lineage	***rpoB* mutation**	**Amino acid change**	**n**	**Source**
Lineage 2	A1198G;C1349T	T400A;S450L	1	(Phelan *et al.*, 2016; Walker *et al.*, 2015)
	C1333T	H445Y	1	(Walker *et al.*, 2015)
	C1349T	S450L	3	(Walker *et al.*, 2015)
Lineage 3	T1289C	L430P	1	(Miotto *et al.*, 2018)
	C1333T	H445Y	1	(Walker *et al.*, 2015)
	C1349T	S450L	1	(Walker *et al.*, 2015)
Lineage 4	A1334T	H445L	1	(Walker *et al.*, 2015)
	G1333C	H445D	1	(Walker *et al.*, 2015)
	C1294G;A1442G	Q432E;E481A	1	(Miotto *et al.*, 2018; Heyckendorf *et al.*, 2017)
	C1333T	H445Y	1	(Walker *et al.*, 2015)
Total			12	

3.5. Discussion

In this study, we classified the countrywide collection of 2,039 MTBC isolates representing 1.6% of all smear positive new and retreatment TB cases notified during 2012 and 2013 in Tanzania. Our findings show that the MTBC strains among the study population are diverse, comprising four main phylogenetic lineages (Lineage 1–4) which occur throughout the country. Specifically, we found that Lineage 4 was the most frequent, followed by Lineage 3 and 1. Despite Lineage 2's recent global dissemination (Cowley *et al.*, 2008), it was the least frequent in our study population. Finally, our analysis on the relationship between MTBC lineages and patients' characteristics revealed associations of Lineage 2 with female sex and retreatment TB cases included in the study population.

Among the 7 human–adapted MTBC lineages, Lineage 4 is the most broadly distributed and occurs at high frequencies in Europe, the Americas and Africa (Demay *et al.*, 2012; Stucki *et al.*, 2016). In our study, we observe that TB epidemics in Tanzania are also predominated by Lineage 4, which is regarded as the most successful of MTBC lineages (Stucki *et al.*, 2016). In general, the wide geographical range of Lineage 4 is postulated to be driven by a combination of its enhanced virulence, high rates of human migration linked to its spread and ultimately its ability to infect different human population backgrounds (Coscolla *et al.*, 2014; Stucki *et al.*, 2016). In contrast, Lineage 1 and 3 are known to be mainly confined to the rim of the Indian Ocean (Gagneux, 2018), which is consistent with our observation that nearly 50% of the MTBC strains included in the study belong to these two lineages. This high frequency of Lineage 1 and 3 likely reflects the long-term migrations between Eastern Africa and the Indian subcontinent (O'Neill *et al.*, 2019). In addition, the distribution and frequency of Lineage 1 and 3 in the mainland subset did not vary from that of the coastal region, suggesting spread via internal migrations. Lineage 1 was proposed to have evolved in East Africa prior disseminating out of the continent (Comas *et al.*, 2013). Based on this, one might expect higher frequencies of Lineage 1 in the region. Instead, the so called "modern" (TbD1–) lineages (4 and 3 in this case) could be dominating in Tanzania despite presumably being introduced into the African continent only after the first European contact (Stucki *et al.*, 2016; Comas *et al.*, 2013). This perhaps illustrates the ability of "modern" lineages to thrive in co-existence with the pre-existing "ancient" (TbD1+) lineages such as Lineage 1 in our case, perhaps because of the comparably higher virulence (Portevin *et al.*, 2011; Stavrum *et al.*, 2014). The neighboring countries of Tanzania on the other hand show comparable MTBC lineage composition to our study (Mbugi *et al.*, 2016; Chihota *et al.*, 2018), suggesting common demographic histories and ongoing exchanges that resulted into distinct MTBC

populations. Our findings would suggest the frequency of Lineage 2–Beijing in Tanzania, like in most parts of the continent except for South Africa (Mbugi *et al.*, 2016; Chihota *et al.*, 2018) to be relatively low, despite the long-standing African-Asian contacts (Chihota *et al.*, 2018). Evidence from recent studies show that Lineage 2–Beijing was only recently introduced into Africa (Cowley *et al.*, 2008; Rutaihwa *et al.*, 2019a).

The burden of TB disease is generally higher in males (Guerra-Silveira *et al.*, 2013; WHO, 2017) rendering male sex as a potential risk factor for TB. Furthermore, the male bias among TB patients is also observed in settings with no obvious sex-based differences in health-seeking behavior (Rhines, 2013). Whilst we show similar trends in this study, our findings reveal that the proportion of females was higher among TB patients infected with Lineage 2. This finding is consistent with several other previous studies conducted in different settings (Holt *et al.*, 2018; Malla *et al.*, 2012; Buu *et al.*, 2009). Social and physiological factors predisposing males to higher risk of TB have been indicated (Nhamoyebonde *et al.*, 2014). On the one hand, these include risk behaviors such as substance abuse (alcoholism, tobacco smoking) and gender specific roles such as risk occupations (e.g., mining) that are male dominated and known to increase the risk for TB. On the other hand, genetic makeup and sex hormones might contribute to the differences in TB susceptibility among females and males, as epidemiological and experimental studies have suggested female sex hormones to be protective (Nhamoyebonde *et al.*, 2014). These observations would propose that the sex imbalance in TB emerges after the onset of puberty. Of note, we observe less sex imbalance in "child" age group (< 15 years) which also corroborates the national notification rates (NTLP, 2013). However, this observation can be confounded by BCG vaccination which might be most effective in this age group. Despite the high prevalence of HIV among young females in sub-Saharan Africa (Hegdahl *et al.*, 2016) and HIV being the strongest risk factor for TB, TB burden remains higher in males. While social and physiological aspects play an important role, findings from this study and others previously conducted in Nepal and Vietnam (Holt *et al.*, 2018; Malla *et al.*, 2012) suggest that bacterial factors could disrupt the trends towards male bias in TB, a finding which warrants further investigation. Our hypothesis is that because of higher virulence, Lineage 2 strains are able to overcome the resistance poised by female sex which could explain the less pronounced sex imbalanced.

In addition to its association with female sex, we found that retreatment TB cases were more likely infected with Lineage 2. A retreatment case in our study population represented recurrent TB case either due to relapse or reinfection. We hypothesized that this observation was possibly linked to drug resistance, given the previous reported association

between Lineage 2 and drug resistance (Borrell *et al.*, 2009). However, we detected only 8.3% (12/144) of strains among the retreatment subset tested to contain mutations conferring resistance to rifampicin, five of which belonged to Lineage 2. These findings would suggest that retreatment cases included in this study are mainly driven by reinfection as opposed to treatment failure or relapse. Finally, based on the age distribution of TB cases in our study, recent or ongoing transmission in high burden countries is implicated as the main contributor to the TB burden rather than disease reactivation (following longer latency periods) (Yates *et al.*, 2016). Additionally, an association with young age has been used as an epidemiological proxy for highly transmissible strains and faster rates of disease progression (Jong *et al.*, 2008; Borgdorff *et al.*, 2013). In this study, we did not detect any differences in median age of TB patients infected with different lineages (Figure A.5), an observation that could speak for high ongoing transmission rates in general, irrespective of lineage.

Our study is limited by focusing on a convenient collection of MTBC clinical isolates that could be retrieved from the culture archives, representing 1.6% of all TB cases notified in 2012 and 2013. Given that our findings are based on a limited number of TB cases, the results particularly those related to associations between lineages and patients' characteristics should be taken with caution as the strength or lack of such associations could likely be affected by the sampling. In addition, most of the geographical zones were underrepresented which could in turn underestimate the respective regional lineage composition and the overall countrywide distribution. Unfortunately, data on drug susceptibility based on other methods such as Xpert MTB/RIF, phenotypic DST and Line Probe Assay (LPA) were unavailable, which could have complemented the drug resistance genotyping performed on a limited subset of the retreatment cases. Systematic sampling would allow for better resolution on the distribution patterns, the frequencies and on epidemiological features of MTBC lineages, which might partially determine the regional specific epidemics.

In conclusion, this study addresses for the first time the countrywide MTBC population structure based on robust SNP-typing. We show that MTBC population in Tanzania is diverse with four of the seven known lineages detected. This study sets the stage for further in depth investigations on epidemiological impact of MTBC lineages in Tanzania.

3.6. Supporting information

Supplementary Figures and Tables are available online under `https://doi.org/10.1371/journal.pone.0206334` and on appendix A of the thesis.

Figure A.1. Flowchart illustrating estimated notified TB cases in 2012 and 2013 (dashed lines) and the study population (solid line).

Figure A.2. Patients' data included in the study. Proportion of observed and missing data for the variables included in the study.

Figure A.3. MTBC lineage proportions. Distribution of MTBC lineages across different regions of Tanzania. Size of the circle is proportional to the number of isolates analyzed from the regions. MTBC lineage proportions.

Figure A.4. Flowchart of genotyped strains for *rpoB* mutations. A subset of MTBC strains from retreatment cases included for *rpoB* drug resistance genotyping.

Figure A.5. Patients' age distribution across MTBC lineages. The age distributions of TB patients grouped by infecting MTBC lineage.

Figure A.6. Spoligotype patterns of a subset of MTBC clinical strains.

Table A.1. Sex distribution across different age groups of TB patients.

Table A.2. Mutations detected in the *rpoB* gene.

Table A.3. Distribution of *rpoB* mutations across the four MTBC lineages.

3.7. Acknowledgments

We would like to thank the National Tuberculosis Leprosy Programme (NTLP) through the Central Tuberculosis Reference Laboratory (CTRL) for permission to use the MTBC isolate collection for this study.

4. Molecular epidemiology of *Mycobacterium tuberculosis* in urban and rural Tanzania

Liliana K. Rutaihwa[1,2,3], Jerry Hella[1,2,3], George Sikalengo[3], Francis Mhimbira[1,2,3], Chloe Loiseau[1,2], Mohamed Sasamalo[3], Hellen Hiza[3], Lujeko Kamwela[3], Miriam Rheinhard[1,2], Julia Feldmann[1,2], Sonia Borrell[1,2], Emilio Letang[1,2,3], Klaus Reither[1,2], Daniela Brites[1,2], Lukas Fenner[1,2,4], Sebastien Gagneux[1,2]

[1] Swiss Tropical and Public Health Institute, Basel, Switzerland
[2] University of Basel, Basel, Switzerland
[3] Ifakara Health Institute, Bagamoyo, Tanzania
[4] Institute of Social and Preventive Medicine, University of Bern, Bern, Switzerland

4.1. Abstract

Human tuberculosis (TB) is primarily caused by seven human–adapted *Mycobacterium tuberculosis* (Mtb) phylogenetic lineages. Together with host and environmental determinants, factors related to Mtb genetic variation are known to modulate the outcome of TB infection and disease. In this study, we assessed for clinical phenotypes associated with Mtb lineages in urban and rural Tanzania, a country with a high TB burden.

We studied 900 pulmonary TB patients recruited at the urban setting of Temeke, Dar es Salaam between November 2013 and June 2018, and 242 pulmonary TB patients recruited at the rural setting of Ifakara, Morogoro between August 2015 and October 2018. We used a combination of single nucleotide polymorphism (SNP)-typing and whole genome sequencing (WGS) to characterize 764 (84.9%) Mtb isolates from the urban setting and 110 (45.5%) Mtb isolates from the rural setting. Sociodemographic and clinical information obtained during patient enrolment was used to investigate the epidemiological relevance of Mtb genetic variation in the two settings.

Our findings revealed four of the seven human-adapted Mtb lineages in the study sample, with Lineage 3 and 4 predominating in both settings, accounting for 367 (48%) and 254 (33.3%) cases in the urban setting, and 55 (50%) and 40, (36.4%) in the rural setting. Similar epidemiological features were observed between the urban and rural settings. However, Lineage 2 was more frequent in patients of young age in the urban setting (adjusted odds ratio [aOR] 0.96; 95% confidence interval [95% CI] 0.93–1, p = 0.03). No evidence for Mtb lineage associations with any other patient characteristic was found in either setting. Further analysis of 515 Mtb genomes from urban patients revealed that ongoing transmission is disproportionally due to strains of the modern lineages, Lineage 2–4.

Our findings provide novel insights into the genetic diversity of Mtb and the influence of pathogen–related factors on the TB epidemic in urban and rural Tanzania.

4.2. Introduction

Approximately a quarter of the 10 million new tuberculosis (TB) cases estimated to have occurred in 2017 were from Africa (WHO, 2018). The burden of TB is highest in sub-Saharan African countries, which include Tanzania. Tanzania is also recognized as one of the 30 high-burden TB countries, with a reported annual average incidence of 129 per 100,000 (NTLP, 2016). Prevalence of TB is particularly high amongst HIV co-infected individuals. About one-third of TB cases is also HIV co-infected, making HIV the most important risk factor for TB (NTLP, 2016). In addition, poor social economics, population growth and urban overcrowding contribute to this high TB burden (Dye *et al.*, 2010; MoHSW, 2013). In Tanzania, prevalence of TB varies across regions, and the TB burden is particularly high in dense populated regions (Figure 4.1). Of note, Dar es Salaam, one of the fastest growing African cities and most densely populated in Tanzania, is the largest contributor of all notified TB cases in the country. By contrast, drug resistance (DR)–TB across the country is low; approximately 1% in new cases and 3% in retreatment cases (Nagu *et al.*, 2015).

Human TB is primarily caused by seven human-adapted phylogenetic lineages (Lineage 1–7) of the *Mycobacterium tuberculosis* (Mtb) complex. Evidence suggests that genetic variation in strains of Mtb can lead to different biological and clinical phenotypes (Coscolla, 2017). Epidemiological measures such as patient sociodemographic and clinical characteristics, disease category (new cases, relapse), disease severity, site of disease (pulmonary, extra pulmonary), treatment outcomes, drug resistance, and transmission potential are used to make such inferences. On the one hand, the so-called "ancient" lineages including those of the *M. africanum* lineages, Lineage 5 and 6, and Lineage 1 are considered to be less virulent in average compared to the "modern" lineages, Lineage 2, 3 and 4 (Jong *et al.*, 2008; Coscolla *et al.*, 2014; Coscolla, 2017). Modern lineages on the other hand, which include Mtb *sensu stricto* lineages, Lineage 2–4 contain more virulent strains which are also more transmissible in average (Holt *et al.*, 2018; Yang *et al.*, 2012), and exhibit enhanced disease progression and severity (Jong *et al.*, 2008; Stavrum *et al.*, 2014).

Several studies have provided insights into the epidemiological consequences of Mtb lineages in different clinical settings, including Africa (Jong *et al.*, 2008; Guerra-Assunção *et al.*, 2015b; Holt *et al.*, 2018). For instance, a study conducted in Malawi showed differences by lineage in their contribution to recent transmission and disease recurrence (Guerra-Assunção *et al.*, 2015b). Based on previous studies, four of the seven known Mtb lineages have been reported in Tanzania (Mfinanga *et al.*, 2014; Mbugi *et al.*, 2016;

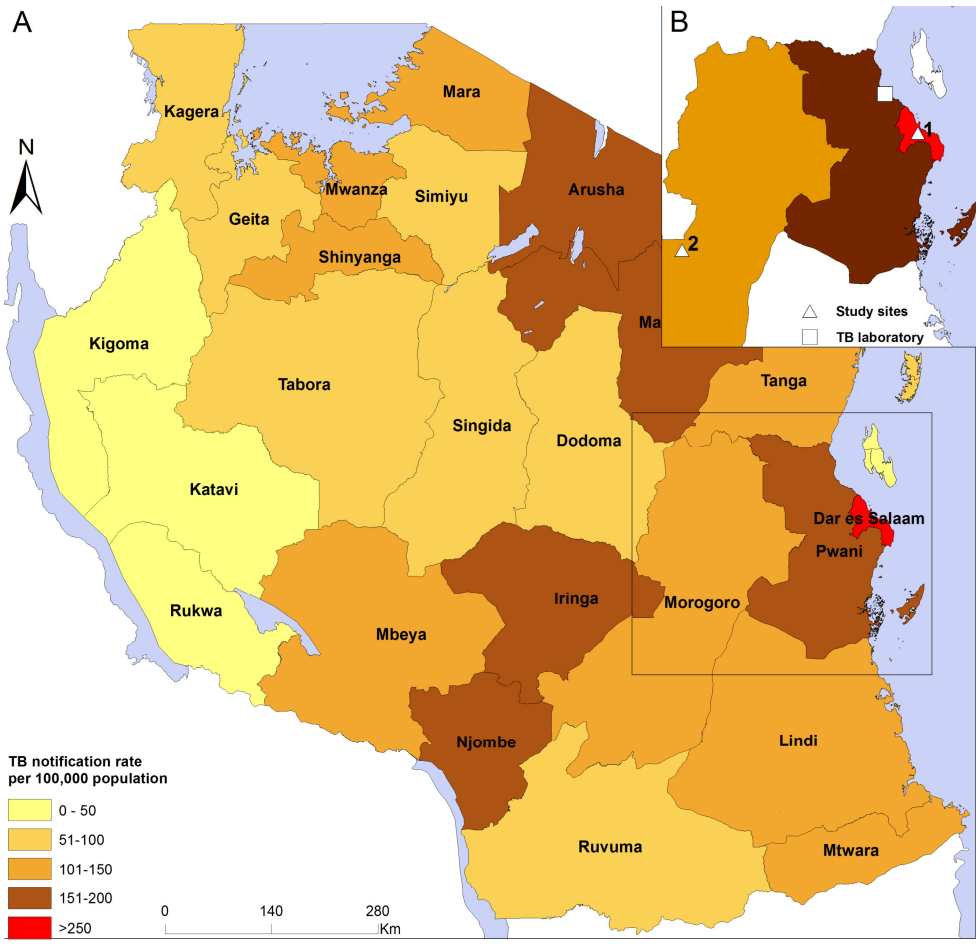


Figure 4.1.: Map of Tanzania illustrating (A) the regional tuberculosis (TB) notification rates and (B) the locations of the study sites (triangles): study site 1 (urban), Temeke District, Dar es Salaam Region; study site 2 (rural), Ifakara, Kilombero District, Morogoro Region; and the TB laboratory (square) in Bagamoyo, Pwani Region (Sikalengo *et al.*, 2018).

Chihota *et al.*, 2018) but only a few studies have assessed for lineage associations with clinical phenotypes in Tanzania (Stavrum *et al.*, 2014; Rutaihwa *et al.*, 2019b).

In this study, we focused on urban setting of Temeke, Dar es Salaam to investigate the epidemiological consequences of Mtb lineages. We collected the most comprehensive clinical and socio–demographic data to date to infer in clinico phenotypes based on single nucleotide polymorphisms (SNPs)-typing complemented with whole genome sequencing (WGS). Further, we examined genetic clustering among the strains to discover potential differential transmission by Mtb strains. Moreover, we searched for differences between the molecular epidemiology of TB between rural and urban settings of Tanzania.

4.3. Methods

4.3.1. Study setting

This study was conducted based on an ongoing prospective cohort that studies the epidemiology of TB (TB-DAR) in the urban setting of Temeke District in Dar es Salaam. Dar es Salaam is the commercial capital of Tanzania and the most populated city (~ 5.5 million in 2016), with a third of its population residing in Temeke. Dar es Salaam has the highest TB case notification rate in the country (Figure 4.1). In 2016, Dar es Salaam had an estimated total of 13,257 TB cases, of which 4,495 (33.9%) cases were notified in Temeke District alone. The TB-DAR cohort enrolls sputum smear positive and Xpert MTB/RIF positive adult TB patients (≥ 18 years of age) who attend the TB clinic at Temeke district hospital since November 2013. Temeke District Hospital is one of three regional referral hospitals in Dar es Salaam and the largest health facility in the district.

For the rural arm of the study, we included TB patients enrolled at the second recruitment site located in Ifakara in the Kilombero District, Morogoro (Figure 4.1). The Kilombero District is much less populated (~ 448,000 in 2016) compared to Temeke. Recruitment of TB patients for the rural site has been ongoing since August 2015 at the Chronic Disease Clinic of Ifakara (CDCI) situated within the St. Francis Referral Hospital.

4.3.2. Study population and study procedures

In this study, we included a total of 900 out of 1,262 TB patients recruited at the urban site (Temeke) between November 2013 and June 2018 (Figure 4.2), and a total of 242 out of 339 TB patients recruited at the rural site (Ifakara) between August 2015 and October 2018 (Figure 4.3). Figures 4.4 and Figure 4.5 summarize the proportion of study patients included by year of recruitment. During recruitment, confirmed TB patients were interviewed and received physical examination as previously described (Mhimbira *et al.*, 2017; Said *et al.*, 2017). Furthermore, clinical data and biological specimen i.e. sputum samples were collected for downstream laboratory analyses. We also obtained geographic coordinates (global positioning system [GPS]) from the patients' residences using Samsung Tab 4 android tablets (Samsung; Suwon, South Korea).

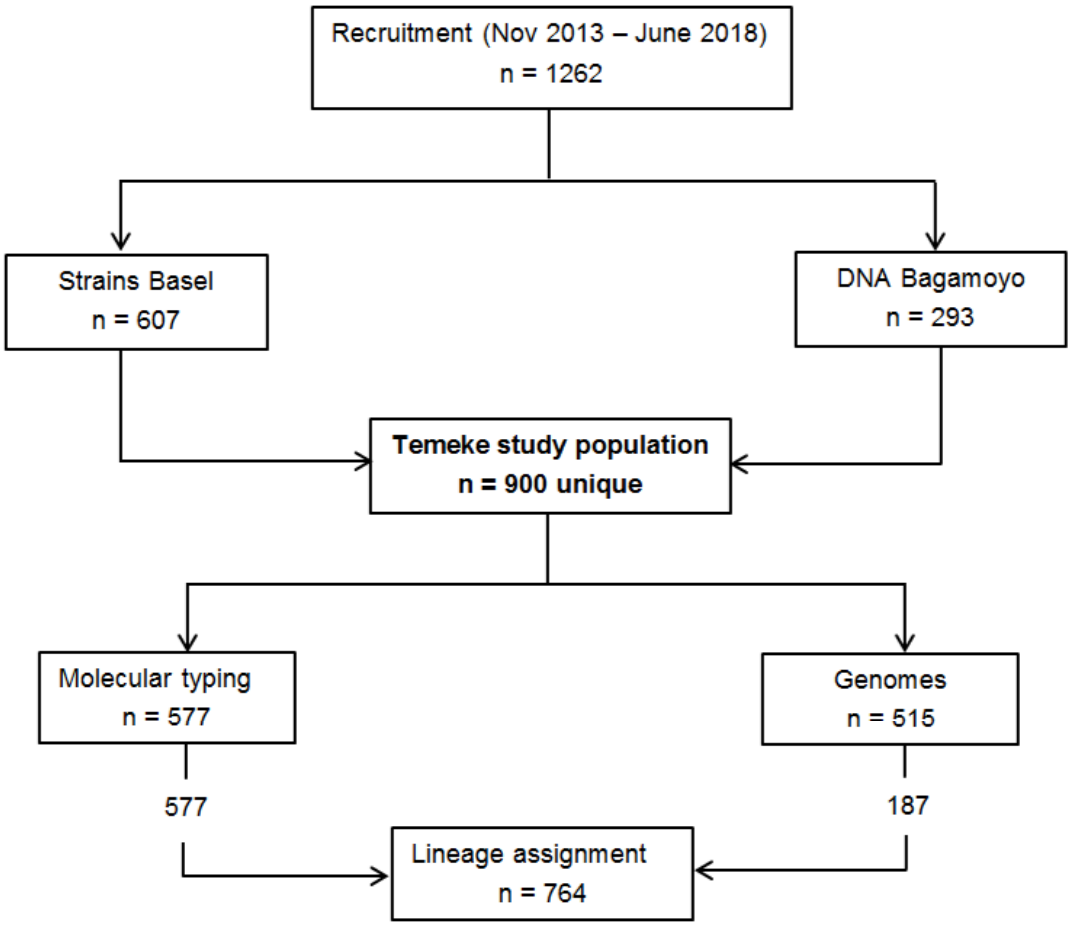


Figure 4.2.: Study population flow chart for Temeke (urban) site. The chart shows total number of TB patients recruited to the proportion that was included in this study.

4.3.3. Data collection and definitions

We collected socio-demographic, clinical and laboratory data (e.g. Ziehl Nielsen sputum smear results and HIV status) from the enrolled TB patients. Data were captured in electronic case report forms via the OpenDataKit (ODK) application (`www.opendatakit.org`) on Android PC tablets. The data were then uploaded and stored to a password protected and a regularly backed-up server and monitored using the eManagement "odk_planner" tool (Steiner *et al.*, 2016).

To calculate clinical severity of TB, we applied the modified 12 points TB score parameters system (Mhimbira *et al.*, 2017) adopted from Wejse and colleagues (Wejse *et al.*, 2008). i) cough ii) hemoptysis iii) dyspnea iv) chest pain v) night sweating vi) observed anemia vii) auscultation viii) high temperature 37.0 °C ix) mid upper arm circumference (MUAC) < 220 x) MUAC < 200 xi) body mass index (BMI) < 18 kg/m^2 xii) BMI < 16 kg/m^2. We then grouped TB score into mild (0–5), moderate (6–7) and severe (≥ 8).

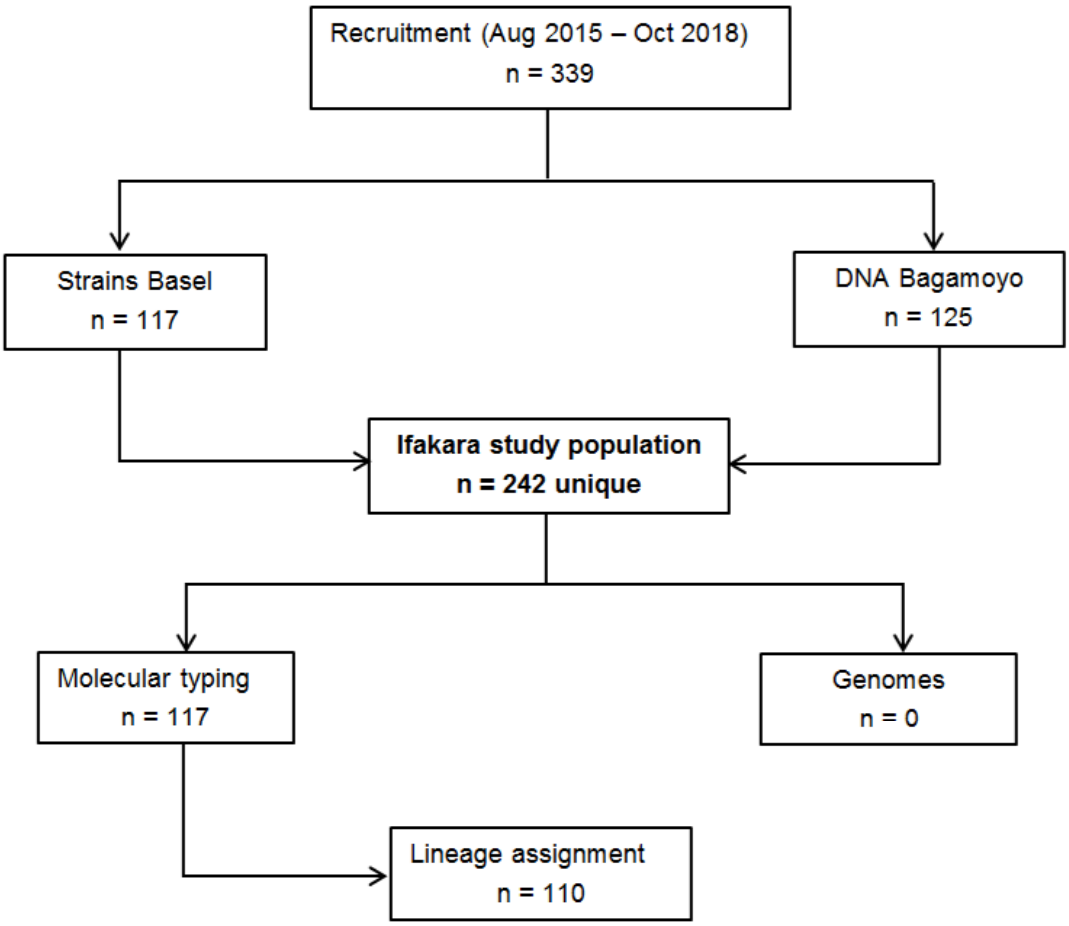


Figure 4.3.: Study population flow chart for Ifakara (rural) site. The chart shows total number of TB patients recruited to the proportion that was included in this study.

4.3.4. Mtb isolates and DNA extraction

Sputum samples collected from smear and Xpert MTB/RIF positive TB patients at the Temeke District Hospital were sent to the biosafety level 2+ TB laboratory at Bagamoyo Research and Training Centre (BRTC) and cultured on Löwenstein-Jensen solid media. In the case of rural site, sputum samples were preserved in cetylpyridinium chloride (CPC) and shipped to BRTC by post and processed for culture accordingly (Hiza *et al.*, 2017). Prior establishment of DNA extraction protocol at the BRTC (in 2017), Mtb isolates stored in glycerol were shipped to Basel for processing in the biosafety level 3 TB laboratory at the Swiss Tropical and Public Health Institute (Swiss TPH). Before processing the Mtb isolates we performed *rpoB* PCR to screen for potential MDR strains. The Mtb isolates were then sub-cultured in 7H9 liquid medium and incubated at 37°C after excluding those containing *rpoB* mutations (3/607). Pure genomic DNA was extracted from liquid culture (in Basel) and solid culture (in Bagamoyo) using the CTAB extraction method

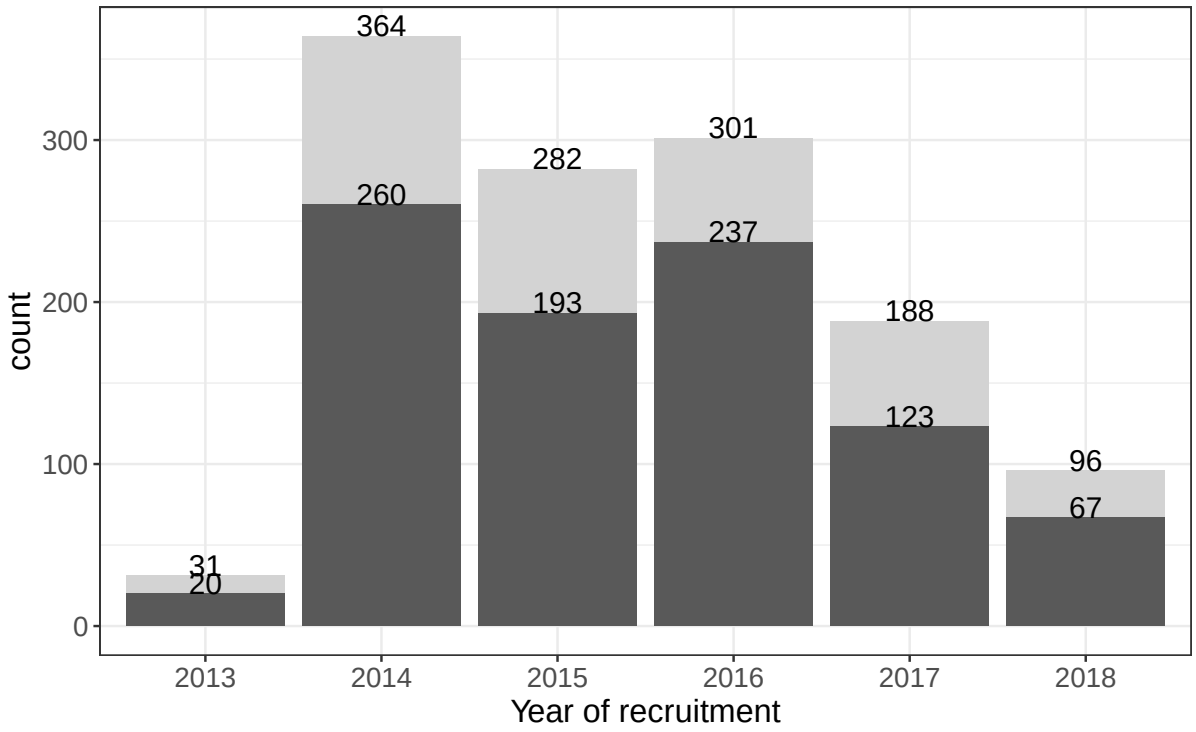


Figure 4.4.: Proportion of study population compared to total patients recruited in Temeke. Study population in dark gray and total number of patients recruited in light gray.

(Embden *et al.*, 1993). Since 2017, mycobacterial DNA is routinely being extracted at BRTC using the same protocol.

4.3.5. Molecular genotyping

Classification of Mtb isolates into main phylogenetic lineages was done using TaqMan real-time PCR protocol (Stucki *et al.*, 2012) on heat inactivated samples (1 hour, 95 °C). SNP-typing was only performed for Mtb isolates shipped to Basel.

4.3.6. Whole genome sequencing and phylogenetic inference

Whole genome sequencing was performed on prepared libraries from purified genomic DNA using Illumina Nextera®XT library and NEBNext®Ultra™ II FS DNA Library Prep Kits (Illumina, San Diego, USA). Sequencing was done at the Department of Biosystems Science and Engineering of ETH Zürich, Basel (D-BSSE) on the Illumina HiSeq 2500 or NextSeq 500 platforms (Illumina, San Diego, USA), generating paired-end 100 or 150 bp sequencing reads. For the whole genome analysis, first Illumina adaptors were

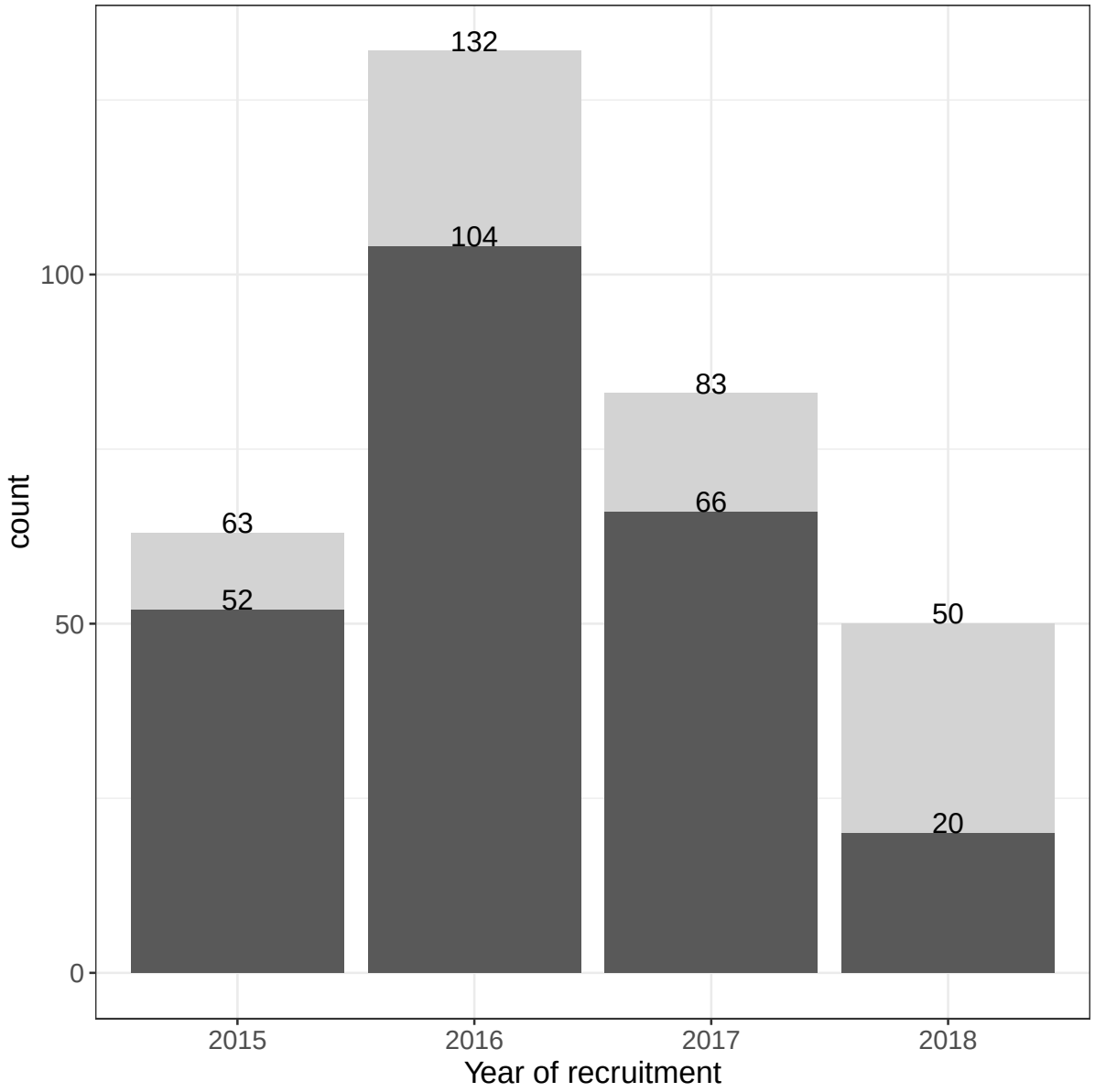


Figure 4.5.: Proportion of study population compared to total number of patients recruited in Ifakara. Study population in dark gray and total number of patients in recruited in light gray.

clipped from raw sequencing reads and low quality reads were trimmed using Trimmomatic 0.33 (SLIDINGWINDOW:5:20) (Bolger *et al.*, 2014). Short reads (< 20 bp) were excluded from the downstream analysis. Overlapping paired-end reads were merged with SeqPrep 1.2 (overlap size = 15) (`https://github.com/jstjohn/SeqPrep`). Reads were then mapped to a reconstructed Mtb complex ancestor sequence (Comas *et al.*, 2013) using BWA 0.7.13 (Li *et al.*, 2009). We used Picard 2.9.1 module (Mark Duplicates) to mark duplicated reads and excluded them afterwards. Pysam 0.9.0 (`https:`

//github.com/pysam-developers/pysam) was used to exclude reads with alignment score lower than (0.93*read_length)-(read_length*4*0.07), consistent with more than 7 miss-matches per 100 bp. SNPs were then extracted with Samtools 1.2 mpileup (Li, 2011) and VarScan 2.4.1 (Koboldt *et al.*, 2012) and only SNPs with minimum mapping quality of 20, minimum base quality at a position of 20, minimum read depth at a position of 7X, minimum percentage of reads supporting the call 90% and maximum strand bias for a position 90% were kept. We excluded genomes with coverage < 20 X and those containing phylogenetic SNPs belonging to different lineages (mixed strains). SNPs were annotated using snpEff 4.11 according to the Mtb H37Rv reference annotation (NC_000962.3). We excluded SNPs within regions such as PPE and PE-PGRS, phages, insertion sequences repetitive regions (at least 50 bp identical to other regions in the genome) (Stucki *et al.*, 2016). SNPs known to confer drug resistance were also excluded for phylogenetic reconstruction. Finally, we inferred a maximum likelihood phylogeny with RAxML 8.3.2 using a general time reversible (GTR) model and 1,000 bootstrap inferences (Stamatakis, 2006).

4.3.7. Lineage assignment

We combined SNP-typing and WGS data to assign the Mtb strains to the corresponding phylogenetic lineages (Figure 4.2). Out of 607 Mtb isolates received in Basel, 577; 95.1% were classified into one of the main lineages (562 by SNP-typing and 15 by WGS). At the time of this analysis, we obtained lineage information for an additional 187 (63.8%) out of 293 Mtb DNA extracted in Bagamoyo and genome sequenced in Basel. We included 764 out of 900 (84.9%) total study population to describe the Mtb lineages in Temeke and their relationship with patients' socio-demographics and clinical variables.

We processed 117 Mtb isolates from the Ifakara site for SNP typing and 110 (94.0%) of these were successfully classified into main lineages (Figure 4.3). We did not have WGS data available from the rural site at the time of analysis. Hence, we used 45.5% (110/242) of the rural study population to search for relationship between Mtb strain diversity and patients' characteristics. Figures 4.6 and 4.7 show the proportion of samples with genotype information compared to the number of TB patients included in the study by recruitment year.

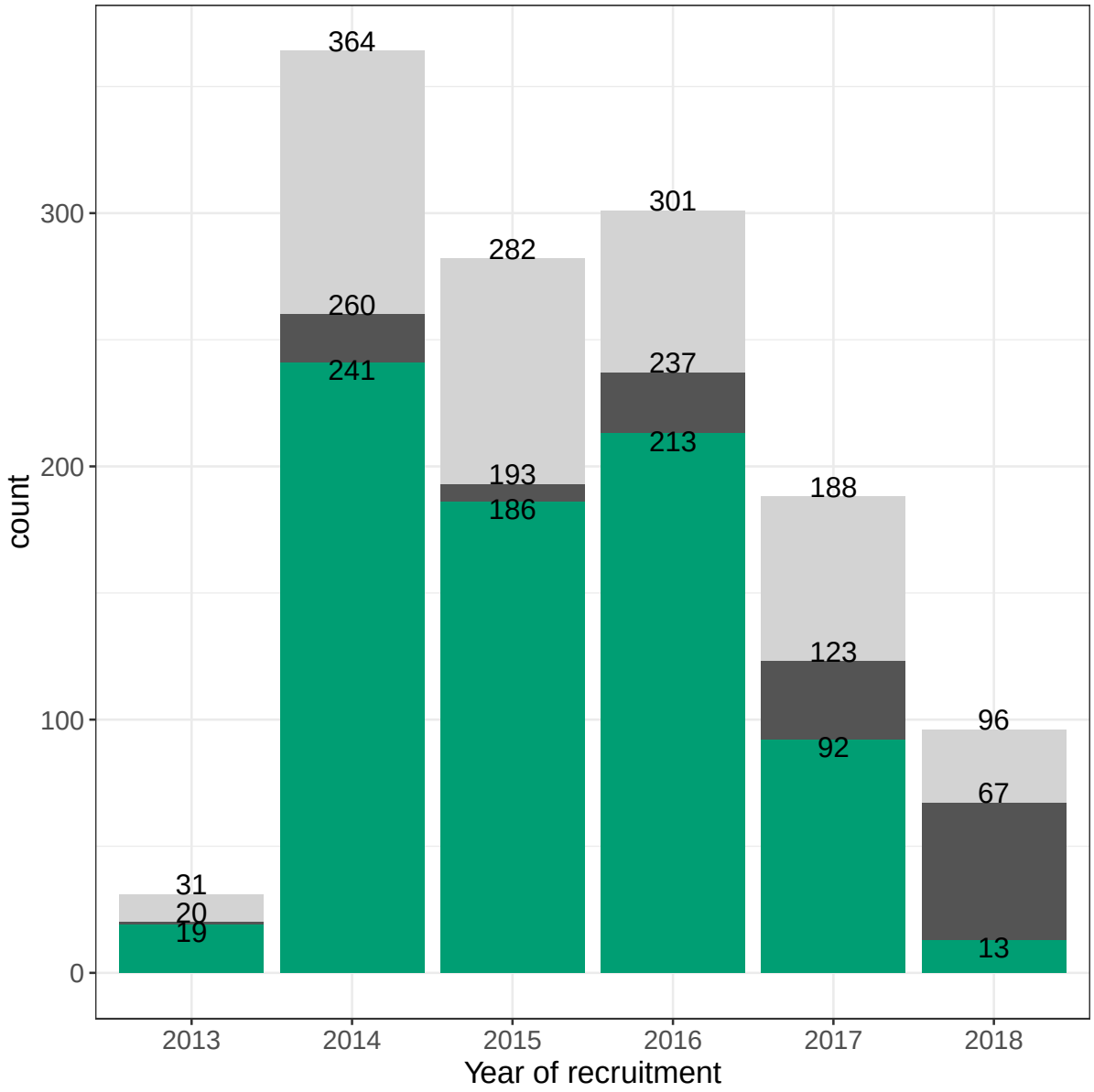


Figure 4.6.: Proportions of samples genotyped among those included in the study and the total patients recruited at the Temeke site. Genotyped (green), study population (dark gray), total patients recruited (light gray).

4.3.8. Statistical analyses

We used descriptive statistics to characterize TB patients in our study population. Kruskal Wallis and Wilcoxon rank-sum tests were used for continuous variables and χ^2 or Fisher's exact (when applicable) for comparison of categorical variables. With a focus on Lineage 1 and 3, we targeted a sample size of 720 TB patients, assuming a proportion of 22% TB patients infected with Lineage 1 and 47% infected with Lineage 3, based on frequencies described previously (Master thesis "Molecular Epidemiology of *Mycobacterium tubercu-*

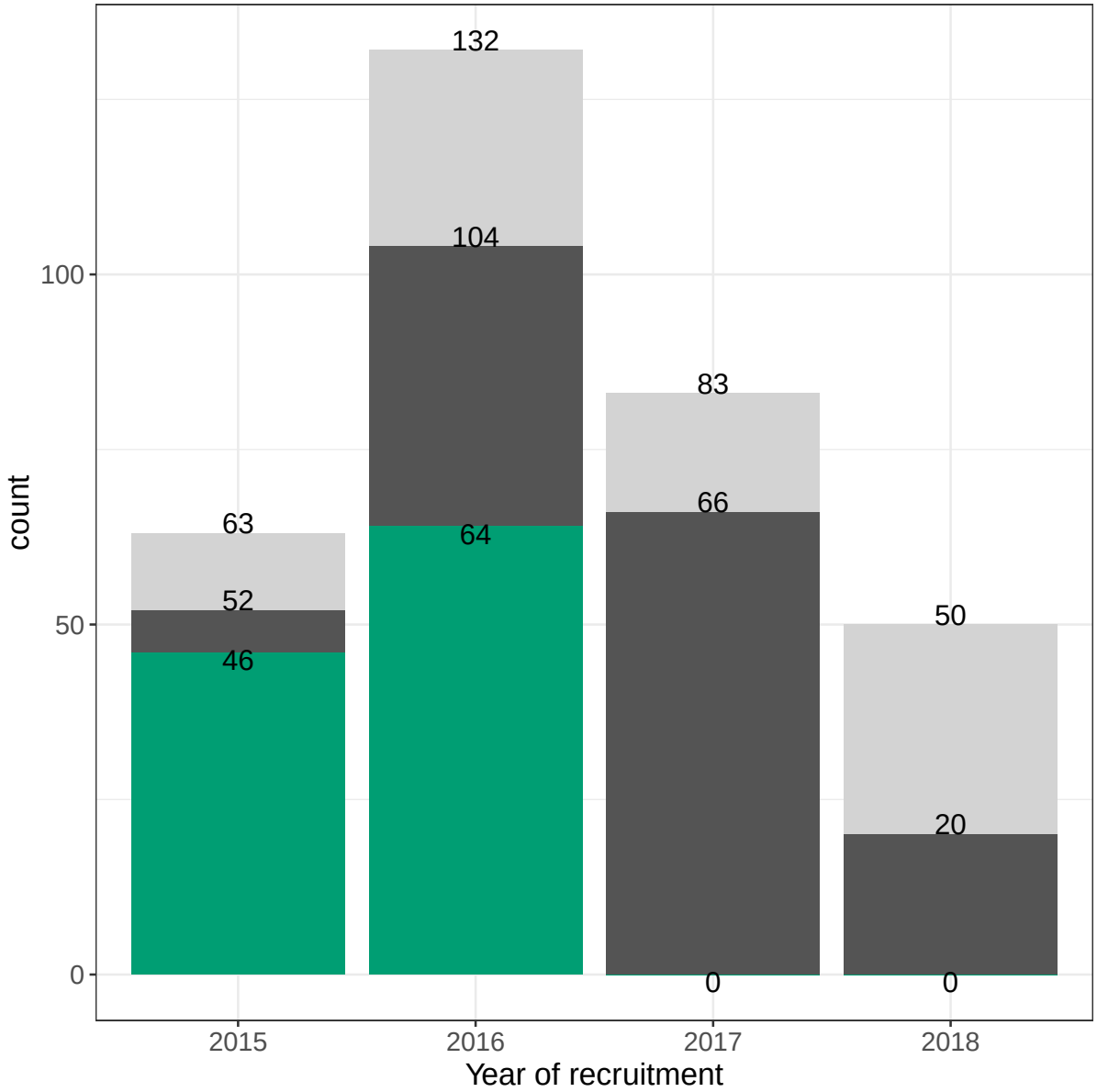


Figure 4.7.: Proportions of samples genotyped among those included in the study and the total patients recruited at the Ifakara site. Genotyped (green), study population (dark gray), total patients recruited (light gray).

losis in Bagamoyo, Tanzania", Jackson Thomas) and a power of 80% at 5% significance level to detect clinical relevance difference of 10% between the two lineages. We used univariate and multivariate logistic regression models to identify lineage associations with patients' demographics and clinical data adjusting for age, sex, HIV and risk factors for TB i.e., smoking, alcohol abuse and recent household contact with TB case. Statistical analyses were performed using R 3.5.0 (R Core Team, 2018).

4.3.9. Ethical approval

This study was approved by the institutional review board of the Ifakara Health Institute (IHI; reference no. IHI/IRB/No 04-2015), the Medical Research Coordinating Committee of the National Institute of Medical Research (NIMR; reference no. NIMR/HQ/R.8c/Vol.I /357) in Tanzania and the ethics committee of Northwestern and Central Switzerland (EKNZ; reference no. UBE-15/42). All patients signed a written informed consent prior enrollment.

4.4. Results

4.4.1. Patients' demographics and clinical characteristics in urban Temeke

We studied a total of 900 TB patients recruited between November 2013 and June 2018 at the urban site of Temeke. Table 4.1 summarizes the socio-demographic and clinical characteristics of the TB patients. The median age of TB patients in the study population was 33 years (interquartile range [IQR]: 27–40 years) and 640 (71.1%) were male patients. The TB patients had a median body mass index (BMI) of 18.2 kg/m^2 (IQR: 16.7–20.0) and 55.3% (498/900) of the patients were underweight (< 18.5 kg/m^2). HIV co-infection was detected in 187/900 (20.8%) of TB patients. More than half of TB patients had mild TB severity score, and only 9.3% showed a severe TB score.

4.4.2. Risk factors in urban Temeke

In addition to HIV, we determined the prevalence of nine other known risk factors for TB including recent contact with a TB case, diabetes, alcohol abuse (defined as regular alcohol drinking–at least three bottles of beer daily), use of steroids, smoking (defined as

Table 4.1.: Socio-demographic and clinical characteristics of TB patients in Temeke

Characteristics	**Total (n = 900)**	**Proportion %**
Age (years), medium (IQR)		
33 (27-40)		
Age groups (years)		
Child age (<15)	0	0.00
Young age (15-24)	158	17.6
Early adult (25-44)	606	67.3
Late adult (45-64)	123	13.7
Old age (>65)	13	1.4
Sex		
Female	260	28.9
Male	640	71.1
BMI (kg/m^2), median (IQR)		
18.2 (16.7-20.0)		
BMI		
categories, (kg/m^2)		
Underweight <18.5	498	55.3
Normal 18.5–24.9	366	40.7
Overweight 25.0–29.9	29	3.2
Obese >30	7	0.8
HIV status		
Negative	701	77.9
Positive	187	20.8
Unknown	12	1.3
TB score		
Mild 0–5	557	61.9
Moderate 6–7	259	28.8
Severe ≥ 8	84	9.3

n, Number; IQR, Interquartile range; BMI, Body mass index; HIV, Human immunodeficiency virus.

current smoking), intravenous drug use, incarceration, silicosis and tumor necrosis factor therapy (Figure 4.8). Among the nine risk factors, smoking was the most prevalent among TB patients (23%) followed by alcohol abuse (19.8%) and recent contact with a TB case via household (10.9%). More than one-third of TB patients lacked diabetes status, and only 1.1% of those who did had diabetes. Less than 1% of TB patients had history of silicosis, steroids and intravenous drug use. None of the TB patients were recorded to have had tumor necrotic factor therapy.

4.4.3. Spatial distribution of TB patients in Temeke

We measured geographical locations of patients' residences using geographical coordinates (longitude and latitude) collected by trained field workers. Figure 4.9 shows the spatial distribution of TB patients included in the study area of Temeke. Temeke is one of three districts in Dar es Salaam located in the southern part and bordered by the Indian Ocean to the east. Temeke District Hospital is situated on the western part of the district and it is intersected with regional and interregional highways. Most of the TB patients were densely distributed in close proximity to the district hospital.

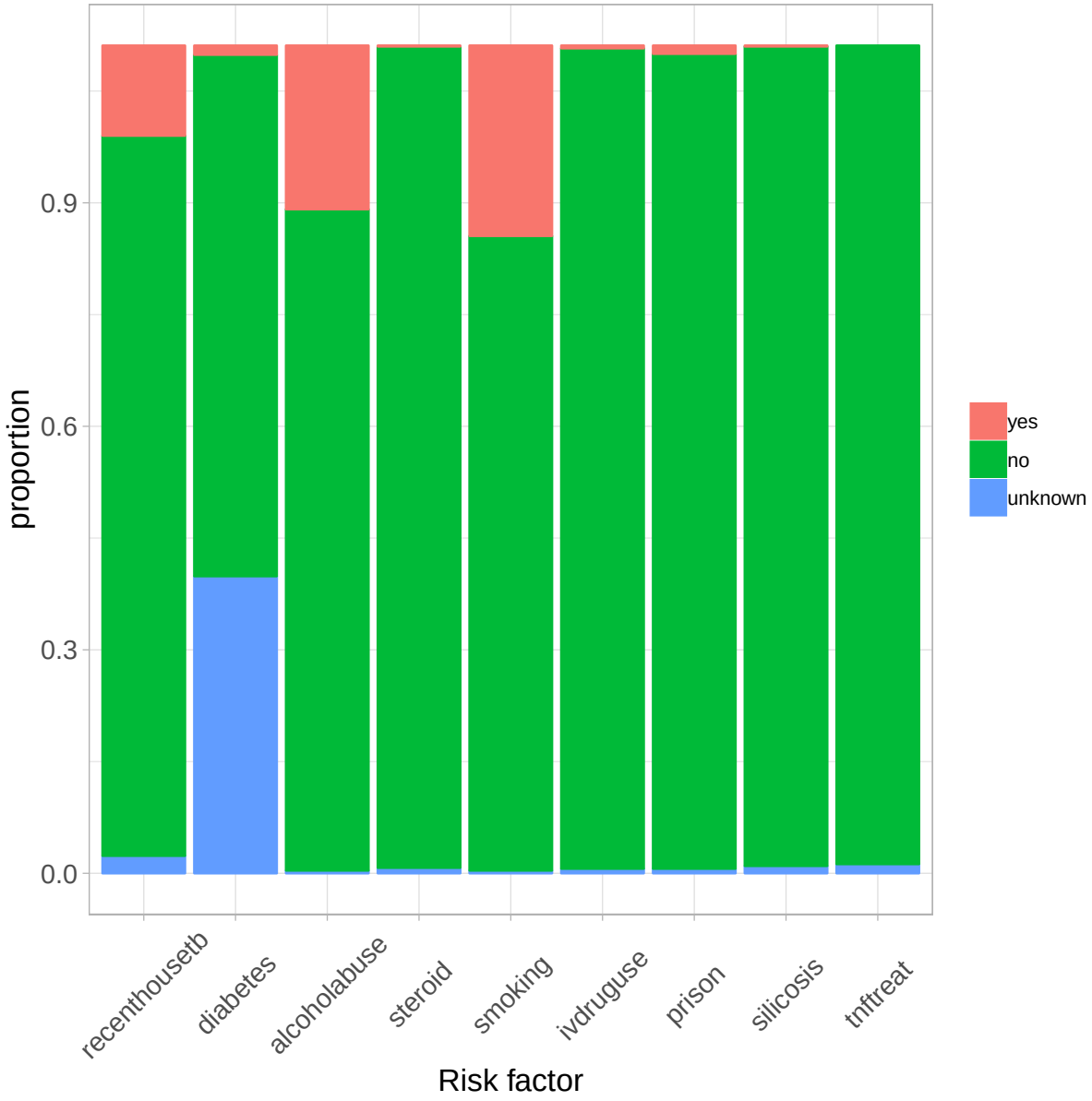


Figure 4.8.: Proportion of risk factors quantified among TB patients in the study population from urban site (Temeke). recenthousetb, recent contact with a TB case; alcoholabuse, alcohol abuse; ivdruguse, intravenous drug use; prison, incarceration; and tnftreat, tumor necrosis factor therapy.

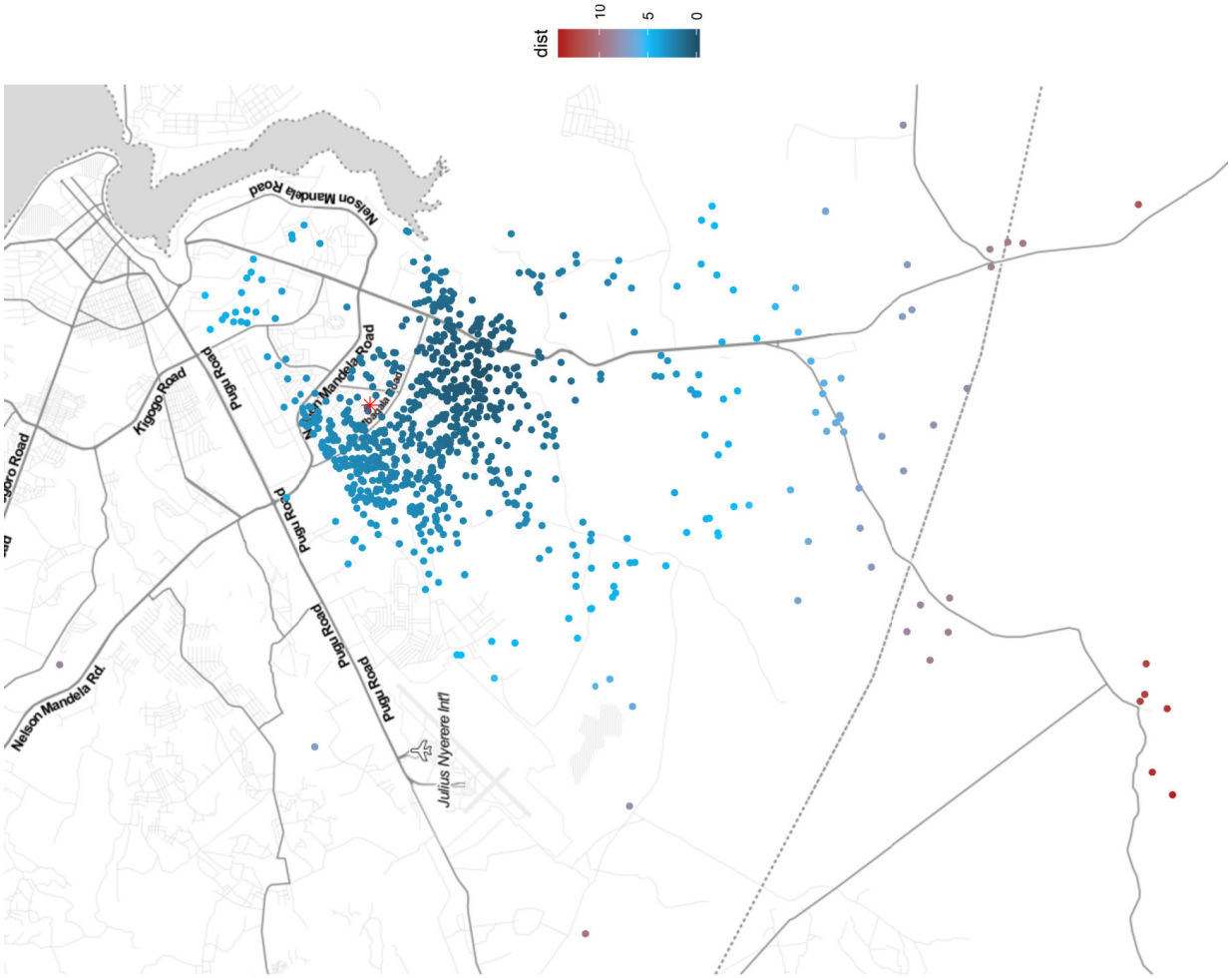


Figure 4.9.: Map of Temeke area illustrating the spatial distribution of TB patients' residences enrolled in the study. The red asterisk indicates the location of the Temeke District Hospital.

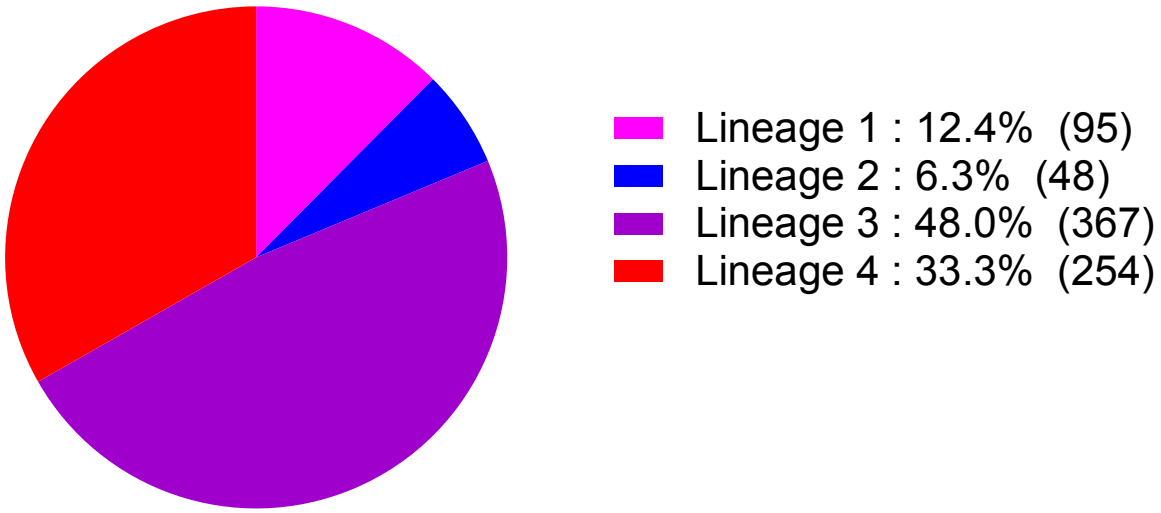


Figure 4.10.: Frequency distribution of Mtb lineages detected in Temeke. Mtb lineage classification of 764 clinical strains.

4.4.4. Mtb lineages in Temeke

We classified Mtb strains isolated from 764 (84.9%) out of 900 TB patients included in the study into the main phylogenetic lineages. Four Mtb lineages were detected (Figure 4.10). Lineage 3 was the most frequent (48%), with the other Mtb strains belonging to Lineage 4 (33.3%), Lineage 1 (12.4%) and Lineage 2 (6.3%). The overall Mtb lineage distribution was fairly constant during the 2014 to 2016 collection years ($p = 0.10$) (Figure 4.11).

4.4.5. Mtb lineages and patients' demographics and clinical characteristics in Temeke

We assessed the relationship of Mtb lineages with the patients' socio-demographics and clinical characteristics using 764 genotypes available (Table 4.2). We observed TB patients infected with Lineage 2 to be younger in average (30 years, IQR: 25–35.3 years) compared to patients infected with other lineages (Table 4.2 and Figure 4.12). This observation was also reflected in a higher proportion of Lineage 2 strains among the "young age" TB patients (22.9%) and a lower proportion in "late adult" TB patients (8.3%) than those of Lineage 1, 3, and 4 (13.7%, 14.7% and 13.8%, respectively). We did not observe any difference between sex distribution, median BMI, HIV status and TB severity scores across the four infecting Mtb lineages (Table 4.2 and Figure 4.13).

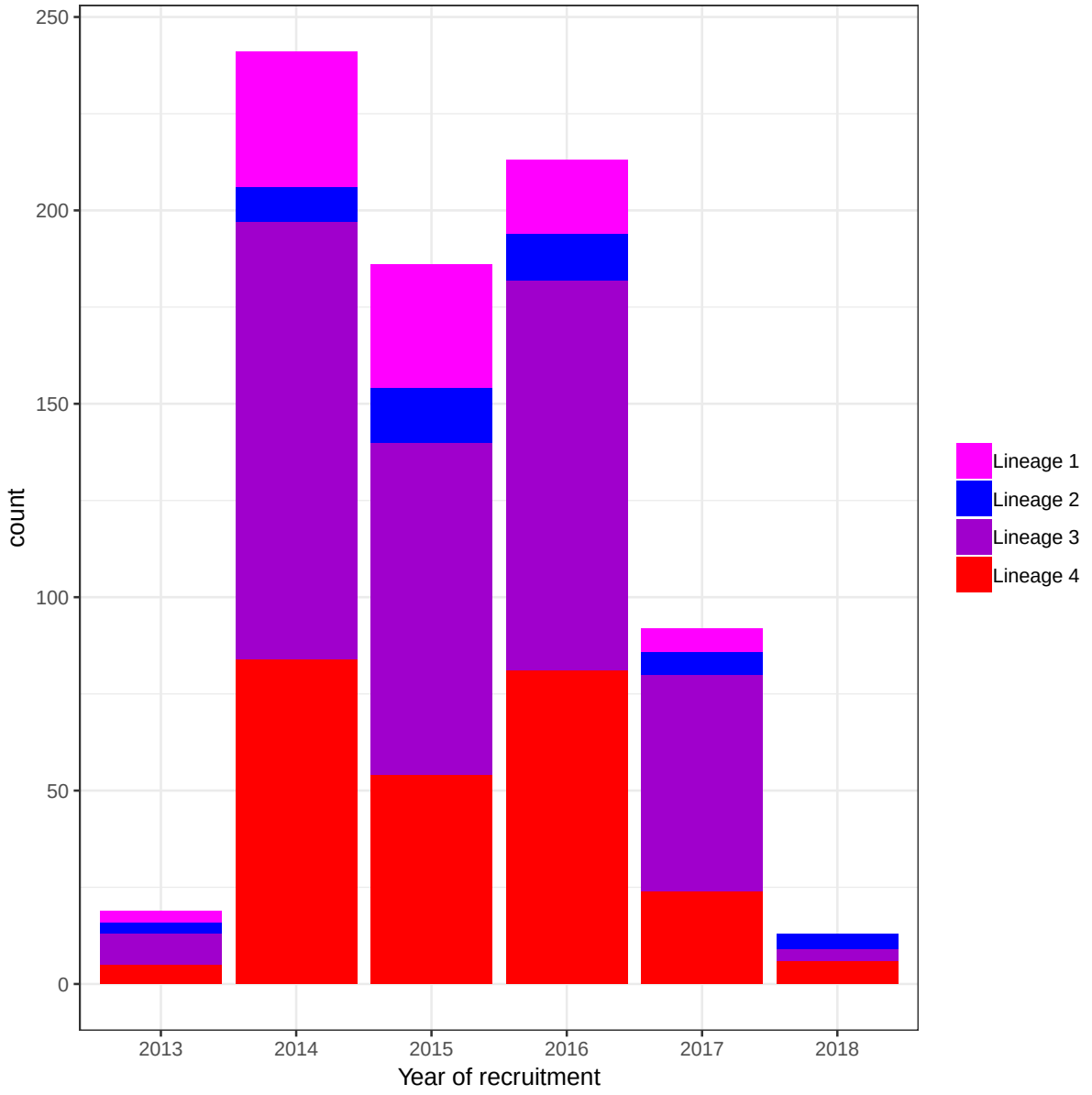


Figure 4.11.: Frequency distribution of the four Mtb lineages circulating in Temeke by recruitment year.

Table 4.2.: Sociodemographic and clinical information of TB patients by infecting Mtb lineage in Temeke

Characteristics	**Lineage 1** n (%)	**Lineage 2** n (%)	**Lineage 3** n (%)	**Lineage 4** n (%)
Age (years), median (IQR)	34 (27–40)	30 (25–35.3)	34 (27.5–41)	33 (27–40)
Age groups (years)				
Child age (<15)	0	0	0	0
Young age (15–24)	16 (16.8)	11 (22.9)	55 (15)	48 (18.9)
Early adult (25–44)	64 (67.4)	33 (68.7)	252 (68.7)	169 (66.5)
Late adult (45–64)	13 (13.7)	4 (8.3)	54 (14.7)	35 (13.8)
Old age (>65)	2 (2.1%)	0 (0)	6 (1.6)	2 (0.8)
Sex				
Female	31 (32.6)	14 (29.2)	105 (28.6)	73 (28.7)
Male	64 (67.4)	34 (70.8)	262 (71.4)	181 (71.3)
BMI (kg/m^2), median (IQR)	18.2 (16.9 –20.2)	18.4 (16.9 –19.8)	18.2 (16.7–20.0)	18.1 (16.7 - 20.1)
BMI categories (kg/m^2)				
Underweight <18.5	52 (54.7)	25 (52.1)	203 (55.3)	140 (55.1)
Normal 18.5–24.9	41 (43.2)	22 (45.8)	148 (40.3)	103 (40.6)
Overweight 25.0–29.9	2 (2.1)	1 (2.1)	14 (3.8)	8 (3.1)
Obese >30	0 (0)	0 (0)	2 (0.5)	3 (1.2)
HIV status				
Negative	67 (70.5)	41 (85.4)	280 (76.3)	203 (79.9)
Positive	26 (27.4)	6 (12.5)	84 (22.9)	49 (19.3)
Unknown	2 (2.1)	1 (2.1)	3 (0.8)	2 (0.8)
TB score				
Mild 0–5	49 (51.6)	32 (66.7)	223 (60.8)	159 (62.6)
Moderate 6–7	38 (40)	12 (25)	104 (28.3)	75 (29.5)
Severe 8	8 (8.4)	4 (8.3)	40 (10.9)	20 (7.9)
Total	95	48	367	254

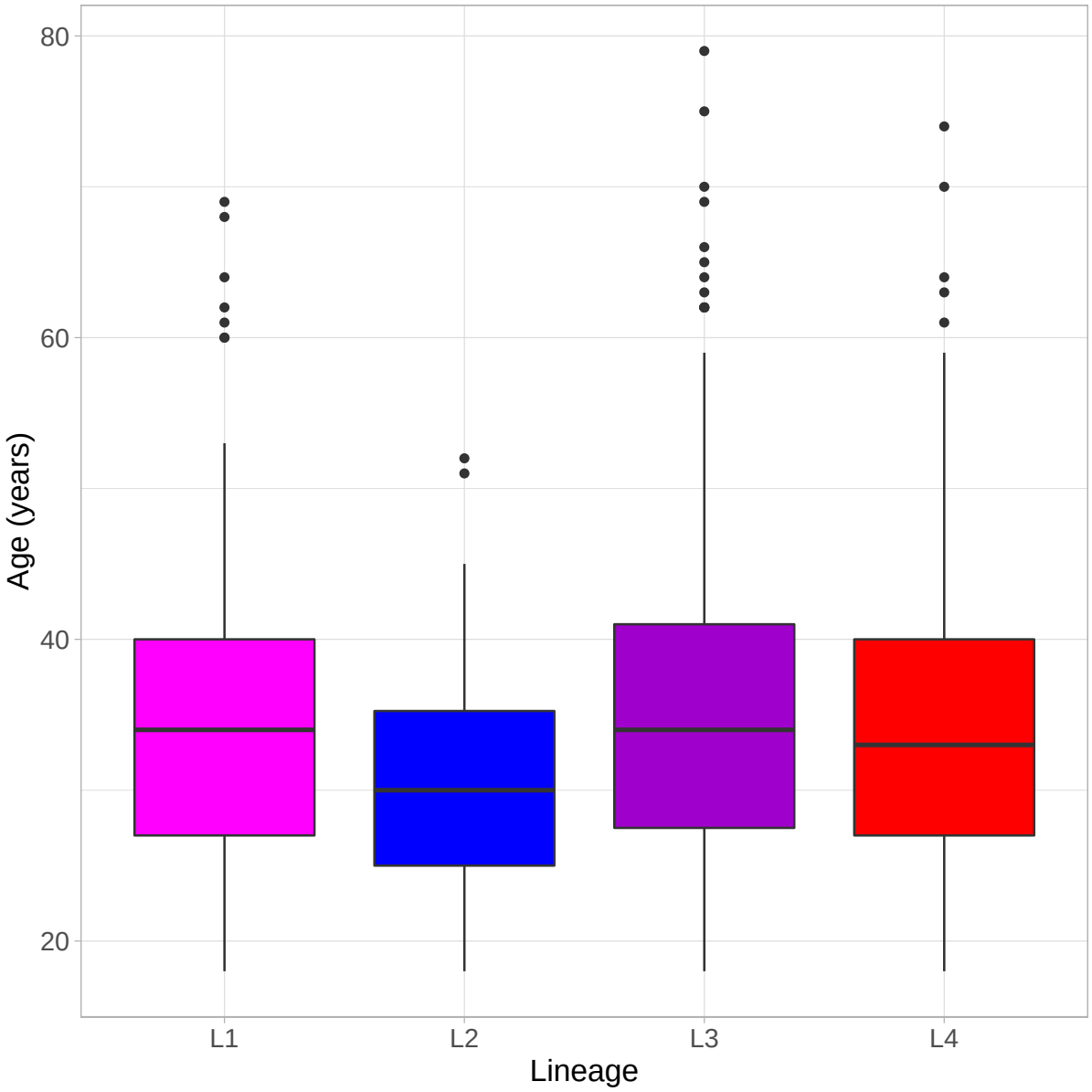


Figure 4.12.: Distribution of age of TB patients by the respective infecting Mtb lineage in Temeke.

4.4.6. Associations between Mtb lineages with patients' sociodemographic and clinical characteristics in Temeke

Given that TB patients infected with Lineage 2 showed a lower median age, we next performed logistic regression analyses to assess for a possible association by comparing Lineage 2 against all other lineages. The multivariate analysis revealed that for every year increase in age there is a 4% decrease chance for being infected with Lineage 2 (adjusted

odds ratio [aOR] 0.96; 95% confidence interval [95% CI] 0.93–1, p = 0.03).

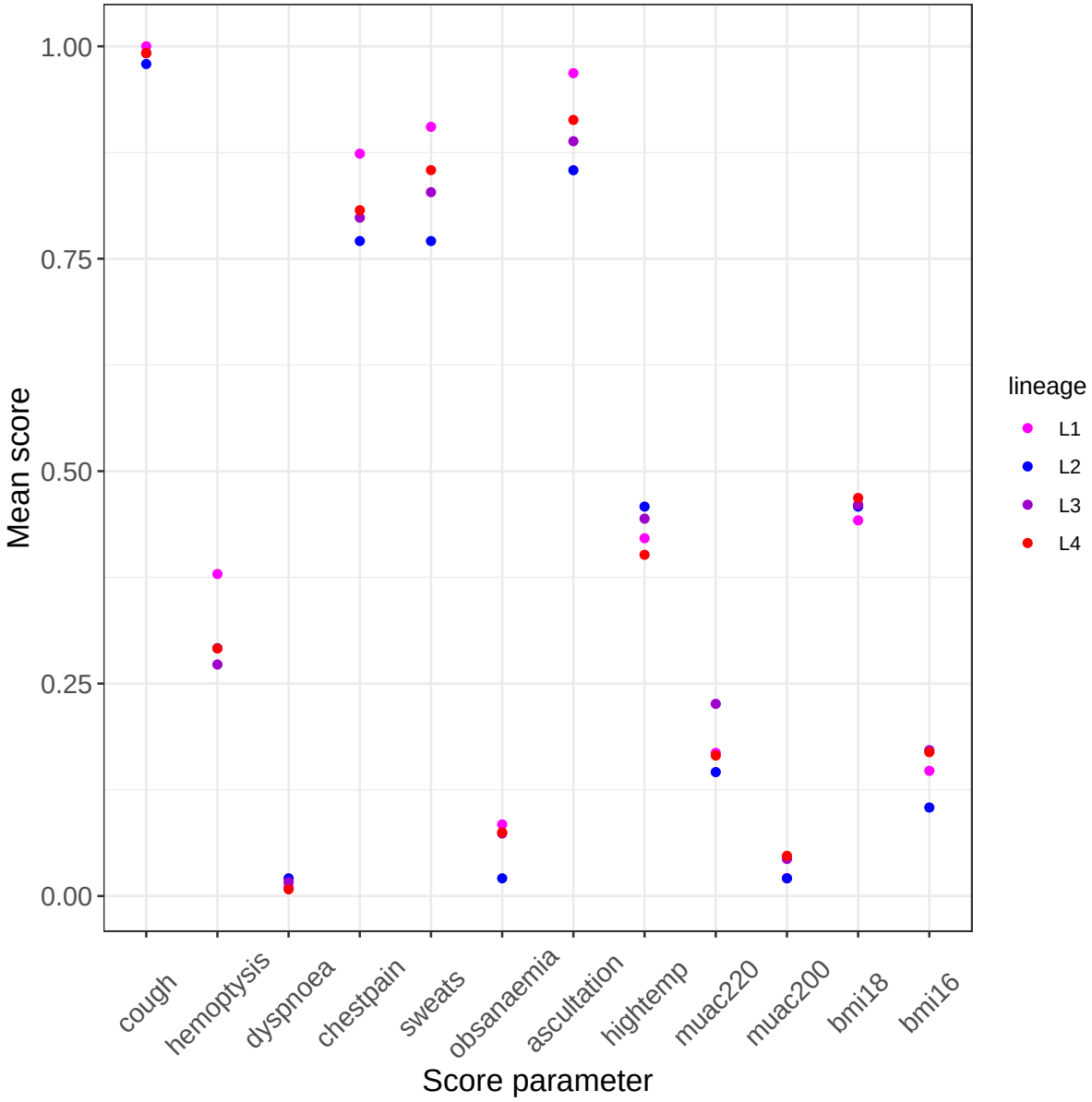


Figure 4.13.: Mean severity TB scores for each of the 12 TB score parameters among the four Mtb lineages.

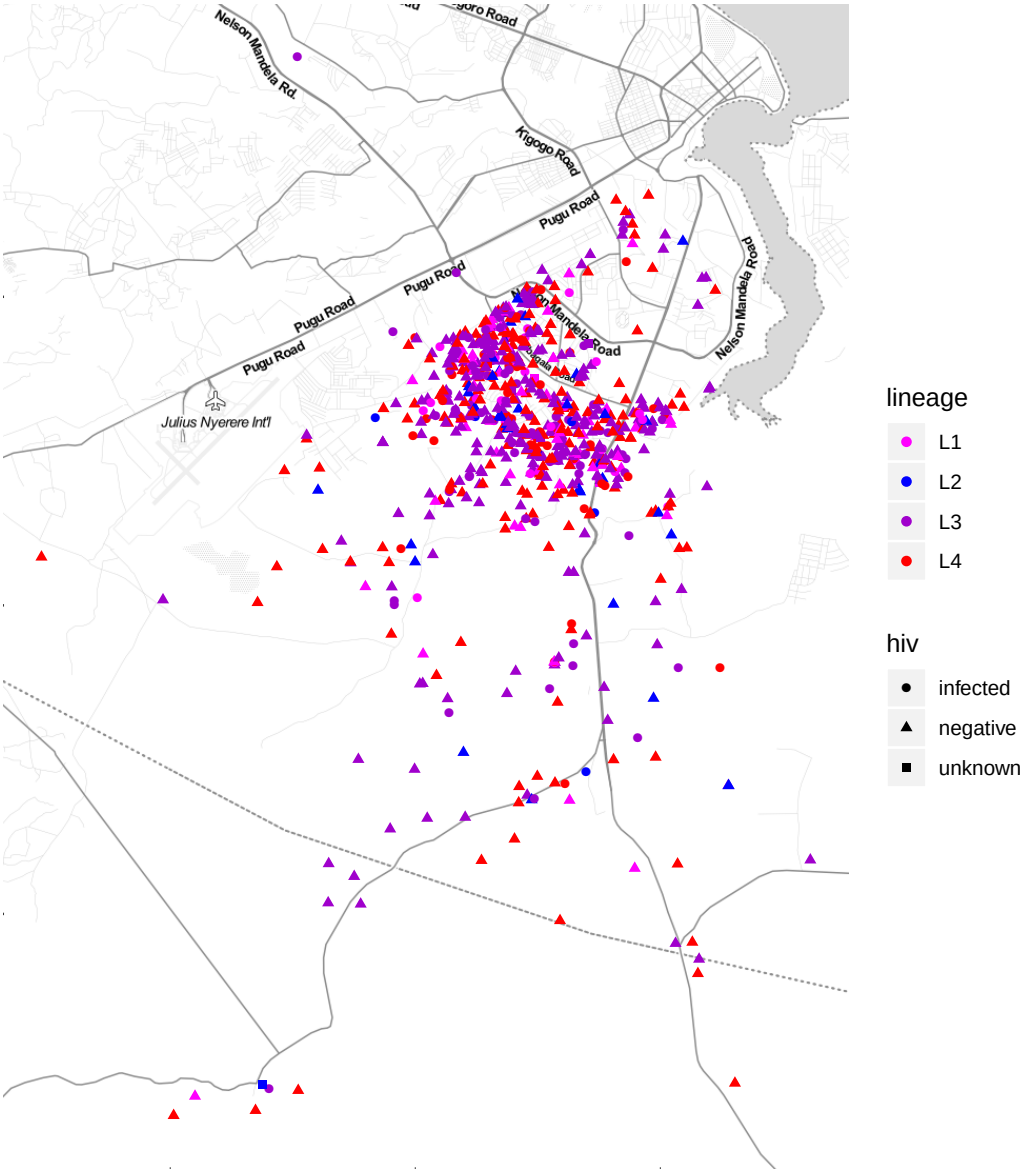


Figure 4.14.: Spatial distribution of TB patients mapped by the infecting Mtb lineage and HIV status.

Figure 4.15.: Spatial distribution of TB patients mapped by the infecting Mtb lineage and HIV status with each lineage plotted separately.

4.4.7. Local diversity of Mtb strains circulating in Temeke

To further characterize the Mtb diversity in Temeke, we analyzed a total of 515 whole genome sequences that were available at the time of analysis. Figure 4.16 shows the phylogenetic tree reconstructed from 23,306 variable single nucleotide positions. The whole genome sequence dataset reflected similar lineage frequencies (Lineage 1, 65; 12.6%, Lineage 2, 38; 7.4%, Lineage 3, 256; 49.7% and Lineage 4 156; 30.3%) to those determined by genotyping (Figure 4.10).

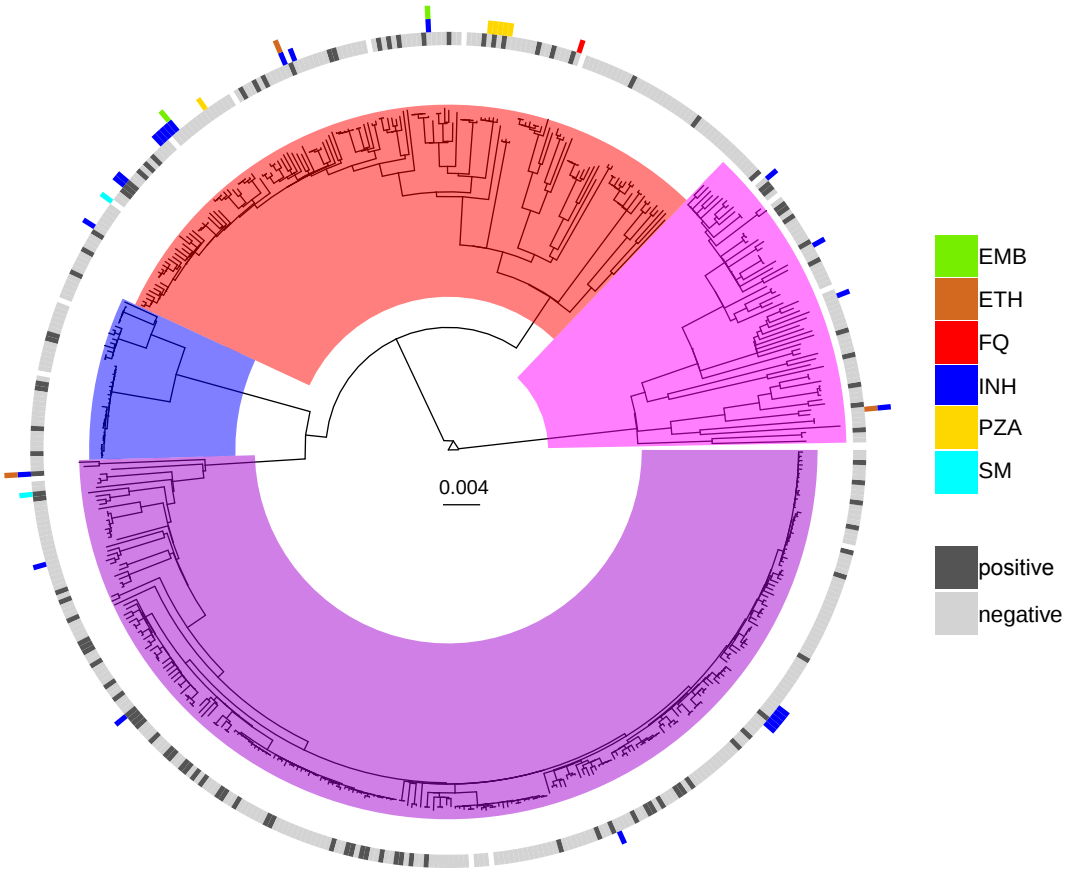


Figure 4.16.: Maximum likelihood phylogeny of 515 Mtb strains from TB patients in Temeke. The backgrounds on the phylogeny are shaded by lineage specific color code (Lineage 1; pink, Lineage 2; blue, Lineage 3; purple and Lineage 4; red). The first outer ring indicates the HIV status of the patients. The second and third outer rings show the presence of drug resistance associated mutations detected in the strains (these are colored by drugs they confer resistance to as indicated on the legend).

4.4.8. Drug resistance

From the whole genome sequences, we further detected mutations conferring resistance to ethambutol (EMB), fluoroquinolone (FQ), isoniazid (INH), pyrazinamide (PZA) and streptomycin (SM). These drug resistance conferring mutations occurred in all Mtb lineages with varying frequencies, except in Lineage 2 (Table 4.3 and Table 4.4). Overall, 13.2% of the Mtb strains contained at least one drug resistance mutation. We found that drug resistance mutations were more frequent in Lineage 4 (12.2%). Despite being the most frequent, only 3.9% of Lineage 3 strains harbored drug resistance mutations. Furthermore, we detected mutations conferring resistance to more drugs among Lineage 4 strains compared to Lineage 1 and 3. We did not detect any mutation conferring resistance to rifampicin (RIF). Of note, we assigned *fabG1* C-15T mutation as conferring resistance to Ethionamide(ETH) as well since the *fabG1-inhA* promoter mutations are known to confer cross resistance to both INH and ETH drugs (Rueda *et al.*, 2015; Cirillo *et al.*, 2017).

Although *ethA* mutations S266R and N345K were previously reported in MDR strains as conferring resistance to ETH (Boonaiam *et al.*, 2010; Brossier *et al.*, 2011), we detect both mutations to be phylogenetic markers. The two mutations were identified among Lineage 2 and Lineage 1 strains corresponding to sublineages Lineage 2–Beijing (L2.2.1) and L1.2.2, respectively (Coll *et al.*, 2014; Coll *et al.*, 2015). These mutations were excluded in the frequency analysis of drug resistance mutations.

Table 4.3.: Frequency of strains containing (any) drug resistance mutation across Mtb lineages

Lineage	**Mutation**		**Total**
	No, n (%)	Yes, n (%)	
Lineage 1	61 (75.4)	4 (6.2)	65
Lineage 2	38 (100)	0 (0)	38
Lineage 3	246 (96.1)	10 (3.9)	256
Lineage 4	137 (87.8)	19 (12.2)	156
Total	447 (86.8)	68 (13.2)	515

Table 4.4.: Drug resistance mutation in Mtb strains

Lineage	**Drug resistance profile**		n
	Drug	Mutation | Gene	
Lineage 1	INH/ETH	C-15T | *fabG1*	1
	INH	S315T | *katG*	3
Lineage 3	INH/ETH	C-15T | *fabG1*	1
	INH	S315T | *katG*	8
	SM	A200E | *gid*	1
Lineage 4	INH/EMB	S315T;D328Y | *katG*; *embB*	1
		S315T;M306I | *katG*; *embB*	1
	INH/ETH	C-15T | *fabG1*	1
	PZA	L172P | *pncA*	5
		V7L | *pncA*	1
	INH	S315T | *katG*	7
		W90R | *katG*	1
	SM	C517T | *rrs*	1
	FQ	D94H | *gyrA*	1

EMB, ethambutol; ETH, ethionamide; FQ, fluoroquinolone; INH, isoniazid; PZA, pyrazinamide; SM, streptomycin.

Table 4.5.: Proportion of clustered and non-clustered strains within the Mtb lineages

Lineage	**Cluster number**	**Cluster members**		**Total**
	No, n (%)	Yes, n (%)		
Lineage 1	0	65 (100)	0 (0)	65
Lineage 2	4	14 (36.8)	24 (63.2)	38
Lineage 3	7	203 (79.3)	53 (20.7)	256
Lineage 4	6	134 (85.9)	22 (14.1)	156
Total	17	416 (80.8)	99 (19.2)	515

4.4.9. Transmission clusters of Mtb strains in Temeke

From clustering analysis we identified a total of 28 clusters containing two or more strain members using a 5 SNPs threshold (Figure 4.17). Overall, we found 23.5% (121/515) of TB patients were in clusters. Of note, these clusters contained TB patients infected with Mtb modern lineages (Lineage 2–4) whereby, 17, 7 and 4 clusters belonged to TB patients infected with Lineage 3, 4 and 2, respectively. No clusters were identified among TB patients infected with Lineage 1 (Figure 4.18). We detected 11 clusters out of the 28 comprising of two patients, 10 belonging to Lineage 3 and one to Lineage 4.

For further analysis we kept clusters with at least three members that resulted into 17 clusters where: 4 clusters belonged to Lineage 2, 7 clusters to Lineage 3 and 6 clusters to Lineage 4. Considering cluster size of three and more, 19.2% (99/515) of the TB patients were found in clusters. The cluster sizes ranged between 3 and 15 patients. Patients infected with Lineage 3 had the largest cluster sizes (maximum 15 members) followed by those infected with Lineage 2 (maximum 8 members) and Lineage 4 had the smallest cluster sizes (maximum 5 members). On quantifying the proportion of strains in clusters within each lineage, we discovered that Lineage 2 strains were more frequently found in clusters compared to Lineage 3 and 4 ($p = <0.0001$). These results were consistent when considering strains in two-member clusters also.

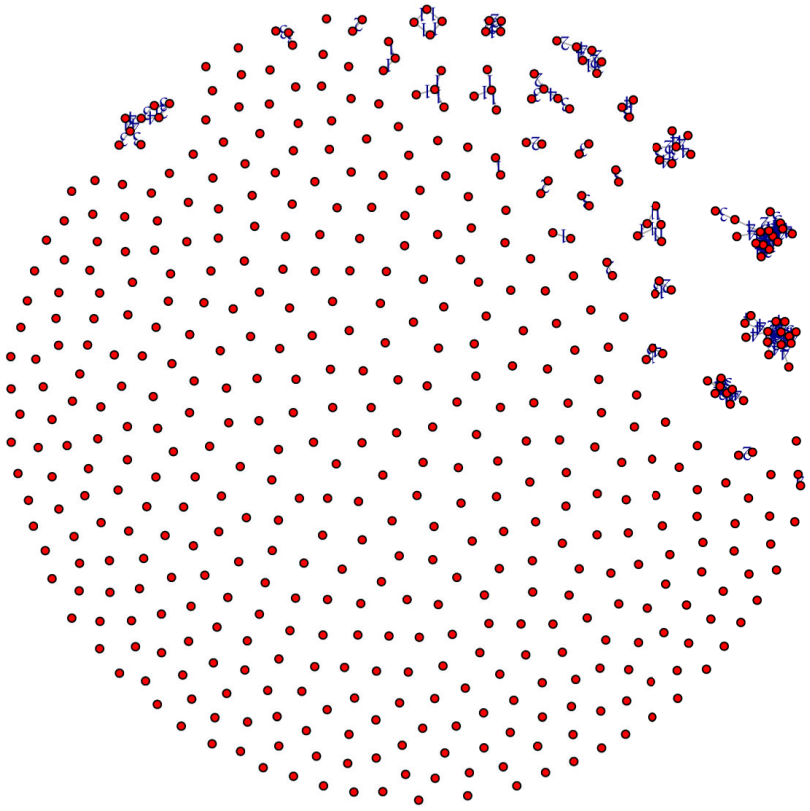


Figure 4.17.: Transmission clusters of the 515 Mtb strains identified based on a 5-SNP threshold.

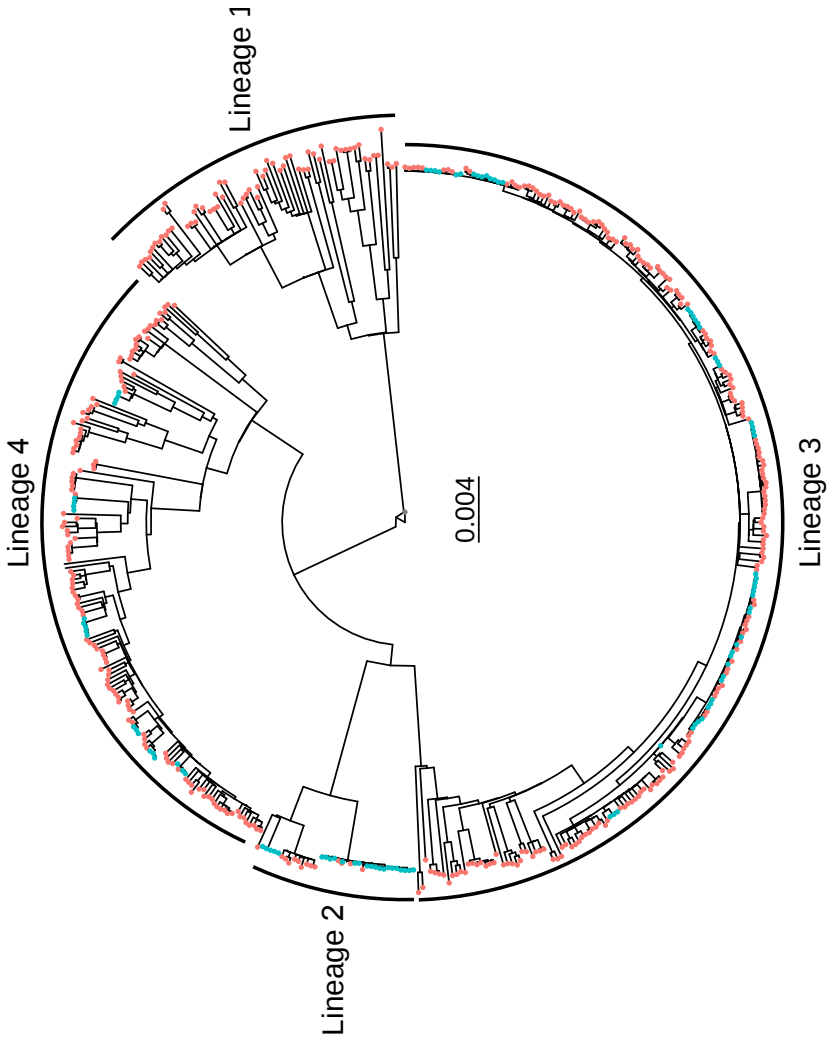


Figure 4.18.: Maximum likelihood phylogeny of 515 Mtb strains from TB patients in Temeke illustrating clustering. Strains found in clusters (green) and those that were not (red). The grey tip point is the outgroup *M. canettii*.

4.4.10. 4.4.11 Patients characteristics and risk factors in rural Ifakara

We studied 242 TB patients recruited at the rural Ifakara site from August 2015 and October 2018. TB patients from the rural site had a median age of 36 years (IQR: 28–43) and two-thirds of the patients were males (161; 66.5%). Out of the 242 TB patients recruited in the rural setting, 58 (24.0%) were HIV co-infected. Table 4.6 summarizes the sociodemographic and clinical characteristics of patients in the rural site.

In addition to sociodemographic and clinical information, we quantified known risk factors for TB in the study population (Figure 4.19). Noteworthy, 20.7% of TB patients enrolled in Ifakara were recorded to have had recent contact with a TB case within a household. Other predominant risk factors detected were alcohol abuse (28.5%) and smoking (22.3%).

Table 4.6.: Socio-demographic and clinical characteristics of TB patients in the Ifakara site

Characteristics	**Total (n = 242)**	**Proportion %**
Age (years), median (IQR)		
36 (28-43)		
Age groups (years)		
Child age (<15)	0	0.00
Young age (15-24)	39	16.1
Early adult (25-44)	146	60.3
Late adult (45-64)	47	19.4
Old age (>65)	10	4.1
Sex		
Female	81	33.5
Male	161	66.5
BMI (kg/m^2), median (IQR)		
17.8 (16.2-19.9)		
BMI categories, (kg/m^2)		
Underweight <18.5	146	60.3
Normal 18.5-24.9	92	38.0
Overweight 25.0-29.9	3	1.2
Obese >30	1	0.4
HIV status		
Negative	177	73.1
Positive	58	24.0
Unknown	7	2.9

n, Number; IQR, Interquartile range; BMI, Body mass index; HIV, Human immunodeficiency virus.

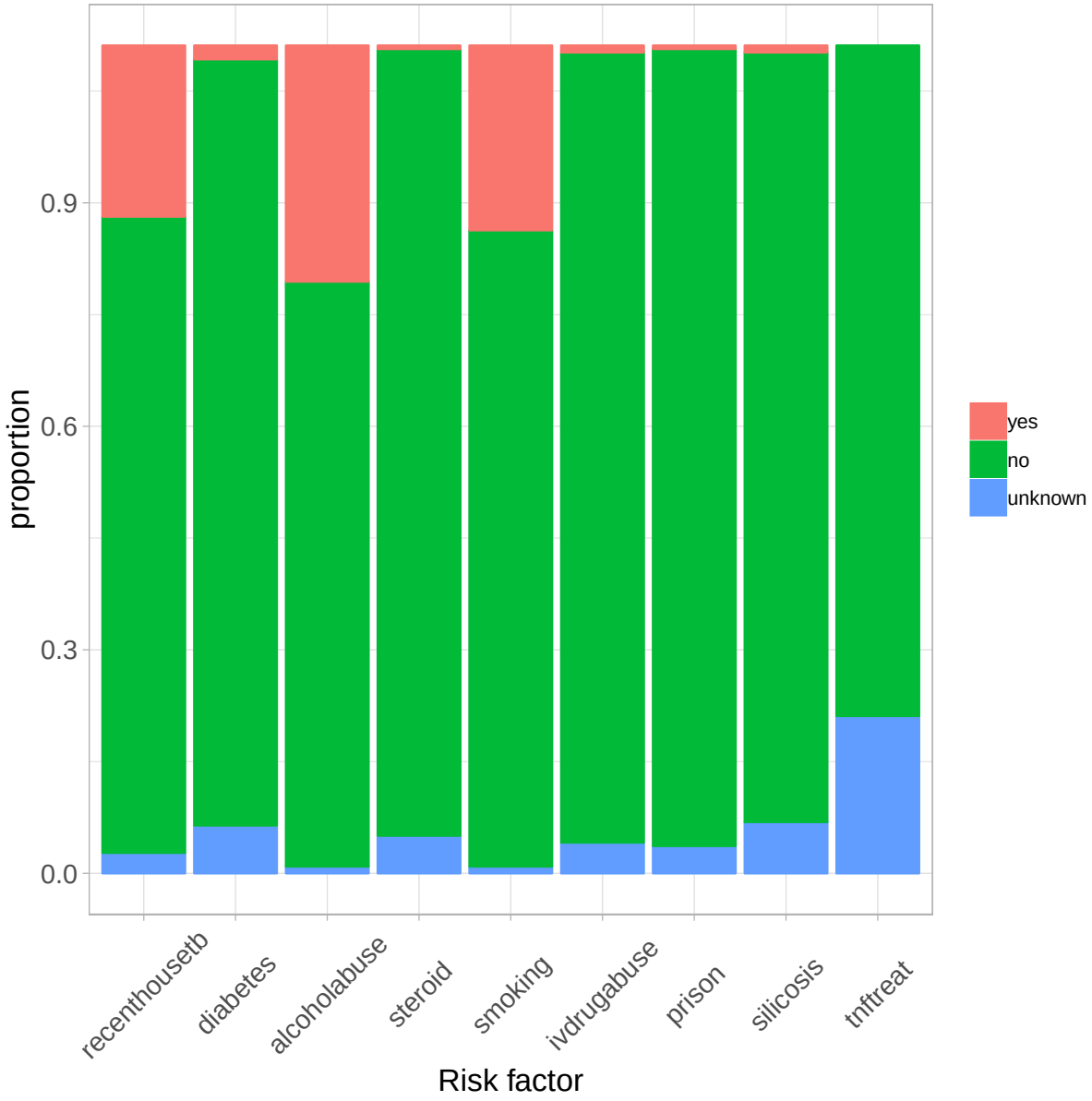


Figure 4.19.: Proportion of risk factors evaluated among TB patients in the study population from the rural site (Ifakara).recenthousetb, recent contact with a TB case; alcoholabuse, alcohol abuse; ivdruguse, intravenous drug use; prison, incarceration; and tnftreat, tumor necrosis factor therapy.

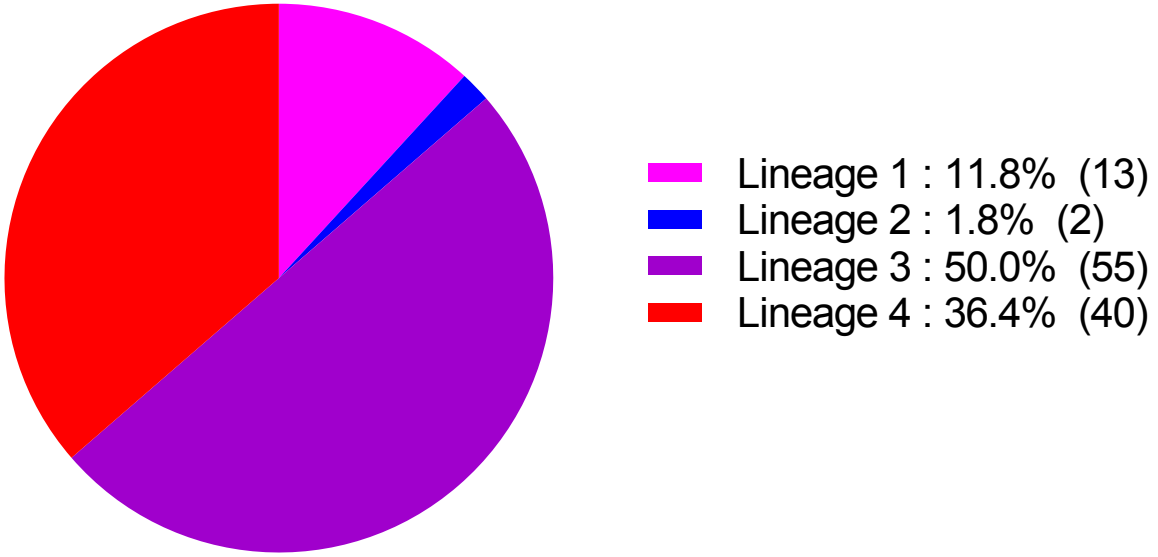


Figure 4.20.: Frequency distribution of Mtb lineages detected in Ifakara. Mtb lineage classification of 110 clinical strains.

4.4.11. Mtb lineages in Ifakara

We classified 110 Mtb strains representing 45.5% of TB patients included in the study from the rural site. Our SNP-typing results revealed the presence of four Mtb lineages (Lineage 1 –4). The most frequent lineages were Lineage 3 with 55 strains (50%) and Lineage 4 with 40 strains (36.4%). Thirteen strains (11.8%) belonged to Lineage 1 and only two (1.8%) to Lineage 2 (Figure 4.20).

4.4.12. Mtb lineages and patients' characteristics in Ifakara

Using lineage information available for the 110 TB patients, we described patients' characteristics by infecting Mtb lineage, excluding Lineage 2 due to the small number of patients infected with this lineage (Table 4.7). We observed a similar median age and median BMI in patients infected with Lineage 1, 3 and 4. Furthermore, sex distribution and HIV status did not differ among the lineages.

4.4.13. Rural and urban comparison

We analyzed 900 patients enrolled at the urban setting in Temeke and 242 patients enrolled at the rural setting in Ifakara. Patients in the urban setting were younger than those from the rural setting (median age of 33 years (IQR: 27–40) compared to 36 years (IQR: 28–43],

p = 0.007). Both patient populations from the urban and rural site had similar median BMI (18.2 [IQR: 16.7–20.0] vs., 17.8 [IQR: 16.2–19.9], p = 0.05). HIV prevalence was higher in the rural setting compared to the urban setting (24% vs., 20.8%) however this difference was not statistically significant. The majority of the patients in both settings were new cases, although the proportion of recurrent TB was higher in rural cases (5.8%; 14/242 vs., 1.8%; 16/900, p = 0.001). In addition, the recurrent TB cases were distributed among Lineage 3 and 4 (data not shown). There were no differences in sex distribution between the two sites. The two sites shared comparable risk factors for TB for instance, smoking, alcohol abuse and recent household contact with a TB case. In the rural setting however, the frequency of patients who had contact history with a TB case within a household was two-fold higher than those reported in the urban site (20.7% vs., 10.9%, p = 0.0003). In both sites, we detected the same four Mtb lineages (Lineage 1–4) occurring in comparable frequencies although Lineage 2 was less frequent in the rural setting.

Table 4.7.: Sociodemographic and clinical information of TB patients by infecting Mtb lineage in Ifakara

Characteristics	**Lineage 1** n (%)	**Lineage 2** n (%)	**Lineage 3** n (%)	**Lineage 4** n (%)
Age (years), median (IQR)	35 (23–40)	62.5 (59.8–65.3)	37 (29–46.5)	33 (27–46.3)
Age group (years)				
Child age (<15)	0	0	0	0
Young age (15–24)	4 (30.8)	0 (0)	5 (9.1)	4 (10.0)
Early adult (25–44)	6 (46.2)	0 (0)	35 (63.6)	25 (62.5)
Late adult (45–64)	3 (23.1)	1 (50.0)	13 (23.6)	9 (22.5)
Old age (>65)	0 (0%)	1 (50.0)	2 (3.6)	2 (5)
Sex				
Female	1 (7.7)	2 (100)	105 (28.6)	17 (30.9)
Male	12 (92.3)	0 (0)	262 (71.4)	38 (69.1)
BMI (kg/m^2), median (IQR)	17.9 (16.7–19.4)	16.0 (15.5–16.6)	17.0 (15.9–19.3)	17.1 (15.6–18.2)
BMI categories (kg/m^2)				
Underweight <18.5	8 (61.5)	2 (100)	36 (65.5)	32 (80.0)
Normal 18.5–24.9	5 (38.5)	0 (0)	18 (32.7)	8 (20.0)
Overweight 25.0–29.9	0 (0)	0 (0)	1 (1.8)	0 (0)
Obese >30	0 (0)	0 (0)	0 (0)	0 (0)
HIV status				
Negative	11 (84.6)	1 (50.0)	36 (65.5)	29 (72.5)
Positive	2 (15.4)	1 (50.0)	17 (30.9)	10 (25.0)
Unknown	0 (0)	0 (0)	2 (3.6)	1 (2.5)
Total	13	2	55	40

4.5. Discussion

We studied 900 urban and 242 rural pulmonary TB patients in Tanzania to test for associations between Mtb lineage and patient characteristics. We found that in both settings, TB was more frequent in young adults, males, and comprised mainly new TB cases. These results corroborate previous findings (Rutaihwa *et al.*, 2019b; Sikalengo *et al.*, 2018). Our findings further showed Lineage 3 and 4 to be the dominant lineages in both the urban and rural locales. Finally, we demonstrate Lineage 2 strains to be highly transmissible in the urban setting, despite being the least frequent.

TB was most prevalent among the "early adult" age group (25-44) both in the urban and rural sites, an observation that is consistent with national estimates (NTLP, 2016; Rutaihwa *et al.*, 2019b). This finding highly speaks for on-going transmissions in the two settings, which was irrespective of infecting Mtb lineage. Furthermore, we found that younger TB patients in the urban setting were more likely to be infected with Mtb Lineage 2 strains. We did not detect additional correlations between genetic characteristics of the infecting Mtb strains with patients' sociodemographic and clinical information.

The emergence of HIV in Tanzania was linked to the rise of TB incidence (Egwaga *et al.*, 2006). More than 20% of TB patients in both study populations were also HIV co-infected. These findings indicate that HIV is still an important risk factor for TB and a major contributor to TB burden, particularly in settings of high TB transmission. HIV has been suggested to have minor overall contribution to TB transmission for reasons such as minimal social contact due to morbidity, short infectious periods due to increased mortality, smear negative and extrapulmonary forms of disease (Corbett *et al.*, 2004; Yates *et al.*, 2015). However, these outcomes could vary depending on the HIV immunosuppression levels of the TB patients and TB/HIV co-infected patients with high CD4 counts are shown to transmit as well as HIV negative individuals. We found Mtb strains from TB/HIV patients were as likely to be found in transmission clusters as strains from HIV negative TB patients. Unfortunately, we did not have CD4 counts for such patients. However, this observation could be driven by the fact that our study populations consisted of TB patients who were smear positive which is the most transmissible form of disease. Overall, HIV individuals could contribute to TB transmission given the scale up of antiretroviral therapy in HIV endemic countries, which restores immunosuppression.

Based on the proportions of TB patients with recent household contact to TB case, our findings suggest that household transmission occurred more frequently in the rural setting than the urban setting. However, given these proportions, still more than 80%

of TB transmission appears to occur in the communities. Moreover, sharing a household with an index case does not necessarily reflect within-household transmission as previous evidence shows multiple TB cases in one household to be infected with dissimilar Mtb strains (Lalor *et al.*, 2017).

Our genetic clustering results indicated that Lineage 2 and 3 strains are likely to transmit more than Lineage 4 strains in the urban setting. One would expect this to be the case for Lineage 3 since it is the predominant lineage, which in general had a higher number of clusters and also formed large cluster sizes. Lineage 2 on the other hand, is the least frequent but showed remarkable propensity to transmit, where more than 50% of its strains were found in clusters. In line with findings from other settings, recent and higher magnitude of transmission in Lineage 2, hence increased virulence is unrelated to drug resistance. Despite being one of the two major lineages, Lineage 4 strains formed the smallest cluster sizes and had the least proportions of strains found in clusters. We found that Lineage 1 formed no transmission clusters in the urban setting. Patients infected with Lineage 1 strains have been shown elsewhere to be less likely due recent transmission and also less likely to cause secondary cases (Guerra-Assunção *et al.*, 2015a; Holt *et al.*, 2018). Our results support the notion that ancient lineages, in this case Lineage 1 is less virulent compared to modern lineages. However, this does neither seem to be related to HIV co-infections nor old age.

We observed 80.5% of the TB patients to have taken at least one type of antibiotic prior to TB diagnosis. The most common antibiotics consumed include amoxicillin, cloxacillin, amplicox, and ampicillin. Although 6% of the patients from the urban site were recorded to have taken ciprofloxacin, which is a quinolone, this proportion did not include the patient identified with a FQ resistant strain. We mostly found mutations conferring resistance to first line drugs where INH resistance was more prominent. Mutations conferring resistance to rifampicin were not detected and this might also be due to excluding such strains prior processing in the BSL3 laboratory in Basel. Drug resistance levels are low in Tanzania particularly among new TB cases (Nagu *et al.*, 2015). Part of the reason could be because Tanzania was the first country in the world to implement the directly observed treatment, short course strategy more than four decades ago. In addition, TB drugs are administered by the national control program. However, in addition to factors like traditional healers, over-counter availability of other antibiotics could contribute to delays in TB diagnosis and treatment which ultimately could have consequences on disease transmission (Said *et al.*, 2017).

Our study has several limitations. First, TB-DAR cohort enrolls smear-positive TB pa-

tients. Therefore, we were unable to assess associations of Mtb strains from smear-negative patients, which are a relevant clinical presentation of TB disease, especially in TB/HIV patients. Hence the proportion of HIV/TB might have been underestimated. Another limitation could be due to convenience sampling as only patients visiting Temeke District Hospital and CDCI hospitals in the respective study sites were recruited. However, patient enrolment was random. Hence we would expect representative study populations. Since the study used a subset of the total patients recruited during the study duration, findings from this study are preliminary and incomplete. For instance, x-ray scores were unavailable at the time of analysis which hindered us from assessing correlations of cavitary disease with Mtb strains and genetic clustering. Cavitary disease is a proxy for enhanced transmission and virulence to reflect more successful Mtb strains (Wampande *et al.*, 2013). Finally, WGS data of Mtb strains from the rural setting were not available at the time of analysis. Therefore, genetic clustering (and drug resistance) among rural Mtb strains could not be evaluated and transmission pattern disparities between urban and rural addressed.

Although several studies have defined Mtb clinical phenotypes with respect to host population in question, the implications of Mtb genetic variation in clinical settings are not entirely clear. The existing inconsistent results or lack of correlations between infecting lineage and patients' characteristics in our case might be contributed by sampling in addition to other factors such as host genetics and the environment. Perhaps future studies should consider addressing the interplay between the "triad" and its influence on TB infection and disease preferably in a population based setting for over a period of time. This could inform us on relevant epidemiological features particularly those related to transmission and help improve control strategies.

5. Phylogeography of *Mycobacterium tuberculosis* Lineage 1 and Lineage 3

Liliana K. Rutaihwa[1,2,3], Fabrizio Menardo[1,2], Chloe Loiseau[1,2], Emilyn Costa Conceição[4], Miriam Rheinhard[1,2], Julia Feldmann[1,2], Serej D Ley[1,2], Bijaya Malla[1,2], Niaina Rakotosamimanana[5], Horng-Yunn Dou[6], Janet Fyfe[7], Iñaki Comas[8,9], Christophe Sola[10,11], Darío García-de-Viedma[12,13,14], Sonia Borrell[1,2], Klaus Reither[1,2], Lukas Fenner[1,2,15], Daniela Brites[1,2], Sebastien Gagneux[1,2]

[1] Swiss Tropical and Public Health Institute, Basel, Switzerland
[2] University of Basel, Basel, Switzerland
[3] Ifakara Health Institute, Bagamoyo, Tanzania
[4] Universidade Federal do Rio de Janeiro, Instituto de Microbiologia Paulo de Goés, Rio de Janeiro, Brazil
[5] Unité des Mycobactéries, Institut Pasteur de Madagascar, Antananarivo, Madagascar
[6] National Institute of Infectious Diseases and Vaccinology, National Health Research Institutes, Zhunan, Miaoli, Taiwan
[7] Victorian Infectious Diseases Reference Laboratory, Melbourne, Victoria, Australia
[8] CIBER en Epidemiología y Salud Pública, Valencia, Spain
[9] Instituto de Biomedicina de Valencia, IBV-CSIC, Valencia, Spain
[10] Institut de Biologie Intégrative de la Cellule (I2BC), CEA, CNRS, Univ. Paris Sud, Université Paris-Saclay, F-91198 Gif-sur-Yvette, France
[11] National Institute for Infectious and arasitic Diseases, Sofia, Bulgaria
[12] Servicio de Microbiología Clínica y Enfermedades Infecciosas, Hospital General Universitario Gregorio Marañón, Madrid, Spain
[13] Instituto de Investigación Sanitaria Gregorio Marañón, Madrid, Spain
[14] CIBER Enfermedades respiratorias, CIBERES, Spain

[15] Institute of Social and Preventive Medicine, University of Bern, Bern, Switzerland

5.1. Abstract

There are 7 human-adapted phylogenetic lineages of *Mycobacterium tuberculosis* (Mtb) complex. With the increasing number of whole genome sequenced strains, it is clear that Mtb strains are also heterogeneous within individual lineages. Whilst several studies have investigated the phylogenetic substructures of Lineage 2 and Lineage 4, Lineage 1 and 3 remain unexplored. Unlike Lineage 2 and 4 which are globally distributed, Lineage 1 and 3 show an intermediate geographical range that mainly rims the Indian Ocean, making the two lineages important drivers of tuberculosis (TB) epidemics in that part of the world.

In this study, we refined the phylogenies and geographical distribution of Lineage 1 and 3 by using the most comprehensive whole genome dataset to date, covering a wide geographical range of clinical strains belonging to these two lineages. We confirmed our previous classifications for describing the global population structure of Lineage 1, which however was not the case for Lineage 3. We also identified Lineage 1 strains in Brazil that were likely introduced multiple times from Africa. Further, our results suggest multiple transfers of Lineage 3 across the endemic geographic regions, indicating the ability of these strains to establish in various locales of different host backgrounds. Finally, the prevalence of Lineage 1 and 3 in the African region are likely reflecting back-to-Africa migrations from South Asia.

5.2. Introduction

Human tuberculosis is predominantly caused by seven human-adapted phylogenetic lineages of *Mycobacterium tuberculosis* (Mtb) complex, Lineage 1–7 (Gagneux, 2018). These lineages are distributed in a phylogeographic manner, a pattern that partly reflects human demographic history and local adaptation of the pathogen to particular human population (Wirth *et al.*, 2008; Gagneux, 2012; Comas *et al.*, 2013). The most likely origin of the Mtb complex overall is in Africa, from where it dispersed around the world following human migrations out of Africa (Wirth *et al.*, 2008; Comas *et al.*, 2013). The present phylogeography of the Mtb complex shows human-adapted lineages to exhibit any of three geographical ranges, where the "generalist" Lineages 2 and 4 are globally distributed, the "specialist" Lineages 5, 6 and 7 are geographically restricted, and Lineage 1 and 3 showing an intermediate geographic range (Coscolla *et al.*, 2014; Stucki *et al.*, 2016).

Lineage 1 has been hypothesized to have originally emerged in Africa, after a first split from the most recent common ancestor (MRCA) of Mtb complex (Comas *et al.*, 2013). Of the three "ancient" lineages, Lineage 1 is the only phylogenetic lineage observed in Eastern Africa and outside of Africa, mainly in regions along the rim of the Indian Ocean, and has thus previously been referred to as the Indo-Oceanic lineage (Gagneux *et al.*, 2006). Lineage 3 on the other hand, emerged after the second split from the MRCA among the three evolutionary modern lineages, Lineage 2–4 (collectively also referred to as "Eurasian" lineages) around the Eastern Mediterranean region. The Eurasian lineages have previously been hypothesized to have expanded together with their respective human host populations in Europa, India and China, respectively, and to have later been reintroduced into Africa (Hershberg *et al.*, 2008). Today, Lineage 3 strains are mainly distributed in the Indian sub-continent, Eastern and Northern Africa and Central Asia (O'Neill *et al.*, 2019).

The recent increases in whole genome data from Mtb clinical strains reveal that strains are heterogeneous within individual lineages. The global diversity and evolutionary histories of lineages 2 and 4 have been extensively studied (Luo *et al.*, 2015; Merker *et al.*, 2015; Stucki *et al.*, 2016; Brynildsrud *et al.*, 2018), and these studies revealed substantial within lineage heterogeneity in phylogenetic structure and geographical distribution. However, similar studies are largely lacking for Lineage 1 and 3, and their substructures remain poorly characterized. These two lineages remain important drivers of TB epidemics, particularly in Southern Asia and the high TB burden settings of Africa.

The currently known phylogenetic substructures of Lineage 1 and 3 were primarily defined

based on spoligotyping where the two lineages are referred to East African Indian (EAI) and Central Asian (CAS), respectively. The spoligotyping subfamilies of Lineage 1 include EAI2-Manila, EAI1-SOM, EAI8-Madagascar, EAI3-INDIA, EAI4-VNM and EAI (6 and 7) - BGD1, reflecting regions they are commonly isolated from (Brudey *et al.*, 2006). Similarly, the CAS1-Delhi and CAS1-Kili are the two common subfamilies within Lineage 3 mainly isolated in the Indian subcontinent (Bhanu *et al.*, 2002; Singh *et al.*, 2004) and in Tanzania (Eldholm *et al.*, 2006; Mbugi *et al.*, 2016), respectively.

Furthermore, the geographic distribution of Lineage1 and 3 might have been underestimated. Until recently, the two lineages were known to be confined to the rim of the Indian Ocean. However, recently, Lineage 1 strains have been reported in Acapulco, Mexico (Nava-Aguilera *et al.*, 2011). These strains were of the EAI2-Manila subtype (a predominant Lineage 1 family in the Philippines), possibly reflecting the historical contacts between Mexico and Philippines during Spanish colonization. Lineage 3 on the other hand was reported in New Zealand and found to be associated with the indigenous Māori ethnic group (Yen *et al.*, 2013). These findings suggest Lineage 1 and 3 distributions that are uncharacteristic, which require further investigation.

We recently used WGS data of 420 Mtb clinical strains (Coscolla, unpublished) and extracted phylogenetically informative SNPs to define sublineages of Lineage 1 and Lineage 3 (Rutaihwa, master's thesis, 2014). These definitions were based on WGS data from 50 and 42 strains, respectively (Figure 5.2 and 5.3), and overlapped with previous classifications (Coll *et al.*, 2014). Our findings showed substantial substructure within Lineage 1 and 3. In the present study, we aimed to revisit the global phylogenies and geographical distributions of Lineage 1 and 3 using the most comprehensive global WGS datasets to date and to infer their evolutionary histories.

Figure 5.1.: **Geographical distribution of Lineage 1 and 3.** An intermediate geographical range Lineage 1 and 3 (Coscolla *et al.*, 2014).

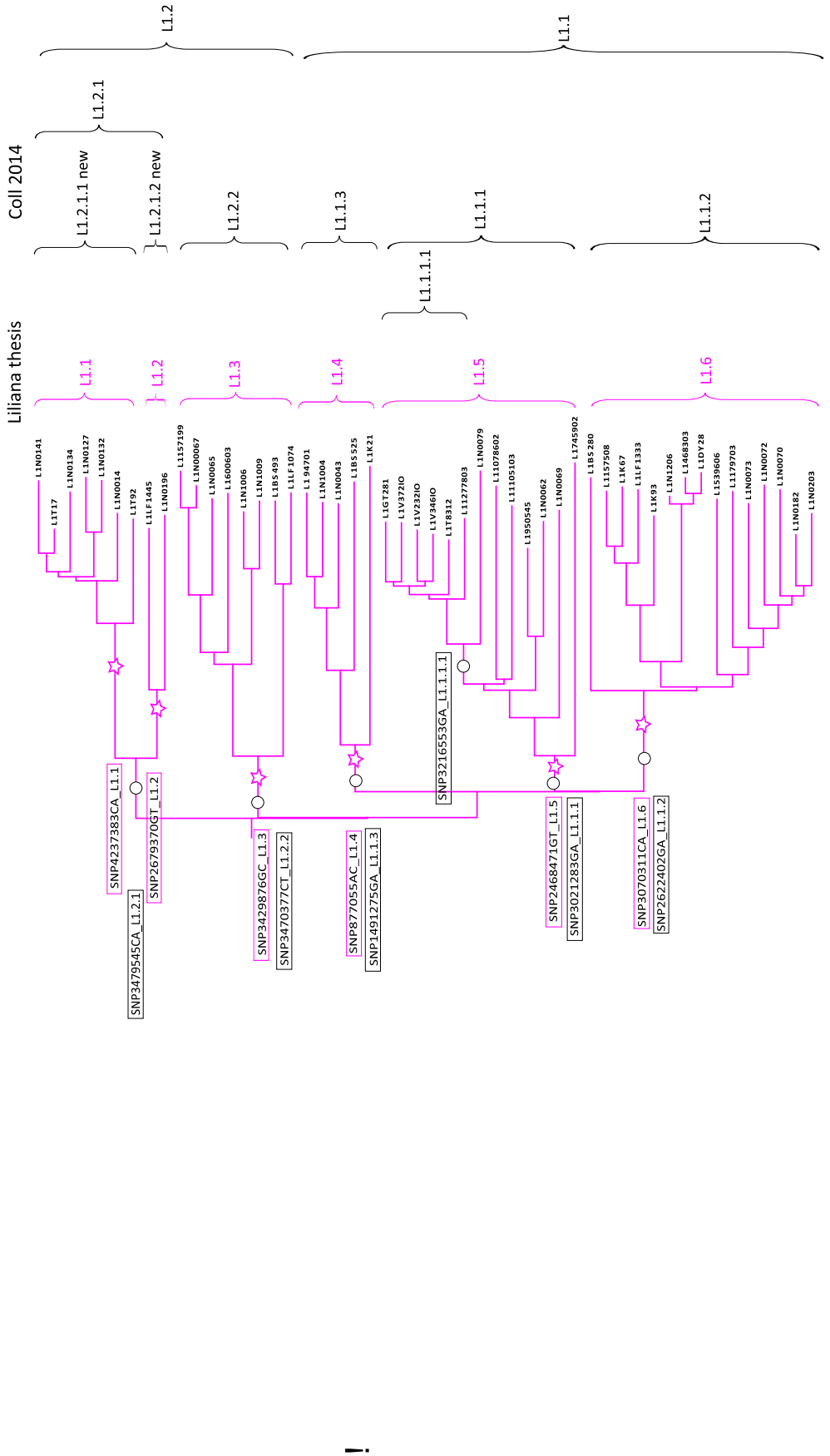


Figure 5.2.: **Phylogenetic tree of whole genome sequences of 50 Lineage 1 strains, mapped with SNP markers defining the sublineages.** In pink are the sublineage definitions by Rutaihwa, master's thesis, 2014 and in black are those defined by Coll *et al.*, 2014

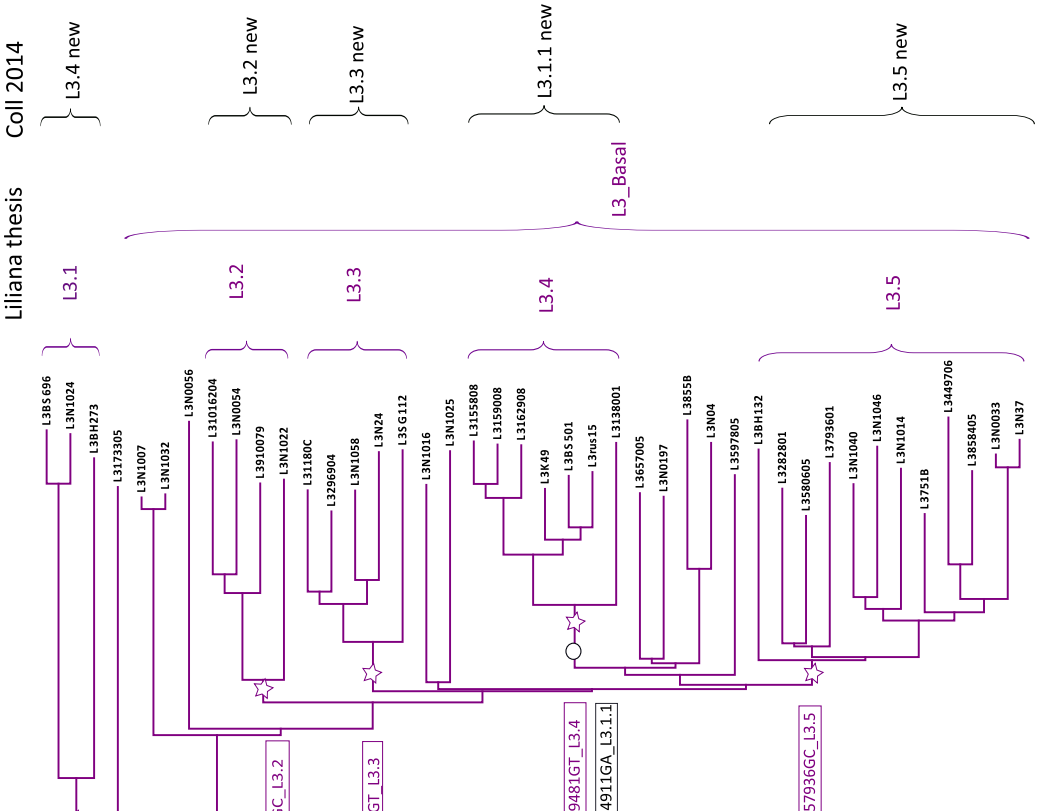


Figure 5.3.: **Phylogenetic tree of whole genome sequences of 42 Lineage 3 strains, mapped with SNP markers defining the sublineages.** In purple are the sublineage definitions by Rutaihwa, master's thesis, 2014 and in black are those defined by Coll *et al.*, 2014.

5.3. Methods

5.3.1. Study sample collection

To refine the global phylogenies and geographical distribution of Lineage 1 and 3, we obtained globally representative strain collections of the two lineages focusing on the geographical locations where these strains are commonly reported. In addition, we actively explored geographical locations where these strains are uncharacteristic, for example Lineage 1 strains from Brazil. The geographical locations mostly included Eastern Africa, Southern Asia for both lineages; South-eastern Asia, Australia, Southern America and Melanesia for Lineage 1; and Central Asia for Lineage 3. In addition, we obtained whole genome sequencing data of Lineage 1 and 3 from previously published studies. In total, we included 1,667 genome sequences of Lineage 1 and 2,104 genome sequences of Lineage 3.

5.3.2. Mtb isolates and DNA extraction

Mtb isolates previously identified as Lineage 1 and 3 either by SNP-typing or spoligotyping, were grown in Middlebrook 7H9 liquid medium supplemented with ADC and incubated at 37°C. Purified genomic DNA was obtained from cultures using the CTAB extraction method (Embden *et al.*, 1993).

5.3.3. Whole genome sequencing, variant calling and phylogenetic inference

For the Lineage 1 and 3 strains that we sequenced in house, whole genome sequencing was performed on prepared libraries from purified genomic DNA using Illumina Nextera®XT library and NEBNext®Ultra™ II FS DNA Library Prep Kits (Illumina, San Diego, USA). Sequencing was performed using the Illumina HiSeq 2500 or NextSeq 500 paired-end technology (Illumina, San Diego, USA). For the whole genome analysis, we first clipped the Illumina adapters from the raw sequencing reads and then trimmed low quality sequencing reads using Trimmomatic 0.33 (SLIDINGWINDOW:5:20) (Bolger *et al.*, 2014). We excluded short reads (< 20 bp) from the downstream analysis and the overlapping paired-end reads were merged with SeqPrep 1.2 (overlap size = 15) (`https://github.com/jstjohn/SeqPrep`). The reads were mapped to a reconstructed Mtb complex ancestor sequence (Comas *et al.*, 2013) using BWA 0.7.13 (Li *et al.*, 2009). Picard 2.9.1 module (Mark Duplicates) was used to mark duplicated reads, which we excluded

afterwards. Pysam 0.9.0 (`https://github.com/pysam-developers/pysam`) was used to exclude reads with alignment score lower than (0.93*read_length)-(read_length*4*0.07), consistent with more than 7 miss-matches per 100 bp. We then extracted SNPs with Samtools 1.2 mpileup (Li *et al.*, 2009) and VarScan 2.4.1 (Koboldt *et al.*, 2012) and kept SNPs with minimum mapping quality of 20, minimum base quality at a position of 20, minimum read depth at a position of 7X, minimum percentage of reads supporting the call 90% and maximum strand bias for a position 90%. Genomes with coverage < 20 X and those containing phylogenetic SNPs belonging to different lineages (mixed strains) were excluded. We annotated the SNPs using snpEff 4.11 according to the Mtb H37Rv reference annotation (NC_000962.3). We excluded SNPs within regions such as PPE and PE-PGRS, phages, insertion sequences repetitive regions (at least 50 bp identical to other regions in the genome) (Stucki *et al.*, 2016). We excluded SNPs known to confer drug resistance from phylogenetic reconstruction. We inferred a maximum likelihood phylogeny with RAxML 8.3.2 using a general time reversible (GTR) model and 1,000 bootstrap inferences (Stamatakis, 2006).

5.3.4. Lineage 1 anad 3 substructures and distribution

Having reconstructed the phylogenies of Lineage 1 and 3, we next examined their geographical distribution by mapping the patients' place of birth as a proxy for origin of the strains. Further, we used the previous sublineage definitions to describe the substructure of Lineage 1.

5.4. Results

5.4.1. Lineage 1 global phylogeny and geographical distribution

To characterize the global diversity of Lineage 1, we analyzed 1,667 whole genome sequences from at least 49 countries, including Africa (Eastern, Southern Central and Western), Asia (Southern, South-eastern, Eastern and Western), America (Southern and Northern), Melanesia, Australia and Europe. We next reconstructed ML phylogeny for Lineage 1 based on 107,703 variable positions (Figure 5.4). As previously demonstrated, our Lineage 1 phylogeny displayed a defined substructure with discrete monophyletic sublineages. We found our previous definitions for the Lineage 1 substructure to be comprehensive and covering the entire diversity of the globally representative strains. That

is to say, most of the Lineage 1 strains were classified into any of the six sublineages we had defined previously (L1.1 – L1.6). However, we detected an additional clade branching off between the L1.1 (EAI2-Manila) and L1.2 sublineages, and the three together formed the L1.2.1 clade previously defined by (Coll *et al.*, 2014). Overall, L1.2.1 is the deepest branching clade of Lineage 1 comprising on the one hand strains from Eastern Africa (i.e. Ethiopia, Djibouti, Somalia and Tanzania), Melanesia (i.e. Papua New Guinea), and Southeastern Asia (i.e. East Timor) in the L1.2 sub clade. On the other hand, L1.1 also embedded within L1.2.1 clade, comprised strains mainly from Southeastern Asia (e.g. the Philippines). The majority of Lineage 1 strains in our dataset came from Eastern African and Southern Asia (i.e. the Indian subcontinent), and these were scattered throughout on the phylogeny into several sublineages. Of note, strains from the Indian subcontinent occupied basal positions in most sublineages (i.e. L1.3, L.4 and L1.6). Our Lineage 1 dataset also included strains from Brazil. We observed such strains to fall into three of the six Lineage 1 sublineages (L1.3, L1.4 and L1.6). These Lineage 1 strains from Brazil clustered with those from African countries including Malawi and Madagascar, indicating multiple introductions from Africa into Brazil, perhaps linked to the European slave trade. Strains belonging to sublineage L1.5 were mostly from Vietnam, which clustered in a monophyletic sub group L1.1.1.1 (Figure 5.2). These strains were reported from a recent study in Vietnam, where little transfer was detected with strains of non-Vietnamese origin (Holt *et al.*, 2018). Of note, the basal clade of L1.5 comprised strains from West Africa.

5.4.2. Lineage 3 global phylogeny and geographical distribution

We used 2,104 whole genome sequences from at least 42 countries including Africa (Eastern, Southern Central and Western), Asia (Southern, South-eastern, Eastern and Western), Northern America and Europe to characterize the global diversity of Lineage 3. We inferred a ML phylogeny of the Lineage 3 strains from 76,319 variable nucleotide positions (Figure 5.5). In contrast to Lineage 1, Lineage 3 showed a less structured phylogeny and as a result the previously established sublineage classifications were not optimal to describe its true phylogenetic substructure. In general, the deep branching clade on the phylogeny encompassed a previously defined L3.1 sublineage, which includes the previously described "pseudo-Beijing" strains (Fenner *et al.*, 2011), with the rest of the strains belonging to the monophyletic "L3Basal" clade (Figure 5.3). Most of the Lineage 3 strains originated from Lineage 3 endemic areas of Eastern Africa and Southern Asia. Although the phylogeny

depicts clustering of strains by geography to a certain degree, these strains still appear to intersperse irrespective of their geographical origin. For instance in Figure 5.5, we highlighted (in red) strains isolated from a clinical setting in Dar es Salaam, Tanzania, which are intermingled with strains from several other locations (in black).

Figure 5.4.: Whole-genome phylogenetic tree of 1,667 Lineage 1 strains mapped with country origin of strains.

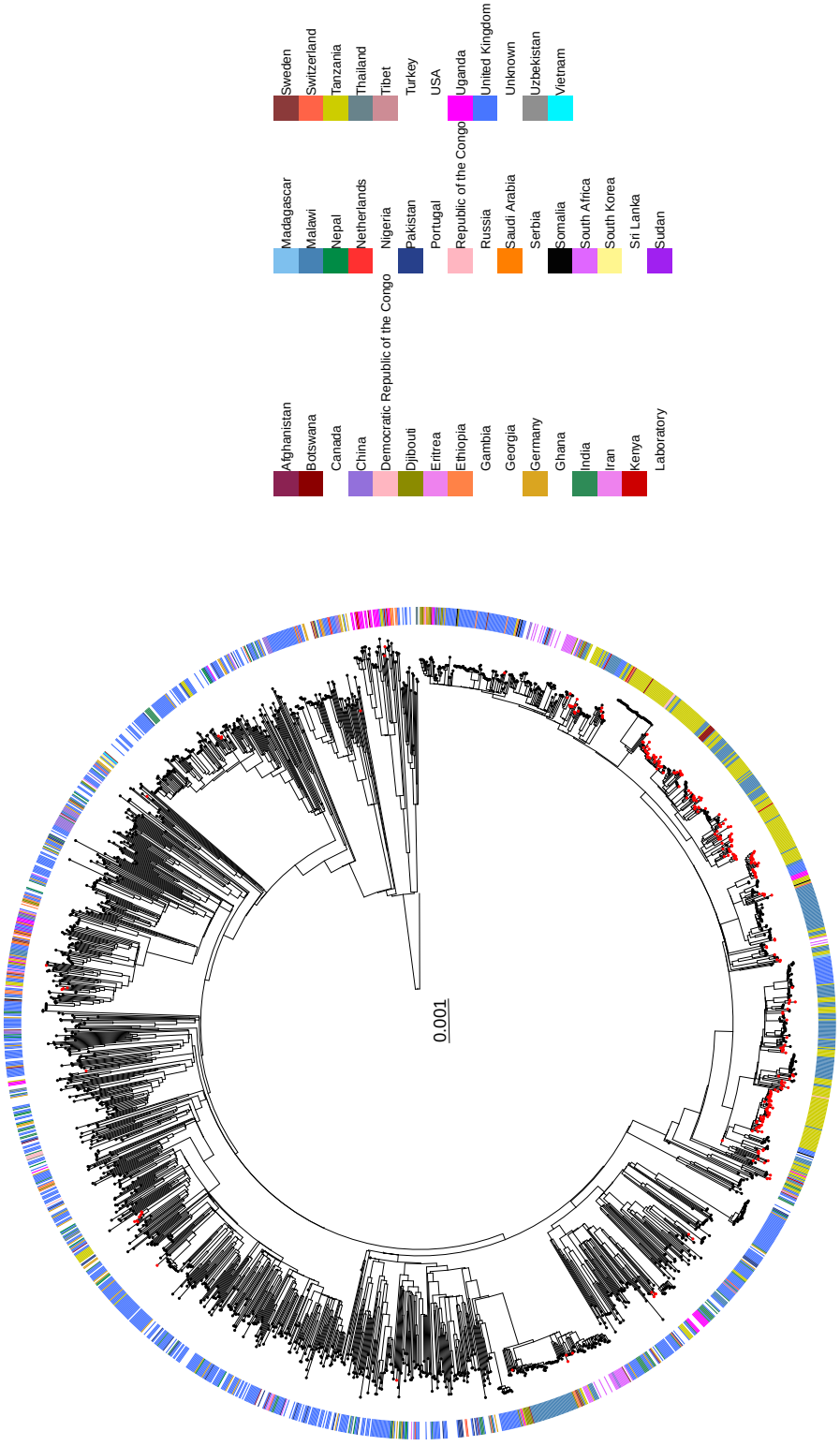


Figure 5.5.: Whole-genome phylogenetic tree of 2,104 Lineage 3 strains mapped with country origin of strains. Red tipopoints are the Lineage 3 genomes from TBDAR dataset (*Chapter 4*) and the rest in black.

5.5. Discussion

We studied the global phylogenies and geographical distribution of Mtb Lineage 1 and 3, using the most comprehensive whole genome datasets available to date. We confirmed the previously described geographical ranges for Lineage 1 and 3, where both lineages span from Eastern Africa and along the Indian Ocean regions to Southern Asia, with Lineage 1 extending further east into South-East Asia and Melanesia. Furthermore, we show that our previously established Lineage 1 sublineage definitions are robust, covering most of the global diversity of Lineage 1 strains. By contrast, Lineage 3 presented a less distinct phylogenetic substructure. Our novel findings revealed the presence of Lineage 1 in the Para state, North of Brazil that likely resulted from multiple introductions. We observed Lineage 3 strains to intermingle geographically, suggesting frequent exchange among the endemic regions. Finally, our study proposes a common evolutionary history regarding the introductions of Lineage 1 and 3 back to Africa.

We characterized Lineage 1 strains in our study using the previously defined six sublineages, L1.1 – L1.6. Based on the phylogenetic positions and geographical distribution of the strains, we hypothesize possible evolutionary scenarios (some previously put forward) to explain the present phylogeography of Lineage 1. We found that the most basal branch of the Lineage 1 phylogeny (L1.2) consisted mainly of strains from Eastern Africa and Melanesia, an observation which supports the emergence of Lineage 1 in Eastern Africa followed by its early dispersal into Melanesia (Hershberg *et al.*, 2008). This dispersal corresponds to human migrations out of Africa that mostly likely arrived to Melanesia for example in Papua New Guinea and Australia then as part of the former Sahul. Although L1.1/EAI2-Manila is commonly isolated in patients of Filipino origin (Douglas *et al.*, 2003) also as evidenced in this study, our results reveal this clade contains strains from other patients' origins such Papua New Guinea and some Southeastern Asian countries. This distribution of L1.1/EAI2-Manila could reflect the Austronesian migration wave into such regions via the Philippines.

In general, we observe strains from Eastern Africa and Southern Asia (i.e. India subcontinent) to be the most diverse. The fact that strains from Southern Asia are spread out and occasionally occupy basal positions on the phylogeny indicates further diversification of Lineage 1 strains in Southern Asia post introductions from Eastern Africa. Therefore, suggesting the current Lineage 1 geographical distribution to partly reflect migrations out of Southern Asia (O'Neill *et al.*, 2019). Throughout history, the Indian Ocean has been vital to interactions between East and West for instance via several trade routes. Such historical and present interactions could have facilitated the spread of Mtb strains along

the Indian Ocean region (Gilbert, 2002).

The presence of Lineage 1 strains in Brazil suggest historical contacts with Eastern Africa as these strains clustered with those from Malawi and Madagascar. One possible scenario to have led to the introduction of Lineage 1 into Brazil is the Trans-Atlantic Ocean slave trade between the 16th and 19th centuries (Vos, 2012). In addition to Western Africa and Mozambique, African slaves were also shipped from Eastern Africa (UNESCO, 1992) which is part of geographical range for Lineage 1. Of note, strains from Brazil were grouped into three different Lineage 1 sublineages, indicating multiple introductions of Lineage 1 into Brazil.

In line with previous observations, Lineage 3 presented less phylogenetic structure. Thus, the previous definitions of Lineage 3 sublineages were inappropriate to describe the current substructure. Similar to Lineage 2, it was proposed that Lineage 3 experienced sudden expansion as a result of human population expansions, albeit with lower rates of migration compared to e.g. Lineage 2 (Comas *et al.*, 2013; Luo *et al.*, 2015; O'Neill *et al.*, 2019). By contrast, Lineage 3 emerged in Southern Asia from where it has solely disseminated to other geographical locations mainly into Eastern and North Africa. Based on our Lineage 3 phylogeny, it appears that there is constant exchange of strains across the Lineage 3 endemic regions. As an illustration, we mapped Lineage 3 strains isolated from TB patients in Dar es Salaam, Tanzania (studied in *Chapter 4* of this thesis). We found these strains to intersperse with others from neighboring countries such as Malawi. This observation would suggest multiple transfers of Lineage 3 between the regions. Again, Lineage 3 strains from India were seen to be scattered throughout the phylogeny, indicating a similar scenario to Lineage 1 strains in the subcontinent. In both phylogenies, we observe further intracontinental spread.

Due to logistical and time constrains, our analysis of this chapter is preliminary and only included partial datasets for the two respective lineages. Therefore, we are aware that our research might have a number of limitations at this stage. To begin with, both Lineage 1 and 3 global phylogenies were still missing important geographical links that was contributed by either difficulty in the accessibility of data or availability of the data at the time of analysis. For instance, at the time of analysis, Lineage 1 genomes from Taiwan, which included strains isolated from Aboriginal population were not yet available. Given that our Lineage 1 collection included strains from other Austronesian-speaking regions such as Madagascar and others from Southeastern Asia, we might have explored dispersal of Lineage 1 linked to the Austronesian expansion (Ko *et al.*, 2014). Unfortunately, we so far were also unsuccessful in obtaining Lineage 1 strains from Australian Aborigine

population. Such strains would have perhaps allowed us to disentangle introductions of Lineage 1 linked to early migration waves from Africa from those related to Austronesian migrations (Main *et al.*, 2001; Reich *et al.*, 2011). An additional setback in our Lineage 1 dataset was the unavailability of L1.1/EAI2-Manila strains from Acapulco, Mexico previously reported (Nava-Aguilera *et al.*, 2011), which hinder us from revisiting the evolutionary spread most likely as result of historical Spanish contacts. Similarly, our Lineage 3 dataset lacked representatives from important geographical zones such as the Eastern Mediterranean and North Africa where the strains are known to be diverse and frequent (Comas *et al.*, 2015; O'Neill *et al.*, 2019). In addition, information on place of origin was missing for several strains of both lineages. This put a setback when inferring the phylogeographies and evolutionary histories of the two lineages. For instance, in the current analysis, most of Lineage 3 strains on the deepest rooted branches lacked information on geographical origin.

Unfortunately even for the data available, due to time constraints, we were unable to perform more extensive downstream analyses following the phylogenetic inferences. First, we were unable to validate the existing sublineage definitions using alternative methods. We propose to validate / confirm the existing sublineage definitions using approaches such as principle component analysis (PCA), pairwise distances, genetic diversity comparison i.e. nucleotide diversity to determine the within lineage separation for robust classifications and revisiting the Lineage 3 substructure all together. Given the historical scenarios put forward, we would also perform coalescent analyses to reconstruct the evolutionary history and spread of the two lineages.

Based on the phylogenetic and geographic distribution, our results suggest initial spread of Lineage 1 out of Africa to Southern Asia, extending further east. These introductions were likely followed by diversification of Lineage 1 mainly in the Indian subcontinent prior to dissemination out of Southern Asia (O'Neill *et al.*, 2019). This is in contrast to Lineage 3, which is proposed to have emerged in Southern Asia. With regard to out-of-India / Southern Asian migration, both Lineage 1 and Lineage 3 were likely distributed via similar human migrations particularly those that link Eastern Africa to Southern Asia before spreading within the African continent and beyond Eastern Africa to Southern America in the case of Lineage 1. Our study highlights the importance of globally representative datasets when studying population structure and geographical distribution of Mtb lineages. This contributes to our understanding of the evolutionary history and spread of Lineage 1 and 3, which are currently underexplored.

6. Multiple introductions of *Mycobacterium tuberculosis* Beijing into Africa over centuries

Liliana K. Rutaihwa[1,2,3], Fabrizio Menardo[1,2], David Stucki[1,2], Sebastian M. Gygli[1,2], Serej D Ley[1,2,4,5], Bijaya Malla[1,2,6], Julia Feldmann[1,2], Sonia Borrell[1,2], Christian Beisel[7], Kerren Middelkoop[8,9], E. Jane Carter[10,11], Lameck Diero[11], Marie Ballif[12], Levan Jugheli[1,2], Klaus Reither[1,2], Lukas Fenner[1,2,12], Daniela Brites[1,2*#], Sebastien Gagneux[1,2*#]

[1] Department of Medical Parasitology and Infection Biology, Swiss Tropical and Public Health Institute, Basel, Switzerland

[2] University of Basel, Basel, Switzerland

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

[4] Infection and Immunity Unit, Papua New Guinea Institute of Medical Research, Goroka, Papua New Guinea

[5] Division of Molecular Biology and Human Genetics, Faculty of Medicine and Health Sciences, DST-NRF Centre of Excellence for Biomedical Tuberculosis Research, South African Medical Research Council Centre for Tuberculosis Research, Stellenbosch University, Cape Town, South Africa

[6] Department of Radiation Oncology, Inselspital, Bern University Hospital, University of Bern, Bern, Switzerland

[7] Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland

[8] Desmond Tutu HIV Centre, Institute of Infectious Disease and Molecular Medicine, Cape Town, South Africa

[9] Department of Medicine, University of Cape Town, Cape Town, South Africa

[10] Division of Pulmonary and Critical Care Medicine, Warren Alpert School of Medicine at Brown University, Providence, RI, United States

[11] Department of Medicine, Moi University School of Medicine, Eldoret, Kenya
[12] Institute of Social and Preventive Medicine, University of Bern, Bern, Switzerland

* Corresponding authors
Email: d.brites@swisstph.ch and sebastien.gagneux@swisstph.ch

# Equal contribution

This paper has been published in *Frontiers in Ecology and Evolution* 2019, 7:112.

6.1. Abstract

The Lineage 2–Beijing (L2–Beijing) sub-lineage of *Mycobacterium tuberculosis* has received much attention due to its high virulence, fast disease progression, and association with antibiotic resistance. Despite several reports of the recent emergence of L2–Beijing in Africa, no study has investigated the evolutionary history of this sub-lineage on the continent. In this study, we used whole genome sequences of 781 L2 clinical strains from 14 geographical regions globally distributed to investigate the origins and onward spread of this lineage in Africa. Our results reveal multiple introductions of L2–Beijing into Africa linked to independent bacterial populations from East- and Southeast Asia. Bayesian analyses further indicate that these introductions occurred during the past 300 years, with most of these events pre-dating the antibiotic era. Hence, the success of L2–Beijing in Africa is most likely due to its hypervirulence and high transmissibility rather than drug resistance.

6.2. Introduction

Tuberculosis (TB) is mainly caused by a group of closely related bacteria referred to as the *Mycobacterium tuberculosis* Complex (MTBC). The MTBC comprises seven phylogenetic lineages adapted to humans and several lineages adapted to different wild and domestic animal species (Gagneux, 2018). The human-adapted lineages of the MTBC show a distinct geographic distribution, with some "generalist" lineages such as Lineage (L)2 and L4 occurring all around the world and others being geographically restricted "specialist" that include L5, L6, and L7 (Coscolla *et al.*, 2014; Stucki *et al.*, 2016). Africa is the only continent which is home to all seven human-adapted lineages, including the three "specialist" lineages exclusively found on the continent. Current evidence suggests that the MTBC overall originated in Africa (Gagneux, 2018) and subsequently spread around the globe following human migratory events (Wirth *et al.*, 2008; Comas *et al.*, 2013). The broad distribution of some of the "generalist" lineages and their presence in Africa has been attributed to past exploration, trade, and conquest. For instance, an important part of the TB epidemics in sub-Saharan Africa is driven by the generalist Latin–American–Mediterranean (LAM) sublineage of L4, which is postulated to have been introduced to the continent post-European contact (Stucki *et al.*, 2016; Brynildsrud *et al.*, 2018).

Among the different human-adapted MTBC lineages, the L2–Beijing sublineage has been of particular interest (Merker *et al.*, 2015). L2–Beijing has expanded and emerged worldwide from East Asia; its most likely geographical origin (Luo *et al.*, 2015; Merker *et al.*, 2015). In some parts of the world, the recent emergence of L2–Beijing has been linked to increased transmission (Yang *et al.*, 2012; Holt *et al.*, 2018) high prevalence of multidrug–resistant TB (MDR–TB) (Borrell *et al.*, 2009), and to social and political instability, resulting into displacement of people and poor health systems (Eldholm *et al.*, 2016). Increasingly, L2–Beijing is also being reported in Africa (Affolabi *et al.*, 2009; Gehre *et al.*, 2016; Mbugi *et al.*, 2016; Bifani *et al.*, 2002), and evidence suggests that L2–Beijing in African regions is becoming more prevalent over time (Cowley *et al.*, 2008; Spuy *et al.*, 2009; Glynn *et al.*, 2010; Bifani *et al.*, 2002). Some authors have hypothesized that the introduction of L2–Beijing into South Africa resulted from the importation of slaves from Southeast Asia during the 17th and 18th centuries and/or the Chinese labor forces arriving in the 1900s (Helden *et al.*, 2002). Alternatively, in West Africa, the presence of L2–Beijing was proposed to reflect more recent immigration from Asia (Affolabi *et al.*, 2009; Gehre *et al.*, 2016). To a certain extent, the recent expansion of L2–Beijing in parts of Africa has been associated with drug resistance (Githui *et al.*, 2004; Klopper

et al., 2013) and higher transmissibility (Guerra-Assunção *et al.*, 2015a). In addition, a study in the Gambia showed a faster progression from latent infection to active TB disease in patient house-hold contacts exposed to L2–Beijing (Jong *et al.*, 2008).

Whilst L2–Beijing seems to be expanding in several regions of Africa, no study has formally investigated the evolutionary history of L2–Beijing on the continent. In this study, we used whole genome sequencing data from a global collection of L2 clinical strains to determine the most likely geographical origin of L2–Beijing in Africa and its spread across the continent.

6.3. Materials and Methods

6.3.1. Identification of Lineage 2 Strains and Whole-Genome Sequencing

We obtained whole-genome sequencing data of L2 strains from the two previously largest studies focusing on the evolutionary history and global spread of L2–Beijing strains (Luo *et al.*, 2015; Merker *et al.*, 2015). We then identified additional published genomes as African and non-African representatives of L2– Beijing strains from other studies (Comas *et al.*, 2010; Comas *et al.*, 2013; Casali *et al.*, 2012; Casali *et al.*, 2014; Kato-Maeda *et al.*, 2012; Zhang *et al.*, 2013; Portevin *et al.*, 2014; Guerra-Assunção *et al.*, 2015a; Koch *et al.*, 2017; Manson *et al.*, 2017). Moreover, we newly sequenced 116 additional L2–Beijing strains using Illumina HiSeq 2000/2500 paired end technology PRJNA488343. In total, we included 781 L2 genome sequences (Figure B.1 and Supplementary Table 1).

6.3.2. Whole Genome Sequence Analysis and Phylogenetic Inference

We used a customized pipeline previously described to map short sequencing reads with BWA 0.6.2 to a reconstructed hypothetical MTBC ancestor used as reference (Comas *et al.*, 2013). SAMtools 0.1.19 was used to call single nucleotide polymorphisms (SNPs), and these SNPs were annotated using ANNOVAR and customized scripts based on the *M. tuberculosis* H37Rv reference annotation (AL123456.2). For downstream analyses, we excluded SNPs in repetitive regions, those annotated in problematic regions such as "PE/PPE/PGRS" and SNPs in drug-resistance associated genes. Small insertions and deletions were also excluded from the analyses. Only SNPs with minimum coverage of

20x and minimum mapping quality of 30 were kept. All SNPs classified by Samtools as having frequencies of the major non-reference allele lower than 100%(AF1 < 1) within each genome were considered to be heterogeneous and were treated as ambiguities and excluded, and were otherwise considered fixed (AF1 = 1). Mixed infections or contaminations were discarded by excluding genomes with more than 1,000 heterogeneous positions and genomes for which the number of heterogeneous SNPs was higher than the number of fixed SNPs. In addition we excluded genomes for which the number of fixed SNPs would fall below Q1–1.5 IQR of all fixed SNPs considering all L2 genomes (Q1 being the first quantile and IQR the inner quantile range as calculated in R 3.5.0). All genomes were typed for lineage and sub-lineage using SNP markers as defined in (Steiner *et al.*, 2016) plus (Coll *et al.*, 2014) and those showing simultaneously more than one marker were excluded. We concatenated fixed SNPs from the variable positions obtained, which yielded a 32,269 bp alignment. The alignment was then used to infer a maximum likelihood phylogeny using RAxML 8.3.2 with a general time reversible (GTR) model in RAxML and 1,000 rapid bootstrap inferences, followed by a thorough maximum-likelihood search (Stamatakis, 2006). The topology was rooted using the reference strain, H37Rv which belongs to Lineage 4.

6.3.3. Phylogeographic Analyses

Reconstruction of the Ancestral Geographic Range

To investigate the likely geographic origin of L2–Beijing strains in Africa, we inferred the historical biogeography of L2 using the RASP software (Yu *et al.*, 2015) on a representative subset of 422 genomes due to software's sample limitation.We achieved this by randomly removing clustered genomes (i.e., those with 12 or less SNP distances) from the same country of origin, using hierarchical clustering implemented in *pvclust* package in R (Suzuki *et al.*, 2006) on a distance matrix of the 781 genome sequences. We then applied a Bayesian binary based method (BBM) in RASP to reconstruct geographical states at the ancestral nodes on the best-scoring ML tree inferred with RAxML using the 422 L2 genomes. We used geographical regions (according to United Nations geoscheme) as proxy for origins of the L2 strains. In total 14 regions were loaded as geographic distributions (indicated in Supplementary Table 1). The ancestral reconstruction was performed with the Proto-Beijing clade as outgroup. We finally ran Bayesian analysis with 10 chains and 50,000 generations.

Stochastic Character Mapping

To determine the number of introduction events of L2–Beijing into African regions, we applied stochastic character mapping as implemented in SIMMAP (Bollback, 2006) on the 781 L2 phylogeny inferred from the best-scoring ML tree rooted on the Proto-Beijing clade, using the *make.simmap* function in phytools package 0.6.60 in R 3.5.0 (Revell, 2012; R Core Team, 2018). Geographical origin of the L2 strains was treated as a discrete trait and modeled onto the phylogeny using ARD model with 100 replicates. This model allows unequal rates of state transition permiting independent region-to-region transfers.We summarized the results of the 100 replicates using the function *summary* in phytools package 0.6.60 in R (Revell, 2012). We referred to the resulting introductions as migration events "M," and discuss only those introductions with 5 or more genomes.

Population Genetic Analyses

Nucleotide diversity (pi)

We calculated the mean pair-wise nucleotide diversity per site (Pi) measured by geographic region. We excluded geographic regions represented by < 20 genomes. Confidence intervals were obtained by bootstrapping through resampling using the sample function in R with replacement and the respective lower and upper confidence levels by calculating 2.5th and 97.5th quartiles. Resampling was additionally done using the smallest size of the geographical regions to account for the effect of different sample sizes.

Pairwise SNP distances

We used *dist.dna* function of ape package implemented in R (Paradis *et al.*, 2004) to calculate pairwise SNP distances with raw mutation counts and pairwise deletions for gaps. Mean pairwise SNP distance to all strains of the same geographic population was calculated per strain and the distribution of the mean SNP pairwise distance for all strains plotted. The mean pairwise SNP distances were assumed not to be normally distributed and we therefore used Wilcoxon rank-sum test to test the differences among geographic regions. Additionally, we calculated pairwise SNP distances within African L2– Beijing populations for migration events with more than 10 genomes each.

Drug Resistance

To distinguish between drug-susceptible and drug-resistant strains, we used genotypic drug resistance molecular markers previously described (Steiner *et al.*, 2014). We categorized strains into: susceptible as having no drug resistance specific mutations; monoresis-

tant as having mutations conferring resistance to a single drug; MDR as having mutations conferring resistance to isoniazid and rifampicin; and extensively drug-resistant (XDR) as having mutations conferring resistance to fluoroquinolones and aminoglycosides in addition to being MDR (Table B.1).

Bayesian Molecular Dating

Data preparation and preliminary analysis

To estimate the historical period in which L2–Beijing was introduced to Africa, we performed a set of Bayesian phylogenetic analyses using tip-calibration (Rieux *et al.*, 2017). Among the 781 studied L2 strains, we had information on the year of sampling for 308. We performed all further analysis on this subset of 308 strains. We excluded all genomic positions that were invariable in this subset and all positions that were undetermined (missing data or deletions) in more than 25% of the strains, and obtained an alignment of 10,769 polymorphic positions. In tip dating analysis it is important to test whether the dataset contains strong enough temporal signal (Rieux *et al.*, 2016). To do this, we performed a tip randomization test (Ramsden *et al.*, 2008) as follows. We used BEAST2 v. 2.4.8 (Bouckaert *et al.*, 2014) to run a phylogenetic analysis with a HKY + GAMMA model (Hasegawa *et al.*, 1985), a constant population size prior on the tree and a strict molecular clock. Additionally, we used the years in which the strains were sampled to time-calibrate the tree, and we modified the extensiblemarkup language (xml) file to specify the number of invariant sites as indicated by the developers of BEAST2 here: `https://groups.google.com/forum/#!topic/beast-users/QfBHMOqImFE` (strict_preliminary.xml). We ran three independent runs (245 million generations in total), and we used Tracer 1.7 (Rambaut *et al.*, 2018) to identify the burn-in (8 million generations), to assess that the different runs converged, and to estimate the effective sample size (ESS) for all parameters, the posterior and the likelihood (ESS >110 for all parameters). We then used TipDatingBeast (Rieux *et al.*, 2017) to generate 20 replicates of the xml file in which the sampling dates were randomly reassigned to different strains. For each replication, we ran the same BEAST2 analysis as for the original (observed) dataset (one run per replicate, 50 million generations, 10% burn-in). We used TipDatingBeast to parse the log files output of BEAST2 and compare the clock rate estimates for the observed data and the randomized replications. The estimates of the molecular clock rate did not overlap between the observed and the randomized dataset, indicating that there is a clear temporal signal and that we could proceed with further analysis (Figure B.2).

Model selection

To identify the clock model that best fits the data, we estimated the marginal likelihood of three different clock models: UCED and UCLD (Drummond *et al.*, 2006), assuming a coalescent constant population size tree prior and the HKY model of nucleotide substitution. We used the Model selection package of BEAST2 to run a path sampling analysis (Lartillot *et al.*, 2006) following the recommendations of the BEAST2 developers (`http://www.beast2.org/path-sampling/`). We used the following settings: 100 steps, 4 million generations per step, alpha = 0.3, pre-burn-in=1million generations, burn-in for each step=40% (*PS.xml). For these analyses, we used proper priors as suggested by (Baele *et al.*, 2012).

UCLD analysis

Since the model selection analysis indicated that the UCLD clock was the best fitting model, we repeated the analysis using the UCLD and the same settings used in the path sampling analysis, sampling every 10,000 generations. We ran three independent runs (800 million generations in total), we used Tracer 1.7 (Rambaut *et al.*, 2018) to identify the burn-in (10 million generations), to assess that the different runs converged and to estimate the effective sample size (ESS) for all parameters, the posterior and the likelihood (ESS >260 for all parameters) (UCLD_final.xml and Supplementary Table 3).

We checked the sensitivity to the priors by running one analysis of 250 million generation sampling from the prior, and compared the parameter estimates with the analysis using the data. We observed the posterior distribution and the prior distribution of all parameters are very distinct (Supplementary Table 4), indicating that the parameter estimates are influenced by the data and not by the priors (Bromham *et al.*, 2018). Additionally we repeated the dating analysis with a coalescent exponential population growth tree prior (UCLD + HKY + exponential growth) and with a GTR model of nucleotide substitution (UCLD + GTR + constant size). All these analyses resulted in similar estimates of the age of the introductions of L2 in Africa, thus showing that our results are not strongly influenced by the tree prior and the nucleotide substitution model (Table B.3). We repeated the tip randomization test with the UCLD model as described above (20 replicates, one run per replicate, 105 million generations per replicate or more, burn-in 10%), and again we found a temporal signal (Figure B.3).

To summarize the results, we sampled the trees from the three runs (5% burn-in corresponding to 10 million generations or more, sampling every 25,000 generation). We then summarized the 31,758 sampled trees, created a maximum clade credibility tree using the software TreeAnnotator fromthe BEAST2 package and used FigTree version 1.4.2

(http://tree.bio.ed.ac.uk/software/ figtree) for visualization (Figure B.4).

6.4. Results

6.4.1. Phylogenetic Inference of L2 Strains

We analyzed a total of 781 L2 genomes originating from 14 geographical regions including Eastern and Southern Africa (Figure B.1 and Supplementary Table 1). We focused on seven geographical regions that had more than 20 genomes each, and assigned the remainder to "Other," including two genomes from Western Africa (Figure 6.1 A). The resulting phylogeny of L2 was divided into two main sublineages: the L2–proto-Beijing and L2– Beijing, supporting previous results (Luo *et al.*, 2015; Shitikov *et al.*, 2017). The L2–proto-Beijing was the most basal L2 sublineage and was restricted to East- and Southeast Asia. L2–Beijing, particularly the "modern" (also known as "typical") sublineage, was geographically widely distributed and included strains from Africa. We further characterized L2–Beijing using the recently described unified classification scheme for L2 (Shitikov *et al.*, 2017).

6.4.2. The Population Structure of L2–Beijing in Eastern and Southern Africa

Our findings showed the population of African L2–Beijing to be heterogeneous (Figures 6.1 B, 6.2 and Table B.2). Most of the African L2–Beijing strains were classified into several groups within the "modern" sublineage, which included primarily the "Asian- African" sublineages (L2.2.4, L2.2.5 and L2.2.7), consistent with previous findings (Merker *et al.*, 2015). We also identified the "ancient" (atypical) strains among the African L2–Beijing. Given that "ancient" L2–Beijing strains (L2.2.1—L2.2.3) are generally uncommon (Luo *et al.*, 2015), it is interesting to observe such strains in both African regions. In several instances, African L2– Beijing strains did not fall into any of the previously defined groups (Figure 6.2). Of the two African regions studied here, East Africa had higher proportion of previously uncharacterized L2–Beijing strains (43/92, 46.7%). In summary, our findings show that African regions harbored distinct L2–Beijing populations. This is unlike Eastern Europe and Central Asia, where L2–Beijing is dominated by a few highly similar strains (Casali *et al.*, 2014; Eldholm *et al.*, 2016). Of note, L2–Beijing strains typical Eastern Europe and Central Asia were completely absent from the African populations (Figure 6.2).

6.4.3. Genetic Diversity of L2–Beijing Strains Across Geographic Regions

The spatial distribution of L2–Beijing sublineages and the prevalence of "ancient" L2–Beijing strains observed in this study and previously (Luo *et al.*, 2015; Merker *et al.*, 2015), suggest that L2–Beijing has expanded worldwide from Asia. This view can further be supported by the measures of genetic diversity of L2–Beijing in the different geographical regions (Figure 6.3). As expected, East- and Southeast Asia contained the most genetically diverse L2 populations, which is consistent with previous results (Luo *et al.*, 2015). Conversely, L2 populations in other geographies were less genetically diverse, suggesting recent dissemination of L2 to these regions. Within Africa, Southern Africa showed a higher diversity in L2–Beijing populations compared to Eastern Africa. The genetic diversity within the African L2–Beijing populations not only reflects the number and variety of source populations but also local patterns of diversification that occurred after their introduction. Therefore, the higher genetic diversity of the L2–Beijing populations in Southern Africa compared to Eastern Africa likely reflects both aspects.

6.4.4. Multiple Introductions of L2–Beijing From Asia Into Africa

Based on our reconstructed phylogeny, African L2–Beijing strains clustered into several unrelated clades indicating multiple introductions into Africa (Figure 6.1 B). We next investigated the most likely geographical origins of those introductions. As anticipated, our ancestral reconstruction using RASP estimated East Asia as the most likely origin of all L2 (posterior probability of 96.1%) and L2–Beijing (posterior probability 92.5%) (Figure B.5). Our data further indicate that L2–Beijing was introduced into Africa from East- and Southeast Asia on multiple occasions independently. Furthermore, we observed both direct introductions from Asia into Africa as well as subsequent dispersal within the continent (Figures B.6, B.7). Using stochastic mapping, we estimated a total of 13 introductions or migration events (M1–M13) into Africa (Figure 6.4). Eight of the African L2–Beijing introductions originated from East Asia and five from Southeast Asia. Out of the 13 introductions, three (M3, M10, and M13) were present in both African regions analyzed here, suggesting initial introductions from Asia followed by subsequent spread within Africa. Overall, our analysis inferred more independent introductions into Southern Africa (n = 7, M1, M4, M7-9, M11 and M12, all of them with extant strains sampled in South Africa) than Eastern Africa (n = 3, M2 sampled in Malawi and Tanzania; M5 and

M6 sampled in Kenya and in Malawi, respectively). Taken together, our data suggest that multiple migration events have shaped the populations of L2–Beijing in Africa.

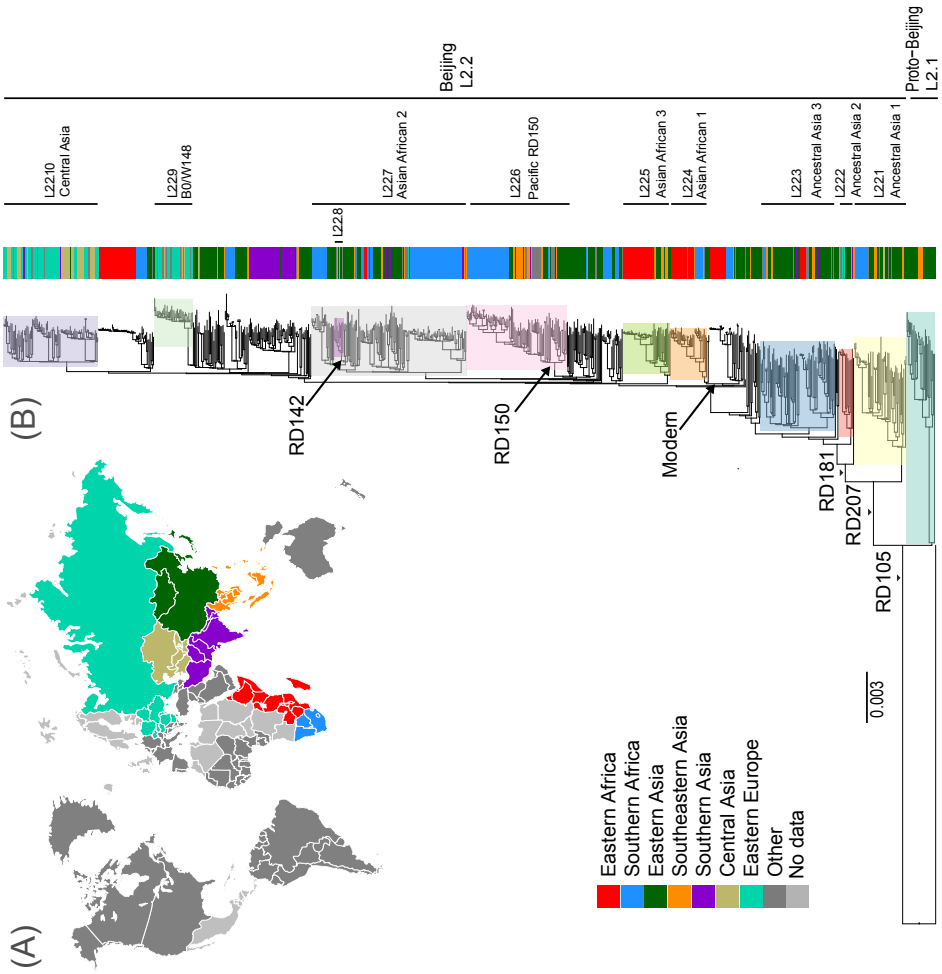


Figure 6.1.: Global phylogeny and geographical distribution of L2 strains. (A) Geographical origin (according to United Nations geoscheme) for the 781 L2 strains. The geographical origins with less than 20 strains are colored dark gray and those with missing data light gray. (B) Maximum likelihood phylogeny inferred from 32,269 variable single nucleotide positions of the 781 strains rooted with the reference strain H37Rv. Taxa are colored according to the geographical origin of the strains and the clades are highlighted according to previously defined sublineages. L2 defining markers i.e., deletions (RD) are also mapped onto the phylogeny.

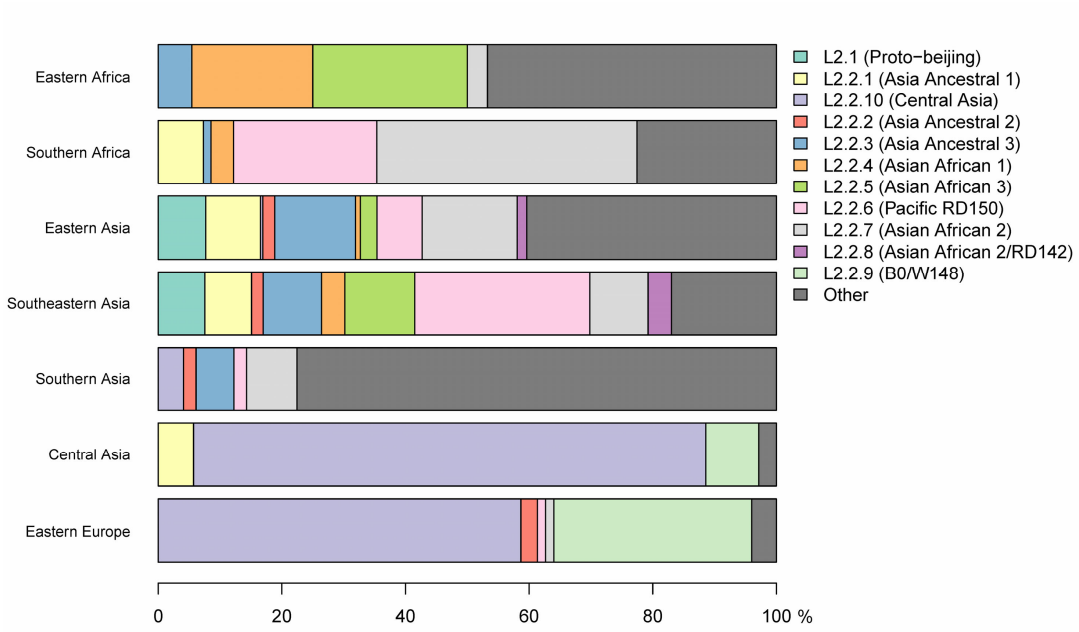


Figure 6.2.: Frequency in proportions of Lineage 2 sub-lineages across seven geographical regions.

6.4.5. Bayesian Molecular Dating

Different hypotheses have been formulated on the possible timing of the introduction of L2–Beijing into Africa (Helden *et al.*, 2002). Here we used tip-calibration to date the phylogenetic tree of L2 and estimate the age of its introduction to Africa. For these analyses, we identified 308 strains among the 781 for which the sampling year was known. These strains were sampled during a period of 19 years; 1995–2014 (Figure B.8), were evenly distributed on the complete phylogenetic tree (Figure B.9) and included 40% members of the African L2–Beijing strains (Figure B.10). Eleven of the 13 African introductions were represented in this dataset (M1-M3 and M6-M13).

We performed a Date Randomization Test with a strict clock and with a relaxed clock. With both models we detected no overlap in the 95% credibility interval of the clock rate estimates of observed and randomized datasets indicating that there was sufficient temporal signal in the dataset to perform inference (see methods, Figures B.2, B.3). Further, we found that the UCLD clock had the highest marginal likelihood and a Bayes Factor of 27 with the second best fitting model, the strict clock (Table 6.1), indicating strong evidence in favor of the UCLD clock (Kass *et al.*, 1995).

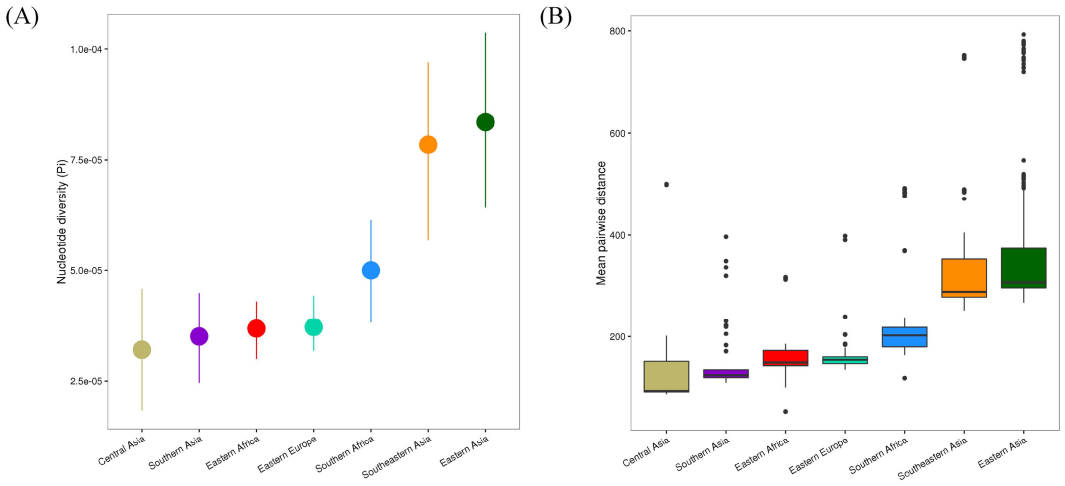


Figure 6.3.: Genetic diversity of L2 strains within geographical regions. (A) Nucleotide diversity (p) per site of L2 strains by geography. Error bars are the 95% confidence intervals obtained by bootstrapping. (B) Pairwise genetic SNP distance of L2 by geography (p-values were obtained from Wilcoxon rank-sum tests). Each box represents the 25% and 75% quartiles and the line denotes the median.

We performed a phylogenetic analysis with BEAST2 using the UCLD clock. Under the UCLD model, the coefficient of variation (COV), which is a summary of the branch rates distribution (standard deviation divided by the mean), gives an indication on the clock-likeness of the data (Drummond *et al.*, 2006). A coefficient of variation of zero indicates that the data fit a strict clock, whilst a greater COV indicates a higher heterogeneity of rates through the phylogeny. We obtained a mean COV of 0.22 (95% credibility interval= 0.1732, 0.2732), indicating a moderate level of rate variation across different branches and thus supporting the results of the path sampling analysis that favored the UCLD model.

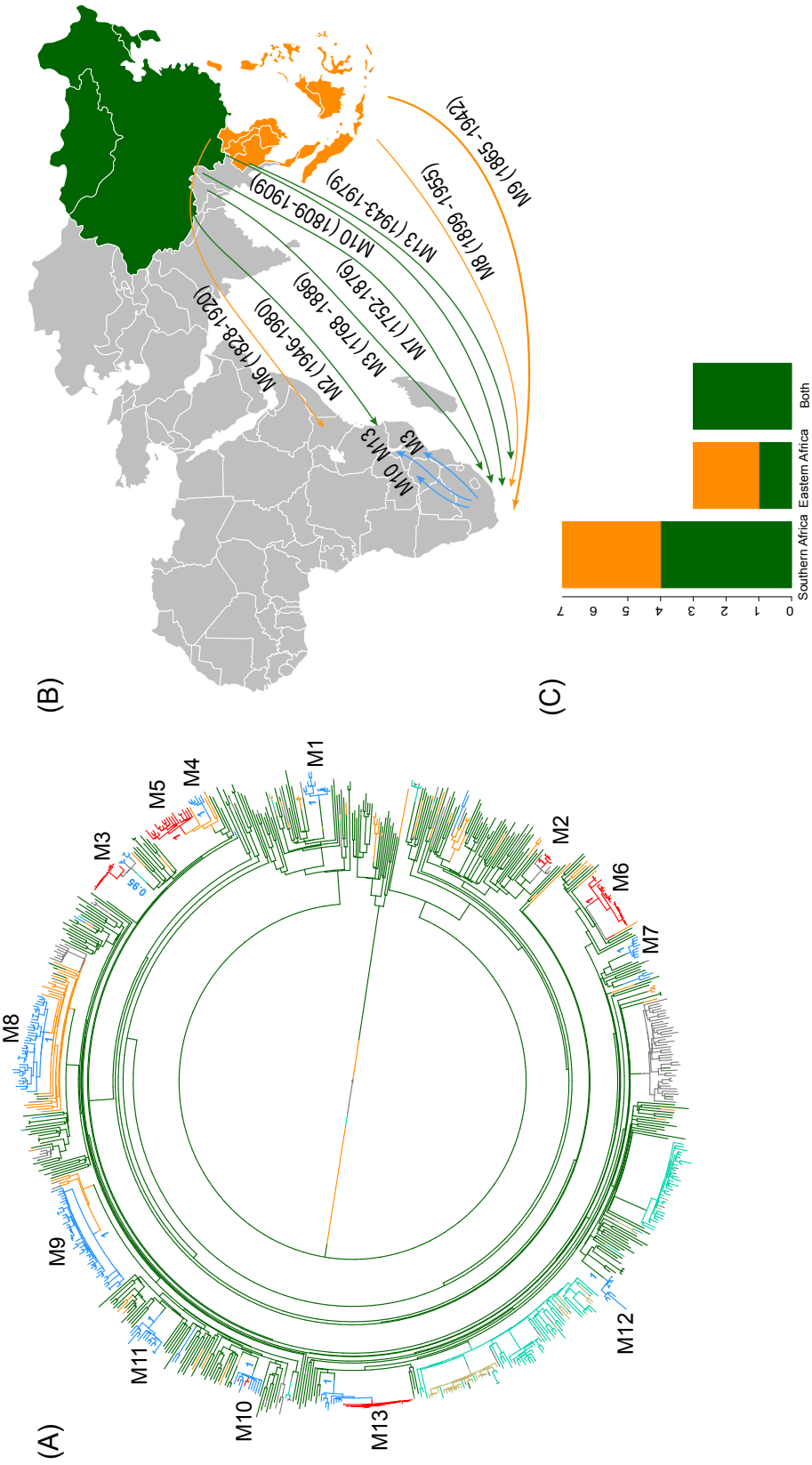


Figure 6.4.: Introductions of L2 strains to Africa. (A) Stochastic mapping of the geographic origin of L2 strain obtained with SIMMAP onto the Maximum Likelihood phylogeny of the 781 MTB Lineage 2 strains. Branches are colored according to the geographical region inferred and one random tree is represented. The 13 migration events to Africa (M1–M13) are indicated and summary posterior probabilities (from 100 runs) of the ancestral states (colored according to their geographical region) are indicated at the nodes defining each M event (B) Proposed scenario for representative events of the multiple introductions of L2–Beijing into Africa. (C) Plot summarizing the number of introduction events to Eastern and Southern Africa from East- and Southeast Asia. Color codes are the same as in Figure 6.3

Table 6.1.: Model selection based on path sampling Log-Marginal Likelihood.

Clock Model	Log(e) Marginal Likelihood	Bayes factor (UCLD vs. model)
UCLD	–5374827	–
Strict	–5374854	27
UCED	–5374897	70

6.4.6. Recent Origins of the African L2–Beijing Clades

We used the UCLD clock model to infer the clock rate and divergent times of the 308 L2 strains with known sampling dates and estimated a mean substitution rate of 1.34 × 107 [95%Highest Posterior Density (HPD), 9.2867×1081.7719× 107]. These estimates are in agreement with previously reported rates from epidemiological studies (Walker *et al.*, 2013; Eldholm *et al.*, 2016). However, the estimated rate by (Eldholm *et al.*, 2016) is relatively higher compared to our estimates, which likely reflects the fact that our dataset included the entire Lineage 2 as opposed to a only a single particular clade in the case of (Eldholm *et al.*, 2016).

We estimated the most recent common ancestor (MRCA) of the extant L2–Beijing of the 308 strains to the year 1225 [95% HPD, 9001519] (Figure B.4). This estimate was slightly younger than the previous estimate for the whole of Lineage 2 by the study of Bos *et al.*, 2014, likely reflecting the difference in methodology. The latter study used ancient *M. pinipeddii* DNA recovered from pre-Columbian human remains to calibrate the phylogeny as opposed to tip-dating based on contemporary sampling dates. For each African clade, we estimated the year of introduction using the 0.975 quantile of the HPD of the age of the MRCA as the upper limit (most recent possible year) and the 0.025 quantile of the HPD of the divergence time between the closest non-African L2–Beijing strain (the closest outgroup) and the African clade of interest as lowest limit (most ancient possible year). This approach produced conservative estimates, while relying only on the age of the MRCA of the African clades would systematically underestimate the age of the introductions. Our estimates placed the earliest introductions of the African L2–Beijing (M1, M3, M7, and M12) in the eighteenth and nineteenth century (Figure 6.5) and Table A.6). Four additional migration events (M6, M9, M10, and M11) were estimated to have occurred between the beginning of the nineteenth century and the first half of the twentieth century. Finally, the three most recent introductions to Africa happened in the second half of the twentieth century (M2, M8, and M13). Diversity patterns of the African clades exclusive to Eastern and Southern Africa could further provide support for the recent introductions of African L2–Beijing. We thus calculated the pairwise SNP

distances within the individual introductions to explore the local patterns of diversification associated with regional epidemics after the introductions. Although strains within Southern African introductions were relatively more distantly related, L2–Beijing strains from both African regions were on average 20 to 40 SNPs apart (Figures B.11, B.12, B.13). The latter thresholds were proposed to correspond to strains involved in transmission clusters of estimated 50 to 100 years (Meehan et al., 2018), supporting the relatively recent introductions of L2–Beijing into the African continent.

Overall, these results indicate that the different introductions of L2–Beijing to Africa occurred over a period of 300 years.While the earliest introduction is unlikely to have happened after 1732– 1874, the most recent is unlikely to have occurred before 1946–1980. However, our 95% HPDs show a wide range of uncertainty, which likely resulted from extrapolating several centuries back in time based on a comparably short calibration interval spanning only 19 years.

6.4.7. Introductions of L2–Beijing Into Africa Unrelated to Drug Resistance

Because of the repeated association of L2–Beijing with antibiotic resistance (Borrell and Gagneux, 2009), the emergence and dissemination of L2–Beijing strains has often been attributed to drug resistance. However, our estimated timing of these introductions suggest that African L2–Beijing strains were introduced prior the discovery of TB antibiotics, and thus must have involved drug-susceptible strains (Figure 6.5). To explore this question further, we assessed the drug resistance profiles of L2–Beijing strains linked to the various introduction events into the two African regions. We found that all the Eastern African populations contained only drug-susceptible strains and that approximately three-quarters of L2–Beijing strains in the Southern African populations were drug-susceptible, with the remaining being either mono–, multi–, or extensively drugresistant (Figure 6.6 and Figure B.14). Taken together, these results suggest that the emergence of L2–Beijing in Africa, particularly in Eastern Africa, was not driven by drug resistance. Moreover, our data indicate independent acquisition of drug resistance for the resistant strains detected in the Southern African L2– Beijing population (Figure 6.6), which might partly contribute to the subsequent spread of L2–Beijing in Southern Africa but not in Eastern Africa.

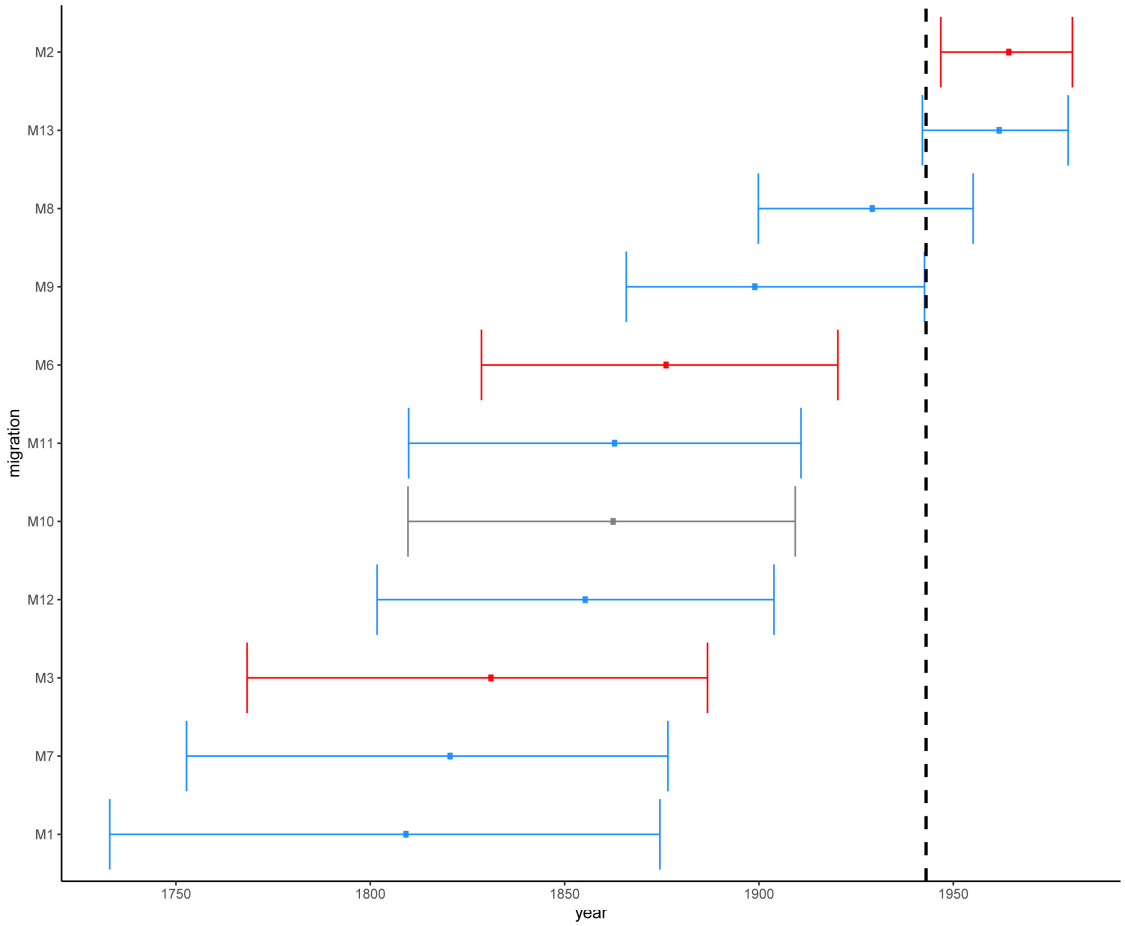


Figure 6.5.: Estimated time in median ages for the introductions of African L2–Beijing (M1-M3 and M6-M13). Introductions to Eastern Africa are colored in red and those to Southern Africa in blue. Migration M10 contained L2–Beijing from both Southern and Eastern Africa. Dotted line marks the year of first anti-TB drug discovery (1943). The error bars correspond to the 95% HPD.

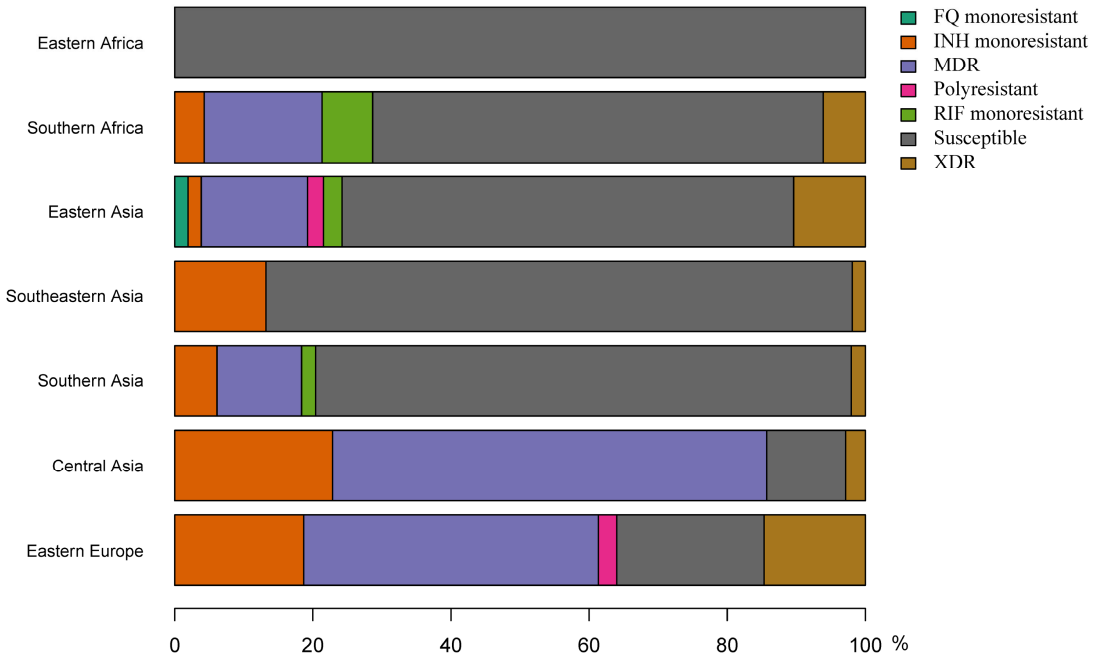


Figure 6.6.: Proportions of drug resistance profiles for L2 strains in seven geographical regions.

6.5. Discussion

This study investigated the most likely geographical origin of the L2–Beijing in Africa. In line with previous findings (Luo et al., 2015; Merker et al., 2015), we identified East Asia as the most likely place of origin of L2 and L2–Beijing. Our findings further revealed multiple independent introductions of L2–Beijing into Africa linked to separate populations originating from both East and Southeast Asia. Some of these introductions were followed by further onward spread of L2–Beijing within African regions. Finally, we demonstrate that most introductions of L2–Beijing on the continent occurred before the age of antibiotics.

L2–Beijing has received much attention given its hypervirulence in infection models (Manca et al., 2001; Ribeiro et al., 2014), faster progression to disease and higher transmission potential in humans (de Jong et al., 2008; Holt et al., 2018), frequent association with drug resistance, and recent emergence in different regions of the world (Bifani et al., 2002; Borrell and Gagneux, 2009; Fenner et al., 2013). Several studies indicate L2–Beijing originated in Asia and spread from there to the rest of the world (Luo et al., 2015; Merker

et al., 2015). Our results support this notion by identifying "Asia" as themost likely geographical origin of both L2 and L2–Beijing based on our ancestral reconstructions and the fact the L2–Beijing populations in Asia are much more diverse than in other regions. In addition, our findings show that L2–Beijing was introduced into Africa multiple times from both East- and Southeast Asia. The presence of L2–Beijing in South Africa has previously been proposed to be due to the importation of slaves from Southeastern Asia by Europeans in the seventeenth and eighteenth centuries followed by the import of Chinese labor-forces in the early 1900s (Helden *et al.*, 2002)(van Helden et al., 2002; Mokrousov et al., 2005). Our Bayesian dating estimates predicted the earliest introductions of L2–Beijing into Africa to have occurred in the eighteenth and nineteenth centuries, concurring with these proposed time periods. However, our findings also point to later introductions of L2–Beijing into the continent in the nineteenth and early twentieth centuries. The timings of the latest three introductions in the second half of the twentieth century coincide with the decolonization and post-colonial period in Africa when investments into infrastructure and other projects by Chinese enterprises substantially increased (Yuan, 2006; Rice, 2011). These activities also brought many Chinese workers to Africa during a time when TB was still very prevalent in China (Murray, 2018). Hence, many of these workers were likely latently infected with L2–Beijing and might have later reactivated (Pescarini et al., 2017). Overall, our findings suggest that L2–Beijing has emerged in Africa over the last 300 years and that these introductions have occurred sporadically ever since. The repeated association of L2–Beijing with drug resistance (Borrell and Gagneux, 2009) has led some to propose that drug resistance is another reason why this sublineage might successfully compete against and eventually replace other *M. tuberculosis* genotypes (Parwati et al., 2010). However, the underlying reason for the association of L2–Beijing with drug resistance remains unclear (Borrell and Trauner, 2017), and it is also far from universal, with several reports from e.g., China and other regions finding no such association (Hanekom et al., 2007; Yang et al., 2012). Our results show that most introduction events of L2–Beijing into Africa pre-date the antibiotic era, and because of that, these introductions were most likely caused by drug-susceptible strains. The notion that the initial emergence of L2–Beijing in Africa was not driven by drug resistance is further supported by our findings that none of L2–Beijing strains from Eastern Africa strains analyzed here were drug-resistant. Of note, our observations suggest that drug resistance in South Africa was acquired via independent events post initial introductions from Asia. This is in sharp contrast to the situation in Eastern Europe and Central Asia, where L2–Beijing is highly prevalent but dominated by few recently expanded drug-resistant clones, which account for up to 60% of the L2–Beijing populations in some of these countries (Casali *et al.*, 2014;

Eldholm *et al.*, 2016). The association of L2–Beijing with drug resistance in these regions were likely favored by the economic and public health crises that followed the collapse of Soviet Union (Luo et al., 2015; Merker et al., 2015).

Based on our finding that the original introductions of L2– Beijing into Africa involved drug-susceptible strains and that the prevalence of drug-resistant L2–Beijing in Africa overall is comparably low (WHO, 2017), we propose that some of the other characteristics of this sub-lineage, in particular its high virulence, high transmissibility and rapid progression from infection to disease, were responsible for the initial competitive success of L2–Beijing in Africa. Until recently, Africa was considered a Virgin Soil for TB and that TB was only introduced following European contact (Comas et al., 2015). However, this notion is incompatible with the current evidence supporting an African origin for the MTBC overall (Hershberg et al., 2008; Wirth et al., 2008; Comas et al., 2013). Indeed, Africa is the only continent to harbor all seven known human-adapted MTBC lineages, three of which are almost exclusively observed there (Gagneux, 2018), as well as several animal-adapted MTBC ecotypes, which affect several wild African mammals (Brites et al., 2018). Moreover, one study showed that TB epidemics on the continent were caused by many different "native" genotypes prior to foreign contacts (Comas et al., 2015), suggesting that TB existed in Africa before the initial European contact. Perhaps the clinical and epidemiological picture of TB in Africa was different at the time, characterized by a low virulent and slow progressing disease, which would have escaped the attention of European colonial officials reporting the absence of TB in Africa, at a time when the TB epidemic was raging in the cities of Europe and North America, killing up to 20% of the adult population (Comas and Gagneux, 2011). We hypothesized previously that rapid co-expansion of MTBC L2, L3, and L4 with their respective human host populations in China, India and Europe might have selected for higher virulence and shorter latency in these "modern" lineages (Hershberg et al., 2008). The emergence and expansion of these "foreign" genotypes including L2–Beijing into Africa as reported here, and as reported earlier for L4 (Stucki et al., 2016; Brynildsrud et al., 2018), demonstrate the ability of these lineages to successfully compete against the existing genotypes on the continent, likely as a result of their high transmissibility and rapid progression to disease. Following their initial establishment, poor TB treatment programs subsequently selected for drug resistance in L2–Beijing but also in other MTBC lineages including L4, which might have facilitated their further spread in countries such as South Africa (Müller et al., 2013).

Finally, the estimates of the TMRCA for L2 reported here are largely consistent with recent reports using similar tipdating analyses based on isolation dates (Eldholm *et al.*,

2016), and support the notion that the MTBC overall is younger than what has been proposed in earlier studies, based on a hypothesized co-divergence of the human-adapted MTBC and modern humans since theirmigration out of Africa (Comas *et al.*, 2013; Luo *et al.*, 2015).

This study is limited by the fact that we analyzed a globally diverse collection of L2 genomes available in public repositories. Hence, these strains might not be fully representative of the respective geographical regions. Moreover, our African L2– Beijing dataset came from convenient sampling and comprised L2–Beijing mainly from Eastern and Southern Africa, as whole genome data of L2–Beijing from the other African regions were unavailable at the time of the study. However, the representation of African L2–Beijing in our sample reflects the overall prevalence of this sub-lineage as recently reported for the continent (Mbugi *et al.*, 2015; Chihota *et al.*, 2018). Moreover, although regions outside of Eastern- and Southern Africa were underrepresented, this is unlikely to invalidate our findings regarding the multiple independents of L2– Beijing into Africa, except by underestimating the number of true introductions.

In conclusion, this is the first study to address the geographical origins of L2–Beijing in Africa using whole genome sequencing data. Our findings indicate multiple independent introductions of L2–Beijing epidemics into Africa from East- and Southeast Asia during the last 300 years that were unrelated to drug resistance. The TB epidemics in Africa have remained fairly stable over the last few decades (WHO, 2017). However, Africa's population growth and increasing urbanization (driven by booming economies) are likely to have an impact on the future of TB in this continent, whether directly by e.g., facilitating transmission or indirectly by promoting new risk factors such as diabetes that increase TB susceptibility (Dye *et al.*, 2010). It is therefore crucial to follow the TB epidemics in the continent very closely, especially those related to hypervirulent strains such as L2–Beijing, as these might take particular advantage of this expanding ecological niche (Cowley *et al.*, 2008).

6.6. Data Availability

The xml files used for this study can be found here `https://github.com/SwissTPH/TBRU_L2Africa`

6.7. Authors Contributions

LR, DB, FM, DS, LF, and SG planned the study. SL, BM, and JF performed the experiments. LR, DB, FM, SMG, SL, BM, CB, SB, KM, MB, LJ, KR, EJC, LD and LF contributed strains and prepared the data. LR, DB, FM, and SG analyzed the data. LR, DB, FM, and SG drafted the manuscript. All authors critically reviewed the manuscript.

6.8. Acknowledgments

We would like to thank Sebastián Duchêne and Yan Yu for their technical support and Linda-Gail Bekker for contributing strains. All bioinformatics analyses were performed at the scientific computing core facility of the University of Basel, sciCORE (`http://scicore.unibas.ch/`). This work was supported by the Swiss National Science Foundation (grants 310030_166687 to SG), the European Research Council (309540-EVODRTB to SG) and SystemsX.ch This research was also partially supported (strain collection) by a funding supplement from the National Institutes of Allergy and Infectious Diseases (NIAID) under award numbers U01 AI069924 (IeDEA Southern Africa) and U01 AI069911 (IeDEA East Africa).

6.9. Supplementary Material

The Supplementary Material for this article can be found online at: `https://www.frontiersin.org/articles/10.3389/fevo.2019.00112/full#supplementary-material` and on the appendix B of the thesis

7. General Discussion

It was only a few decades ago that Mtb complex was considered a "clone" (Sreevatsan *et al.*, 1997; Musser *et al.*, 2000). Despite the high sequence identity among the Mtb strains, substantial genetic variation exists (Hershberg *et al.*, 2008; Comas *et al.*, 2010). Over the years, we have witnessed the advancements in different molecular markers and genotyping techniques (Embden *et al.*, 1993; Kamerbeek *et al.*, 1997; Hershberg *et al.*, 2008; Supply *et al.*, 2000; Stucki *et al.*, 2012) and lately whole genome sequencing technologies (Comas *et al.*, 2010; Roetzer *et al.*, 2013; Walker *et al.*, 2013). Without a doubt, these developments have given us invaluable insights into the micro- and macroevolution of Mtb, the basis of its disease pathogenesis and fundamentally, the phenotypic consequences of Mtb variation in TB infection and disease.

In this PhD thesis, we first aimed to gain insights into the national and local diversity of Mtb in Tanzania using a combination of SNP-typing and WGS techniques and to assess for the associated clinical phenotypes. Secondly, we aimed to study the evolutionary history of the prevalent Lineages 1 and 3, and the recently emerged Lineage 2-Beijing in Africa. We show that TB epidemic in Tanzania is caused by four of the seven human-adapted Mtb lineages and identified associated clinical phenotypes related to enhanced transmission and unusual higher female-to-male ratio in TB patients. Using the most comprehensive whole genome datasets to date, we were able to reveal that prevalence of Lineage 1 and 3 in Eastern Africa and the recent emergence of Lineage 2 on the African continent reflect multiple introductions back-to-Africa from Asia.

In this chapter, we summarize the key findings of the research conducted in this PhD thesis and the implications of molecular epidemiology and evolutionary history studies of Mtb in high TB endemic settings of sub-Sahara Africa.

7.1. Mtb lineages in Africa

Africa is the only continent where all the seven human-adapted Mtb lineages are found (Gagneux, 2018). Remarkably, the Mtb lineages show a geographical distribution within the continent with the "specialist" Lineage 5 and 6 confined to Western and Lineage 7 to the Horn of Africa. The globally distributed "generalist" Lineage 4 occurs in high frequencies throughout the continent (Stucki *et al.*, 2016) whilst Lineage 2 has only recently emerged and is found at low frequencies in most regions (Affolabi *et al.*, 2009; Gehre *et al.*, 2016; Chihota *et al.*, 2018; Rutaihwa *et al.*, 2019a) except Southern Africa (Chihota *et al.*, 2018; Cowley *et al.*, 2008; Spuy *et al.*, 2009). On the other hand, the "intermediate" Lineage 1 and 3 are prevalent in the Eastern and Northern Africa (Comas *et al.*, 2015; O'Neill *et al.*, 2019; Rutaihwa *et al.*, 2019b) and less frequent elsewhere in the continent (Chihota *et al.*, 2018).

Reconstruction of the evolutionary history of Mtb revealed the African origin prior its dispersal particularly of the evolutionary "ancient" Lineage 1 out of the continent (Comas *et al.*, 2013). Mtb was also brought back-to-Africa as evidenced by the presence of the evolutionary "modern" Lineage 2–4 that emerged outside of Africa (Hershberg *et al.*, 2008; Stucki *et al.*, 2016; O'Neill *et al.*, 2019). Our study on the phylogeography of the "intermediate" lineages (described in *Chapter 5* of this thesis) revealed the re-introduction and introductions of heterogeneous populations of Lineage 1 and Lineage 3, respectively, into Africa from Southern Asia. Moreover, in *Chapter 6* we showed recent introductions of Lineage 2–Beijing into Africa related to heterogeneous bacterial populations from Eastern- and Southeastern Asia. These introductions were followed by subsequent spread within the continent.

The phylogeography of human-adapted Mtb lineages like that of other geographically structured human-pathogens (Falush *et al.*, 2003; Monot *et al.*, 2005) reflects human demographic history and propose local adaptation of the Mtb lineages to particular human populations (Reed *et al.*, 2009), suggesting long–standing host-pathogen relationships (Woolhouse *et al.*, 2002). This is further demonstrated by Mtb lineage–human population combinations, for instance in metropolitan settings where different Mtb and human populations co-exist, sympatric associations are favored over allopatric ones (Hirsh *et al.*, 2004; Gagneux *et al.*, 2006). In Africa, typical sympatric host-pathogen associations are reflected by "specialist" lineages such as Lineage 5 in Ewe ethnicity (Asante-Poku *et al.*, 2015; Asante-Poku *et al.*, 2016) and Lineage 7 in Ethiopia (Firdessa *et al.*, 2013) as well as by "specialist" sublineages such as Lineage 4–Uganda and Lineage 4–Cameroon in Uganda and Cameroon, respectively (Stucki *et al.*, 2016). However, a large proportion of

TB epidemic in Africa is caused by other genotypes of Lineage 4 and 3 variants like we observed in Tanzania (in *Chapter 3* and *4* in this thesis). This could suggest that sympatric/allopatric relationships are not as clear-cut, particularly in the African settings. On the one hand, perhaps the TB/HIV co-infections could disrupt such relationships (Fenner *et al.*, 2013), but the prevalence of HIV-coinfection is only about 30% in TB patients. On the other hand, the frequency of "foreign" genotypes in Africa demonstrates their successful establishment and expansion in the presence of "local" genotypes (Comas *et al.*, 2015). The fact that these genotypes co-exist show their ability to transmit within African populations, some better than others. Indeed, we show from our preliminary molecular clustering analysis in *Chapter 4* that on-going transmission in urban Tanzania is disproportionally contributed by modern lineages, Lineage 2–4. Hence, the current TB epidemic in Africa could be largely shaped by "virulent" strains of modern lineages in combination with factors related to demographic changes such as overcrowding due to increased host density and urbanization across the African continent (Dye *et al.*, 2010).

Prior the two studies conducted at local and national level in Tanzania (*Chapter 3* and *4* in this thesis), no such study in the country had addressed the Mtb diversity to that extent using the complementary SNP-typing and WGS technologies and only a few studies have been conducted elsewhere in Africa (Guerra-Assunção *et al.*, 2015b; Guerra-Assunção *et al.*, 2015a). This calls for the opportunity to employ such tools to investigate the complex interactions of the TB infection and disease determinants in high-endemic settings.

7.2. Molecular epidemiology of Mtb in high TB burden settings

It is now widely accepted that genetic variation in Mtb plays role in determining the fate of TB infection and disease (Coscolla, 2017). This is relevant especially when genetic variation translates into relevant phenotypes. Of particular interest would be Mtb clinical phenotypes related to transmission, taking into account the undesirable magnitude of Mtb transmission in high-endemic TB settings at present (Yates *et al.*, 2016). Understanding transmission is among the research priorities in TB where the focus lies on its interruption in order to reduce TB incidence. Social and environmental aspects contributing to transmission are often evaluated and particularly in the urban and rural settings studied in *Chapter 4*. A few studies have assessed health seeking behavior in order to determine pathways to care taken by TB patients (Said *et al.*, 2017; Sikalengo *et al.*, 2018), so as to minimize delays in diagnosis and treatment which contribute to on-going transmission

by extending the duration of infectiousness (Golub *et al.*, 2006). Another study looked at potential transmission hot-spots using indoor carbon dioxide measures that would suggest ways to minimize Mtb exposure through contaminated air (Hella *et al.*, 2017).

Given the high transmission levels and the co-existence of Mtb lineages, it would appear that transmission occurs independently of Mtb background. Our findings based on the age of TB patients at the national level in Tanzania (*Chapter 3*) would suggest that to be case. Even though findings from the urban setting in Tanzania (*Chapter 4*) suggest otherwise as Lineage 2 was more frequent in patients' of young age, Lineage 2 is found in low frequencies. One obvious reason could be due to the recent introductions of Lineage 2 into the African regions as established in *Chapter 6* of this thesis. Indeed, in another African setting, the Gambia, Mtb lineages showed similar transmission rates but the rates of progression to disease were higher in contacts exposed to Lineage 2–Beijing (Jong *et al.*, 2008). What this portrays is that the faster "progressors" would end up contributing much sooner into the transmission cycle. Yet still other "local" Mtb genotypes persist e.g. Lineage 6 in the Gambia and Lineage 1 in Tanzania. The two lineages constitute the evolutionary "ancient" lineages which emerged in Africa and therefore have co-existed with their respective hosts for much longer creating stable host-pathogen relations. These lineages are hypothesized to have adapted to low human population densities characterized by slow progression to active disease. Certainly, contacts exposed to Lineage 6 were less likely to progress to disease (Jong *et al.*, 2008) and Lineage 1 has been found to be frequent in patients of older ages, which is proxy for reactivation of long term latent infection (Holt *et al.*, 2018).

Knowledge on circulating Mtb genotypes in a clinical setting gives us insights into transmission patterns and helps identify transmission chains. Although our study conducted in urban Tanzania (*Chapter 4*) was not population-based, we could still detect some level of Mtb genetic clustering suggesting on-going transmission, the magnitude of which was however likely underestimated. In a perfect scenario, following the detection of clustered Mtb strains, such strains could be tracked by screening the new incoming TB cases, and better yet, contact tracing in households and communities. WGS provides the opportunity to identify highly transmissible strains in communities in "real-time", which in theory could guide contact tracing approaches and thus interrupt transmission. The reality however is different as discussed below in section 7.5 of this chapter.

Apart from transmission, other epidemiological features correlated with Mtb lineage include disease severity, although when considering parameters such as cavitation and bacterial load, severity can be extrapolated to transmission. We pooled 12 parameters to score

disease severity of TB patients in the urban setting (*Chapter 4*) with the aim to assess for clinical phenotypes related to TB disease severity. While, no evidence was found for such an association, a recent study showed the influence of severity on treatment duration and outcome of TB patients (Imperial *et al.*, 2018). Hence, it remains necessary to investigate severity of disease among patients and to determine the pathogen influence.

7.3. Sex bias in TB and the role of pathogen

TB affects generally more males than females (WHO, 2018; Guerra-Silveira *et al.*, 2013). The higher male-to-female ratio has almost universally been reported in high burden settings with few exceptions like Afghanistan (WHO, 2018). We observed a similar trend in both nationwide and local settings of Tanzania (*Chapter 3* and *4*). One could speculate the observed differences to be influenced by biased notification, diagnosis or health seeking behavior, however evidence shows those factors not to be confounding (Rhines, 2013). The consistent global trends toward male bias in TB have been explained by gender-related social and physiological differences (Nhamoyebonde *et al.*, 2014). Social factors tied to risk behavior such as substance abuse e.g. alcoholism, smoking and drug use and those related to risk occupations e.g. mining are frequent in males thus placing them at a much higher risk for TB. Aside from social determinants, biological differences involving sex hormones and genetic makeup have been implicated to influence sex-specific TB susceptibility (Nhamoyebonde *et al.*, 2014). Epidemiological and experimental studies suggest female sex hormones to be protective, an observation which might explain the apparent male-female bias in adult TB patients. What is striking is that, the sex bias is maintained even though the most important risk factor for TB, HIV, is more frequent in females (Hegdahl *et al.*, 2016). We (in *Chapter 3*) and a few others previously (Holt *et al.*, 2018; Malla *et al.*, 2012) reveal a higher trend in female-male-ratio in TB patients infected with Lineage 2. These observations could suggest a possible pathogen role in Lineage 2 "naïve" populations where female "resistance" to TB is overcome. However, we could not replicate these findings at the local level (*Chapter 4*), which could be due to differential sampling in the national and local settings. In addition, gender factors influenced by health seeking behavior could also contribute to this inconsistence. For instance, a recent study in our setting revealed that men sought more formal care than women (Said *et al.*, 2017). Future studies should further investigate bacterial related factors that influence this bias and taking the potential confounders into account.

7.4. WGS application and challenges

In recent years, high-throughput sequencing technologies have significantly advanced in terms of sequence amount and quality, costs and turnaround time (Loman *et al.*, 2012). These developments have led us to the era of genomic epidemiology including in the field of TB, where we can tackle transmission questions at finer scales compared to conventional molecular typing techniques (Comas, 2017). So far, application of WGS for clinical microbiology practices has been primarily in low-endemic settings, with only a handful of examples existing in high-endemic settings.

Most of genomic epidemiology studies in TB have been conducted retrospectively (Roetzer *et al.*, 2013; Walker *et al.*, 2013; Stucki *et al.*, 2015). In cases where WGS is applied prospectively, the nature of TB outbreaks is unusual, given the complex infection outcome in TB (Koul *et al.*, 2011), where patients in the same transmission links can be identified over a period of many years (Stucki *et al.*, 2015). This fact raises concerns whether WGS is an ideal "real-time" epidemiological tool and whether it would have an impact on TB infection control, which is crucial in high-endemic settings (Comas, 2017). However, with the integration of social links, WGS can disentangle transmission events that are inclined to last over years in a given setting (Gardy *et al.*, 2011). This could ultimately inform control strategies and guide intervention approaches such as targeted contact tracing in order to interrupt transmission. In addition, WGS has been successfully applied in high-endemic setting such that it was possible to assess transmissibility within Mtb lineages and distinguish disease recurrence vs., recent transmissions in population-based approach (Guerra-Assunção *et al.*, 2015b).

Although challenges concerning ease on the generation of sequencing data might be well in-check, including "onsite" in high-endemic settings, analytical and technical challenges remain. For starters, WGS requires high quality DNA as template, which involves culturing of Mtb isolates. While culturing capacity might be feasible in a research setting, it is not the case for health care facilities which in fact attend to TB patients. Even if culturing facilities were widely available, the turnaround time for culture is much longer. One way to circumvent such limitation is to invest on culture-independent methods for instance, sequencing directly from sputum samples. Other challenges involve the rapid and large amounts of data generated by the technology, creating considerable concerns on data management and storage infrastructure, lack of skilled and trained personnel to utilize such data in a meaningful way.

Despite the challenges that come with WGS technologies, it is still worth to discover

its full potential in high-endemic settings. Beyond sequencing of clinical isolates, high-throughput sequencing provide the opportunity to study complex microbial communities related to human health and disease via approaches such as metagenomics and sequence profiling (Qin *et al.*, 2010; Hess *et al.*, 2011). Overall, high-throughput sequencing could become an invaluable epidemiological tool in high-endemic settings as well, provided the willingness to tackle the challenges and limitations that tag along with the technology.

7.5. Conclusions

Sub-Saharan African countries are highly burdened with TB. The recent advances in genotyping techniques and gradually WGS provide the opportunity to study the molecular epidemiology of TB in such settings. From one arm of this thesis, we provided insights into nationwide and local diversity of Mtb lineages in Tanzania using cost-effective SNP-typing complemented with WGS. This allowed us to further investigate epidemiological features associated with Mtb lineages using patients' sociodemographic and clinical information and explored molecular clustering of Mtb strains in urban Tanzania. We revealed four of seven human-adapted lineages showing similar epidemiological features in the rural and urban setting. By contrast, Lineage 2 was more frequent in patients of young age in the urban setting. Our analysis of Mtb genomes isolated from urban TB patients showed a transmission bias towards strains of the modern lineages, Lineage 2–4.

From the other arm of the thesis, we demonstrate the value of using large whole genome datasets to study the phylogeography and evolutionary histories of Mtb lineages. We revisited the global phylogenies and geographical distributions of Lineage 1 and 3 where we showed substantial substructures within these two lineages, which have most likely been partially shaped by back-to-Africa migrations from South Asia. Secondly, we show that Lineage 2–Beijing has only recently emerged in Africa via multiple introductions from Eastern- and Southeastern Asia. Further, the estimated timing indicated most of these Lineage 2–Beijing introductions pre-date the antibiotic era, and therefore suggest its success to be linked to higher virulence and enhanced transmission rather than to drug resistance.

Overall, our findings highlight epidemiological features related to the pathogen in Tanzania, at local and national levels, improving our knowledge on the molecular epidemiology of TB in the country. Finally, our results reveal the origins of the recently emerged Lineage 2–Beijing and the prevalent Lineage 1 and 3 across Africa, which were until now

largely underexplored.

8. Bibliography

Achtman, M (2008) Evolution, population structure, and phylogeography of genetically monomorphic bacterial pathogens. *Annual review of microbiology*, **62** 53–70.

Affolabi, D, Faïhun, F, Sanoussi, N, Anyo, G, Shamputa, IC, Rigouts, L, Kestens, L, Anagonou, S, Portaels, F (2009) Possible outbreak of streptomycin-resistant *Mycobacterium tuberculosis* Beijing in Benin. *Emerging infectious diseases*, **15**(7): 1123–5.

Alexander, KA, Laver, PN, Michel, AL, Williams, M, Helden, PD van, Warren, RM, Gey van Pittius, NC (2010) Novel *Mycobacterium tuberculosis* Complex Pathogen, *M. mungi*. *Emerging Infectious Diseases*, **16**(8): 1296–1299.

Asante-Poku, A, Otchere, ID, Osei-Wusu, S, Sarpong, E, Baddoo, A, Forson, A, Laryea, C, Borrell, S, Bonsu, F, Hattendorf, J, Ahorlu, C, Koram, KA, Gagneux, S, Yeboah-Manu, D (2016) Molecular epidemiology of *Mycobacterium africanum* in Ghana. *BMC infectious diseases*, **16** 385.

Asante-Poku, A, Yeboah-Manu, D, Otchere, ID, Aboagye, SY, Stucki, D, Hattendorf, J, Borrell, S, Feldmann, J, Danso, E, Gagneux, S (2015) *Mycobacterium africanum* is associated with patient ethnicity in Ghana. *PLoS neglected tropical diseases*, **9**(1): e3370.

Baele, G, Li, WLS, Drummond, AJ, Suchard, MA, Lemey, P (2012) Accurate Model Selection of Relaxed Molecular Clocks in Bayesian Phylogenetics. *Molecular Biology and Evolution*, **30**(2): 239–243.

Bhanu, NV, Soolingen, D van, Embden, JDA van, Dar, L, Pandey, RM, Seth, P (2002) Predominace of a novel *Mycobacterium tuberculosis* genotype in the Delhi region of India. *Tuberculosis*, **82**(2-3): 105–12.

Bifani, PJ, Mathema, B, Kurepina, NE, Kreiswirth, BN (2002) Global dissemination of the *Mycobacterium tuberculosis* W-Beijing family strains. *Trends in Microbiology*, **10**(1): 45–52.

Bolger, AM, Lohse, M, Usadel, B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**(15): 2114–20.

Bollback, JP (2006) SIMMAP: stochastic character mapping of discrete traits on phylogenies. *BMC bioinformatics*, **7** 88.

Boonaiam, S, Chaiprasert, A, Prammananan, T, Leechawengwongs, M (2010) Genotypic analysis of genes associated with isoniazid and ethionamide resistance in MDR-TB isolates from Thailand. *Clinical microbiology and infection*, **16**(4): 396–9.

Borgdorff, MW, Soolingen, D van (2013) The re-emergence of tuberculosis: what have we learnt from molecular epidemiology? *Clinical microbiology and infection*, **19**(10): 889–901.

Borrell, S, Gagneux, S (2009) Infectiousness, reproductive fitness and evolution of drug-resistant *Mycobacterium tuberculosis*. *The international journal of tuberculosis and lung disease*, **13**(12): 1456–66.

Bos, KI, Harkins, KM, Herbig, A, Coscolla, M, Weber, N, Comas, I, Forrest, SA, Bryant, JM, Harris, SR, Schuenemann, VJ, Campbell, TJ, Majander, K, Wilbur, AK, Guichon, RA, Wolfe Steadman, DL, Cook, DC, Niemann, S, Behr, MA, Zumarraga, M, Bastida, R, Huson, D, Nieselt, K, Young, D, Parkhill, J, Buikstra, JE, Gagneux, S, Stone, AC, Krause, J (2014) Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. *Nature*, **514**(7523): 494–497.

Bouckaert, R, Heled, J, Kühnert, D, Vaughan, T, Wu, CH, Xie, D, Suchard, MA, Rambaut, A, Drummond, AJ (2014) BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLoS Computational Biology*, **10**(4): e1003537.

Bromham, L, Duchêne, S, Hua, X, Ritchie, AM, Duchêne, DA, Ho, SYW (2018) Bayesian molecular dating: opening up the black box. *Biological Reviews*, **93**(2): 1165–1191.

Brosch, R, Gordon, SV, Marmiesse, M, Brodin, P, Buchrieser, C, Eiglmeier, K, Garnier, T, Gutierrez, C, Hewinson, G, Kremer, K, Parsons, LM, Pym, AS, Samper, S, Soolingen, D van, Cole, ST (2002) A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. *Proceedings of the National Academy of Sciences of the United States of America*, **99**(6): 3684–9.

Brossier, F, Veziris, N, Truffot-Pernot, C, Jarlier, V, Sougakoff, W (2011) Molecular investigation of resistance to the antituberculous drug ethionamide in multidrug-resistant clinical isolates of *Mycobacterium tuberculosis*. *Antimicrobial agents and chemotherapy*, **55**(1): 355–60.

Brudey, K, Driscoll, JR, Rigouts, L, Prodinger, WM, Gori, A, Al-Hajoj, SA, Allix, C, Aristimuño, L, Arora, J, Baumanis, V, Binder, L, Cafrune, P, Cataldi, A, Cheong, S, Diel, R, Ellermeier, C, Evans, JT, Fauville-Dufaux, M, Ferdinand, S, Garcia de Viedma, D, Garzelli, C, Gazzola, L, Gomes, HM, Guttierez, MC, Hawkey, PM, Helden, PD van, Kadival, GV, Kreiswirth, BN, Kremer, K, Kubin, M, Kulkarni, SP, Liens, B, Lillebaek,

T, Ho, ML, Martin, C, Martin, C, Mokrousov, I, Narvskaïa, O, Ngeow, YF, Naumann, L, Niemann, S, Parwati, I, Rahim, Z, Rasolofo-Razanamparany, V, Rasolonavalona, T, Rossetti, ML, Rüsch-Gerdes, S, Sajduda, A, Samper, S, Shemyakin, IG, Singh, UB, Somoskovi, A, Skuce, RA, Soolingen, D van, Streicher, EM, Suffys, PN, Tortoli, E, Tracevska, T, Vincent, V, Victor, TC, Warren, RM, Yap, SF, Zaman, K, Portaels, F, Rastogi, N, Sola, C (2006) *Mycobacterium tuberculosis* complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology. *BMC microbiology*, **6** 23.

Brynildsrud, OB, Pepperell, CS, Suffys, P, Grandjean, L, Monteserin, J, Debech, N, Bohlin, J, Alfsnes, K, Pettersson, JOH, Kirkeleite, I, Fandinho, F, Silva, MA da, Perdigao, J, Portugal, I, Viveiros, M, Clark, T, Caws, M, Dunstan, S, Thai, PVK, Lopez, B, Ritacco, V, Kitchen, A, Brown, TS, Soolingen, D van, O'Neill, MB, Holt, KE, Feil, EJ, Mathema, B, Balloux, F, Eldholm, V (2018) Global expansion of *Mycobacterium tuberculosis* lineage 4 shaped by colonial migration and local adaptation. *Science advances*, **4**(10): eaat5869.

Buu, TN, Huyen, MN, Lan, NTN, Quy, HT, Hen, NV, Zignol, M, Borgdorff, MW, Cobelens, FGJ, Soolingen, D van (2009) The Beijing genotype is associated with young age and multidrug-resistant tuberculosis in rural Vietnam. *The international journal of tuberculosis and lung disease*, **13**(7): 900–6.

Buu, TN, Soolingen, D van, Huyen, MNT, Lan, NTN, Quy, HT, Tiemersma, EW, Kremer, K, Borgdorff, MW, Cobelens, FGJ (2012) Increased transmission of *Mycobacterium tuberculosis* Beijing genotype strains associated with resistance to streptomycin: a population-based study. *PloS one*, **7**(8): e42323.

Casali, N, Nikolayevskyy, V, Balabanova, Y, Harris, SR, Ignatyeva, O, Kontsevaya, I, Corander, J, Bryant, J, Parkhill, J, Nejentsev, S, Horstmann, RD, Brown, T, Drobniewski, F (2014) Evolution and transmission of drug-resistant tuberculosis in a Russian population. *Nature genetics*, **46**(3): 279–86.

Casali, N, Nikolayevskyy, V, Balabanova, Y, Ignatyeva, O, Kontsevaya, I, Harris, SR, Bentley, SD, Parkhill, J, Nejentsev, S, Hoffner, SE, Horstmann, RD, Brown, T, Drobniewski, F (2012) Microevolution of extensively drug-resistant tuberculosis in Russia. *Genome Research*, **22**(4): 735–745.

Casanova, JL, Abel, L (2002) Genetic dissection of immunity to mycobacteria: the human model. *Annual review of immunology*, **20**(1): 581–620.

Chihota, VN, Niehaus, A, Streicher, EM, Wang, X, Sampson, SL, Mason, P, Källenius, G, Mfinanga, SG, Pillay, M, Klopper, M, Kasongo, W, Behr, MA, Gey van Pittius, NC,

Helden, PD van, Couvin, D, Rastogi, N, Warren, RM (2018) Geospatial distribution of *Mycobacterium tuberculosis* genotypes in Africa. *PloS one*, **13**(8): e0200632.

Cirillo, DM, Miotto, P, Tortoli, E (2017) Evolution of Phenotypic and Molecular Drug Susceptibility Testing. *In: Gagneux S. (eds) Strain Variation in the Mycobacterium tuberculosis Complex: Its Role in Biology, Epidemiology and Control. Advances in Experimental Medicine and Biology*, **1019** 221–246.

Colditz, GA, Berkey, CS, Mosteller, F, Brewer, TF, Wilson, ME, Burdick, E, Fineberg, HV (1995) The efficacy of bacillus Calmette-Guérin vaccination of newborns and infants in the prevention of tuberculosis: meta-analyses of the published literature. *Pediatrics*, **96** 29–35.

Coll, F, McNerney, R, Guerra-Assunção, JA, Glynn, JR, Perdigão, J, Viveiros, M, Portugal, I, Pain, A, Martin, N, Clark, TG (2014) A robust SNP barcode for typing *Mycobacterium tuberculosis* complex strains. *Nature communications*, **5**(1): 4812.

Coll, F, McNerney, R, Preston, MD, Guerra-Assunção, JA, Warry, A, Hill-Cawthorne, G, Mallard, K, Nair, M, Miranda, A, Alves, A, Perdigão, J, Viveiros, M, Portugal, I, Hasan, Z, Hasan, R, Glynn, JR, Martin, N, Pain, A, Clark, TG (2015) Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. *Genome Medicine*, **7**(1): 51.

Comas, I (2017) Genomic Epidemiology of Tuberculosis. *In: Gagneux S. (eds) Strain Variation in the Mycobacterium tuberculosis Complex: Its Role in Biology, Epidemiology and Control. Advances in Experimental Medicine and Biology*, **1019** 79–93.

Comas, I, Chakravartti, J, Small, PM, Galagan, J, Niemann, S, Kremer, K, Ernst, JD, Gagneux, S (2010) Human T cell epitopes of *Mycobacterium tuberculosis* are evolutionarily hyperconserved. *Nature genetics*, **42**(6): 498–503.

Comas, I, Coscolla, M, Luo, T, Borrell, S, Holt, KE, Kato-Maeda, M, Parkhill, J, Malla, B, Berg, S, Thwaites, G, Yeboah-Manu, D, Bothamley, G, Mei, J, Wei, L, Bentley, S, Harris, SR, Niemann, S, Diel, R, Aseffa, A, Gao, Q, Young, D, Gagneux, S (2013) Out-of-Africa migration and Neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans. *Nature genetics*, **45**(10): 1176–82.

Comas, I, Gagneux, S (2009) The past and future of tuberculosis research. *PLoS pathogens*, **5**(10): e1000600.

Comas, I, Hailu, E, Kiros, T, Bekele, S, Mekonnen, W, Gumi, B, Tschopp, R, Ameni, G, Hewinson, RG, Robertson, BD, Goig, GA, Stucki, D, Gagneux, S, Aseffa, A, Young, D, Berg, S (2015) Population Genomics of *Mycobacterium tuberculosis* in Ethiopia Contradicts the Virgin Soil Hypothesis for Human Tuberculosis in Sub-Saharan Africa. *Current biology : CB*, **25**(24): 3260–6.

Corbett, EL, Charalambous, S, Moloi, VM, Fielding, K, Grant, AD, Dye, C, De Cock, KM, Hayes, RJ, Williams, BG, Churchyard, GJ (2004) Human Immunodeficiency Virus and the Prevalence of Undiagnosed Tuberculosis in African Gold Miners. *American Journal of Respiratory and Critical Care Medicine*, **170**(6): 673–679.

Coscolla, M (2017) Biological and Epidemiological Consequences of MTBC Diversity. *In: Gagneux S. (eds) Strain Variation in the Mycobacterium tuberculosis Complex: Its Role in Biology, Epidemiology and Control. Advances in Experimental Medicine and Biology*, **1019** 95–116.

Coscolla, M, Gagneux, S (2014) Consequences of genomic diversity in *Mycobacterium tuberculosis*. *Seminars in immunology*, **26**(6): 431–44.

Coscolla, M, Lewin, A, Metzger, S, Maetz-Rennsing, K, Calvignac-Spencer, S, Nitsche, A, Dabrowski, PW, Radonic, A, Niemann, S, Parkhill, J, Couacy-Hymann, E, Feldman, J, Comas, I, Boesch, C, Gagneux, S, Leendertz, FH (2013) Novel *Mycobacterium tuberculosis* complex isolate from a wild chimpanzee. *Emerging infectious diseases*, **19**(6): 969–76.

Cousins, DV, Peet, RL, Gaynor, WT, Williams, SN, Gow, BL (1994) Tuberculosis in imported hyrax (Procavia capensis) caused by an unusual variant belonging to the *Mycobacterium tuberculosis* complex. eng. *Veterinary microbiology*, **42**(2-3): 135–45.

Cowley, D, Govender, D, February, B, Wolfe, M, Steyn, L, Evans, J, Wilkinson, RJ, Nicol, MP (2008) Recent and rapid emergence of W-Beijing strains of *Mycobacterium tuberculosis* in Cape Town, South Africa. *Clinical infectious diseases*, **47**(10): 1252–9.

Daniel, TM (2006) The history of tuberculosis. *Respiratory medicine*, **100**(11): 1862–70.

Demay, C, Liens, B, Burguière, T, Hill, V, Couvin, D, Millet, J, Mokrousov, I, Sola, C, Zozio, T, Rastogi, N (2012) SITVITWEB–a publicly available international multimarker database for studying *Mycobacterium tuberculosis* genetic diversity and molecular epidemiology. *Infection, genetics and evolution*, **12**(4): 755–66.

Douglas, JT, Qian, L, Montoya, JC, Musser, JM, Van Embden, JDA, Van Soolingen, D, Kremer, K (2003) Characterization of the Manila Family of *Mycobacterium tuberculosis*. *Journal of Clinical Microbiology*, **41**(6): 2723–2726.

Drummond, AJ, Ho, SYW, Phillips, MJ, Rambaut, A (2006) Relaxed Phylogenetics and Dating with Confidence. *PLoS Biology*, **4**(5): e88.

Dye, C, Williams, BG (2010) The population dynamics and control of tuberculosis. *Science*, **328**(5980): 856–61.

Egwaga, SM, Cobelens, FG, Muwinge, H, Verhage, C, Kalisvaart, N, Borgdorff, MW (2006) The impact of the HIV epidemic on tuberculosis transmission in Tanzania. *AIDS*, **20**(6): 915–21.

Eldholm, V, Matee, M, Mfinanga, SGM, Heun, M, Dahle, UR (2006) A first insight into the genetic diversity of *Mycobacterium tuberculosis* in Dar es Salaam, Tanzania, assessed by spoligotyping. *BMC microbiology*, **6**(1): 76.

Eldholm, V, Pettersson, JHO, Brynildsrud, OB, Kitchen, A, Rasmussen, EM, Lillebaek, T, Rønning, JO, Crudu, V, Mengshoel, AT, Debech, N, Alfsnes, K, Bohlin, J, Pepperell, CS, Balloux, F (2016) Armed conflict and population displacement as drivers of the evolution and dispersal of *Mycobacterium tuberculosis*. *Proceedings of the National Academy of Sciences*, **113**(48): 13881–13886.

Embden, JD van, Cave, MD, Crawford, JT, Dale, JW, Eisenach, KD, Gicquel, B, Hermans, P, Martin, C, McAdam, R, Shinnick, TM (1993) Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology. *Journal of clinical microbiology*, **31**(2): 406–9.

Falush, D, Wirth, T, Linz, B, Pritchard, JK, Stephens, M, Kidd, M, Blaser, MJ, Graham, DY, Vacher, S, Perez-Perez, GI, Yamaoka, Y, Mégraud, F, Otto, K, Reichard, U, Katzowitsch, E, Wang, X, Achtman, M, Suerbaum, S (2003) Traces of human migrations in Helicobacter pylori populations. *Science*, **299**(5612): 1582–5.

Fenner, L, Egger, M, Bodmer, T, Furrer, H, Ballif, M, Battegay, M, Helbling, P, Fehr, J, Gsponer, T, Rieder, HL, Zwahlen, M, Hoffmann, M, Bernasconi, E, Cavassini, M, Calmy, A, Dolina, M, Frei, R, Janssens, JP, Borrell, S, Stucki, D, Schrenzel, J, Böttger, EC, Gagneux, S, Swiss HIV Cohort and Molecular Epidemiology of Tuberculosis Study Groups (2013) HIV infection disrupts the sympatric host-pathogen relationship in human tuberculosis. *PLoS genetics*, **9**(3): e1003318.

Fenner, L, Malla, B, Ninet, B, Dubuis, O, Stucki, D, Borrell, S, Huna, T, Bodmer, T, Egger, M, Gagneux, S (2011) "Pseudo-Beijing": evidence for convergent evolution in the direct repeat region of *Mycobacterium tuberculosis*. *PloS one*, **6**(9): e24737.

Fine, PE (1995) Variation in protection by BCG: implications of and for heterologous immunity. *Lancet*, **346**(8986): 1339–45.

Firdessa, R, Berg, S, Hailu, E, Schelling, E, Gumi, B, Erenso, G, Gadisa, E, Kiros, T, Habtamu, M, Hussein, J, Zinsstag, J, Robertson, BD, Ameni, G, Lohan, AJ, Loftus, B, Comas, I, Gagneux, S, Tschopp, R, Yamuah, L, Hewinson, G, Gordon, SV, Young, DB, Aseffa, A (2013) Mycobacterial lineages causing pulmonary and extrapulmonary tuberculosis, Ethiopia. *Emerging infectious diseases*, **19**(3): 460–3.

Gagneux, S (2012) Host-pathogen coevolution in human tuberculosis. *Philosophical transactions of the Royal Society of London*, **367**(1590): 850–9.

Gagneux, S (2018) Ecology and evolution of *Mycobacterium tuberculosis*. *Nature reviews. Microbiology*, **16**(4): 202–213.

Gagneux, S, DeRiemer, K, Van, T, Kato-Maeda, M, Jong, BC de, Narayanan, S, Nicol, M, Niemann, S, Kremer, K, Gutierrez, MC, Hilty, M, Hopewell, PC, Small, PM (2006) Variable host-pathogen compatibility in *Mycobacterium tuberculosis*. *Proceedings of the National Academy of Sciences*, **103**(8): 2869–2873.

Gardy, JL, Johnston, JC, Ho Sui, SJ, Cook, VJ, Shah, L, Brodkin, E, Rempel, S, Moore, R, Zhao, Y, Holt, R, Varhol, R, Birol, I, Lem, M, Sharma, MK, Elwood, K, Jones, SJM, Brinkman, FSL, Brunham, RC, Tang, P (2011) Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. *The New England journal of medicine*, **364**(8): 730–9.

Gehre, F, Kumar, S, Kendall, L, Ejo, M, Secka, O, Ofori-Anyinam, B, Abatih, E, Antonio, M, Berkvens, D, Jong, BC de (2016) A Mycobacterial Perspective on Tuberculosis in West Africa: Significant Geographical Variation of *M. africanum* and Other *M. tuberculosis* Complex Lineages. *PLoS neglected tropical diseases*, **10**(3): e0004408.

Getahun, H, Gunneberg, C, Granich, R, Nunn, P (2010) HIV infection-associated tuberculosis: the epidemiology and the response. *Clinical infectious diseases*, **50**(s3): S201–7.

Gilbert, E (2002) Coastal East Africa and the Western Indian Ocean: Long-Distance Trade, Empire, Migration, and Regional Unity, 1750-1970. *The History Teacher*, **36**(1): 7.

Githui, WA, Jordaan, AM, Juma, ES, Kinyanjui, P, Karimi, FG, Kimwomi, J, Meme, H, Mumbi, P, Streicher, EM, Warren, R, Helden, PD van, Victor, TC (2004) Identification of MDR-TB Beijing/W and other *Mycobacterium tuberculosis* genotypes in Nairobi, Kenya. *The international journal of tuberculosis and lung disease*, **8**(3): 352–60.

Glynn, JR, Alghamdi, S, Mallard, K, McNerney, R, Ndlovu, R, Munthali, L, Houben, RM, Fine, PEM, French, N, Crampin, AC (2010) Changes in *Mycobacterium tuberculosis* Genotype Families Over 20 Years in a Population Based Study in Northern Malawi. *PLoS ONE*, **5**(8): e12259.

Golub, JE, Bur, S, Cronin, WA, Gange, S, Baruch, N, Comstock, GW, Chaisson, RE (2006) Delayed tuberculosis diagnosis and tuberculosis transmission. *The international journal of tuberculosis and lung disease*, **10**(1): 24–30.

Gonzalo-Asensio, J, Malaga, W, Pawlik, A, Astarie-Dequeker, C, Passemar, C, Moreau, F, Laval, F, Daffé, M, Martin, C, Brosch, R, Guilhot, C (2014) Evolutionary history of tuberculosis shaped by conserved mutations in the PhoPR virulence regulator. *Proceedings of the National Academy of Sciences of the United States of America*, **111**(31): 11491–6.

Guerra-Assunção, JA, Crampin, AC, Houben, RMGJ, Mzembe, T, Mallard, K, Coll, F, Khan, P, Banda, L, Chiwaya, A, Pereira, RPA, McNerney, R, Fine, PEM, Parkhill, J, Clark, TG, Glynn, JR (2015a) Large-scale whole genome sequencing of *M. tuberculosis* provides insights into transmission in a high prevalence area. *eLife*, **4**

Guerra-Assunção, JA, Houben, RMGJ, Crampin, AC, Mzembe, T, Mallard, K, Coll, F, Khan, P, Banda, L, Chiwaya, A, Pereira, RPA, McNerney, R, Harris, D, Parkhill, J, Clark, TG, Glynn, JR (2015b) Recurrence due to relapse or reinfection with *Mycobacterium tuberculosis*: a whole-genome sequencing approach in a large, population-based cohort with a high HIV infection prevalence and active follow-up. *The Journal of infectious diseases*, **211**(7): 1154–63.

Guerra-Silveira, F, Abad-Franch, F (2013) Sex bias in infectious disease epidemiology: patterns and processes. *PloS one*, **8**(4): e62390.

Gutierrez, MC, Brisse, S, Brosch, R, Fabre, M, Omaïs, B, Marmiesse, M, Supply, P, Vincent, V (2005) Ancient origin and gene mosaicism of the progenitor of *Mycobacterium tuberculosis*. *PLoS pathogens*, **1**(1): e5.

Hanekom, M, Spuy, GD van der, Streicher, E, Ndabambi, SL, McEvoy, CRE, Kidd, M, Beyers, N, Victor, TC, Helden, PD van, Warren, RM (2007) A recently evolved sublineage of the *Mycobacterium tuberculosis* Beijing strain family is associated with an increased ability to spread and cause disease. *Journal of clinical microbiology*, **45**(5): 1483–90.

Hasegawa, M, Kishino, H, Yano, T (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of molecular evolution*, **22**(2): 160–74.

Hegdahl, HK, Fylkesnes, KM, Sandøy, IF (2016) Sex Differences in HIV Prevalence Persist over Time: Evidence from 18 Countries in Sub-Saharan Africa. *PloS one*, **11**(2): e0148502.

Helden, PD van, Warren, RM, Victor, TC, Spuy, G van der, Richardson, M, Hoal-van Helden, E (2002) Strain families of *Mycobacterium tuberculosis*. *Trends in microbiology*, **10**(4): 167–8.

Hella, J, Morrow, C, Mhimbira, F, Ginsberg, S, Chitnis, N, Gagneux, S, Mutayoba, B, Wood, R, Fenner, L (2017) Tuberculosis transmission in public locations in Tanzania: A novel approach to studying airborne disease transmission. *The Journal of infection*, **75**(3): 191–197.

Hershberg, R, Lipatov, M, Small, PM, Sheffer, H, Niemann, S, Homolka, S, Roach, JC, Kremer, K, Petrov, DA, Feldman, MW, Gagneux, S (2008) High functional diversity in *Mycobacterium tuberculosis* driven by genetic drift and human demography. *PLoS biology*, **6**(12): e311.

Hess, M, Sczyrba, A, Egan, R, Kim, TW, Chokhawala, H, Schroth, G, Luo, S, Clark, DS, Chen, F, Zhang, T, Mackie, RI, Pennacchio, LA, Tringe, SG, Visel, A, Woyke, T, Wang, Z, Rubin, EM (2011) Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science*, **331**(6016): 463–7.

Heyckendorf, J, Andres, S, Köser, CU, Olaru, ID, Schön, T, Sturegård, E, Beckert, P, Schleusener, V, Kohl, TA, Hillemann, D, Moradigaravand, D, Parkhill, J, Peacock, SJ, Niemann, S, Lange, C, Merker, M (2017) What Is Resistance? Impact of Phenotypic versus Molecular Drug Resistance Testing on Therapy for Multi- and Extensively Drug-Resistant Tuberculosis. *Antimicrobial Agents and Chemotherapy*, **62**(2): 1550–1567.

Hirsh, AE, Tsolaki, AG, DeRiemer, K, Feldman, MW, Small, PM (2004) Stable association between strains of *Mycobacterium tuberculosis* and their human host populations. *Proceedings of the National Academy of Sciences of the United States of America*, **101**(14): 4871–6.

Hiza, H, Doulla, B, Sasamalo, M, Hella, J, Kamwela, L, Mhimbira, F, Reither, K, Gagneux, S, Jugheli, L, Fenner, L (2017) Preservation of sputum samples with cetylpyridinium chloride (CPC) for tuberculosis cultures and Xpert MTB/RIF in a low-income country. *BMC infectious diseases*, **17**(1): 542.

Holt, KE, McAdam, P, Thai, PVK, Thuong, NTT, Ha, DTM, Lan, NN, Lan, NH, Nhu, NTQ, Hai, HT, Ha, VTN, Thwaites, G, Edwards, DJ, Nath, AP, Pham, K, Ascher, DB, Farrar, J, Khor, CC, Teo, YY, Inouye, M, Caws, M, Dunstan, SJ (2018) Frequent transmission of the *Mycobacterium tuberculosis* Beijing lineage and positive selection for the EsxW Beijing variant in Vietnam. *Nature genetics*, **50**(6): 849–856.

Houben, RMGJ, Dodd, PJ (2016) The Global Burden of Latent Tuberculosis Infection: A Re-estimation Using Mathematical Modelling. *PLoS medicine*, **13**(10): e1002152.

Imperial, MZ, Nahid, P, Phillips, PPJ, Davies, GR, Fielding, K, Hanna, D, Hermann, D, Wallis, RS, Johnson, JL, Lienhardt, C, Savic, RM (2018) A patient-level pooled analysis of treatment-shortening regimens for drug-susceptible pulmonary tuberculosis. *Nature medicine*, **24**(11): 1708–1715.

Ingen, J van, Rahim, Z, Mulder, A, Boeree, MJ, Simeone, R, Brosch, R, Soolingen, D van (2012) Characterization of *Mycobacterium orygis* as *M. tuberculosis* complex subspecies. *Emerging infectious diseases*, **18**(4): 653–5.

Jong, BC de, Hill, PC, Aiken, A, Awine, T, Antonio, M, Adetifa, IM, Jackson-Sillah, DJ, Fox, A, Deriemer, K, Gagneux, S, Borgdorff, MW, McAdam, KPWJ, Corrah, T, Small, PM, Adegbola, RA (2008) Progression to active tuberculosis, but not transmission, varies by *Mycobacterium tuberculosis* lineage in The Gambia. *The Journal of infectious diseases*, **198**(7): 1037–43.

Kamerbeek, J, Schouls, L, Kolk, A, Agterveld, M van, Soolingen, D van, Kuijper, S, Bunschoten, A, Molhuizen, H, Shaw, R, Goyal, M, Embden, J van (1997) Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. *Journal of clinical microbiology*, **35**(4): 907–14.

Kass, RE, Raftery, AE (1995) Bayes Factors. *Journal of the American Statistical Association*, **90**(430): 773–795.

Kato-Maeda, M, Metcalfe, JZ, Flores, L (2011) Genotyping of *Mycobacterium tuberculosis*: application in epidemiologic studies. *Future microbiology*, **6**(2): 203–16.

Kato-Maeda, M, Shanley, CA, Ackart, D, Jarlsberg, LG, Shang, S, Obregon-Henao, A, Harton, M, Basaraba, RJ, Henao-Tamayo, M, Barrozo, JC, Rose, J, Kawamura, LM, Coscolla, M, Fofanov, VY, Koshinsky, H, Gagneux, S, Hopewell, PC, Ordway, DJ, Orme, IM (2012) Beijing Sublineages of *Mycobacterium tuberculosis* Differ in Pathogenicity in the Guinea Pig. *Clinical and Vaccine Immunology*, **19**(8): 1227–1237.

Kibiki, GS, Mulder, B, Dolmans, WMV, Beer, JL de, Boeree, M, Sam, N, Soolingen, D van, Sola, C, Zanden, AGM van der (2007) *M. tuberculosis* genotypic diversity and drug susceptibility pattern in HIV-infected and non-HIV-infected patients in northern Tanzania. *BMC microbiology*, **7**(1): 51.

Klopper, M, Warren, RM, Hayes, C, Pittius, NCG van, Streicher, EM, M??ller, B, Sirgel, FA, Chabula-Nxiweni, M, Hoosain, E, Coetzee, G, Helden, PD van, Victor, TC, Trollip, AP (2013) Emergence and spread of extensively and totally drug-resistant tuberculosis, South Africa. *Emerging Infectious Diseases*, **19**(3): 449–455.

Ko, AMS, Chen, CY, Fu, Q, Delfin, F, Li, M, Chiu, HL, Stoneking, M, Ko, YC (2014) Early Austronesians: into and out of Taiwan. *American journal of human genetics*, **94**(3): 426–36.

Koboldt, DC, Zhang, Q, Larson, DE, Shen, D, McLellan, MD, Lin, L, Miller, CA, Mardis, ER, Ding, L, Wilson, RK (2012) VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome research*, **22**(3): 568–76.

Koch, AS, Brites, D, Stucki, D, Evans, JC, Seldon, R, Heekes, A, Mulder, N, Nicol, M, Oni, T, Mizrahi, V, Warner, DF, Parkhill, J, Gagneux, S, Martin, DP, Wilkinson, RJ (2017) The Influence of HIV on the Evolution of *Mycobacterium tuberculosis*. *Molecular Biology and Evolution*, **34**(7): 1654–1668.

Koul, A, Arnoult, E, Lounis, N, Guillemont, J, Andries, K (2011) The challenge of new drug discovery for tuberculosis. *Nature*, **469**(7331): 483–490.

Kwan, CK, Ernst, JD (2011) HIV and tuberculosis: a deadly human syndemic. *Clinical microbiology reviews*, **24**(2): 351–76.

Lalor, MK, Anderson, LF, Hamblion, EL, Burkitt, A, Davidson, JA, Maguire, H, Abubakar, I, Thomas, HL (2017) Recent household transmission of tuberculosis in England, 2010-2012: retrospective national cohort study combining epidemiological and molecular strain typing data. *BMC medicine*, **15**(1): 105.

Lartillot, N, Philippe, H (2006) Computing Bayes Factors Using Thermodynamic Integration. *Systematic Biology*, **55**(2): 195–207.

Li, H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, **27**(21): 2987–93.

Li, H, Durbin, R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**(14): 1754–60.

Loman, NJ, Constantinidou, C, Chan, JZM, Halachev, M, Sergeant, M, Penn, CW, Robinson, ER, Pallen, MJ (2012) High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. *Nature reviews. Microbiology*, **10**(9): 599–606.

Luo, T, Comas, I, Luo, D, Lu, B, Wu, J, Wei, L, Yang, C, Liu, Q, Gan, M, Sun, G, Shen, X, Liu, F, Gagneux, S, Mei, J, Lan, R, Wan, K, Gao, Q (2015) Southern East Asian origin and coexpansion of *Mycobacterium tuberculosis* Beijing family with Han Chinese. *Proceedings of the National Academy of Sciences of the United States of America*, **112**(26): 8136–41.

Main, P, Attenborough, R, Chelvanayagam, G, Bhatia, K, Gao, X (2001) The peopling of New Guinea: evidence from class I human leukocyte antigen. *Human biology*, **73**(3): 365–83.

Malla, B, Stucki, D, Borrell, S, Feldmann, J, Maharjan, B, Shrestha, B, Fenner, L, Gagneux, S (2012) First insights into the phylogenetic diversity of *Mycobacterium tuberculosis* in Nepal. *PloS one*, **7**(12): e52297.

Manson, AL, Cohen, KA, Abeel, T, Desjardins, CA, Armstrong, DT, Barry Iii, CE, Brand, J, TBResist Global Genome Consortium, Chapman, SB, Cho, SN, Gabrielian, A, Gomez, J, Jodals, AM, Joloba, M, Jureen, P, Lee, JS, Malinga, L, Maiga, M, Nordenberg, D, Noroc, E, Romancenco, E, Salazar, A, Ssengooba, W, Velayati, AA, Winglee, K, Zalutskaya, A, Via, LE, Cassell, GH, Dorman, SE, Ellner, J, Farnia, P, Galagan, JE, Rosenthal, A, Crudu, V, Homorodean, D, Hsueh, PR, Narayanan, S, Pym, AS, Skrahina, A, Swaminathan, S, Walt, M der, Alland, D, Bishai, WR, Cohen, T, Hoffner, S, Birren, BW, Earl, AM (2017) Genomic analysis of globally diverse *Mycobacterium tuberculosis* strains provides insights into the emergence and spread of multidrug resistance. *Nature genetics*, **49** 395–402.

Mbugi, EV, Katale, BZ, Streicher, EM, Keyyu, JD, Kendall, SL, Dockrell, HM, Michel, AL, Rweyemamu, MM, Warren, RM, Matee, MI, Helden, PD van, Couvin, D, Rastogi, N (2016) Mapping of *Mycobacterium tuberculosis* Complex Genetic Diversity Profiles in Tanzania and Other African Countries. *PloS one*, **11**(5): e0154571.

Mbugi, EV, Katale, BZ, Siame, KK, Keyyu, JD, Kendall, SL, Dockrell, HM, Streicher, EM, Michel, AL, Rweyemamu, MM, Warren, RM, Matee, MI, Helden, PD van (2015) Genetic diversity of *Mycobacterium tuberculosis* isolated from tuberculosis patients in the Serengeti ecosystem in Tanzania. *Tuberculosis*, **95**(2): 170–8.

McEvoy, CRE, Falmer, AA, Gey van Pittius, NC, Victor, TC, Helden, PD van, Warren, RM (2007) The role of IS6110 in the evolution of *Mycobacterium tuberculosis*. *Tuberculosis*, **87**(5): 393–404.

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

Mfinanga, SGM, Warren, RM, Kazwala, R, Kahwa, A, Kazimoto, T, Kimaro, G, Mfaume, S, Chonde, T, Ngadaya, E, Egwaga, S, Streicher, EM, Van Pittius, GNC, Morkve Odd, M, Cleaveland, S (2014) Genetic profile of *Mycobacterium tuberculosis* and treatment outcomes in human pulmonary tuberculosis in Tanzania. en. *Tanzania journal of health research*, **16**(2): 58–69.

Mhimbira, F, Hella, J, Said, K, Kamwela, L, Sasamalo, M, Maroa, T, Chiryamkubi, M, Mhalu, G, Schindler, C, Reither, K, Knopp, S, Utzinger, J, Gagneux, S, Fenner, L (2017) Prevalence and clinical relevance of helminth co-infections among tuberculosis patients in urban Tanzania. *PLoS neglected tropical diseases*, **11**(2): e0005342.

Migliori, GB, De Iaco, G, Besozzi, G, Centis, R, Cirillo, DM (2007) First tuberculosis cases in Italy resistant to all tested drugs. eng. *Euro surveillance : European communicable disease bulletin*, **12**(5): E070517.1.

Miotto, P, Cabibbe, AM, Borroni, E, Degano, M, Cirillo, DM (2018) Role of Disputed Mutations in the rpoB Gene in Interpretation of Automated Liquid MGIT Culture

Results for Rifampin Susceptibility Testing of *Mycobacterium tuberculosis*. *Journal of Clinical Microbiology*, **56**(5): e01599–17.

MoHSW (2013) First tuberculosis prevalence survey in the United Republic of Tanzania.

Monot, M, Honoré, N, Garnier, T, Araoz, R, Coppée, JY, Lacroix, C, Sow, S, Spencer, JS, Truman, RW, Williams, DL, Gelber, R, Virmond, M, Flageul, B, Cho, SN, Ji, B, Paniz-Mondolfi, A, Convit, J, Young, S, Fine, PE, Rasolofo, V, Brennan, PJ, Cole, ST (2005) On the origin of leprosy. *Science*, **308**(5724): 1040–2.

Musser, JM, Amin, A, Ramaswamy, S (2000) Negligible genetic diversity of *Mycobacterium tuberculosis* host immune system protein targets: evidence of limited selective pressure. *Genetics*, **155**(1): 7–16.

Nagu, TJ, Aboud, S, Mwiru, R, Matee, M, Fawzi, W, Mugusi, F (2015) Multi drug and other forms of drug resistant tuberculosis are uncommon among treatment naïve tuberculosis patients in Tanzania. *PloS one*, **10**(4): e0118601.

Nathanson, E, Lambregts-van Weezenbeek, C, Rich, ML, Gupta, R, Bayona, J, Blöndal, K, Caminero, JA, Cegielski, JP, Danilovits, M, Espinal, MA, Hollo, V, Jaramillo, E, Leimane, V, Mitnick, CD, Mukherjee, JS, Nunn, P, Pasechnikov, A, Tupasi, T, Wells, C, Raviglione, MC (2006) Multidrug-resistant tuberculosis management in resource-limited settings. *Emerging infectious diseases*, **12**(9): 1389–97.

Nava-Aguilera, E, López-Vidal, Y, Harris, E, Morales-Pérez, A, Mitchell, S, Flores-Moreno, M, Villegas-Arrizón, A, Legorreta-Soberanis, J, Ledogar, R, Andersson, N (2011) Clustering of *Mycobacterium tuberculosis* cases in Acapulco: Spoligotyping and risk factors. *Clinical & developmental immunology*, **2011** 408375.

Nhamoyebonde, S, Leslie, A (2014) Biological differences between the sexes and susceptibility to tuberculosis. *The Journal of infectious diseases*, **209**(3): S100–S106.

NTLP (2013) Annual Report 2013, 1689–1699.

NTLP (2016) Annual Report 2016, 1–48.

O'Neill, MB, Shockey, A, Zarley, A, Aylward, W, Eldholm, V, Kitchen, A, Pepperell, CS (2019) Lineage specific histories of *Mycobacterium tuberculosis* dispersal in Africa and Eurasia. *Molecular ecology*, **28**(13): 3241–3256.

Paradis, E, Claude, J, Strimmer, K (2004) APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics*, **20**(2): 289–290.

Parsons, SDC, Drewe, JA, Gey van Pittius, NC, Warren, RM, Helden, PD van (2013) Novel cause of tuberculosis in meerkats, South Africa. *Emerging infectious diseases*, **19**(12): 2004–7.

Phelan, J, Coll, F, McNerney, R, Ascher, DB, Pires, DEV, Furnham, N, Coeck, N, Hill-Cawthorne, GA, Nair, MB, Mallard, K, Ramsay, A, Campino, S, Hibberd, ML, Pain,

A, Rigouts, L, Clark, TG (2016) *Mycobacterium tuberculosis* whole genome sequencing and protein structure modelling provides insights into anti-tuberculosis drug resistance. *BMC Medicine*, **14**(1): 31.

Portevin, D, Gagneux, S, Comas, I, Young, D (2011) Human macrophage responses to clinical isolates from the *Mycobacterium tuberculosis* complex discriminate between ancient and modern lineages. *PLoS pathogens*, **7**(3): e1001307.

Portevin, D, Sukumar, S, Coscolla, M, Shui, G, Li, B, Guan, XL, Bendt, AK, Young, D, Gagneux, S, Wenk, MR (2014) Lipidomics and genomics of *Mycobacterium tuberculosis* reveal lineage-specific trends in mycolic acid biosynthesis. *MicrobiologyOpen*, **3**(6): 823–35.

Qin, J, Li, R, Raes, J, Arumugam, M, Burgdorf, KS, Manichanh, C, Nielsen, T, Pons, N, Levenez, F, Yamada, T, Mende, DR, Li, J, Xu, J, Li, S, Li, D, Cao, J, Wang, B, Liang, H, Zheng, H, Xie, Y, Tap, J, Lepage, P, Bertalan, M, Batto, JM, Hansen, T, Le Paslier, D, Linneberg, A, Nielsen, HB, Pelletier, E, Renault, P, Sicheritz-Ponten, T, Turner, K, Zhu, H, Yu, C, Li, S, Jian, M, Zhou, Y, Li, Y, Zhang, X, Li, S, Qin, N, Yang, H, Wang, J, Brunak, S, Doré, J, Guarner, F, Kristiansen, K, Pedersen, O, Parkhill, J, Weissenbach, J, MetaHIT Consortium, Bork, P, Ehrlich, SD, Wang, J (2010) A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, **464**(7285): 59–65.

R Core Team (2018) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.

Rambaut, A, Drummond, AJ, Xie, D, Baele, G, Suchard, MA (2018) Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Systematic Biology*, **67**(5): 901–904.

Ramsden, C, Melo, FL, Figueiredo, LM, Holmes, EC, Zanotto, PM (2008) High Rates of Molecular Evolution in Hantaviruses. *Molecular Biology and Evolution*, **25**(7): 1488–1492.

Reed, MB, Domenech, P, Manca, C, Su, H, Barczak, AK, Kreiswirth, BN, Kaplan, G, Barry, CE (2004) A glycolipid of hypervirulent tuberculosis strains that inhibits the innate immune response. *Nature*, **431**(7004): 84–7.

Reed, MB, Gagneux, S, Deriemer, K, Small, PM, Barry, CE (2007) The W-Beijing lineage of *Mycobacterium tuberculosis* overproduces triglycerides and has the DosR dormancy regulon constitutively upregulated. *Journal of bacteriology*, **189**(7): 2583–9.

Reed, MB, Pichler, VK, McIntosh, F, Mattia, A, Fallow, A, Masala, S, Domenech, P, Zwerling, A, Thibert, L, Menzies, D, Schwartzman, K, Behr, MA (2009) Major *Mycobacterium tuberculosis* lineages associate with patient country of origin. *Journal of clinical microbiology*, **47**(4): 1119–28.

Reich, D, Patterson, N, Kircher, M, Delfin, F, Nandineni, MR, Pugach, I, Ko, AMS, Ko, YC, Jinam, TA, Phipps, ME, Saitou, N, Wollstein, A, Kayser, M, Pääbo, S, Stoneking, M (2011) Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. *American journal of human genetics*, **89**(4): 516–28.

Revell, LJ (2012) phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, **3**(2): 217–223.

Rhines, AS (2013) The role of sex differences in the prevalence and transmission of tuberculosis. *Tuberculosis*, **93**(1): 104–7.

Rieder, HL (1999) *Epidemiologic Basis of Tuberculosis Control First edition 1999*. Tech. rep.

Rieux, A, Balloux, F (2016) Inferences from tip-calibrated phylogenies: a review and a practical guide. *Molecular Ecology*, **25**(9): 1911–1924.

Rieux, A, Khatchikian, CE (2017) TIPDATINGBEAST: an R package to assist the implementation of phylogenetic tip-dating tests using BEAST. *Molecular Ecology Resources*, **17**(4): 608–613.

Roetzer, A, Diel, R, Kohl, TA, Rückert, C, Nübel, U, Blom, J, Wirth, T, Jaenicke, S, Schuback, S, Rüsch-Gerdes, S, Supply, P, Kalinowski, J, Niemann, S (2013) Whole genome sequencing versus traditional genotyping for investigation of a *Mycobacterium tuberculosis* outbreak: a longitudinal molecular epidemiological study. *PLoS medicine*, **10**(2): e1001387.

Rose, G, Cortes, T, Comas, I, Coscolla, M, Gagneux, S, Young, DB (2013) Mapping of genotype-phenotype diversity among clinical isolates of *Mycobacterium tuberculosis* by sequence-based transcriptional profiling. *Genome biology and evolution*, **5**(10): 1849–62.

Rueda, J, Realpe, T, Mejia, GI, Zapata, E, Rozo, JC, Ferro, BE, Robledo, J (2015) Genotypic Analysis of Genes Associated with Independent Resistance and Cross-Resistance to Isoniazid and Ethionamide in *Mycobacterium tuberculosis* Clinical Isolates. *Antimicrobial agents and chemotherapy*, **59**(12): 7805–10.

Rutaihwa, LK, Menardo, F, Stucki, D, Gygli, SM, Ley, SD, Malla, B, Feldmann, J, Borrell, S, Beisel, C, Middelkoop, K, Carter, EJ, Diero, L, Ballif, M, Jugheli, L, Reither, K, Fenner, L, Brites, D, Gagneux, S (2019a) Multiple Introductions of *Mycobacterium tuberculosis* Lineage 2–Beijing Into Africa Over Centuries. *Frontiers in Ecology and Evolution*, **7** 112.

Rutaihwa, LK, Sasamalo, M, Jaleco, A, Hella, J, Kingazi, A, Kamwela, L, Kingalu, A, Malewo, B, Shirima, R, Doetsch, A, Feldmann, J, Reinhard, M, Borrell, S, Brites, D, Reither, K, Doulla, B, Fenner, L, Gagneux, S (2019b) Insights into the genetic diversity of *Mycobacterium tuberculosis* in Tanzania. *PloS one*, **14**(4): e0206334.

Said, K, Hella, J, Mhalu, G, Chiryankubi, M, Masika, E, Maroa, T, Mhimbira, F, Kapalata, N, Fenner, L (2017) Diagnostic delay and associated factors among patients with pulmonary tuberculosis in Dar es Salaam, Tanzania. *Infectious diseases of poverty*, **6**(1): 64.

Shitikov, E, Kolchenko, S, Mokrousov, I, Bespyatykh, J, Ischenko, D, Ilina, E, Govorun, V (2017) Evolutionary pathway analysis and unified classification of East Asian lineage of *Mycobacterium tuberculosis*. *Scientific Reports*, **7**(1): 9227.

Sikalengo, G, Hella, J, Mhimbira, F, Rutaihwa, LK, Bani, F, Ndege, R, Sasamalo, M, Kamwela, L, Said, K, Mhalu, G, Mlacha, Y, Hatz, C, Knopp, S, Gagneux, S, Reither, K, Utzinger, J, Tanner, M, Letang, E, Weisser, M, Fenner, L (2018) Distinct clinical characteristics and helminth co-infections in adult tuberculosis patients from urban compared to rural Tanzania. *Infectious diseases of poverty*, **7**(1): 24.

Singh, UB, Suresh, N, Bhanu, NV, Arora, J, Pant, H, Sinha, S, Aggarwal, RC, Singh, S, Pande, JN, Sola, C, Rastogi, N, Seth, P (2004) Predominant tuberculosis spoligotypes, Delhi, India. *Emerging infectious diseases*, **10**(6): 1138–42.

Soolingen, D van, Hoogenboezem, T, Haas, PE de, Hermans, PW, Koedam, MA, Teppema, KS, Brennan, PJ, Besra, GS, Portaels, F, Top, J, Schouls, LM, Embden, JD van (1997) A novel pathogenic taxon of the *Mycobacterium tuberculosis* complex, Canetti: characterization of an exceptional isolate from Africa. eng. *International journal of systematic bacteriology*, **47**(4): 1236–45.

Spuy, GD van der, Kremer, K, Ndabambi, SL, Beyers, N, Dunbar, R, Marais, BJ, Helden, PD van, Warren, RM (2009) Changing *Mycobacterium tuberculosis* population highlights clade-specific pathogenic characteristics. *Tuberculosis*, **89**(2): 120–5.

Spuy, GD van der, Warren, RM, Richardson, M, Beyers, N, Behr, MA, Helden, PD van (2003) Use of genetic distance as a measure of ongoing transmission of *Mycobacterium tuberculosis*. *Journal of clinical microbiology*, **41**(12): 5640–4.

Sreevatsan, S, Pan, X, Stockbauer, KE, Connell, ND, Kreiswirth, BN, Whittam, TS, Musser, JM (1997) Restricted structural gene polymorphism in the *Mycobacterium tuberculosis* complex indicates evolutionarily recent global dissemination. *Proceedings of the National Academy of Sciences of the United States of America*, **94**(18): 9869–74.

Staden, R (1996) The Staden sequence analysis package. *Molecular biotechnology*, **5**(3): 233–41.

Stamatakis, A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, **22**(21): 2688–90.

Stavrum, R, PrayGod, G, Range, N, Faurholt-Jepsen, D, Jeremiah, K, Faurholt-Jepsen, M, Krarup, H, Aabye, MG, Changalucha, J, Friis, H, Andersen, AB, Grewal, HMS

(2014) Increased level of acute phase reactants in patients infected with modern *Mycobacterium tuberculosis* genotypes in Mwanza, Tanzania. *BMC infectious diseases*, **14**(1): 309.

Steiner, A, Hella, J, Grüninger, S, Mhalu, G, Mhimbira, F, Cercamondi, CI, Doulla, B, Maire, N, Fenner, L (2016) Managing research and surveillance projects in real-time with a novel open-source eManagement tool designed for under-resourced countries. *Journal of the American Medical Informatics Association : JAMIA*, **23**(5): 916–23.

Steiner, A, Stucki, D, Coscolla, M, Borrell, S, Gagneux, S (2014) KvarQ: targeted and direct variant calling from fastq reads of bacterial genomes. *BMC Genomics*, **15**(1): 881.

Stucki, D, Ballif, M, Bodmer, T, Coscolla, M, Maurer, AM, Droz, S, Butz, C, Borrell, S, Längle, C, Feldmann, J, Furrer, H, Mordasini, C, Helbling, P, Rieder, HL, Egger, M, Gagneux, S, Fenner, L (2015) Tracking a Tuberculosis Outbreak Over 21 Years: Strain-Specific Single-Nucleotide Polymorphism Typing Combined With Targeted Whole-Genome Sequencing. *The Journal of Infectious Diseases*, **211**(8): 1306–1316.

Stucki, D, Brites, D, Jeljeli, L, Coscolla, M, Liu, Q, Trauner, A, Fenner, L, Rutaihwa, L, Borrell, S, Luo, T, Gao, Q, Kato-Maeda, M, Ballif, M, Egger, M, Macedo, R, Mardassi, H, Moreno, M, Tudo Vilanova, G, Fyfe, J, Globan, M, Thomas, J, Jamieson, F, Guthrie, JL, Asante-Poku, A, Yeboah-Manu, D, Wampande, E, Ssengooba, W, Joloba, M, Henry Boom, W, Basu, I, Bower, J, Saraiva, M, Vaconcellos, SEG, Suffys, P, Koch, A, Wilkinson, R, Gail-Bekker, L, Malla, B, Ley, SD, Beck, HP, Jong, BC de, Toit, K, Sanchez-Padilla, E, Bonnet, M, Gil-Brusola, A, Frank, M, Penlap Beng, VN, Eisenach, K, Alani, I, Wangui Ndung'u, P, Revathi, G, Gehre, F, Akter, S, Ntoumi, F, Stewart-Isherwood, L, Ntinginya, NE, Rachow, A, Hoelscher, M, Cirillo, DM, Skenders, G, Hoffner, S, Bakonyte, D, Stakenas, P, Diel, R, Crudu, V, Moldovan, O, Al-Hajoj, S, Otero, L, Barletta, F, Jane Carter, E, Diero, L, Supply, P, Comas, I, Niemann, S, Gagneux, S (2016) *Mycobacterium tuberculosis* lineage 4 comprises globally distributed and geographically restricted sublineages. *Nature genetics*, **48**(12): 1535–1543.

Stucki, D, Malla, B, Hostettler, S, Huna, T, Feldmann, J, Yeboah-Manu, D, Borrell, S, Fenner, L, Comas, I, Coscollà, M, Gagneux, S (2012) Two new rapid SNP-typing methods for classifying *Mycobacterium tuberculosis* complex into the main phylogenetic lineages. *PloS one*, **7**(7): e41253.

Supply, P, Marceau, M, Mangenot, S, Roche, D, Rouanet, C, Khanna, V, Majlessi, L, Criscuolo, A, Tap, J, Pawlik, A, Fiette, L, Orgeur, M, Fabre, M, Parmentier, C, Frigui, W, Simeone, R, Boritsch, EC, Debrie, AS, Willery, E, Walker, D, Quail, MA, Ma, L,

Bouchier, C, Salvignol, G, Sayes, F, Cascioferro, A, Seemann, T, Barbe, V, Locht, C, Gutierrez, MC, Leclerc, C, Bentley, SD, Stinear, TP, Brisse, S, Médigue, C, Parkhill, J, Cruveiller, S, Brosch, R (2013) Genomic analysis of smooth tubercle bacilli provides insights into ancestry and pathoadaptation of *Mycobacterium tuberculosis*. *Nature genetics*, **45**(2): 172–9.

Supply, P, Mazars, E, Lesjean, S, Vincent, V, Gicquel, B, Locht, C (2000) Variable human minisatellite-like regions in the *Mycobacterium tuberculosis* genome. *Molecular microbiology*, **36**(3): 762–71.

Suzuki, R, Shimodaira, H (2006) Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*, **22**(12): 1540–1542.

Trunz, BB, Fine, P, Dye, C (2006) Effect of BCG vaccination on childhood tuberculous meningitis and miliary tuberculosis worldwide: a meta-analysis and assessment of cost-effectiveness. *Lancet*, **367**(9517): 1173–80.

UNESCO (1992) *General history of Africa-V Africa from the Sixteenth to the Eighteenth Century*. Tech. rep.

Vos, J (2012) From Africa to Brazil: Culture, Identity, and an Atlantic Slave Trade, 1600-1830. By Walter Hawthorne (New York: Cambridge University Press, 2010. xxi plus 259 pp. $25.99). *Journal of Social History*, **46**(1): 261–264.

Walker, TM, Ip, CLC, Harrell, RH, Evans, JT, Kapatai, G, Dedicoat, MJ, Eyre, DW, Wilson, DJ, Hawkey, PM, Crook, DW, Parkhill, J, Harris, D, Walker, AS, Bowden, R, Monk, P, Smith, EG, Peto, TEA (2013) Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study. *The Lancet Infectious diseases*, **13**(2): 137–46.

Walker, TM, Kohl, TA, Omar, SV, Hedge, J, Del Ojo Elias, C, Bradley, P, Iqbal, Z, Feuerriegel, S, Niehaus, KE, Wilson, DJ, Clifton, DA, Kapatai, G, Ip, CLC, Bowden, R, Drobniewski, FA, Allix-Béguec, C, Gaudin, C, Parkhill, J, Diel, R, Supply, P, Crook, DW, Smith, EG, Walker, AS, Ismail, N, Niemann, S, Peto, TEA (2015) Whole-genome sequencing for prediction of *Mycobacterium tuberculosis* drug susceptibility and resistance: a retrospective cohort study. *The Lancet Infectious Diseases*, **15**(10): 1193–1202.

Wampande, EM, Mupere, E, Debanne, SM, Asiimwe, BB, Nsereko, M, Mayanja, H, Eisenach, K, Kaplan, G, Boom, HW, Gagneux, S, Joloba, ML (2013) Long-term dominance of *Mycobacterium tuberculosis* Uganda family in peri-urban Kampala-Uganda is not associated with cavitary disease. *BMC infectious diseases*, **13**(1): 484.

Wejse, C, Gustafson, P, Nielsen, J, Gomes, VF, Aaby, P, Andersen, PL, Sodemann, M (2008) TBscore: Signs and symptoms from tuberculosis patients in a low-resource

setting have predictive value and may be used to assess clinical course. *Scandinavian journal of infectious diseases*, **40**(2): 111–20.

Weyer, K, Carai, S, Nunn, P (2011) Viewpoint TB diagnostics: what does the world really need? *The Journal of infectious diseases*, **204**(4): S1196–202.

WHO (2017) *Global tuberculosis report. Geneva: World Health Organization.*

WHO (2018) *Global tuberculosis report. Geneva: World Health Organization.*

Wirth, T, Hildebrand, F, Allix-Béguec, C, Wölbeling, F, Kubica, T, Kremer, K, Soolingen, D van, Rüsch-Gerdes, S, Locht, C, Brisse, S, Meyer, A, Supply, P, Niemann, S (2008) Origin, spread and demography of the *Mycobacterium tuberculosis* complex. *PLoS pathogens*, **4**(9): e1000160.

Woolhouse, MEJ, Webster, JP, Domingo, E, Charlesworth, B, Levin, BR (2002) Biological and biomedical implications of the co-evolution of pathogens and their hosts. *Nature genetics*, **32**(4): 569–77.

Yang, C, Luo, T, Sun, G, Qiao, K, Sun, G, DeRiemer, K, Mei, J, Gao, Q (2012) *Mycobacterium tuberculosis* Beijing strains favor transmission but not drug resistance in China. *Clinical infectious diseases*, **55**(9): 1179–87.

Yates, TA, Abubakar, I, Tanser, F (2015) HIV infection and the transmission of tuberculosis. *The Journal of infectious diseases*, **211**(9): 1510.

Yates, TA, Khan, PY, Knight, GM, Taylor, JG, McHugh, TD, Lipman, M, White, RG, Cohen, T, Cobelens, FG, Wood, R, Moore, DAJ, Abubakar, I (2016) The transmission of *Mycobacterium tuberculosis* in high burden settings. *The Lancet. Infectious diseases*, **16**(2): 227–38.

Yen, S, Bower, JE, Freeman, JT, Basu, I, O'Toole, RF (2013) Phylogenetic lineages of tuberculosis isolates in New Zealand and their association with patient demographics. *The international journal of tuberculosis and lung disease*, **17**(7): 892–7.

Young, DB, Perkins, MD, Duncan, K, Barry, CE (2008) Confronting the scientific obstacles to global control of tuberculosis. *The Journal of clinical investigation*, **118**(4): 1255–65.

Yu, Y, Harris, A, Blair, C, He, X (2015) RASP (Reconstruct Ancestral State in Phylogenies): A tool for historical biogeography. *Molecular Phylogenetics and Evolution*, **87** 46–49.

Zhang, H, Li, D, Zhao, L, Fleming, J, Lin, N, Wang, T, Liu, Z, Li, C, Galwey, N, Deng, J, Zhou, Y, Zhu, Y, Gao, Y, Wang, T, Wang, S, Huang, Y, Wang, M, Zhong, Q, Zhou, L, Chen, T, Zhou, J, Yang, R, Zhu, G, Hang, H, Zhang, J, Li, F, Wan, K, Wang, J, Zhang, XE, Bi, L (2013) Genome sequencing of 161 *Mycobacterium tuberculosis* isolates from

China identifies genes and intergenic regions associated with drug resistance. *Nature genetics*, **45**(10): 1255–1260.

List of Figures

List of Tables

A. Supplementary Chapter 3

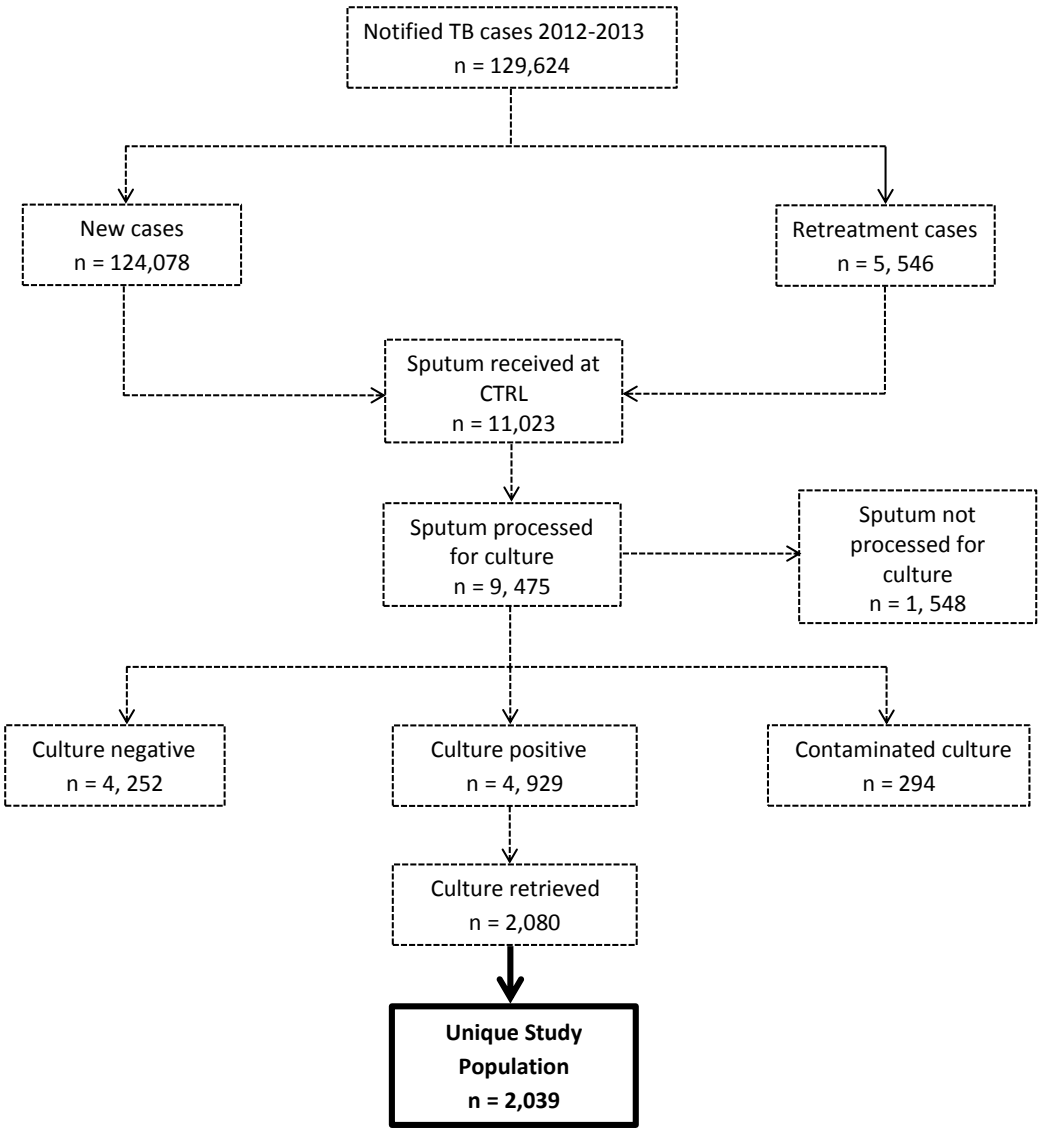


Figure A.1.: Flowchart illustrating estimated notified TB cases in 2012 and 2013 (dashed lines) and the study population (solid line)

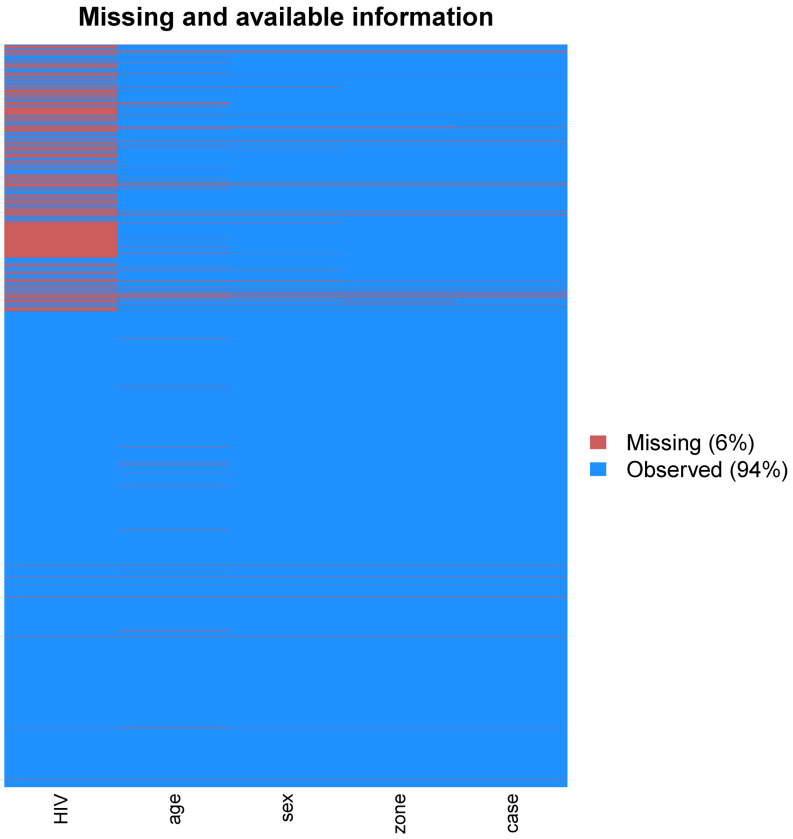


Figure A.2.: Patients' data included in the study. Proportion of observed and missing data for the variables included in the study.

Table A.1.: Sex distribution across the age groups

Age group		Sex	**Total**
	F	M	
Child age (<15)	82 (42.5)	111 (57.5)	193
Young age (15 - 24)	208 (35.9)	372 (64.1)	580
Early adult (25 - 44)	290 (30.9)	647 (69.1)	937
Late adult (45 – 64)	46 (23.5)	150 (76.5)	196
Old age (>65)	13 (27.1)	35 (72.9)	48
Total	639 (32.7)	1315 (67.3)	1954

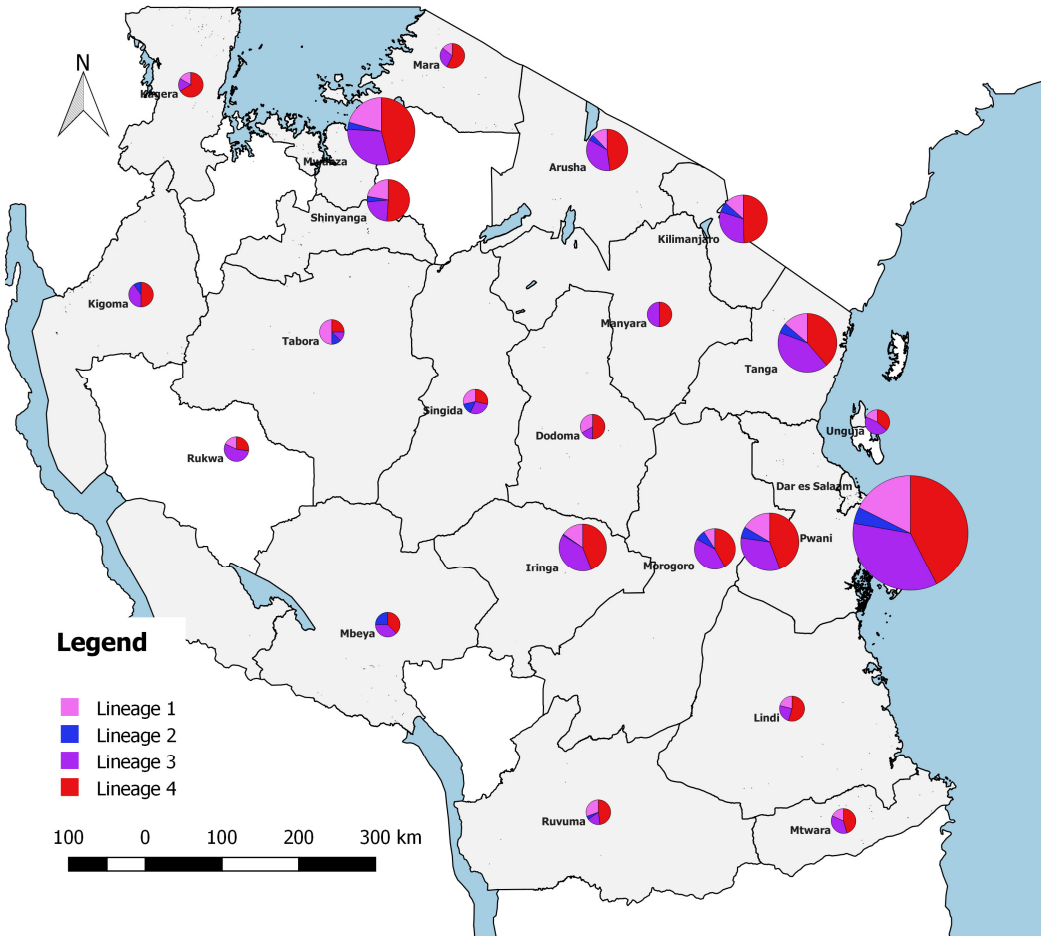


Figure A.3.: MTBC lineage proportions. Distribution of MTBC lineages across different regions of Tanzania. Size of the circle is proportional to the number of isolates analyzed from the regions. MTBC lineage proportions.

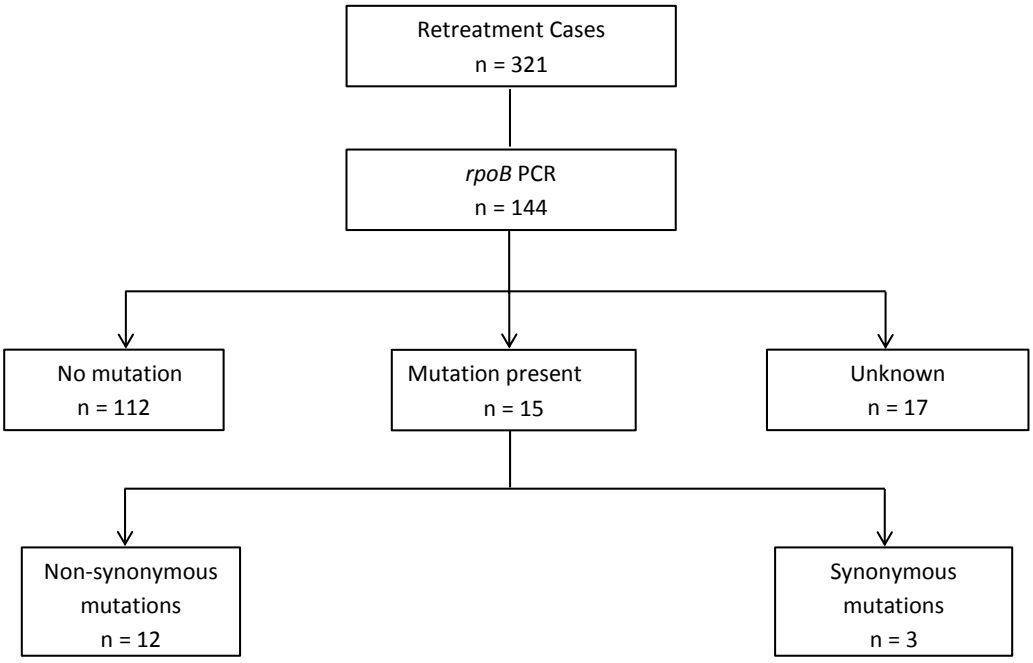


Figure A.4.: Flowchart of genotyped strains for *rpoB* mutations. A subset of MTBC strains from retreatment cases included for *rpoB* drug resistance genotyping.

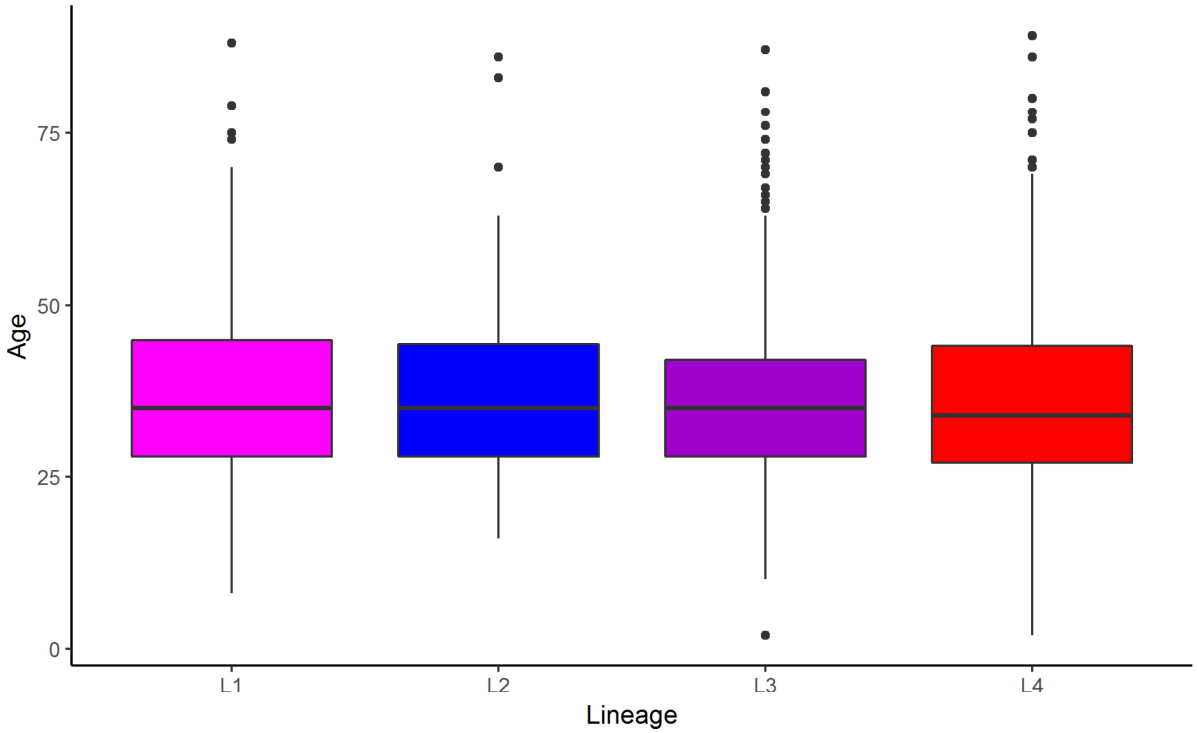
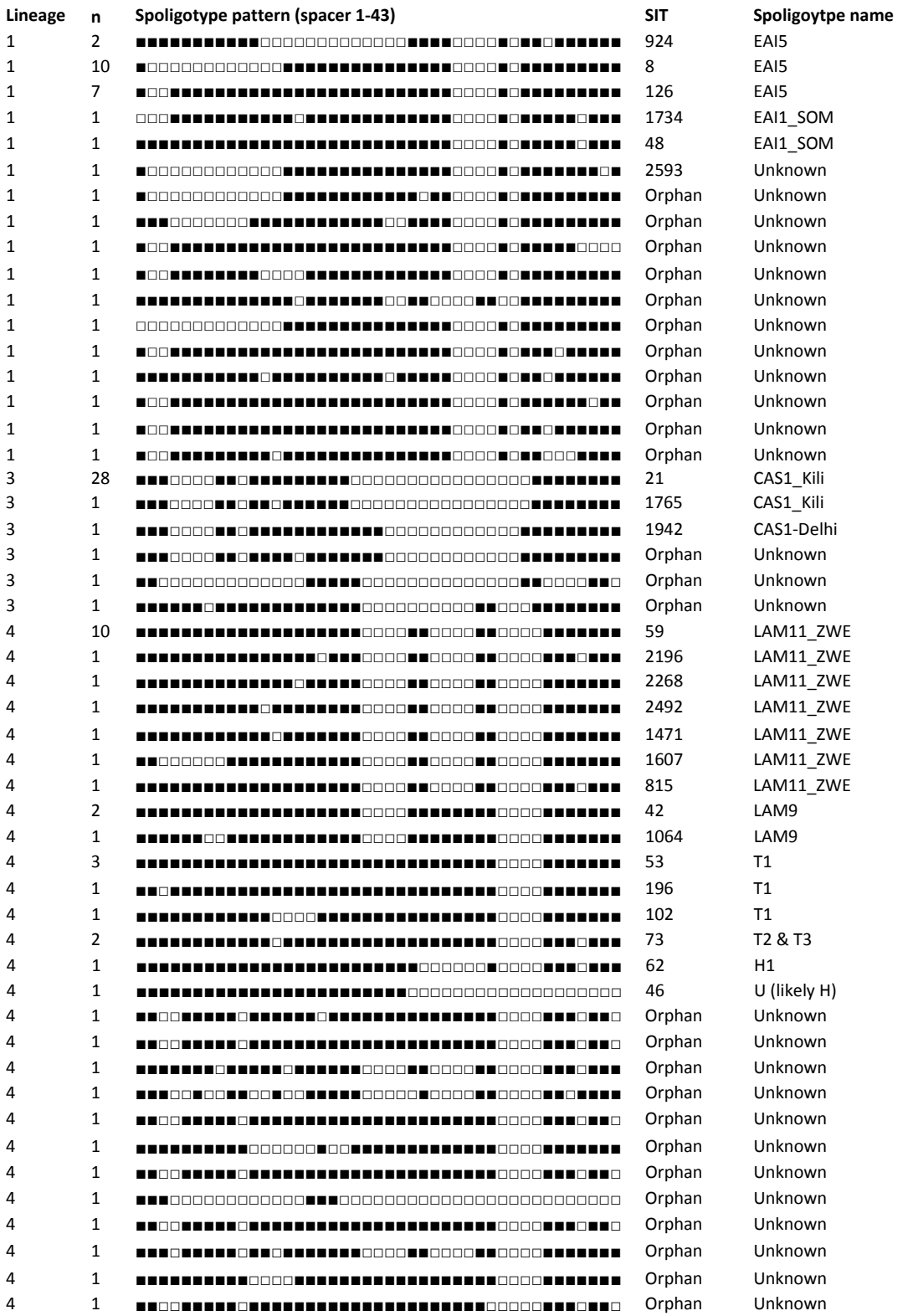


Figure A.5.: Patients' age distribution across MTBC lineages. The age distributions of TB patients grouped by infecting MTBC lineage

Table A.2.: Mutations detected in the *rpoB* gene.

Patient	**Lineage**	**rpoB mutation**	**Amino acid change**
315	L2	A1198G;C1349T	T400A;S450L
446	L3	C1386T	No change
470	L4	G1683A	No change
620	L4	A1334T	H445L
626	L4	G1333C	H445D
718	L2	C1333T	H445Y
719	L4	G1683A	No change
720	L2	C1349T	S450L
1649	L3	T1289C	L430P
1816	L2	C1349T	S450L
1822	L3	C1333T	H445Y
1834	L2	C1349T	S450L
1843	L3	C1349T	S450L
1927	L4	C1294G;A1442G;G1683A	Q432E;E481A; No change
1949	L4	C1333T	H445Y



Lineage	n	Spoligotype pattern (spacer 1-43)	SIT	Spoligoytpe name
1	2	■■■■■■■■■■■□□□□□□□□□□□□□□■■■■□□□□■□■■□■■■■■■	924	EAI5
1	10	■□□□□□□□□□□□□□■■■■■■■■■■■■■■□□□□■□■■■■■■■■■	8	EAI5
1	7	■□□■■■■■■■■■■■■■■■■■■■■■■■□□□□■□■■■■■■■■■	126	EAI5
1	1	□□□■■■■■■■■■■■□■■■■■■■■■■■■■□□□□■□■■■■■■□■■■	1734	EAI1_SOM
1	1	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■□■■■■■■□■■■	48	EAI1_SOM
1	1	■□□□□□□□□□□□□□■■■■■■■■■■■■■■□□□□■□■■■■■■■□■	2593	Unknown
1	1	■□□□□□□□□□□□□□■■■■■■■■■■■■□■■□□□□■□■■■■■■■■■	Orphan	Unknown
1	1	■■■□□□□□□□■■■■■■■■■■■■□□■■■■□□□□■□■■■■■■■■■	Orphan	Unknown
1	1	■□□■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■□■■■■■■□□□□	Orphan	Unknown
1	1	■□□■■■■■■■■□□□□■■■■■■■■■■■■■□□□□■□■■■■■■■■■■	Orphan	Unknown
1	1	■■■■■■■■■■■■■□■■■■■■■□□■■□□□□■■□□■■■■■■■■■■	Orphan	Unknown
1	1	□□□□□□□□□□□□□□■■■■■■■■■■■■■■□□□□■□■■■■■■■■■■	Orphan	Unknown
1	1	■□□■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■□■■■□■■■■■	Orphan	Unknown
1	1	■■■■■■■■■■■□■■■■■■■■■□■■■■■□□□□■□■■□■■■■■■■	Orphan	Unknown
1	1	■□□■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■□■■■■■■■□■■	Orphan	Unknown
1	1	■□□■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■□■■□■■■■■■■	Orphan	Unknown
1	1	■□□■■■■■■■■■□■■■■■■■■■■■■■■■□□□□■□■■□□□■■■■■	Orphan	Unknown
3	28	■■■□□□□■■□■■■■■■■■□□□□□□□□□□□□□□□□□□■■■■■■■■	21	CAS1_Kili
3	1	■■■□□□□■■□■■□■■■■■■□□□□□□□□□□□□□□□□□□■■■■■■■■	1765	CAS1_Kili
3	1	■■■□□□□■■□■■■■■■■■■■■■□□□□□□□□□□□□□□■■■■■■■■■	1942	CAS1-Delhi
3	1	■■■□□□□■■□■■■■□■■■■■■■□□□□□□□□□□□□□□■■■■■■■■■	Orphan	Unknown
3	1	■■□□□□□□□□□□□□□■■■■■□□□□□□□□□□□□□□□□■■□□□□■■□	Orphan	Unknown
3	1	■■■■■■■□■■■■■■■■■■■□□□□□□□□□□□■■□□□■■■■■■■■■	Orphan	Unknown
4	10	■■■■■■■■■■■■■■■■■■■■□□□□■■□□□□■■□□□□■■■■■■■■	59	LAM11_ZWE
4	1	■■■■■■■■■■■■■■■□■■■□□□□■■□□□□■■□□□□■■■□■■■■	2196	LAM11_ZWE
4	1	■■■■■■■■■■■■■■□■■■■□□□□■■□□□□■■□□□□■■■■■■■■	2268	LAM11_ZWE
4	1	■■■■■■■■■■■□■■■■■■■□□□□■■□□□□■■□□□□■■■■■■■■	2492	LAM11_ZWE
4	1	■■■■■■■■■■■■□■■■■■■■□□□□■■□□□□■■□□□□■■■■■■■■	1471	LAM11_ZWE
4	1	■■□□□□□□■■■■■■■■■■■□□□□■■□□□□■■□□□□■■■■■■■■	1607	LAM11_ZWE
4	1	■■■■■■■■■■■■■■■■■■■□□□□■■□□□□■■□□□□■■■□■■■■	815	LAM11_ZWE
4	2	■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■□□□□■■■■■■■■	42	LAM9
4	1	■■■■■■□□■■■■■■■■■■■□□□□■■■■■■■□□□□■■■■■■■■	1064	LAM9
4	3	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■■	53	T1
4	1	■■□■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■■	196	T1
4	1	■■■■■■■■■■■□□□□■■■■■■■■■■■■■□□□□■■■■■■■■	102	T1
4	2	■■■■■■■■■■■□■■■■■■■■■■■■■■■□□□□■■■□■■■	73	T2 & T3
4	1	■■■■■■■■■■■■■■■■■■■■■■■□□□□□□■□□□□■■■□■■■	62	H1
4	1	■■■■■■■■■■■■■■■■■■■■■■■□□□□□□□□□□□□□□□□□□□	46	U (likely H)
4	1	■■□□■■■■■■□■■■■■■□■■■■■■■■■■■■■■□□□□■■■□■■□	Orphan	Unknown
4	1	■■□□■■■■■■□■■■■■■■■■■■■■■■■■■■■■□□□□■■■□■■□	Orphan	Unknown
4	1	■■■■■■■■□■■■■■■□■■■■■■□□□□■■□□□□■■□□□□■■■□■■■	Orphan	Unknown
4	1	■■■□□■□□■■□□■□□■■■■■□□□□□■□□□□■■□□□□■■□■■■■	Orphan	Unknown
4	1	■■□□■■■■■■□■■■■■■■■■■■■■■■■■■■■■□□□□■■■□■■□	Orphan	Unknown
4	1	■■■■■■■■■■□□□□□□■□□■■■■■■■■■■■■□□□□■■■■■■■■	Orphan	Unknown
4	1	■■□□■■■■■■□■■■■■■■■■■■■■■■■■■■■■□□□□■■■□■■□	Orphan	Unknown
4	1	■■■□□□□□□□□□□□□■■■□□□□□□□□□□□□□□□□□□□□□□□□	Orphan	Unknown
4	1	■■□□■■■■■■□■■■■■■■■■■■■■■■■■■■■■□□□□■■■□■■□	Orphan	Unknown
4	1	■■■□■■■■■■□■■□■■■■■■■□□□□■■□□□□■■□□□□■■■■■■■	Orphan	Unknown
4	1	■■■■■■■■■■□□□□■■■■■■■■■■■■■■■■■□□□□■■■■■■■■	Orphan	Unknown
4	1	■■□□■■■■■■□■■■■■■■■■■■■■■■■■■■■■□□□□■■■□■■□	Orphan	Unknown

Figure A.6.: Spoligotype patterns of a subset of MTBC clinical strains.

Table A.3.: Distribution of *rpoB* mutations across the four *M. tuberculosis* lineages

lineage	*rpoB* mutation		***Total***
	no	yes	
L1	7 (100)	0 (0)	7
L2	17 (77.3)	5 (22.7)	22
L3	51 (94.4)	3 (5.6)	54
L4	39 (90.7)	4 (9.3)	43
Total	114	12	126

B. Supplementary Chapter 6

Multiple Introductions of *Mycobacterium tuberculosis* Lineage 2–Beijing into Africa over centuries

Supplementary Figures 1-14

Supplementary Tables 2, 5 and 6. Table 1, 3 and 4 are available online `https://www.frontiersin.org/articles/10.3389/fevo.2019.00112/full#supplementary-material`

Supplementary Figure B.1
Flowchart showing selection of whole genome sequence dataset included in this study.

Supplementary Figure B.2
Tip randomization test on clock rate with strict molecular clock model.

Supplementary Figure B.3
Tip randomization test on rate mean with the UCLD model.

Supplementary Figure B.4
Dated phylogeny of the 308 L2–Beijing strains.

Supplementary Figure B.5
Results of the ancestral reconstruction of 422 L2–Beijing strains using RASP.

Supplementary Figure B.6
Screen shots of the RASP reconstruction to show examples of estimated most likely geographic origins of Lineage 2-Beijing in Eastern Africa.

Supplementary Figure B.7
Screen shots of the RASP reconstruction to show examples of estimated most likely geographic origins of Lineage 2-Beijing in Southern Africa.

Supplementary Figure B.8
Temporal distribution of the 308 samples with information on isolation dates.

Supplementary Figure B.9
Phylogenetic tree showing the distribution of samples with isolation dates.

Supplementary Figure B.10
Proportion of samples with information on isolation dates across the seven geographical regions.

Supplementary Figure B.11
Pairwise SNP distance of Lineage 2 strains linked to introduction events to Eastern Africa (M5 and M6).

Supplementary Figure B.12
Pairwise SNP distance of Lineage 2 strains linked to introduction events to Africa followed by dispersal within the region (M3, M10 and M13).

Supplementary Figure B.13
Pairwise SNP distance of Lineage 2 strains linked to introduction events to Southern

Africa (M1, M8, M9 and M11).

Supplementary Figure B.14

Phylogeny of the 781 Lineage 2 strains mapped with strains' drug resistance status.

Supplementary Table 1 – Provided as a separate file online
List of 781 Lineage 2 strains used in this study with the associated metadata.

Supplementary Table B.1
Drug resistance status of Lineage 2 samples across the seven geographical regions.

Supplementary Table 3 – Provided as a separate file online
Parameter estimates obtained from the UCLD clock model.

Supplementary Table 4 – Provided as a separate file online
Comparison of the prior and posterior distribution for all parameter estimates.

Supplementary Table B.2
Lineage 2 sublineage proportions across the seven geographical regions.

Supplementary Table B.3
Estimate for time to the MRCA of the African Lineage 2-Beijing clades using different nucleotide substitution and demographic models.

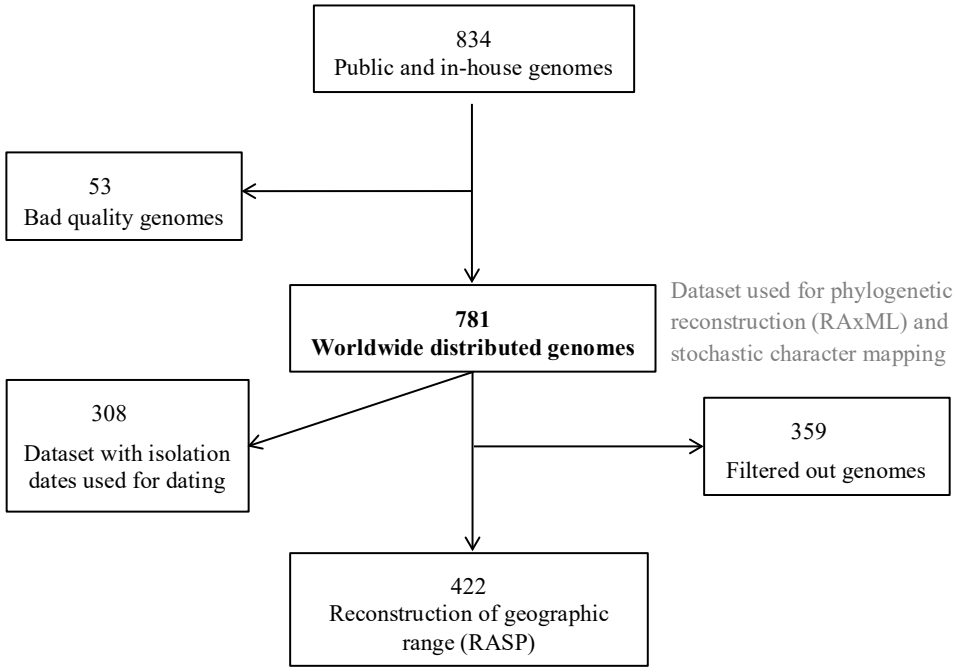


Figure B.1.: Flow chart showing the selection of whole genome sequence dataset from previous studies and current study.

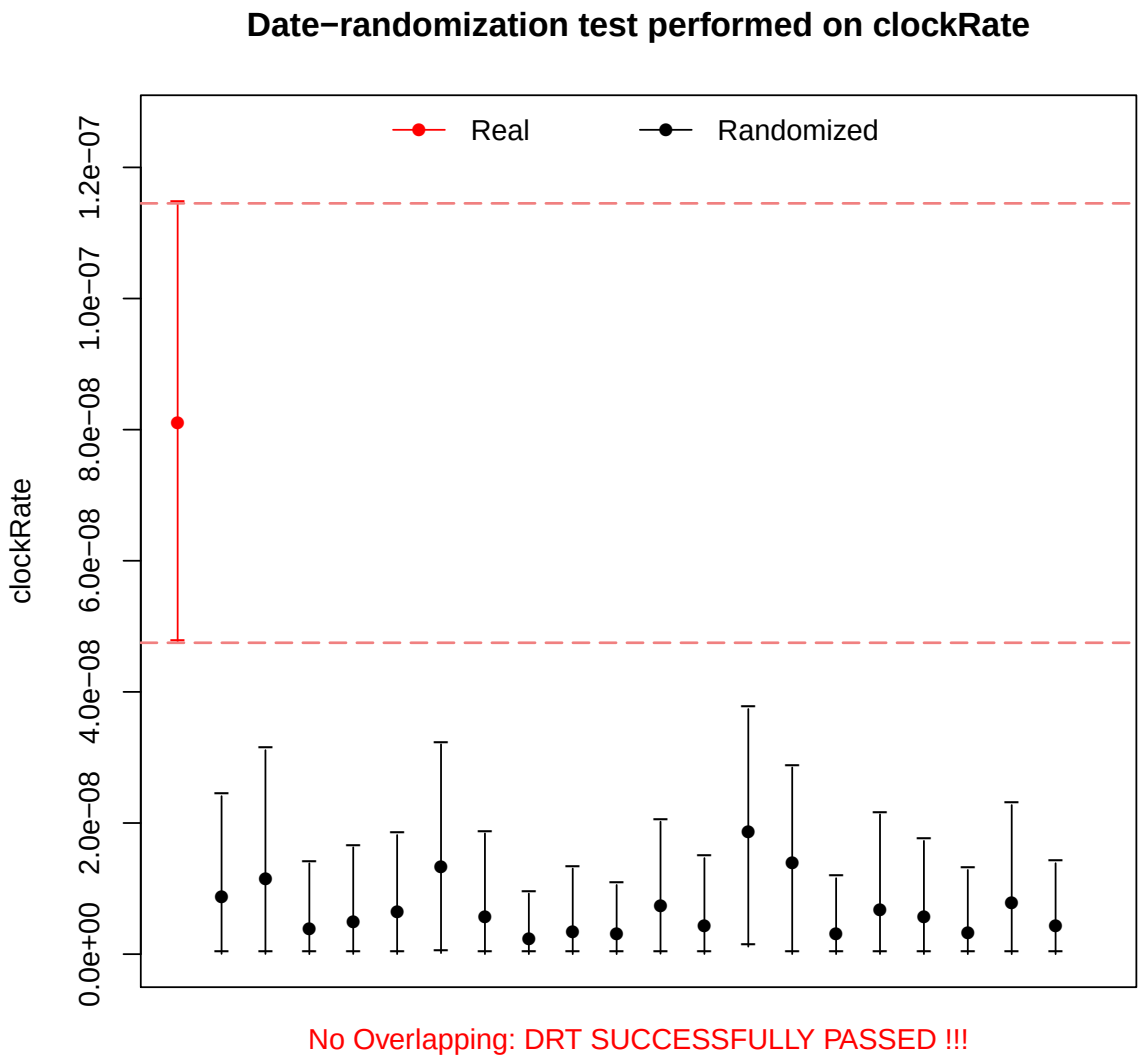


Figure B.2.: Tip randomization test on clock rate with strict molecular clock model. Error bars are the 95% HPD.

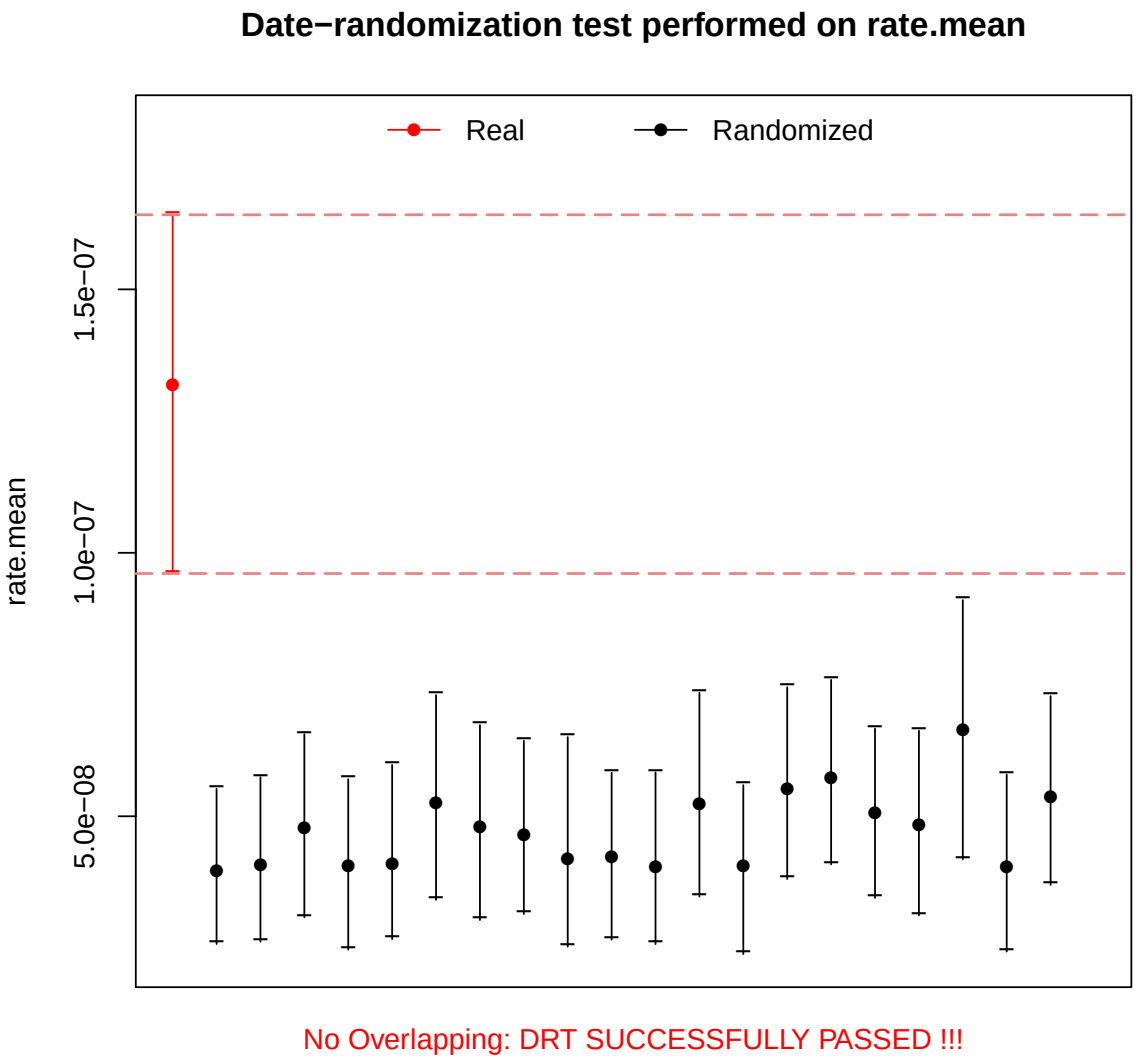


Figure B.3.: Tip randomization test on rate mean with the UCLD model. Error bars are the 95% HPD.

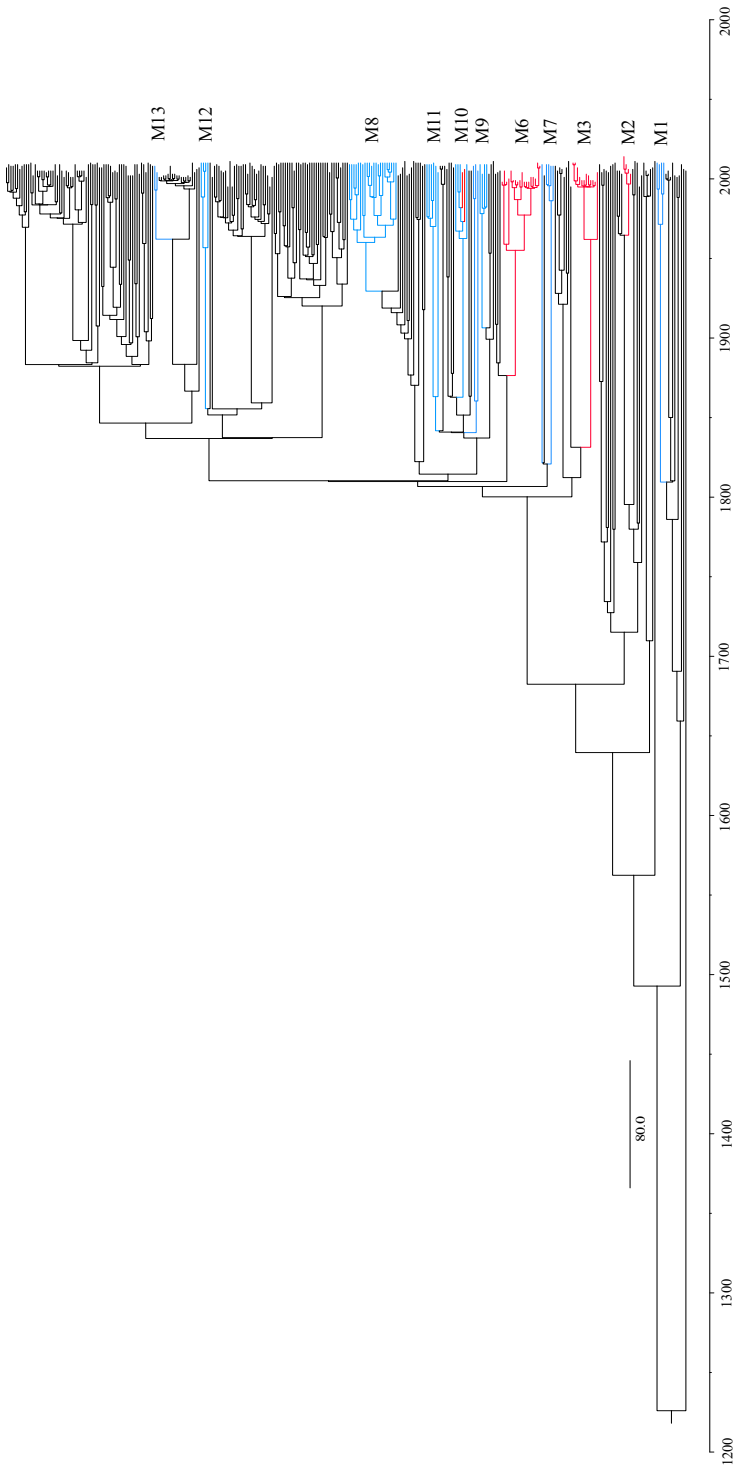


Figure B.4.: Dated phylogeny of the 308 L2–Beijing strains with African clades highlighted. Color codes are the same as described in Figure 6.1.

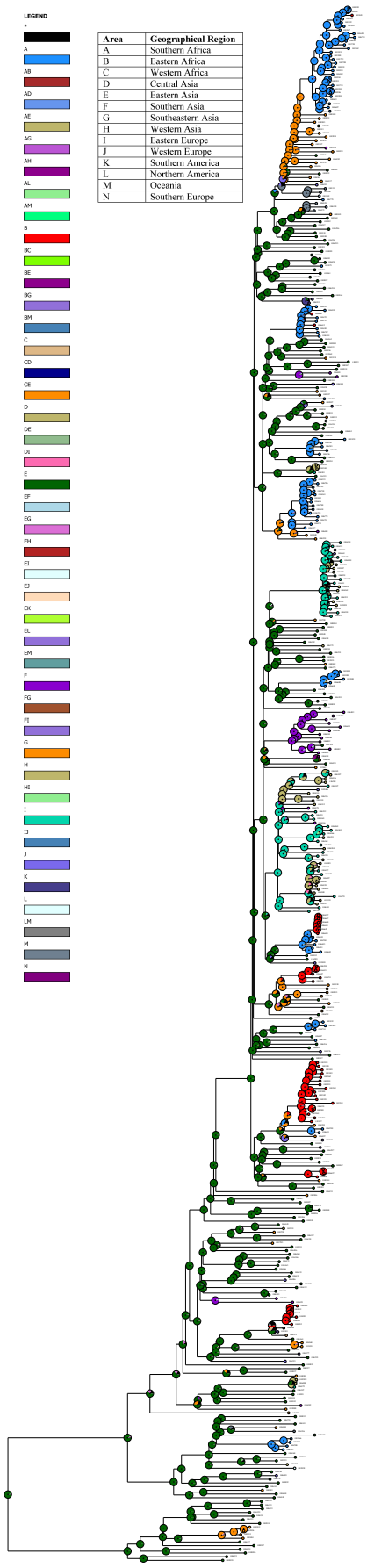


Figure B.5.: Ancestral reconstruction of 422 L2–Beijing strains using RASP software.

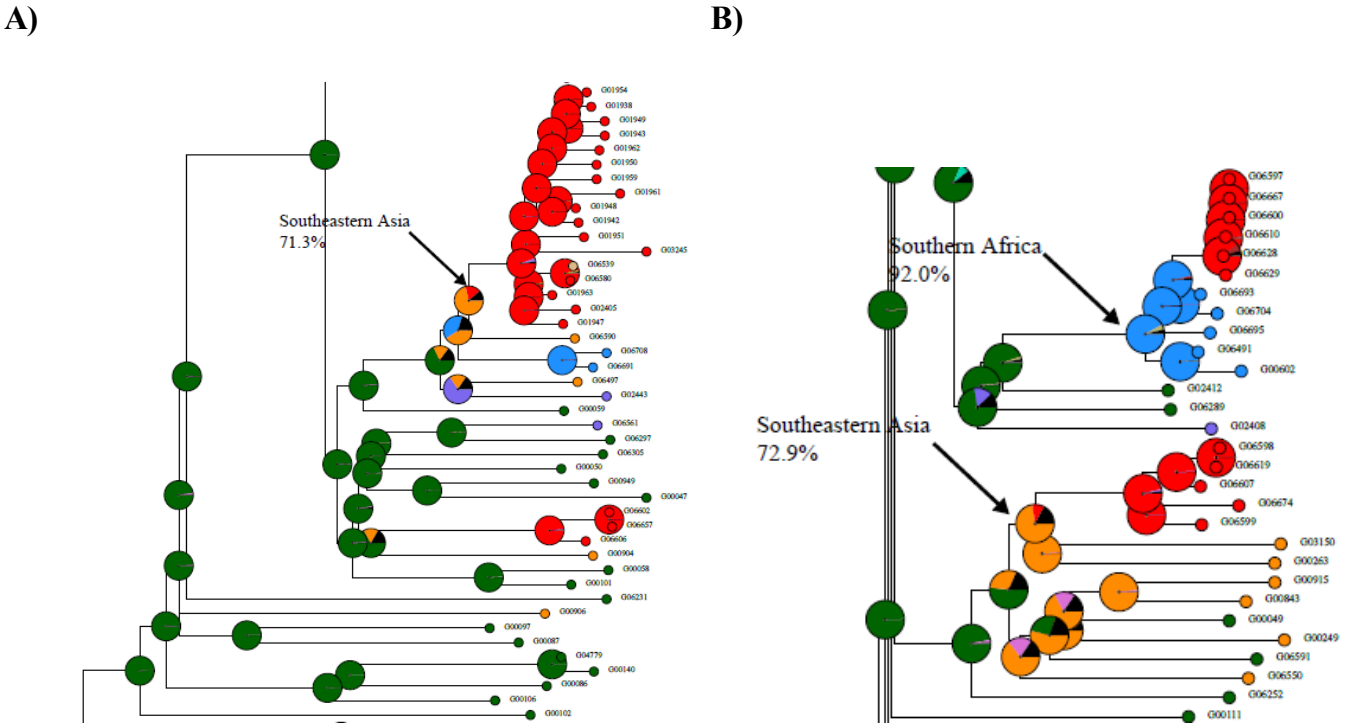


Figure B.6.: Screen shots of the RASP reconstruction to show examples of estimated most likely geographic origins of Lineage 2-Beijing in Eastern Africa A) Direct introduction and (B) dispersal within African region. Pie-charts indicate the reconstructed ancestral geographical range at the nodes. Color codes are the same as described in Figure 6.1.

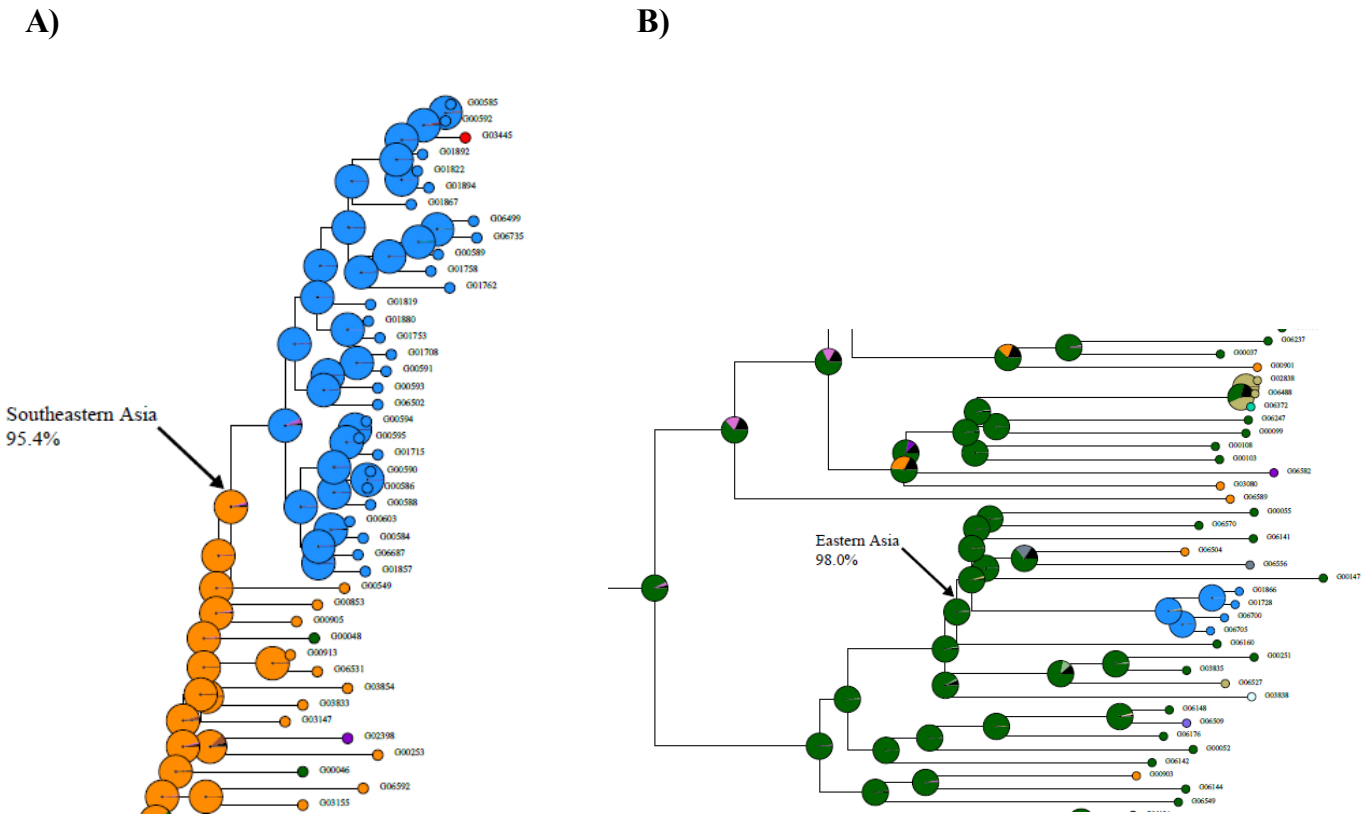


Figure B.7.: Screen shots of the RASP reconstruction to show examples of estimated most likely geographic origins of Lineage 2-Beijing in Southern Africa. Direct introductions from A) Southeastern Asia and (B) Eastern Asia. Pie-charts indicate the reconstructed ancestral geographical range at the nodes. Color codes are the same as described in Figure 6.1.

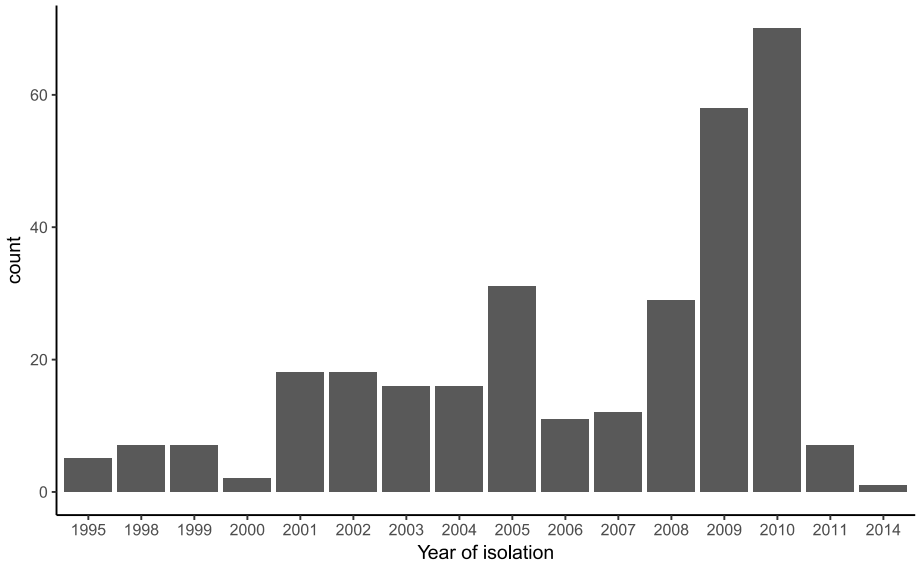


Figure B.8.: Temporal distribution of the 308 samples with information on isolation dates.

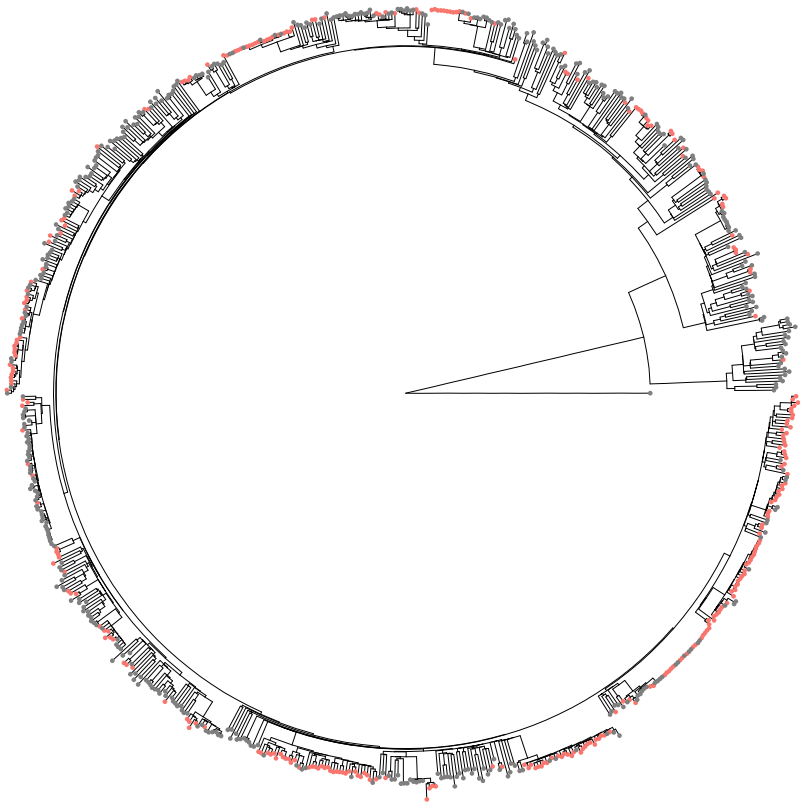


Figure B.9.: Phylogenetic tree of the 781 Lineage 2 showing the distribution of samples with isolation dates. Taxa labeled red are samples with isolation date information and those in grey are samples with missing date information.

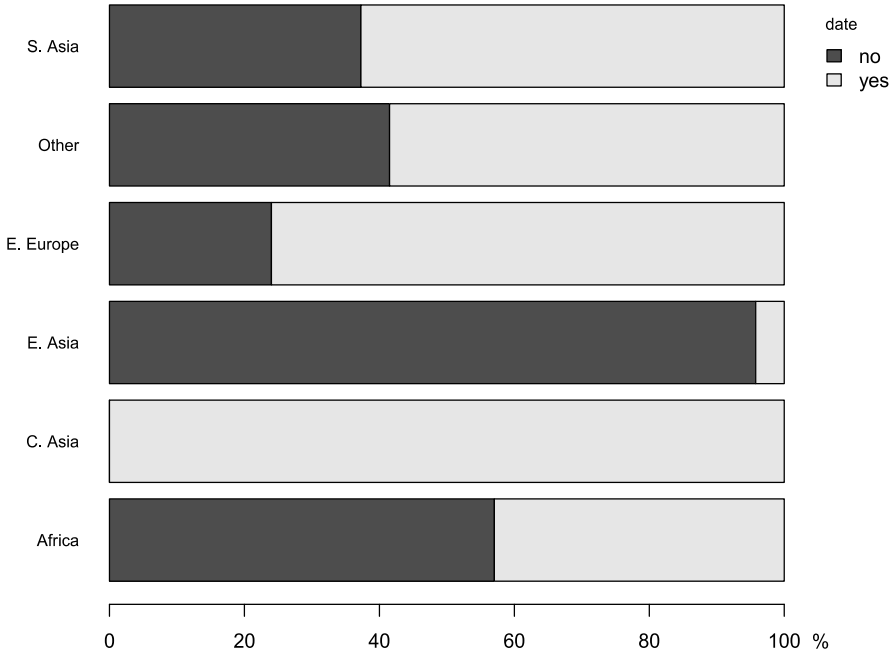


Figure B.10.: Sample proportion with information on isolation dates across the seven geographical regions.

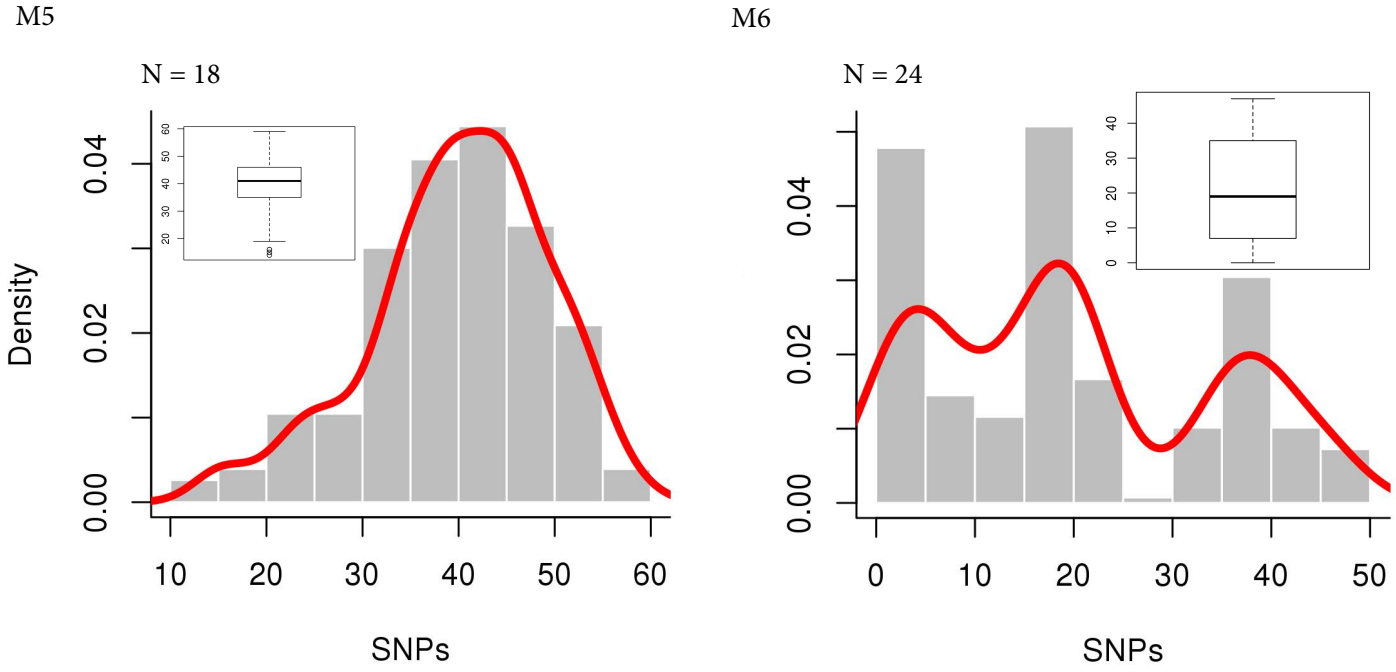


Figure B.11.: Pairwise SNP distance of Lineage 2 strains linked to Eastern Africa introductions (M5 and M6).

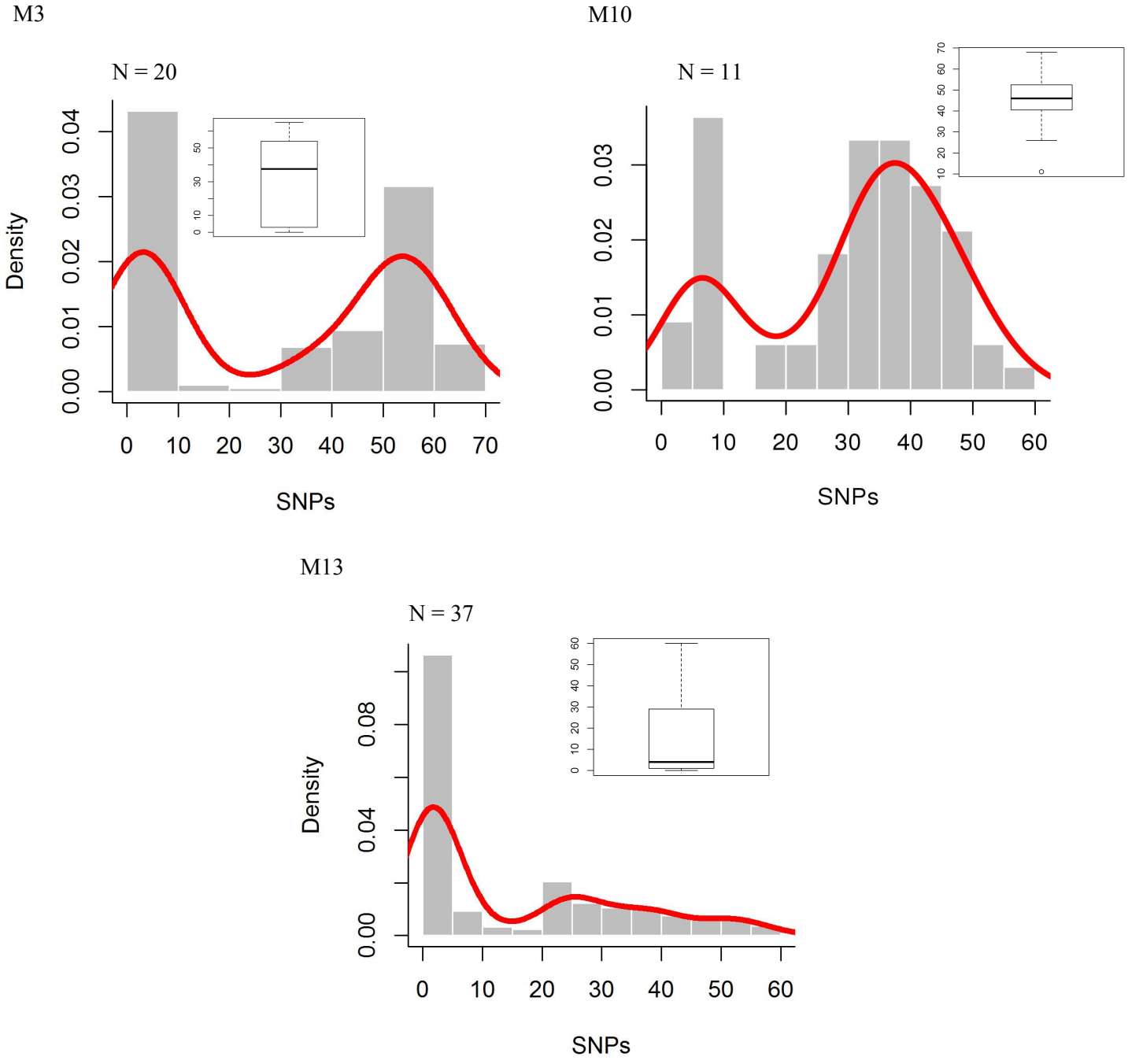


Figure B.12.: Pairwise SNP distance of Lineage 2 strains linked to introductions and dispersal in Africa (M3, M10 and M13).

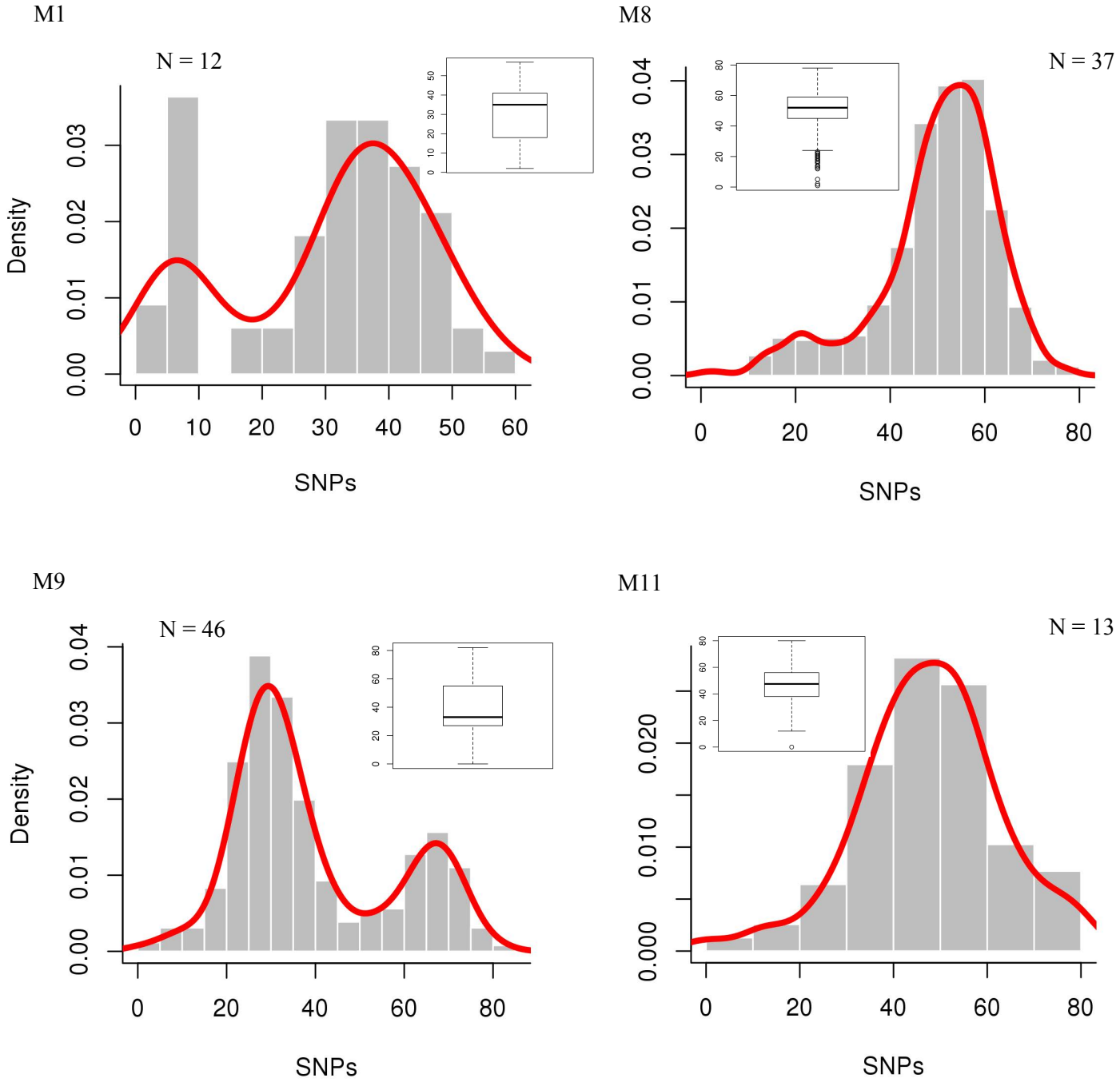


Figure B.13.: Pairwise SNP distance of Lineage 2 strains linked to Southern Africa introductions (M1, M8, M9 and M11).

Figure B.14.: Phylogeny of the 781 Lineage 2 strains mapped with strains' drug resistance status.

Table B.1.: Drug resistance status of Lineage 2 samples across the seven geographical regions.

Drug resistance profile	E. Africa	S. Africa	E. Asia	S.E Asia	S. Asia	C. Asia	E. Europe
FQ_monoresistant	0	0	5	0	0	0	0
INH_monoresistant	0	7	5	7	3	8	14
mdr	0	28	40	0	6	22	32
polyresistant	0	0	6	0	0	0	2
RIF_monoresistant	0	12	7	0	1	0	0
susceptible	92	107	170	45	38	4	16
xdr	0	10	27	1	1	1	11
TOTAL	92	164	260	53	49	35	75

Table B.2.: Lineage 2 sublineage proportions across the seven geographical regions.

Sublineage	Eastern Africa	Southern Africa	Eastern Asia	S.eastern Asia	Southern Asia	Central Asia	Eastern Europe
L2.1 (Proto-beijing)	0	0	20	4	0	0	0
L2.2.1 (Asia Ancestral 1)	0	12	23	4	0	2	0
L2.2.2 (Asia Ancestral 2)	0	0	5	1	1	0	2
L2.2.3 (Asia Ancestral 3)	5	2	34	5	3	0	0
L2.2.4 (Asian African 1)	18	6	2	2	0	0	0
L2.2.5 (Asian African 3)	23	0	7	6	0	0	0
L2.2.6 (Pacific RD150)	0	38	19	15	1	0	1
L2.2.7 (Asian African 2)	3	69	40	5	4	0	1
L2.2.8 (Asian African 2/RD142)	0	0	4	2	0	0	0
L2.2.9 (B0/W148)	0	0	0	0	0	3	24
L2.2.10 (Central Asia)	0	0	1	0	2	29	44
Other	43	37	105	9	38	1	3
TOTAL	92	164	260	53	49	35	75

Table B.3.: Estimates for time to the MRCA of the African Lineage 2-Beijing clades using different nucleotide substitution and demographic models.

Migration	**TMRCA**			**95% HPD**		
	UCLD	GTR	Exp	UCLD	GTR	Exp
M1	1809	1779	1786	1732-1874	1683-1863	1694-1863
M2	1964	1958	1956	1946-1980	1936-1978	1934-1976
M3	1831	1804	1806	1768-1886	1723-1874	1727-1872
M6	1876	1856	1857	1828-1920	1796-1910	1798-1909
M7	1820	1791	1793	1752-1876	1712-1869	1715-1867
M8	1929	1917	1915	1899-1955	1880-1951	1879-1948
M9	1906	1891	1889	1865-1942	1839-1936	1840-1933
M10	1862	1840	1842	1809-1909	1777-1902	1778-1898
M11	1863	1840	1842	1809-1910	1775-1901	1777-1900
M12	1855	1832	1833	1801-1903	1765-1896	1766-1890
M13	1962	1955	1954	1942-1979	1931-1978	1929-1976

C. List of Publications

Haraka, F, **Rutaihwa LK**, Battegay M, Reither K. (2012) '*Mycobacterium intracellulare* infection in non-HIV infected patient in a region with a high burden of tuberculosis', Case Reports, 2012(may07 1), pp. bcr0120125713–bcr0120125713. doi: 10.1136/bcr.01.2012.5713.

Mhalu, G, Hella J, Doulla B, Mhimbira F, Mtutu H, Hiza H, Sasamalo M, **Rutaihwa LK**. et al. (2015) 'Do Instructional Videos on Sputum Submission Result in Increased Tuberculosis Case Detection? A Randomized Controlled Trial', PLOS ONE. Edited by T. M. Doherty, 10(9), p. e0138413. doi: 10.1371/journal.pone.0138413.

Pohl, C, **Rutaihwa LK**. et al. (2016) 'Limited value of whole blood Xpert® MTB/RIF for diagnosing tuberculosis in children', Journal of Infection, 73(4), pp. 326–335. doi: 10.1016/j.jinf.2016.04.041.

Stucki D, Brites D, Jeljeli L, Coscolla M, Liu Q, Trauner A, Fenner L, **Rutaihwa LK**. et al. (2016) '*Mycobacterium tuberculosis* lineage 4 comprises globally distributed and geographically restricted sublineages.', Nature genetics, 48(12), pp. 1535–1543. doi: 10.1038/ng.3704.

Conceição EC, Guimarães AES, Lopes ML, Furlaneto IP, Rodrigues YC, Conceição ML, Barros WA, Cardoso NC, Sharma A, Limab LNGC, Gomes HM, Duartea RS, Frota C, **Rutaihwa LK**. et al. (2018) 'Analysis of potential household transmission events of tuberculosis in the city of Belem, Brazil', Tuberculosis. Churchill Livingstone, 113, pp. 125–129. doi: 10.1016/j.tube.2018.09.011.

Menardo F, Loiseau C, Brites D, Coscolla M, Gygli SM, **Rutaihwa LK**. et al. (2018) 'Treemmer: a tool to reduce large phylogenetic datasets with minimal loss of diversity', BMC Bioinformatics, 19(1), p. 164. doi: 10.1186/s12859-018-2164-8.

Sikalengo G, Hella J, Mhimbira F, **Rutaihwa LK**. et al. (2018) 'Distinct clinical characteristics and helminth co-infections in adult tuberculosis patients from urban compared to rural Tanzania.', Infectious diseases of poverty, 7(1), p. 24. doi: 10.1186/s40249-018-0404-9.

Rutaihwa, LK., Sasamalo, M., et al. (2019) 'Insights into the genetic diversity of *Mycobacterium tuberculosis* in Tanzania.', PloS one. Edited by R. Manganelli. Public Library of Science, 14(4), p. e0206334. doi: 10.1371/journal.pone.0206334.

Rutaihwa, LK., Menardo, F., et al. (2019) 'Multiple Introductions of *Mycobacterium tuberculosis* Lineage 2–Beijing Into Africa Over Centuries', Frontiers in Ecology and Evolution. Frontiers, 7, p. 112. doi: 10.3389/fevo.2019.00112.