UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Genomic insights into the global evolution and antibiotic resistance of the *Mycobacterium tuberculosis* complex

by

Dillon Muzondiwa

Submitted in fulfilment of the requirements for the degree
Doctor of Philosophy (Bioinformatics)

in the

Faculty of Natural and Agricultural Sciences

UNIVERSITY OF PRETORIA

February 2025

© University of Pretoria

SUMMARY

GENOMIC INSIGHTS INTO THE GLOBAL EVOLUTION AND ANTIBIOTIC RESISTANCE OF THE *MYCOBACTERIUM TUBERCULOSIS* COMPLEX

by

Dillon Muzondiwa

Supervisor: Prof. Oleg Reva

Department: Department of Biochemistry, Genetics and Microbiology

University: University of Pretoria

Degree: Doctor of Philosophy (Bioinformatics)

Keywords: *Mycobacterium tuberculosis*, SNP, Maximum Likelihood, Bayesian, Phylogeny, Lineage, BEAST, Drug resistance, Phylogeography, Mutations, MDR-TB, XDR-TB

Mycobacterium tuberculosis (Mtb) recently reclaimed its status as the leading global cause of death from a single infectious agent after three years of COVID-19 holding the top position. We used Mtb whole genome sequencing data (WGS) to explore the diversity of human-adapted lineages of Mtb. Using publicly available datasets, we curated and characterised a large global WGS dataset of more than 9000 Mtb strains sampled across the globe. Based on the distribution of single nucleotide polymorphisms, we performed lineage classification, drug resistance predictions and molecular clock estimations to characterise the global evolution of Mtb and create a non-redundant global reference dataset. Our data suggested that public Mtb WGS datasets are highly redundant, and redundancy minimisation is required before analysing large datasets.

We next sought to explore the evolutionary dynamics that shaped the genetic landscape of the African continent which has been suggested as the origins of Mtb. We demonstrate that Lineage 2 and Lineage 4 are the most dominant on the continent. Using Maximum Likelihood and Bayesian phylogenetic techniques, we mapped identified drug resistance-

© University of Pretoria

associated mutations to time-resolved phylogenies. We estimated that drug resistance on the continent emerged at multiple events, with the earliest emergence of drug resistance occurring in the mid-20th century. We also identified the presence of resistance mutations associated with recently introduced drugs in isolates that were sampled prior to the use of these drugs. Using Bayesian skyline coalescent inference, we observed an expansion in the Mtb population in Africa in timelines that coincided with increased migration from Europe and Asia into Africa. We also inferred a population expansion of Mtb at the time when HIV prevalence was at its peak on the continent.

We next sought to understand the evolutionary dynamics of Lineage 2 and Lineage 4 Mtb in the Southern Africa Development Community (SADC) region, a part of the continent which carries the highest burden of HIV/TB coinfection. We demonstrate that the heterogeneity of Mtb Lineage 2 diversity in the SADC region is under-characterised. We identify 13 sublineages of Lineage 2 in the region from our analysis. To explore the origins of SADC Lineage 2 and Lineage 4, we employed two phylogeographic approaches and both of them place East Asia as the origin of Lineage 2 and Europe as the origin of Lineage 4. We also infer that the two lineages were introduced through multiple introduction events with South Africa as a central hub for the dispersion of the lineages northwards. Taken together, our phylogeographic analysis and our Bayesian skyline results suggest that migration and colonialism played a role in shaping the diversity of the two Lineages in SADC.

Lastly, using mathematical models, drug susceptibility testing data and genomic data, we sought to model the epistatic dynamics that govern drug resistance in Mtb. We obtained co-dependency estimates that represent the probability of one mutation emerging after another mutation. We then created networks and traced the trajectories from drug susceptibility status to pre-XDR-TB status.

© University of Pretoria

LIST OF ABBREVIATIONS

TB	Tuberculosis
MTBC	*Mycobacterium tuberculosis* complex
Mtb	*Mycobacterium tuberculosis*
SNP	Single Nucleotide Polymorphism
MDR	Multidrug resistance
XDR	Extensive drug resistance
LRS	Long read sequencing
RRDR	Rifampin resistance-determining region
ML	Maximum Likelihood
HIV	Human Immunodeficiency Virus
LCR	Low confidence region
MIC	Maximum Inhibitory concentration

© University of Pretoria

DECLARATION

I Dillon Muzondiwa declare that the thesis, which I hereby submit for the degree Doctor of Philosophy (Bioinformatics) at the University of Pretoria, is my own work and has not previously been submitted by me for a degree at this or any other tertiary institution.

ETHICS STATEMENT

The author, whose name appears on the title page of this thesis, has obtained, for the research described in this work, the applicable research ethics approval.

The author declares that he has observed the ethical standards required in terms of the University of Pretoria's Code of Ethics for Researchers and the Policy guidelines for responsible research.

Signature DMuzondiwa
Student name Dillon Muzondiwa
Month Year February 2025

© University of Pretoria

ACKNOWLEDGEMENTS

This project could not have been completed without the divine strength from the Creator himself.

I would like to thank my Supervisor, Prof Oleg Reva for the support and guidance he has given me over the years. I would also like to thank Prof Fourie Joubert and Dr Rian Pierneef for patiently showing me the ropes, especially when I was first introduced to the field of Bioinformatics. Thanks also to all the colleagues at the Centre for Bioinformatics and Computational Biology at the University of Pretoria for the moral support.

I would also like to appreciate Mr Johann Swart for the tremendous effort he puts to ensure that the required computational resources are always available.

I thank my mother for being the source of comfort through this journey. Finally, special thanks to my " Old Mutual " SH for the emotional support and encouragement throughout this tough journey, I appreciate you.

© University of Pretoria

TABLE OF CONTENTS

CHAPTER 1 TUBERCULOSIS: MOTIVATION 1

1.1 A GLOBAL OVERVIEW OF THE CURRENT SITUATION 2

1.1.1 TB incidence 2

1.1.2 TB mortality 2

1.1.3 Drug-resistant TB 3

1.1.4 Global TB after COVID-19 4

1.2 A HISTORICAL PERSPECTIVE ON TB RESEARCH 4

1.3 HUMAN TB ORIGINS 6

1.4 EVOLUTIONARY HISTORY FROM THE PATHOGEN PERSPECTIVE 7

1.5 THE ROLE OF WGS IN EXPLORING MTBC GENOMIC DIVERSITY 12

1.5.1 WGS and antimicrobial resistance (AMR) 13

1.5.2 WGS and the development of novel anti-TB drugs 13

1.5.3 Challenges in using NGS in TB applications 14

1.6 DRUG RESISTANCE AND TB 15

1.6.1 Mechanism of resistance to anti-TB drugs 17

1.7 MTBC STRAINS AND LINEAGES IN-DEPTH 26

1.7.1 Lineage 1 29

1.7.2 Lineage 2 31

1.7.3 Lineage 3 35

1.7.4 Lineage 4 36

1.7.5 Lineage 5, 6, 7, 8, 9 (*M. africanum*) 40

1.8 PHYLOGENOMICS 42

1.8.1 Neighbour-joining (NJ) 42

1.8.2 Parsimony 43

1.8.3 Maximum likelihood (ML) 43

1.8.4 Rooting and dating of phylogenies 45

1.8.5 Nucleotide Substitution models 47

1.9 BAYESIAN INFERENCE IN PHYLOGENETICS 48

1.9.1 Markov chain Monte Carlo (MCMC) 52

© University of Pretoria

1.10 PHYLOGEOGRAPHY 55
1.11 PROJECT RATIONALE 57
1.12 THESIS AIM AND OBJECTIVES 58

CHAPTER 2 SCALABLE APPROACHES FOR COMPUTATIONAL PHYLOGENOMICS OF LARGE MTB DATASETS 59
2.1 ABSTRACT 61
2.2 INTRODUCTION 61
2.3 MATERIALS AND METHODS 64
2.3.1 Data collection 64
2.3.2 Raw data processing and variant calling 65
2.3.3 Lineage classification, drug resistance profiling 66
2.3.4 Evolutionary analysis and redundancy assessment 66
2.4 RESULTS 67
2.4.1 Lineage distribution 67
2.4.2 Evolutionary analysis 70
2.4.3 Insilico drug resistance prediction 72
2.5 DISCUSSION 76
2.6 CONCLUSION 77

CHAPTER 3 PREDICTING ANTIBIOTIC RESISTANCE AND POPULATION DYNAMICS OF MYCOBACTERIUM TUBERCULOSIS IN SUB-SAHARAN AFRICA: AN IN SILICO APPROACH 79
3.1 ABSTRACT 81
3.2 INTRODUCTION 81
3.3 MATERIALS AND METHODS 84
3.3.1 Data collection, processing and variant calling 84
3.3.2 Evolutionary analysis 85
3.3.3 Bayesian evolutionary analysis 86
3.3.4 Dynamics of effective population size 87
3.4 RESULTS 87
3.4.1 Initial data curation and lineage classification 87
3.4.2 Phylogenetic analysis 88

© University of Pretoria

3.4.3 Lineage distribution and antibiotic resistance profiling 92
3.4.4 Dynamics of Mtb lineage expansion in Africa 97
3.5 DISCUSSION 99
3.6 CONCLUSION 101

CHAPTER 4 PHYLOGEOGRAPHY AND GENETIC DIVERSITY OF MTB LINEAGE 4 AND 2 IN THE SADC REGION 103
4.1 ABSTRACT 105
4.2 INTRODUCTION 105
4.3 MATERIALS AND METHODS 107
4.3.1 Collection of genome sequences of MTBC 107
4.3.2 Sub-lineage classification and Insilico drug resistance profiling 108
4.3.3 Phylogenetic reconstruction 108
4.3.4 Bayesian Molecular Clock Dating 109
4.3.5 Phylogeography 110
4.4 RESULTS 111
4.4.1 Classification and Phylogenetic Inference 111
4.4.2 Phylogeography 118
4.5 DISCUSSION 128

CHAPTER 5 THE EPISTATIC LANDSCAPE OF ANTIBIOTIC RESISTANCE OF DIFFERENT LINEAGES OF MTB 132
5.1 ABSTRACT 133
5.2 INTRODUCTION 133
5.3 MATERIALS AND METHODS 136
5.3.1 Data sourcing 136
5.3.2 Functional associations between Mtb mutations 136
5.3.3 Phylogenetic drug resistance inference 137
5.4 RESULTS 139
5.4.1 Lineage 1 139
5.4.2 Lineage 2 141
5.4.3 Lineage 4 142
5.5 DISCUSSION 144

© University of Pretoria

CHAPTER 6 GENERAL DISCUSSION AND CONCLUSIONS 146
6.1 INTRODUCTION 146
6.2 KEY CONCLUSIONS AND RECOMMENDATIONS 147
6.2.1 Data curation 147
6.2.2 Evolution of Mtb in Africa 149
6.2.3 The genetic diversity of Lineage 2 and Lineage 4 Mtb in the SADC region 151
6.2.4 Epistatic landscape of drug-resistant TB 152
6.3 FUTURE DIRECTIONS AND RESEARCH NEEDS 153
6.4 CLOSING REMARKS 154

CHAPTER 7 REFERENCES 156

CHAPTER 8 APPENDIX A 183
8.1.1 Data and code availability 183

List of tables

Table 1: Summary of the molecular clock parameter estimates in the phylogenetic analysis for 8301 Mtb isolates. The substitution rate is estimated at 7.45 x 10^{-7} substitutions /site/year with a standard deviation of 1 x 10^{-7} indicating the rate of evolutionary change across the sequences. The intercept of -0.00141 represents the baseline divergence at the root of the tree, used in the clock model regression. 71
Table 2: Covariance matrix of the intercept and substitution rate parameters from the molecular clock model. The diagonal elements represent the variances of the intercept (1.007 x 10^{-14}) and substitution rate (4.08x10^{-8}), while the off-diagonal elements (-2.03 x 10^{-11}) represent the covariance between these two parameters, indicating the degree of correlation between the intercept and rate estimates. 71
Table 3: Drug resistance-associated mutations identified in this study across the first-line anti-TB drug. 72

© University of Pretoria

Table 4: Drug resistance-associated mutations identified in this study across the second-line anti-TB drugs. 73
Table 5: Drug resistance-associated mutations identified in this study across the novel anti-TB drugs clofazimine, delamanid, linezolid and bedaquiline. 74
Table 6: Comparison of estimated time of emergence of resistance-associated variants in different lineages of Mtb strains isolated on the African continent. The results show that resistance-associated mutations were acquired chronologically across all lineages with isoniazid resistance-associated mutations emerging before rifampicin resistance-associated mutations 94
Table 7: Mascot analysis results for SADC Lineage 2. Effective sample size (ESS) values for all parameters showed that the analysis produced a large number of independent samples from the posterior (high reliability) 121
Table 8: TreeTime inferred phylogeographic parameters for Lineage 4 Mtb across Southern African regions. The equilibrium frequencies (***πi***) represent the long-term likelihood of isolates being assigned to specific regions. Symmetrized rates (W_{ij}) show bidirectional movement between regions, while directional rates (Qij) indicate the rates of migration from one region to another. Regions correspond to the following countries: A = Botswana, B = Eswatini, C = Madagascar, D = Malawi, E = Mozambique, F = South Africa, G = Tanzania, H = Zimbabwe. 123
Table 9 Mascot analysis results for SADC Lineage 4.3. Effective sample size (ESS) values for all parameters showed that the analysis produced many independent samples from the posterior indicating higher reliability of our analysis. 127
Table 10: Structure of input file required by *augur sequence traits.* 139

List of Figures

Figure 1: A schema of the different *M. tuberculosis* marker types and typing technologies. 8
Figure 2: A timeline of the development of *M. tuberculosis* genotyping approaches. Used from Atavliyeva et .al [44] under the terms of the Creative Commons Attribution Licence (CC BY). 8
Figure 3: Maximum Parsimony tree of *M. tuberculosis* with branch colours representing the main lineages. The main clades are labelled according to the geographical regions from which the lineage dominates. Used from Heshberg *et al.*[50] under the terms of the Creative Commons Attribution Licence (CC BY). 11

© University of Pretoria

Figure 4: Summarises the mechanisms of action of some of the main anti-TB drugs currently in clinical use. These can be classified into four groups mainly: i) inhibition of cell wall biosynthesis; ii) disruption of cell membranes synthesis and energetics; iii) inhibition of RNA synthesis; and iv) inhibition of protein synthesis [74, 89]. 17
Figure 5: Maximum Likelihood tree of the MTBC complex. The branch colours correspond to the main MTBC lineages with *M. canettii* as an outgroup. Used from Coscolla *et al.* [61, 171] under the terms of the Creative Commons Attribution Licence (CC BY)..... 28
Figure 6: Cartogram showing the global TB burden by *M. tuberculosis* lineage. Countries are scaled to indicate TB incidence in 2021. Pie charts show the distributions of the *M. tuberculosis* lineage Lineage 1-Lineage 9 as well as the animal-adapted lineages *M. bovis, M. caprae* and *M. orygis* from clinical strains by geographic region according to Napier *et al.*'s scheme [61, 172]. Used from Napier *et al.*[172] under the terms of the Creative Commons Attribution Licence (CC BY) http://creativecommons.org/licenses/by/4.0/..... 29
Figure 7: A Bayesian phylogenetic tree showing the sublineages of lineage 1. Used from Palittapongarnpim *et al*. [181] under the terms of the Creative Commons Attribution Licence (CC BY) http://creativecommons.org/licenses/by/4.0/..... 31
Figure 8: A maximum likelihood tree showing the sub-lineages of lineage 2. Used from Thawornwattana *et al.*[182] under the terms of the Creative Commons Attribution Licence (CC BY) http://creativecommons.org/licenses/by/4.0/. 34
Figure 9: A Mid-point rooted maximum likelihood (ML) phylogenetic tree of Lineage 3. Used from Shuaib *et al.* [194] under the terms of the Creative Commons Attribution Licence (CC BY) http://creativecommons.org/licenses/by/4.0/. 36
Figure 10: (a) shows ten sub-lineages of Lineage 4 according to Stucki *et al.* [170] based on the Coll *et al.*[177] scheme and supported by spoligotyping data. Specialist sublineages are represented by black triangles, while the black circles represent the generalists. (b) The global distribution of each sublineage, ND represents undefined strains. Used with permission from Springer Nature Stucki *et al.*, copyright 2016 [170]..... 38
Figure 11:(a) Global distribution of the different sublineages of Lineage 4. Lineage 4.1.2(Haarlem), Lineage 4.3(LAM) and Lineage 4.10(PGG3) are found at high frequencies globally and are classified as generalists. (b)The specialists' sublineages Lineage 4.1.3 (Ghana), Lineage 4.5, Lineage 4.6.1 (Uganda) and Lineage 4.6.2 (Cameroon) were only restricted to specific geographical regions. Used with permission from Springer Nature Stucki *et al.*, [170] copyright 2016. 39

© University of Pretoria

Figure 12: Geographical ancestral distribution of Lineage 5, Lineage 6 and Lineage 9. (a) Reconstruction of the ancestral region using a Bayesian approach on the ML tree with circles corresponding to the likelihood of ancestral ranges and the most probable ancestry areas are colour-coded. (b)The four geographical regions of Lineage 5, Lineage 6 and Lineage 9 with stars showing the most probable ancestry region for each lineage. The arrows indicate the movements of strains. Used from Coscolla *et al.* [171] under the terms of the Creative Commons Attribution Licence (CC BY) http://creativecommons.org/licenses/by/4.0/..... 41
Figure 13: Flow chart summarizing a phylogenetic inference protocol (green box). The process starts with study data collection which is often combined with sequence data from other relevant strains e.g. strains that can serve as an outgroup. Multiple sequence alignment is performed to make the positions in different sequences comparable. A model of sequence evolution is carefully selected, one that is appropriate to the data. Finally, different approaches to phylogenetic inference are chosen as explained in detail in the text. Used with permission from Springer Nature Holder *et al.*[227] copyright 2003. 44
Figure 14: An example of a Beta-Binomial model showing three different choices of priors. The illustration shows a non-informative uniform prior, scale-invariant Jefferys' prior, and highly informative prior (Left to Right). Used from Banner *et al.*[249] under the terms of the Creative Commons Attribution Licence (CC BY) http://creativecommons.org/licenses/by/4.0/. 50
Figure 15: Difference between marginal and joint estimation. Both a and b represent the likelihood profiles for two phylogenies versus a nuisance parameter. The x-axis represents some nuisance parameter e.g. (the shape parameter of the gamma distribution). The y-axis represents the likelihood (in the context of ML) or the posterior probability density (in the context of Bayesian inference). The (AUC) area under the likelihood curve for tree A is shown in light blue. Orange represents the AUC for tree B. The AUC for both trees is coloured in mauve. A joint estimation of x and the phylogeny will choose tree A (highest peak coloured blue for both a and b). Marginal estimation over x will favour tree B (auc in the orange area is greater than in the blue area.). Used with permission from Springer Nature Holder *et al.*[227] copyright 2003. 51
Figure 16: Trace plot from sampling a posterior distribution using an inefficient MCMC chain with poor mixing. Despite the chain running for long, some important parameters remained low. 54

© University of Pretoria

Figure 17: Trace plot from sampling a posterior distribution using an efficient MCMC chain with good mixing. The chain ran long enough and all-important parameters had ESS values > 200. 55
Figure 18: A workflow summarising the data collection process from sourcing data from public databases to curating a high-quality reference dataset. The workflow highlights the key steps, including data sourcing, quality control, redundancy reduction, and metadata validation, ensuring the dataset is robust and representative of global MTB diversity..... 64
Figure 19: Global distribution of Mtb lineages based on the 9073 isolates used in this study (A) The phylogeographic structure of the 9,073 global samples used in this study. (B) Maximum likelihood phylogenetic tree of the global 8301 Mtb isolates. The tree was constructed based on 148555 genome-wide SNPs. Different coloured blocks are used to represent different lineages. (C) A barplot showing the distribution of the sublineages of lineage 4 which was the most prevalent lineage in this study. (D) Relative tree length (RTL) decay plot for the 9073 Mtb isolates, trimming the original tree to 99% of the original diversity preserves 8565 isolates as represented by the blue dashed lines. The red dashed lines show that 90% of the original diversity is still maintained even if half of the leaves are trimmed indicating high levels of redundancy. 69
Figure 20: A heat map showing the presence (1) or absence (0) of drug resistance-associated mutations for the most used anti-TB drug across all lineages in this study. Resistance was more prevalent for the first-line drugs isoniazid and rifampicin compared to the newly introduced drugs, clofazimine and Linezolid. 75
Figure 21 Sampling scheme, lineage distribution redundancy analysis, temporal signal analyis and antibiotic resistance emergence timing of 5,745 African Mtb strains. (A) Map of Africa showing the countries represented in our original dataset as well as the number of isolates for each country. The map is coloured according to several sequences isolated per country. (B) Unrooted phylogenetic showing the main lineages of Mtb identified in our African dataset. (C)The decay curve shows the high redundancy levels in our starting dataset, the red arrows show that 99% diversity of the dataset can still be preserved after pruning more than 2000 leaves from the phylogenetic tree. (D) Root-to-tip regression plot showing a weak but positive temporal signal. The red dots represent identified outliers. (E) the estimated time of emergence in Africa for some of the known drug resistance mutations. 90
Figure 22: Maximum likelihood tree of the 5745 African Mtb strains. The outer bands show the distribution of the main lineages, with Lineage 4 being the most dominant. The branches

© University of Pretoria

are coloured by regional location. The phylogeographical structure of the Mtb is visible with Lineage 3 (peach band) prevalent in East Africa (blue branches), and Lineage 5 and 6 (orange and grey bands) mostly dominate West Africa (red branches). 91
Figure 23: Bayesian skyline plots (BSPs) showing the inferred effective population size over time for Lineage 2 and 4. The y-axis scales vary between panels to accommodate differences in population size estimates across lineages or datasets. (A) Bayesian skyline plot inferred from 245 Mtb isolates sub-sampled from Lineage 4 strains showing a relative stability between ~1500-1700, there was a graduation population growth starting around 1700-1800. This was followed by a decline in ~1900. Around 2000, the population size rapidly increases. (B) Lineage 2 showed stability between ~1600-1800 followed by a slight expansion between 1800-1900. Towards the year 2000, there was a fluctuation which was followed by a drop. 98
Figure 24: Global diversity of the Lineage 2 strains in this study. (A) Maximum Clade Credibility tree of the combined 413 clinical isolates analysed in this study. The outer bands show the 16 sublineages of Lineage 2 identified in this study. The inner bands shown are coloured according to the region of isolation. (B) The distribution of the 16 sublineages per region. 113
Figure 25: Distribution of sublineages of Lineage 2 Mtb in the SADC region. Sublineage Lineage 2.2.M2.2 was the most prevalent (Top bar plot). South Africa had the highest diversity of Lineage 2 in the region (bottom map). 114
Figure 26: Alignment views showing the three different RD regions in strain SRR1167282. RD105 (A) and RD207(B) were present, while RD181(C) was absent. 115
Figure 27: Genetic diversity of Lineage 4 Mtb in the Southern Africa region. (A) Maximum likelihood tree inferred from 990 Lineage 4 isolates sourced from 8 countries in the SADC region of Africa the outer ring is coloured according to sublineage. (B) map showing the sublineage distribution of Lineage 4 samples in each country included in this analysis. Lineage 4.3 was predominant across the region. 117
Figure 28: Bayesian phylogenetics and phylogeography of Lineage 2 strains isolated in the SADC region. (A) Maximum clade credibility tree inferred from Lineage 2 strains isolated in the SADC region. (B) A phylogenetic tree inferred from the Mascot package in Beast 2 places the root in East Asia (purple) with an 86% probability. Different regions are colour-coded as follows: Europe (red), Southern Africa (turquoise), and East Asia (green). 120

© University of Pretoria

Figure 29: A simplified graphical representation of the possible introduction of Lineage 2 in the SADC region. The red arrows show Southeast Asia as the most likely origin of the MRCA of Lineage 2 strains in the region. The green arrows show the spread of Lineage 2 from South Africa upwards. The orange and purple lines represent the two introductions of Lineage 2 into Zimbabwe one from Botswana (Orange) and the second one from Mozambique (Purple). 122
Figure 30: Spatial dynamics of Lineage 4.3 strains sampled in the Southern Africa region. (A) A phylogenetic tree inferred from the Mascot package in Beast 2 placing the root of Lineage 4.3 isolates sampled in the SADC region in Europe (blue) with a 90% posterior probability. Different regions are colour-coded as follows: Europe (blue), Southern Africa (green), and the Americas (red). (B)A simplified graphical representation of the possible introduction of Lineage 4 in the SADC region. The red arrows show Europe as the most likely origin of the MRCA of Lineage 4 strains in the region. The green arrows show the spread of Lineage 4 from South Africa to different parts of the world in the last century). 126
Figure 31: Network of DR-associated mutations inferred from Lineage 1. The well-characterised DR mutations are coloured blue, while the prerequisite inferred compensatory mutations are highlighted in green. Arrows illustrate relations between prerequisite mutations and the well-characterised DR mutations. 140
Figure 32: Network of DR-associated mutations inferred from Lineage 2. The well-characterised DR mutations are coloured blue, while the prerequisite inferred compensatory mutations are highlighted in green. Arrows illustrate relations between prerequisite mutations and the well-characterised DR mutations. *The gyrA A90V* mutation was identified as a key mutation in the trajectory towards pre-XDR-TB. 142
Figure 33: Network of DR-associated mutations inferred from Lineage 4. The well-characterised DR mutations are coloured blue, while the prerequisite inferred compensatory mutations are highlighted in green. Arrows illustrate relations between prerequisite mutations and the well-characterised DR mutations. 143

© University of Pretoria

CHAPTER 1 TUBERCULOSIS: MOTIVATION

Despite being one of the oldest diseases known to man, tuberculosis (TB) remains a global health threat [1]. The causative agents of TB are several species of highly clonal gram-positive bacteria known as the *Mycobacterium tuberculosis* complex (MTBC) [2]. Traditionally, researchers showed little interest on the evolutionary processes in highly clonal bacteria, and this has resulted in a knowledge gap in the processes that drive evolution in strictly clonal organisms [3]. The unavailability of useful biological data from earlier decades necessitates the need for analytical strategies that help us explore the events that gave rise to the MTBC as we know it. Researchers have previously cited the unpredictable nature of horizontal gene transfer (HGT) in prokaryotes as a major challenge to evolutionary modelling in bacteria and archaea [3, 4]. The MTBC is associated with high clonality, such that HGT is absent [5]. The MTBC genome has characteristics that make it a good model for clonal evolution in that it has a genome size of 4.4 Mb characterised by a high GC content, low homoplastic mutations and a high proportion of nonsynonymous polymorphisms [3]. Despite the high clonality, members of the MTBC show incredible diversity even though individual strains differ by not more than a maximum of ca. 2400 SNPs genome-wide [3, 4]. With genomes of thousands of MTBC strains having been sequenced over the past decade, it is now undeniable that the genetic diversity of Mtb is a factor that influences different biomedical manifestations of TB disease. Characterising Mtb genetic diversity has implications in applications such as TB drug and vaccine development, epidemiology and antibiotic resistance [6, 7].

© University of Pretoria

1.1 A GLOBAL OVERVIEW OF THE CURRENT SITUATION

1.1.1 TB incidence

Approximately 10.8 million people (95% UI: 10.1-11.7 million) fell ill with TB in 2023, indicating a small increase from the 10.7 million cases reported in 2022 (95% UI: 9.6-11 million) and 10.0 million (95% UI 9.4-10.7 million) recorded in 2020 [1, 8]. There was also a 0.2% increase in the global TB incidence rate (new cases per 100,000 population) in 2023 compared with 2022, with a TB incidence rate of 134 (95% UI 125-145) in 2023. Thirty high-TB burden countries (see Appendix) contributed to 87% of the global TB incidence in 2023 [1]. Five countries in Asia contributed to 56% of the global total in 2023: India (26%), Indonesia (10%), China (6.8%), the Philippines (6.8%) and Pakistan (6.3%). 55% of the people who developed TB in 2023 were men, 33% were women, and 12% were children and young adolescents [1]. TB remains a classical disease of poverty, with most 2023 cases recorded in the WHO regions of Southeast Asia(45%), Africa (24%) and the Western Pacific (17%). The Eastern Mediterranean (8.6%), the Americas (3.2%) and Europe (2.1%) contributed the rest of the global TB cases [1, 8] . However, countries like China and India are among the high burden countries despite their growing economies. Encouragingly, the WHO African region achieved good inroads towards the 2025 milestones for reducing TB incidence and TB mortality [1].

The WHO End TB Strategy set a milestone of a 50% reduction in the TB incidence rate by 2025 [9]. However, a net reduction of only 8.3% was achieved from 2015 to 2023 [1]. On a country level, only 8 countries were responsible for more than two thirds of global TB cases namely: India (27%), Indonesia (10%), China (6.8%), the Philippines (6.8%), Pakistan (6.3%), Nigeria (4.6%), Bangladesh (3.5%) and the Democratic Republic of the Congo (3.1%).

1.1.2 TB mortality

There was a decrease in the total global deaths caused by TB in 2023 compared to the estimated 1.32 million in 2022, 1.42 million in 2021, 1.40 million in 2020 and the pre-

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

pandemic estimates of 1.34 million in 2019 [1]. Concerningly, despite the reduction, TB might have reclaimed the top spot as the world's leading cause of death from a single infectious agent, replacing COVID-19. An estimated 1.25 million people (95% uncertainty interval [UI]: 1.13-1.37 million) died of TB globally in 2023, of these deaths, 1.09 million deaths were among HIV-negative people, while 161,000 deaths were among people living with HIV. The mortality figures officially classified as caused by TB in 2023 (1.09 million) were almost double the HIV/AIDS mortality rates (0.63 million) [1, 8].

This also threatens the WHO End TB Strategy goal of reducing the total number of deaths by 75% between 2015 and 2025, especially given that a net reduction of only 23% was achieved between 2015 and 2023 [1, 8, 9]. The highest number of deaths caused by TB among HIV-negative people in 2023 (81%) was recorded in the WHO regions of Southeast Asia(53%) and Africa (27%). India alone contributed to 29% of such deaths [1]. The rest of the regions were as follows: Western Pacific (8.5%), the Eastern Mediterranean (7.2%), the Americas (2.1%), and Europe (1.5%). The African region recorded the most estimated deaths (68%) among people with HIV. Globally, most of the deaths caused by TB among HIV-negative people in 2023 were adult men (52%), 32% were women, and 15% were children (aged < 15 years) [8]. The disparity between the estimated number of cases in children (12%) and the estimated deaths (16%) indicates poorer access to diagnosis and treatment. Of the HIV-positive people who died of TB in 2023, men were the highest (48%), and 36% were women and 16% were children.

1.1.3 Drug-resistant TB

TB is a curable disease; however, the inappropriate use of TB medicines through poor quality drugs, wrong prescription of drugs by healthcare workers, or premature stopping of treatment from patients leads to the emergence of drug-resistant TB (DR-TB) [10]. Multi-drug-resistant tuberculosis (MDR-TB) is a serious form of TB that is caused by bacterial strains that do not respond to the first-line drugs isoniazid and rifampicin, which are the most effective drugs for treating drug-susceptible TB. Second-line drugs are used to treat and cure TB [11]. However, managing MDR-TB is complex and expensive as the required drugs are also costly and toxic [12]. Extensive resistant TB (XDR-TB) can also develop when MDR-

© University of Pretoria

TB patients do not respond to Second-line therapy. MDR-TB and rifampicin-resistant TB (MDR/RR-TB) continue to threaten humanity, with only 40% of people with DR-TB accessing treatment in 2022 [8, 13]. Globally, there was relative stability in the annual estimates for the number of people who developed MDR/RR-TB between 2020 and 2023, following a spiralling trend between 2015 and 2020 [1, 8, 14]. The 2023 estimate sat at 400,000 (95% uncertainty [UI]: 360,000 – 440,000) with an estimated 160,000 (95% UI 98,000 – 220 000) deaths attributed to MDR/RR-TB. A more detailed review of DR-TB will be discussed in the following sections.

1.1.4 Global TB after COVID-19

Disruptions to diagnosis and treatment during the COVID-19 pandemic led to an increase in TB incidence in 2021 and 2022. According to the WHO, newly diagnosed TB cases dropped from 7.1 million in 2019 to 5.8 million and 6.4 million in 2021 [8]. These findings have far-reaching implications because an increase in the number of undiagnosed and untreated cases also results in increased TB transmission and with time, an increase in the number of people developing TB disease. The effect of COVID-19 disruptions was more apparent on TB mortality, with an estimated close to 500,000 excess deaths from TB between 2020 and 2022. This contrasts with the trend observed with HIV/AIDS where mortality rates declined between 2019 and 2022 [8, 15]. Earlier projections had estimated that a reversal of the global increase in TB incidence caused by the COVID-19 pandemic would be achieved in 2023 or 2024 but this may take longer than that [8].

1.2 A HISTORICAL PERSPECTIVE ON TB RESEARCH

In humans, TB is mainly caused by *M. tuberculosis* and *M. africanum* [16, 17]. In addition to these 2 obligate human pathogen species, the MTBC also consists of other members which have adapted to other animal species [18]. As mentioned earlier, TB is a classical disease of poverty and improvements in quality of life and better living conditions during the 20th century saw a steady decline in the incidence of TB in the developed world [19]. This notable

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

event also had unforeseen and unwanted consequences as this resulted in the drying up of funds and resources for TB research and control [20]. This complacency would later prove to be dangerous as the early 1990s saw a resurgence of TB fuelled by the ongoing HIV/AIDS pandemic [21]. This public health crisis led to governments committing funds and resources to TB control. However, years of neglect in the field of research meant that the only weapons to fight the TB war included the unreliable diagnostic method of smear microscopy [22], an "outdated" vaccine (Bacille Calmette-Guerin) popularly known as BCG [23] as well as old antibiotics whose efficacy has been compromised by years of resistance development [24].

It is important to note that despite the technological advancements of the 21st century, these are the only resources available in most developing countries to this date despite carrying most of the brunt of the global TB burden. Years of neglect in TB research also meant that scientists and other stakeholders in the 1990s had to be guided by several obsolete dogmas. Traditional dogma was rooted in the premise of MTBC infection homogeneity, and it also downplayed the role of pathogen factors in the clinical manifestation of TB [19]. Advancements in molecular technologies such as DNA fingerprinting were a significant turning point in MTBC research. Figure 2 shows a timeline of some of the technological advancements in TB diversity studies developed in the last 3 decades. In 1993, van Embden *et al.* published a method of identifying MTBC strains using DNA fingerprinting [25]. Their method, which is based on the *Mycobacterium* insertion sequence 110 (IS6110) [26] went on to become the first international gold standard for MTBC genotyping [19, 25]. This study also shaped global public health policy in terms of managing disease outbreaks and transmission tracking, as well as shedding insights on the effects of treatment failure in disease relapse [27]. This study also spurred further TB research work that dispelled some of the traditional dogmas that governed TB research. However, most of the low-income countries which carry the highest global TB burden are under-resourced, which means that important research data from such settings remains scarce.

The improvement of technology in scientific research also provided scientists with insights into the genetic diversity of MTBC, which was previously underreported [28]. Epidemiological studies also showed certain strains of MTBC were more prevalent in particular demographics [19]. These observations invoked an interest in studying the role of MTBC variation in TB epidemiology.

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

The year 1998 saw another turning point in TB research when the completion of the whole MTBC genome was announced [29]. The high resolution of whole genome sequencing has also been exploited to track transmission dynamics in TB through the analysis of single nucleotide polymorphisms (SNPs) in MTBC. The advent of WGS meant that genomes of different strains of MTBC could be sequenced and compared. Analysis of the data from such studies would later be used to reconstruct MTBC phylogeny and thus added to the knowledge base on the global distribution of MTBC. Tsolaki *et al.* revealed that genomic deletions in MTBC are an important source of genomic plasticity in MTBC [19, 30]. These genomic deletions, also known as large sequence polymorphisms (LSPs), have been used to challenge the traditional dogma on the origins of MTBC. The absence of horizontal gene transfer in MTBC meant that LSPs could be used as lineage markers in inferring the phylogeny of the pathogen. Initially, it was hypothesised that humans acquired MTBC from animals during the Neolithic age [31] when early humans adopted the domestication of animals. However, comparative genomic studies have disproved this theory with evidence rather suggesting that members of MTBC (both human and animal adapted) share a common ancestor, which might have infected humans pre-Neolithic transition [32].

1.3 HUMAN TB ORIGINS

The study of the origins of human MTBC has been a hot topic for years, especially interest in what influences the highly phylo-geographical population structure of MTBC. A 50 000-year-old fossil of *Homo erectus* was found with lesions associated with TB [33]. This finding generated controversy because it suggested that TB predated the origin of modern humans [28]. A recent study by Comas *et al.* sought to infer a WGS-based global phylogeny of human-adapted MTBC [34]. The study concluded that Africa was the only region that had all 6 (known at that time) of the human-adapted MTBC lineages, further highlighting the capacity of human-adapted MTBC diversity the continent harboured [35]. Based on these findings, some researchers have suggested that approximately 70,000 years ago, two "ancient" lineages of MTBC were left in Africa when human migrations seeded the other lineages to the rest of the world [36]. The next section discusses some of the studies and techniques that have been employed to try and characterise the evolutionary mechanisms

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

that shape strain diversity in MTBC. These techniques are also summarized in Figure 1 below with Figure 2 showing the timeline to development of the techniques.

1.4 EVOLUTIONARY HISTORY FROM THE PATHOGEN PERSPECTIVE

The use of LSPs as lineage markers of MTBC gave useful insights into the evolutionary history of MTBC. However, a major limitation in their use is that they fail to reflect the actual genetic distances and relative genetic diversities between different MTBC lineages [19, 37]. This limitation can be observed in a study by Brudey *et al.* who used spoligotyping to infer the global population structure of MTBC [38]. However, their approach uses the presence or absence of the "direct repeat region of MTB" to index genetic diversity [37]. This region is a hotspot for convergent evolution, which limits the use of this technique in phylogenetic studies.

DNA sequence-based techniques have been successfully used to provide insights into the mechanisms governing the evolution of bacterial populations. Baker *et al.* used multi-locus sequence typing (MLST) to study MTB diversity [39]. Their technique used fragments of seven structural genes for each strain to define the genetic population structure of MTBC [19]. However, this study was limited by the small number of informative SNPs that were identified. Subsequent studies by Filliol *et al.* and Gutacker *et al.* employed a similar approach and sequenced 89 complete genes in 108 strains, covering 1.5% of the genome of each strain [40, 41]. One major concern with this type of study is what is known as phylogenetic discovery bias and distorted phylogenetic inference [41-43]. Despite this, the phylogenetic tree derived from this SNP-based study was consistent with the results that had been previously reported in studies that used LSPs. A major key point of this study was that it provided better resolution in the relationship between animal-based MTBC strains and those adapted to humans. Further phylogeographic studies based on these findings further supported the suggested African origin of the MTBC hypothesis [37]. As described earlier, it is widely believed that MTBC is most likely to have its origins in Africa and that the human-adapted MTBC strains were introduced to different parts of the world because of early human migration. It is important to note that a subset of MTBC strains called the "smooth tubercle bacilli" have only been reported in the Horn of Africa region [31, 32]. The

© University of Pretoria

geographical distribution of these strains which are the closest relative of human MTBC further boosts the African origin hypothesis.

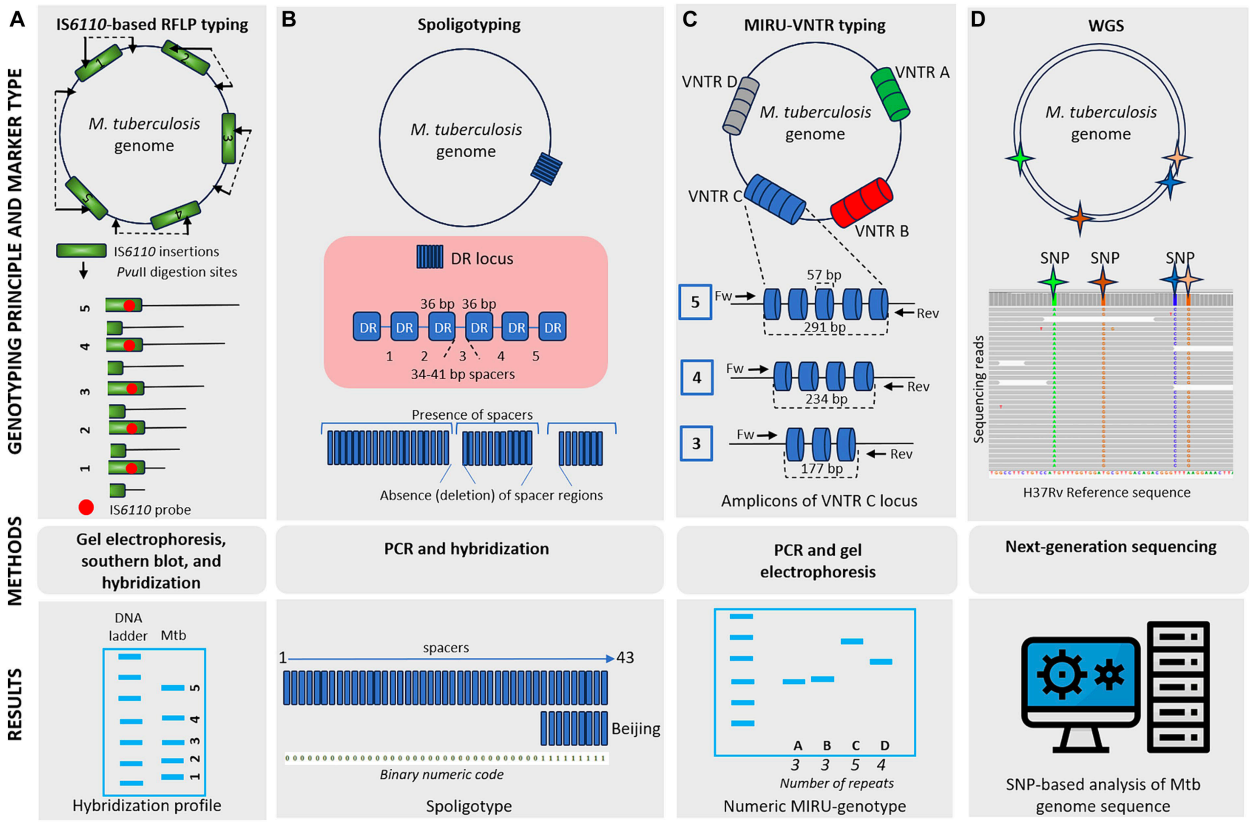


Figure 1: A schema of the different *M. tuberculosis* marker types and typing technologies.

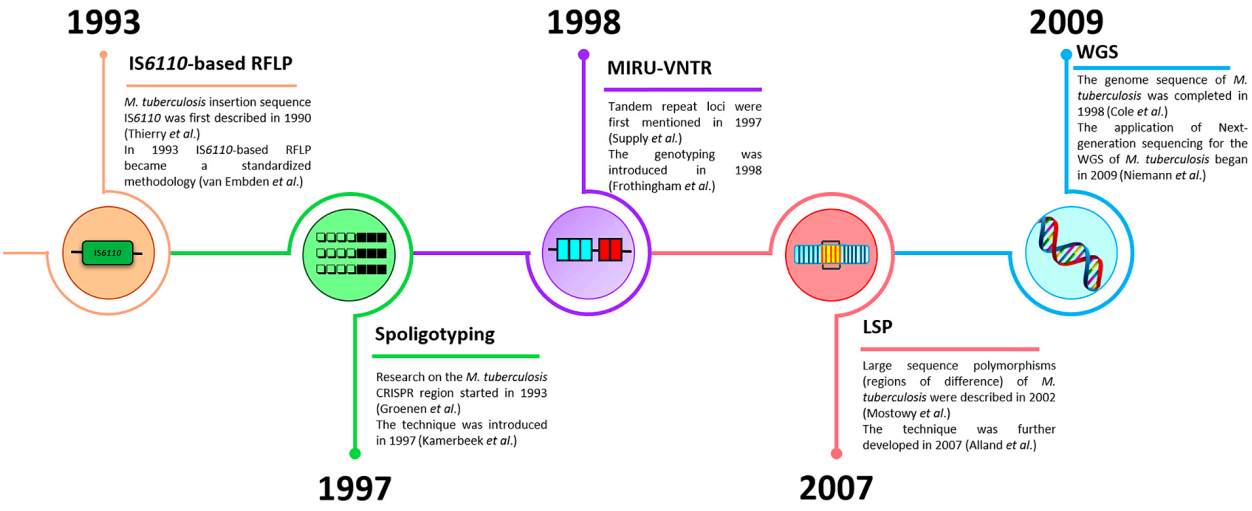


Figure 2: A timeline of the development of *M. tuberculosis* genotyping approaches. Used from Atavliyeva et .al [44] under the terms of the Creative Commons Attribution Licence (CC BY).

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

An MLST-based study by Hershberg *et al.* also looked at the genetic diversity of human-adapted strains of MTBC [37] . This study proposed that the modern lineages, as shown in Figure 3 below, seeded human-adapted MTBC in the Eurasia region [19]. This coincided with dramatic and disproportionate increases in the human population leading to the clonal expansion of these 3 lineages. The past few centuries saw this region dominated by events such as human migration, commerce, conflict, and conquest which facilitated the spreading of these lineages of MTBC to other parts of the globe [37]. Advances in shipbuilding technology also saw this recent spreading of human MTBC mainly along the water routes as opposed to the land routes used in the ancient MTBC spread [37]. This study by Hershberg *et al.* also looked at the evolutionary mechanisms governing the population structure of MTBC. They investigated the role of purifying selection in the evolution and genetic diversity of MTBC. By using the (dN/dS) to calculate the ratio of nonsynonymous and synonymous changes, they reported that the ratio was higher than previously anticipated in MTBC. This indicates a reduction in selection pressure. The team hypothesised that increased genetic drift might be the driving force generating diversity in MTBC.

A review by Gagneux *et al.*[32] further suggests that if the proposed evolutionary history of MTBC with humans did occur, it would have been during the hunter-gatherer times which were associated with low human populations. This suggestion supports several studies that suggest that the latency period by human-adapted MTBC, which is sometimes followed by reactivation and active disease years later might have been an evolutionary adaption by MTBC to allow for access to new birth cohorts of susceptible hosts [32]. An overly virulent strain in low-population settings would not be ideal as it poses the risk of decimating all susceptible hosts. The success of the highly virulent modern strains can be attributed to the high population densities as European cities industrialized which meant that access to susceptible hosts was no longer a limiting factor.

Despite the significant insights that were derived from the “Out-of-and-back-to-Africa” hypothesis proposed by Ganeux *et al.*[45] and Comas *et al.*[34], several mysteries remain unsolved. Traditionally, it was accepted that the Americas were TB-free before European contact [46]. Several studies have reported on the high susceptibility to TB observed in Aboriginal Americans. Researchers suggested a ‘Virgin soil’ hypothesis, which assumed that the Native American populations had not built any form of immunity to TB due to lack

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

of prior exposure [28, 47]. Recently, several inter-disciplinary studies have suggested that TB existed in pre-Columbian America [48]. Ancient DNA studies have managed to isolate mycobacterial DNA in pre-Columbian fossils [49]. However, DNA analysis has not been able to resolve the exact causative agent of TB during that time [48]. The absence of *M. africanum* in the Americas despite years of trans-Atlantic slave trade remains another mystery. Presently, the dominating cause of TB in the Americas is the Euro-American lineage, even though the first colonizers of the region came from Asia through the Bering Strait [28, 47]. Future studies are needed to determine what led to the success of the Euro-American lineage in the region despite possible prior exposure from other lineages such as *M. africanum* and other Asian strains. The possibility of a "now extinct' uncharacterised MTBC lineage cannot be ignored.

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

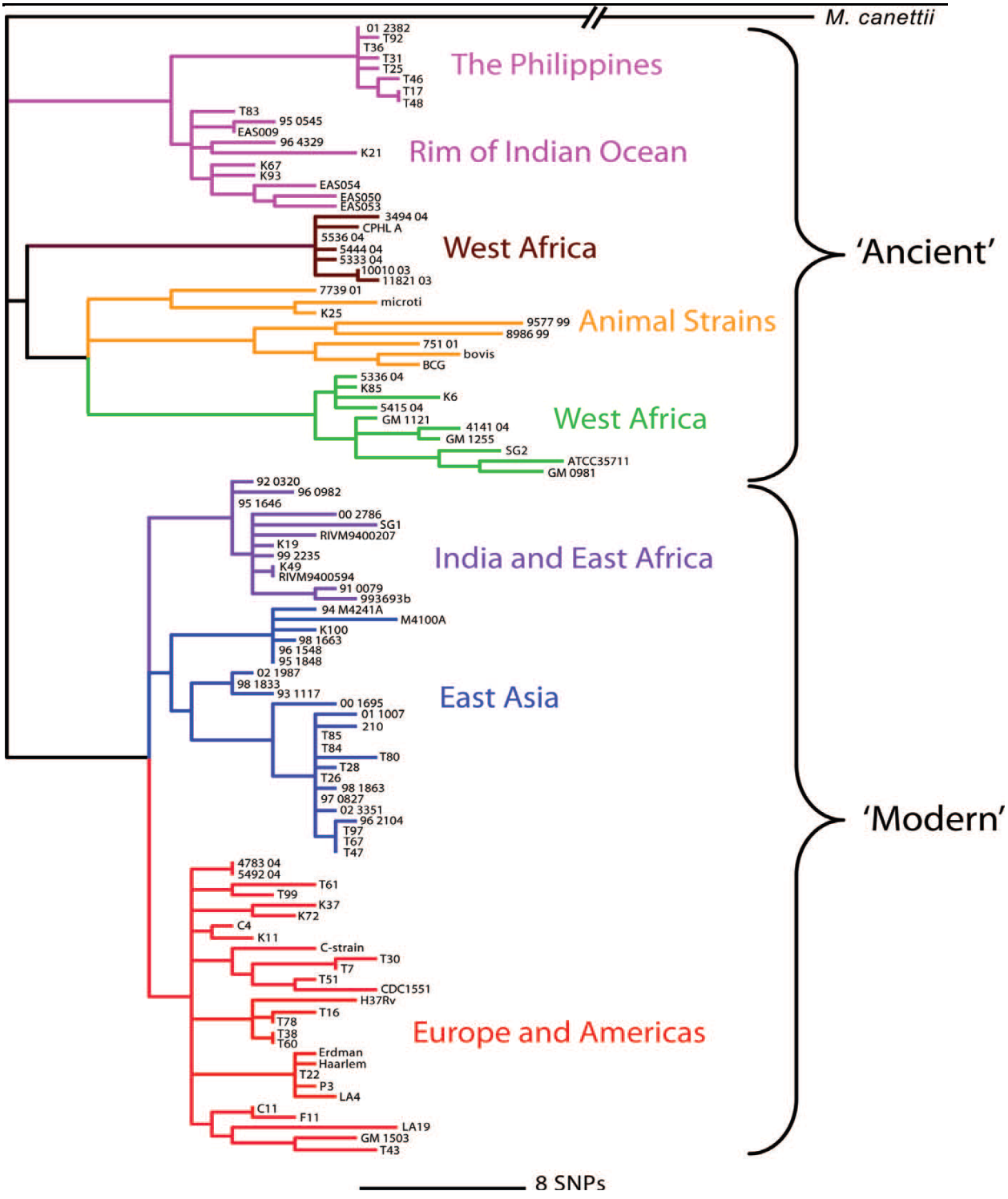


Figure 3: Maximum Parsimony tree of *M. tuberculosis* with branch colours representing the main lineages. The main clades are labelled according to the geographical regions from which the lineage dominates. Used from Heshberg *et al.*[50] under the terms of the Creative Commons Attribution Licence (CC BY).

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

1.5 THE ROLE OF WGS IN EXPLORING MTBC GENOMIC DIVERSITY

The previous section delved into some of the techniques that shaped the study of MTBC diversity, including the limitations associated with their use. The year 1998 saw a turning point in TB research when Stewart Cole at the Institute Pasteur and collaborators at the Sanger Centre in the United Kingdom reported on the first-ever complete genome sequence of Mtb [51]. The genome was sequenced from the well-characterised laboratory reference strain H37Rv. The advent of WGS technologies was a game changer as it made available many whole genome sequences of MTBC clinical strains from around the globe [52]. Analysis of the high-throughput data from WGS has led to the successful reconstruction of the MTBC phylogeny [24].

Of all the genotyping techniques discussed in this review, comparative WGS has proven to be superior in classifying strains of MTBC as well as unmasking the genetic intricacies of individual strains in detail [53]. The robustness of WGS also allows us to index MTBC genomic diversity at different levels, i.e. be it at the whole population level, during household transmission, within single hosts, or during in vitro evolution [52, 54]. WGS allows quantifying the amount of genetic diversity as it is possible to measure phylogenetic distances [55]. The use of genome-wide associated studies (GWAS) also allows us to study in-depth the links between phenotype and genotype through the discovery of new polymorphisms in particular strains [56]. While the cost of WGS has declined rapidly over the years, the adoption and use of WGS-based methods in TB control have been limited by the skill requirements of WGS data analysis [24]. Most of the tools designed for WGS data analysis often require proficiency in using the command line which often discourages would-be end-users. Hence, there is a need to develop more user-friendly analysis tools [24].

Several tools and databases have been designed over the years to address such challenges. These platforms often integrate multiple MTBC resources such as downloadable sequence data, maps, assemblies, annotations, bibliographies, and references [52]. An example of such a platform is the GMTV database which contains information on more than 1000 MTBC strains obtained from different sources. Although the first of these databases such as

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Tuberculist was built around the genome of the reference H37RV strain [57], several groups have used bacterial GWAS studies to identify some novel SNPs which are then incorporated into tools that predict lineages of MTBC. These tools such as KvarQ, Resistance Sniffer and TBProfiler also use SNP catalogues to predict the lineage and drug resistance profile of an MTBC strain [24, 58].

1.5.1 WGS and antimicrobial resistance (AMR)

The WHO lists AMR in the top ten of global health threats as it comes with clinical, financial as well as public health challenges WHO and 25 years paper [59]. WGS has shown great promise in drug susceptibility predictions in Mtb. WGS has also improved our understanding of the genetic basis of drug resistance in Mtb and has allowed us to create a global catalogue of known resistance-associated variants (RAVs) [59, 60]. Current methods of Mtb drug susceptibility testing (DST) are associated with challenges such as slow turnaround time to results (weeks after initial culturing) and low resolution as they are more robust for first-line drugs only remaining costly [59]. Low-incidence and high-income countries have explored the use of WGS in DST [61]. The UK was the first region to implement routine WGS for lineage classification and DST of positive Mtb cultures in 2017 [61, 62]. WGS-based drug susceptibility testing (DST) has the potential for success in regions with high burden of MDR- and XDR-TB and has been piloted in both Kyrgyzstan and South Africa [63, 64].

1.5.2 WGS and the development of novel anti-TB drugs

Over the past decade, WGS played a key role in the successful development of new anti-TB drugs as evidenced by the introduction of bedaquiline which was the first new anti-TB drug in over four decades, and other novel drugs such as delamanid and pretomanid. WGS was highly instrumental at all the different stages of the drug development lifecycle i.e. from initial laboratory elucidation to formulating and calibration of clinical regimens [61]. Genomic sequencing of *M. smegmatis* identified a mutation in the *atpE* gene which was associated with bedaquiline resistance in vitro. This finding revealed that the drug uniquely

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

targeted the ATP biosynthesis pathways, which meant that it could be potent even against dormant bacteria [61, 65, 66]. Broader tests in other mycobacterial species demonstrated potency against several other non-tuberculosis mycobacteria except for *M. xenopi.* Although the resistance profiling was successful in the development of bedaquiline, the same approach was less apparent for the mode of action of the nitroimidazoles: Pretomanid and Delamanid [61]. Earlier work linked mutations in the *fdg1*, an F420-dependent enzyme to resistance in this class of TB drugs. However, WGS soon identified other resistance mutations. Initial studies had shown that pretomanid inhibited ketomycolic acid synthesis, however, in vitro studies had also proved the effectiveness of the drug even in non-replicating culture conditions when production of the cell wall component is not much of a priority to the bacilli [61, 67, 68]. WGS revealed that another redox enzyme, *ddn*, cyclically complements *fdg1* in interacting with the F420 redox cofactor thereby activating Pretomanid to form bactericidal nitric oxide [61, 69, 70]. Elucidating the genetic basis of drug resistance using WGS identified the function of previously unknown gene functions for both bedaquiline and nitroimidazoles. The next section delves deeper into DR-TB and how resistance develops for the commonly used anti-TB drugs.

1.5.3 Challenges in using NGS in TB applications

While WGS has proved to be an invaluable tool in every facet of the fight against TB, it is important to note that there are also limitations associated with the current sequencing methodology that can lead to the misinterpretation of the data. Most of the applications such as phylogenetic inference and DR-TB prediction are SNP-based (rely on the identification of SNPs). The presence of novel contigs or the absence of sequencing reads relative to the reference genome can be easily interpreted as INDELs [31, 71-74]. The variability among mycobacteria has been attributed to the loss of entire genes [71, 74]. However, identifying variants/ polymorphisms in repetitive regions, gene duplication chromosomal rearrangements, and copy number changes of tandem repeats can be challenging when using the widely used short-read-based NGS methods and this can significantly impact the interpretation of the underlying biological data especially when using single read technology [75]. The use of paired-end sequencing technology does improve some of the results as it gives a better resolution for these problematic regions [76]. However, the Mtb genome is

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

littered with several repetitive regions that can cause problems when using short-read sequencing technology. These low-confidence regions are often masked prior to analysis, however, some of these regions harbour several important genes critical in Mtb biology such as the *pncA*, and *esx* genes. Detecting genomic duplications in Mtb using short-read sequencing technology is also complicated by genome assembly ambiguity. Some Lineage 2 strains harbour a large ~350kb duplication in the Rv3128-Rv3427 gene region which also harbours the transcriptional regulator *DosR,* which is involved in hypoxia regulation [71, 74, 77]. Short-read genome assembly would result in several ways of generating contigs for this region thus complicating the identification of essential polymorphisms in this region. Advances in third-generation sequencing technologies, e.g. Oxford Nanopore Technologies (ONT) or PacBio allow us to generate longer reads (LRS), which can overcome some of the challenges associated with short read technologies [74, 78, 79].

1.6 DRUG RESISTANCE AND TB

The emergence of drug-resistant TB (DR-TB) continues to challenge the gains achieved in the war against TB. Since the introduction of antibiotics to the management of TB, drug resistance has remained a barrier to successful treatment [80]. This can mainly be attributed to the bacterial population's mutational capacity, which is a function of both the mutation rate and the size of the bacterial population [74]. Initially, clinicians adopted a monotherapy approach to TB treatment but soon replaced it with a more effective approach - multi-drug therapy due to the rapid development of resistance [81].

The WHO recently updated the definitions of different manifestations of DR-TB [82]. MDR-TB still refers to TB caused by a strain of Mtb that is resistant to rifampicin and isoniazid. However, the definition of XDR-TB was updated to mean TB disease caused by a strain of Mtb that is resistant to rifampicin (may also be MDR-TB) and that is also resistant to at least one fluoroquinolone (levofloxacin or moxifloxacin) and to at least one other "Group A" drug (bedaquiline or linezolid) [11]. Treatment of DR-TB remains a major threat, Gandhi *et al.*[83] reported that rapid progression to death was recorded in 98% of pre-XDR-TB (based on the updated definition) patients during an outbreak in KZN South Africa [83]. Another

© University of Pretoria

term, totally drug-resistant TB (TDR-TB) has been suggested to describe TB isolates that do not respond to any of the currently available drugs although there is no agreed definition of TDR-TB yet [84].

Compared to drug-susceptible TB management, treatment of drug-resistant TB is complicated, with global success rates of less than 50% for MDR-TB [74]. The treatment process requires costly medication and specialised hospital care and is often associated with poor outcomes. The medication prescribed are also highly toxic and can lead to severe side effects such as permanent deafness and psychiatric disorders [74, 85]. These often lead to poor compliance with the treatment regimen and this in turn reduces the cure rates and can even lead to the amplification of resistance [84, 86]. Robust drug susceptibility testing is also crucial in the improvement of treatment outcomes as it ensures that only the appropriate anti-TB drugs are prescribed thus reducing exposure to ineffective and toxic drugs [84].

The key to the control of MDR-TB lies in early diagnosis and treatment, incorrect treatment of TB can be risky both at the patient and population level [74, 87]. Misdiagnosis and inadequate treatment of MDR-TB or XDR-TB potentially leads to the positive selection of resistant subpopulations and hence the creation of resistant strains de novo [74]. This can also render additional drugs ineffective against an already M/XDR-TB strain [87].

Unlike in most other bacterial pathogens, resistance plasmids and horizontal gene transfer play no role in the acquisition of drug resistance in *M. tuberculosis* [74, 88]. The main cause of drug resistance in *M. tuberculosis* is the accumulation of point mutations and indels in genes coding for drug targets or converting enzymes [74, 84]. Recent studies have shown that alterations to these genes and their interactions are just the first step in a longer and more complex process. These mutations are now known to interact with mutations in other genes. These interactions known as epistasis, play an important role in the development of the drug-resistant phenotype as they are known to compensate for the fitness cost that is incurred when resistance is acquired. There is a need for more knowledge in the understanding and accurate detection of these processes that lead to the resistant phenotype [74].

© University of Pretoria

1.6.1 Mechanism of resistance to anti-TB drugs

Figure 4 below summarises the mechanisms of action for some of the most commonly used TB drugs

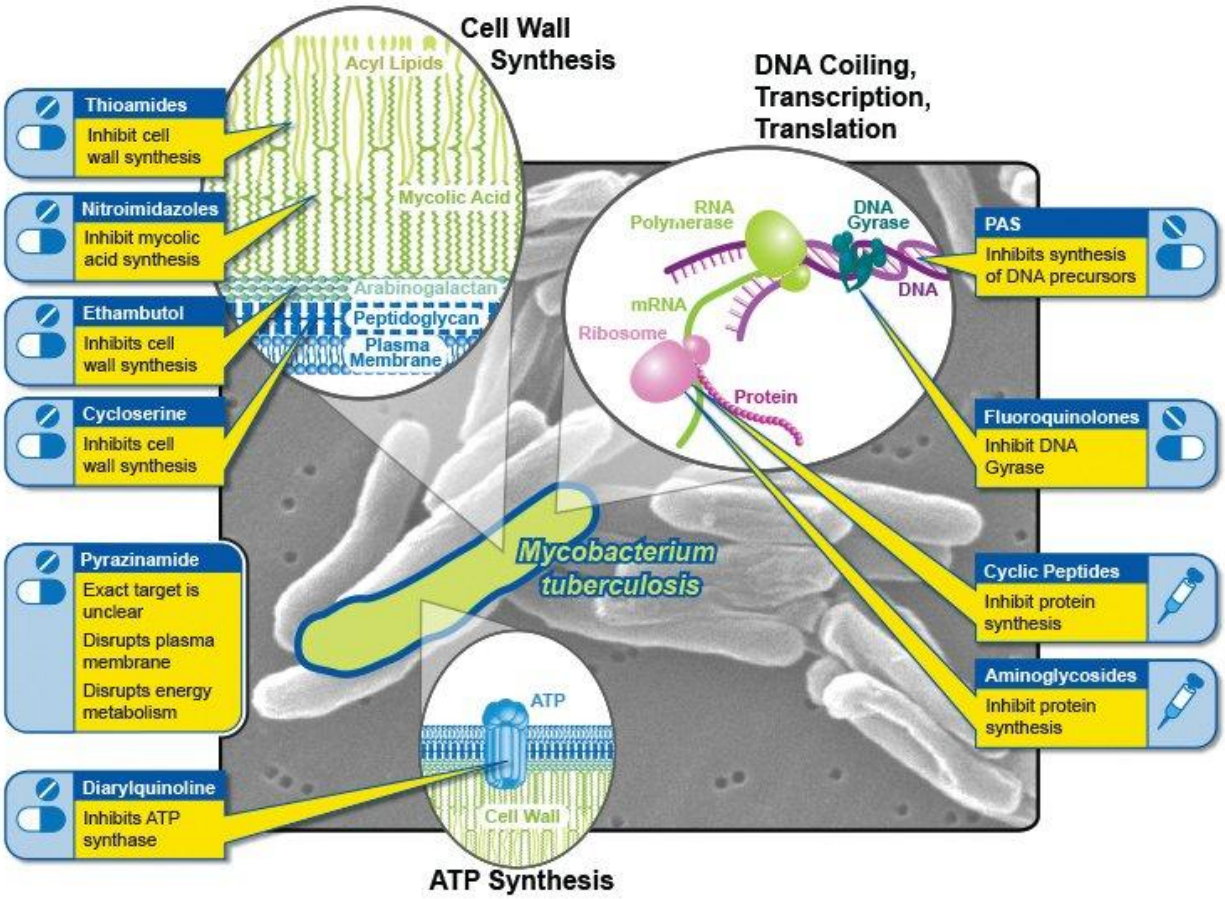


Figure 4: Summarises the mechanisms of action of some of the main anti-TB drugs currently in clinical use. These can be classified into four groups mainly: i) inhibition of cell wall biosynthesis; ii) disruption of cell membranes synthesis and energetics; iii) inhibition of RNA synthesis; and iv) inhibition of protein synthesis [74, 89].

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

1.6.1.1 First-line drugs

1.6.1.1.1 Rifampicin (RIF)

Persistent infection with *M. tuberculosis* has been associated with low metabolic activity as well as non-replication [90-93]. In contrast to most antibiotics currently in use, which target mostly actively growing bacteria, rifampicin also known as rifampin is one of the most powerful drugs in TB treatment as it is highly effective against actively metabolizing and slow metabolizing bacilli [94-98]. RIF's mode of action involves the disruption of mRNA elongation by binding to the beta subunit of RNA polymerase. Mutations grouped in codons 507-533 of the *rpoB* gene which codes for the beta subunit of RNA polymerase have been associated with RIF resistance [99]. The rifampicin resistance determining region (RRDR) is known to shelter variants that contribute to 96% of rifampicin-resistant cases and has been targeted in the development of modern molecular-based diagnostic assays, particularly codons 526 and 531 [94]. Notably, RIF resistance has been detected in some isolates where *rpoB* alterations are absent or that carry mutations outside the RRDR [100]. Genotypic assays that detect RIF resistance have been targeted on the RRDR such that the role of non-RRDR mutations in RIF resistance remains elusive. Extensive research is also needed to elucidate the role of the environmental conditions that favour the emergence of specific *rpoB* mutations in MTB [93]. RIF mono-resistance is a re and almost all RIF-resistant Mtb isolates are also resistant to at least one more anti-TB drug [95]. This has made RIF resistance an important surrogate marker for MDR-TB over the years [101].

1.6.1.1.2 Isoniazid

Isoniazid (INH) is another key pillar of current first-line regimens used in the treatment of drug-susceptible TB. In contrast to RIF, INH is only potent against metabolically active replicating bacilli targeting mycolic acid biosynthesis [95]. It is a prodrug that is activated by the catalase /peroxidase enzyme encoded by the *katG* gene [102]. Once activated, INH forms a bond with NADH-dependent enoyl-acyl carrier protein reductase which is encoded by the *inhA* gene. The binding is known to inhibit mycolic acid synthesis in addition to disrupting several essential metabolic pathways. Mutations in the *katG*, *inhA,* and its promoter regions have been associated with resistance to INH, particularly the *kat*G S315I,

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

S315N, and S315T mutations which have been identified as the most common mechanisms of INH resistance. Mutations in the *inhA*/*mabA* promoter regions have also been associated with drug titration therefore causing varying levels of INH resistance [59]. Of these mutations, the ones located at position 15 of the *inhA* promoter region are the most common and have been associated with low-levels of INH monoresistance as well as cross- resistance with ethionamide (ETA) which is a structural analog of INH. Alterations to the active site of *inhA* lead to a reduced affinity for the INH-NAD product [94, 103, 104]. Mutations in the *kasA, furA, oxyR-ahpC,* and *ndh* genes which are involved in mycolic acid synthesis pathways have also been associated with INH resistance [30]. It is important to note that 84% of global phenotypic INH resistance can be linked to these mutations [94, 105]. Previous research has also suggested that INH resistance mediated by the *katG* S315T mutation is a precursor of RIF resistance which makes the mutation an important marker in the diagnosis and tracking of MDR-TB [106]. These findings highlight the importance of early detection of antibiotic resistance in the management of TB.

1.6.1.1.3 Ethambutol

Ethambutol (EMB) is another key first-line drug that complements RIF and INH in TB treatment due to its mechanism of action. The drug targets multiple pathways of actively multiplying bacilli, specifically those that are involved in arabinogalactan biosynthesis in the cell wall [107, 108]. Arabinan polymerization inhibition aids in facilitating the permeability of other anti-TB drugs leading to better treatment outcomes. Resistance to EMB has been associated with mutations in the *embCAB* operon which encodes the mycobacterial arabinosyl transferase enzyme [94, 107, 108]. Codon 306 mutations in the *embB* gene were initially linked to EMB resistance and several studies have implicated mutations in this gene as a stepping stone to poly-drug resistance in MTB [109]. However, a study by Hazbon *et al.*[110] concluded that mutations in *embB*306 were not necessarily determinants of EMB resistance but rather act as markers of strains that are predisposed to develop resistance to multiple drugs. Findings from other independent studies supported this finding [74, 95, 110]. Safi *et al.* performed several allelic exchange experiments that have highlighted the complexity of EMB resistance development in Mtb [111]. Their work reported that contrary to earlier suggestions, the development of high-level EMB resistance was a multi-step

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

process [88]. Intriguingly, ~30% of EMB-resistant isolates do not carry any *embB* mutations, suggesting the possibility of alternative mechanisms of EMB resistance [112]. The gene *ubiA* has also been linked to high-levels of EMB resistance in some African lineages [74, 94]. This gene encodes for decaprenyl-phosphate 5-phosphoribosyltransferase synthase, an important enzyme in cell wall synthesis pathways [74, 113].

1.6.1.1.4 Pyrazinamide

The introduction of the nicotinamide analogue pyrazinamide (PZA) to the first-line anti-TB arsenal was a game-changer as it resulted in the reduction of treatment durations to just six months [94]. The drug is key to the development and evaluation of new anti-TB regimens as it is included in Phase II and III of all DS-TB and DR-TB clinical trials [94, 114]. The acidic conditions in TB lesions allow for the survival of semi-dormant bacilli as they disrupt the potency of commonly used antibiotics [115]. However, PZA remains active even at low pH levels, which makes it a critical drug in the effective treatment of TB [116]. The enzyme pyrazinamidase/nicotinamidase (Pzase) activates PZA to pyrazinoic acid which in turn disrupts the proton motive force, which is essential for membrane transport [94, 95, 116]. Research has also linked pyrazinoic acid and its n-propyl ester to fatty acid synthase I inhibition in Mtb [94, 117]. It was initially believed that PZA was dependent on low pH environments, however, some studies have shown that low pH and intra-bacterial acidification are not necessarily a prerequisite for PZA activity [118, 119]. However, this cannot be interpreted as if PZA generally works at neutral pH, irrespective of the bacilli's metabolic state [120]. It rather highlights the role of the bacterial metabolic state of the bacilli in determining the activity of the drug. Interestingly, PZA would still be active at neutral pH in dormant persister Mtb even though it will be expected to show more activity at lower pH [74, 120].

The mechanism of PZA resistance is still poorly understood. However, mutations in the gene that encodes for Pzase (*pncA*), and its promoter regions have been associated with PZA resistance. Mutations in this highly polymorphic gene have been suggested to be responsible for 72% - 99% of PZA resistance [121, 122]. DST for PZA is often highly inaccurate and complex as it requires low pH conditions for culturing. The ribosomal protein I (*rpsA*) was

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

initially suggested as a target of PZA, then the drug was found to inhibit trans-translation in Mtb [120]. However, this finding was disproved by Dillon *et al.*[74, 119]. Studies have also shown that PZA inhibits the aspartate decarboxylase (*panD*) gene [74, 123, 124]. This enzyme is involved in pantothenate and co-enzyme A synthesis, which plays a role in energy production. Initially, researchers were sceptical about *panD* as a drug target for pyrazinamide but a study by Gopal *et al.* confirmed the gene as a PZA drug target [74, 125].

1.6.1.1.5 Streptomycin

Streptomycin (SM) was the first antibiotic treatment for TB in 1944 [95, 120]. The aminocyclitol drug disrupts protein biosynthesis in slow-growing bacteria by irreversibly binding to the ribosomal protein S12 and 16S rRNA and therefore interfering with the binding of formyl-methionyl-tRNA to the 30S subunit of the bacterial ribosome. Resistance to SM soon developed due to its uncontrolled use in the previous century [126]. SM resistance has been attributed to mutations in the *rpsL* and *rrs* genes. The *rpsL* gene encodes for the ribosomal protein S12 while the 16S ribosomal subunit is encoded by the *rrs* gene. Mutations in the 7-methyl guanosine methyltransferase gene (*gidB*) have been linked to low-level streptomycin resistance. *rpsL* mutations account for 50% of SM resistance while *rrs* mutations have been attributed to 15% of resistance to the antibiotic [95, 120]. Mutations in *gidB* gene which encodes 7-methylguanosine methyltransferase have also been linked to low-level SM resistance. These mutations inhibit 16S ribosomal methylation and thus reduce the affinity between SM and the 16S ribosomal RNA (rRNA)-binding site [120, 127, 128].

1.6.1.2 Second-line drugs

1.6.1.2.1 Injectable drugs

Kanamycin (KAN), amikacin (AMK) and capreomycin (CM) are the Second-line injectable drugs currently used in the management of TB [74, 95]. Although the two aminoglycosides AMK and KAN and the cyclic polypeptide CM have different modes of action, the three drugs exert function by disrupting protein synthesis pathways. The aminoglycosides are known to bind to the 16S rRNA in the 30S subunit of the MTB ribosome, while CM targets translation pathways specifically by inhibiting the synthesis of proteins by MTB ribosomes

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

[129]. Mutations in the *rrs* gene have been linked to high-level resistance to these drugs [130, 131]. Mutations in the rRNA methyltransferase *tlyA* have also been linked to CM resistance. These mutations are known to inhibit ribosome methylation [95, 132]. The A1401G and G1484T mutations in the *rrs* gene are classified as high-confidence markers of resistance to the three antibiotics, with the former linked to 70%-80% of CM and AMK resistance and 60% KAN resistance [94, 133]. Studies have also reported cross-resistance between the three drugs; however, this is not surprising since the drugs share the same drug target. Low-level resistance to KAN has been attributed to mutations at position -10 and -35 of the *eis* promoter [95]. Up to 80% of clinical isolates exhibiting low-level KAN resistance carry these mutations [134, 135].

1.6.1.2.2 Fluoroquinolones

Fluoroquinolones (FQ) are some of the most important class of antibiotics currently prescribed as second-line treatment drugs for DR-TB. Newer generation FQs, such as levofloxacin and moxifloxacin are now integral components of DR-TB treatment regimens [74, 136]. FQs exert their effect by inhibiting bacterial DNA replication and transcription through binding to DNA gyrase and topoisomerase IV, enzymes critical for maintaining DNA supercoiling and stability [137]. The tetramer type II topoisomerase consists of two subunits α and β encoded by the *gyrA* and *gyrB* genes, which specialise in DNA supercoiling [138]. Genetic mutations in the two genes have been implicated in FQ resistance in Mtb, with most resistance-associated mutations found in the *gyrA* gene. The quinolone resistance determining regions (QRDRs) are located between codons 74 and 113 and codons 500 and 538 [139, 140]. Codons 88, 90, 91 and 94 of the *gyrA* QRDR are known to carry most of the common FQ-resistance mutations. Mutations in *gyrB* have also been linked with low levels of FQ resistance, although they are not as common as the other mutations. Intriguingly, position 95 in *gyrA* can also carry mutations not associated with FQ resistance since they can be present in susceptible Mtb isolates [141]. An interesting finding by Aubry *et al.* showed that the simultaneous occurrence of T80A and A90G in *gyrA* led to hyper-susceptibility to several quinolones [95, 142]. Resistance to FLQ is one of the important markers of the development of XDR-TB, thus, it is important to fully understand the full

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

mechanism of FQ resistance, as low-level resistance has been recently reported in some of the new generation FQs.

1.6.1.2.3 Ethionamide

Ethionamide (ETA), a derivative of isonicotinic acid, is a structural analogue of INH [74, 94]. The prodrug inhibits cell wall synthesis, specifically enoyl-acyl carrier protein reductase binding after it is activated by the mono-oxygenase enzyme. ETA resistance has been attributed to mutations in the *ethA* gene, which encodes the mono-oxygenase enzyme [143, 144]. Mutations in the transcriptional repressor gene *ethR* as well as *inhA* have also been linked to ETA resistance. Several studies have also identified the role of *inhA* mutations in ETA/INH co-resistance. Mutations in these two genes are known to account for 70% of ETA resistance [74]. Further investigations are required to fully understand the mechanism of resistance in ETA.

1.6.1.2.4 Para-aminosalicylic acid (PAS)

PAS was one of the first drugs to be introduced in TB clinical management together with streptomycin. Elucidating the mechanism of PAS resistance has been challenging until recently due to its gastrointestinal toxicity and low potency when compared to RIF and PZA [74]. PAS an analogue of para-aminobenzoic acid, is a competitive inhibitor of dihydropteroate synthase, which is involved in the folate synthesis pathway [95]. The inhibition disrupts bacterial folate metabolism and, subsequently its ability to synthesize nucleotides. PAS has also been implicated in interfering with iron acquisition mechanisms, though its exact role in siderophore metabolism remains under investigation[145]. Mutations in the *thyA* gene have been identified as the main mechanism of resistance, accounting for 40% of PAS resistance [94, 146]. Mutations in the *folC* gene which encodes dihydrofolate synthase were also detected in clinically resistant strains [147]. The usage of PAS as an anti-TB drug has been limited due to its high toxicity levels and side effects. More research is needed to fully understand the mechanism of PAS resistance in Mtb [74].

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

1.6.1.3 New and repurposed drugs

1.6.1.3.1 Bedaquiline

Bedaquiline (BDQ) is classified under the diaryl quinolones, a new class of compounds used in the treatment of TB [65]. This drug, which has been a success in high-burden countries, targets mycobacterial ATP synthase to inhibit bacterial respiration [65]. The drug is highly potent against dormant bacilli which makes it highly effective in treating both DS-TB and MDR-TB, especially when prescribed together with PZA [65]. BDQ functions by inhibiting mycobacterial Atp synthesis as it binds to the C subunit of the F0 complex of ATP synthase encoded by *atpE* [120, 148]. In 2005, in vitro studies by Andries *et al.* implicated mutations in the *atpE* gene with BDQ resistance [65]. The gene encodes for the bacterial F_1F_0 proton synthase, which is involved in ATP synthesis as well as membrane energetics [94, 149]. Mutations in the Rv0678 and *pepQ* genes have also been linked to BDQ resistance as well as cross-resistance with clofazimine. Cross-resistance of both drugs has been attributed to the upregulation of the *mmpL5* encoding MmpL5 efflux pump [120]. The lack of DST data for new and repurposed antibiotics is still a challenge that needs to be addressed if we are to determine the complete mechanism of resistance to BDQ.

1.6.1.3.2 Clofazimine

Clofazimine (CFZ), a riminophenazine antibiotic, was discovered in 1954 and was initially used for the treatment of leprosy as it was ineffective against TB [150, 151]. Studies later showed that CFZ is active against MDR-TB, which led to the WHO recommending the drug as part of the new standardized short-course regimen for DR-TB [151]. It must be noted that although the drug was approved by the FDA in 1986, its use in the treatment of DR-TB has not been approved by any stringent regulatory body and it is therefore prescribed “off-label” [74, 152]. The exact mechanism of CFZ action is still being studied, however, research performed in *M. smegtmatis* suggests that it is a prodrug activated by NAD dehydrogenase, to release reactive oxygen species upon re-oxidation by oxygen [74, 94, 153]. Mutations in Rv0678 have been linked to CFZ resistance together with mutations in *pepQ* and Rv1979c genes [154, 155]. Resistance to CFZ has also been associated with BDQ as explained above. The mechanism of action of CFZ is poorly understood due to its limited use as it is associated

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

with adverse side effects. More research is needed to fully exploit the capabilities of this drug in TB management.

1.6.1.3.3 Linezolid

Linezolid was the first oxazolidinone to get approval for the treatment of TB [156]. The antibiotic is a protein synthesis inhibitor which functions by binding to the V domain of the 50S subunit of the mycobacterium ribosome [94, 157]. Linezolid resistance has been linked to mutations in the 23S (*rrL)* gene [94]. G2061T and G2572T mutations in the *rrL* gene have been implicated in causing the development of high-level linezolid resistance using in vitro studies [158]. The study further reported that no alterations to the *rrL* gene were identified in isolates that exhibited low-level linezolid resistance. The role of *rrL* mutations in linezolid resistance was further supported by Bloemberg *et al.,* who detected A2572C and G2576T mutations in a patient with linezolid-resistant TB [159]. Studies by Zimekov *et al.*have also linked a mutation in the *rplC* gene to the development of linezolid [160]. The gene encodes for the L3 protein on the 50S ribosome and the C154R mutation is the most frequent among linezolid resistant isolates [161]. Linezolid has been shown to improve treatment outcomes in complicated MDR-TB as well as XDR-TB cases thus it will be of great benefit to decode its mechanisms of resistance [161].

1.6.1.3.4 Pretomanid and Delamanid

Pretomanid and Delamanid belong to the nitroimidazole class of antibiotics [67, 162]. Both compounds are prodrugs activated by deazaflavin-dependent nitro-reductase, which is encoded by the *ddn* gene [94]. The enzyme breaks down the prodrug into des-nitro-imidazole and two other unstable metabolites. Des-nitro-imidazole produces reactive nitrogen species as well as nitric oxide, which may promote host macrophages acting against Mtb [69, 70, 163]. Delamanid, formerly known as OPC-67683, is known to disrupt the synthesis of bacterial cell walls by inhibiting the synthesis of methoxy-mycolic and keto-mycolic acids [68]. Mutations in *ddn* and *fgd1* genes which are involved in the activation of the prodrug as well as mutations in *fbiA, fbiB* and *fbiC* have been linked to resistance to both drugs [69, 94]. The *fbi* genes encode for proteins involved in the F420 biosynthetic pathway.

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

1.7 MTBC STRAINS AND LINEAGES IN-DEPTH

The MTBC is made up of various closely related bacterial species and sub-species, which include the human-adapted *M. tuberculosis* sensu stricto and *M. africanum* strains. In addition, various animal-adapted forms exist, which include *M. bovis* (cattle pathogen), *M. caprae* (sheep and goats), *M. microti* (voles), *M. pinnipedii* (sea lions and seals), *M. origys, M. mungi, M. surricate*, the chimpanzee bacillus and the dassie bacillus [28, 53]. *M. bovis* used to be a significant source of TB infection in humans, especially in children fed on raw bovine milk diet. However, improvements in food processing, particularly milk pasteurization, resulted in a significant drop in human infection [164]. It is important to note that human-to-human transmission of *M. bovis* is very rare [28]. Likewise, animal-to-animal transmission of *M. tuberculosis* or *M. africanum* has not been reported, which may indicate that different MTBC strains are best adapted to their hosts.

The MTBC also consists of *M. canettii* and other "smooth tuberculosis bacilli" (STB), which differ in several ways from the classical MTBC [28]. This group of tuberculosis bacilli, first reported in the 1960s, shows strong evidence of HGT and harbours higher levels of genetic diversity compared to classical MTBC [52]. These enigmatic organisms have a smooth colony morphology distinct from the rough colony morphology associated with classical MTBC [28, 165]. To date, only 60 isolates of STB have been reported with the majority of those isolated from TB patients in Djibouti, or individuals who had been to the Horn Africa region. The lack of HGT and human-to-human transmission suggests that *M. canettii* and other STB might be environmental opportunistic microbes occasionally infecting humans [28]. Epidemiological and phylogeographical evidence also suggests that their environmental reservoir might be somewhere in the Horn of Africa region [166]. On the other hand, *M. tuberculosis* and *M. africanum* have no known animal reservoir and have limited survival outside the human body, which makes them obligate pathogens [28]. Transmission to secondary hosts is only possible when the pathogen causes acute disease. This direct link between virulence and transmission is not common in many other pathogens, where transmission is independent of the disease current.

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Comparative WGS studies have broken down MTBC into 9 human-adapted lineages (Figure 1). The modern clade forms a monophyletic group made up of Lineages 2, 3 and 4. These lineages harbour a deletion in the TbD1 genomic region [31]. The presence or absence of this deletion was what led to MTBC strains being classified as either evolutionary ancient or modern. On the other hand, the 'ancient' strains do not form a single phylogenetic group and hence they are paraphyletic [52]. As previously mentioned, human MTBC follows a phylogeographical population structure with some lineages more widely distributed around the globe than others [52]. Lineage 2 and Lineage 4 are more widespread compared to the other lineages, with the former (also known as the East Asian strain) predominantly present in East Asia and the latter (also known as the Euro-American lineage) frequenting Asia, Europe, Africa and America. Lineage 1 and Lineage 3 have a less widespread distribution that is restricted to East Africa, Central, South and Southeast Asia[52]. Lineages 5, 6 and 7 are much more geographically restricted to some specific regions in Africa [52]. Lineages 5 and 6, also known as *M. africanum* West Africa 1 and West Africa 2 respectively, are mostly found exclusively in the Western Africa region or from recent immigrants of West African origin and are attributed to causing up to 40%-60% of TB infections in that region [167] [52]. The Lineage 5 strains are predominantly found further to the East in regions bordering the Gulf of Guinea [52, 168]. Lastly, the Lineage 7 strains are mostly restricted to Ethiopians and recent immigrants from that region. The reason why these three Lineages are exclusive to Africa is still elusive [52]. Figures 5 and 6 show the global distribution of human-adapted MTBC with the colours in the latter corresponding to those in the former. Aside from these broad categories, the modern lineages can be further divided into sub-lineages that also follow a phylogeographical pattern [61, 169, 170]. Another layer of classification can be applied to these sub-lineages to distinguish between "specialist" and "generalist" sub-lineages with the latter referring to lineages which are widely distributed worldwide and the former referring to lineages which are only restricted to a specific geographical region [61].

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

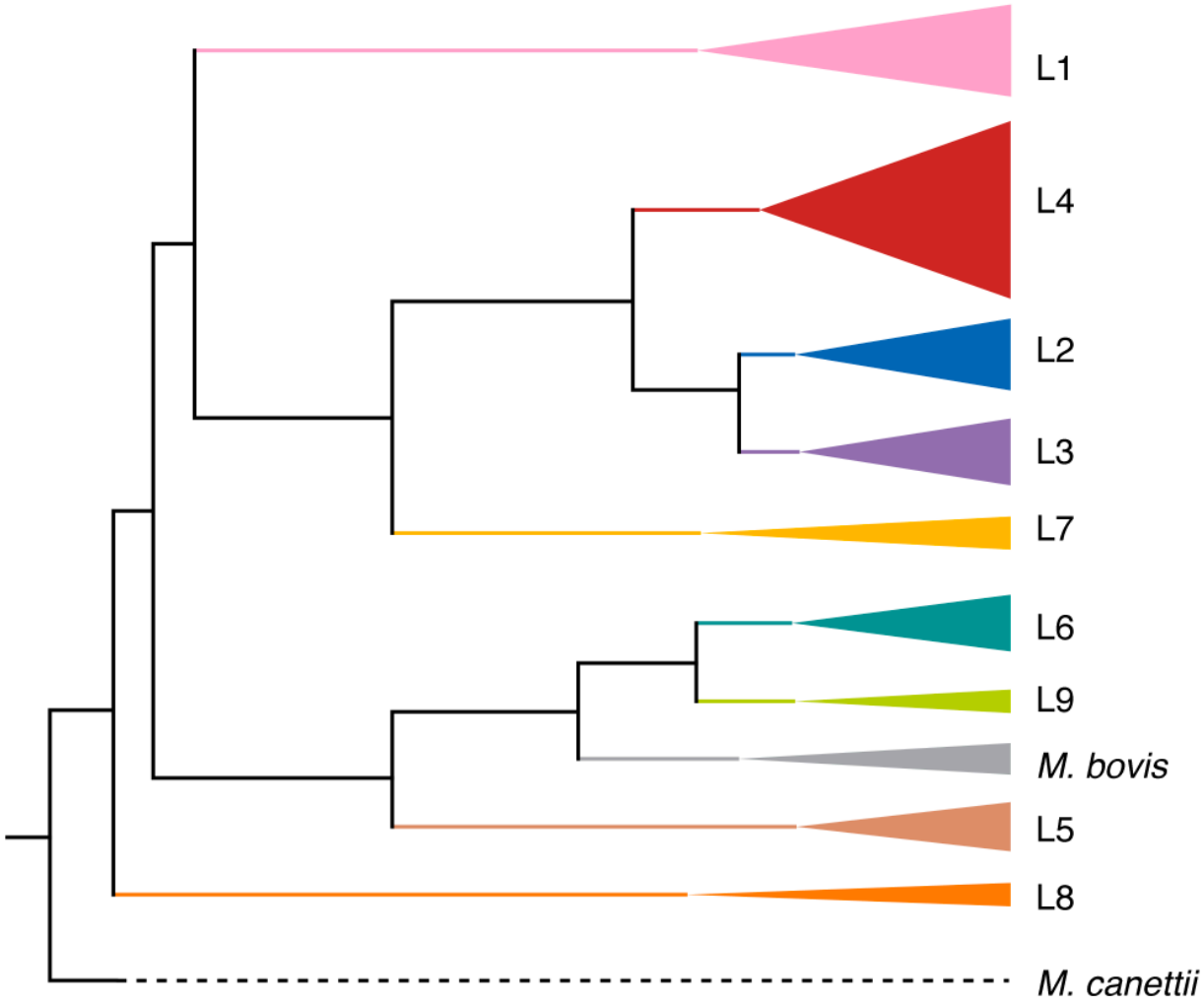


Figure 5: Maximum Likelihood tree of the MTBC complex. The branch colours correspond to the main MTBC lineages with *M. canettii* as an outgroup. Used from Coscolla *et al.* [61, 171] under the terms of the Creative Commons Attribution Licence (CC BY).

© University of Pretoria

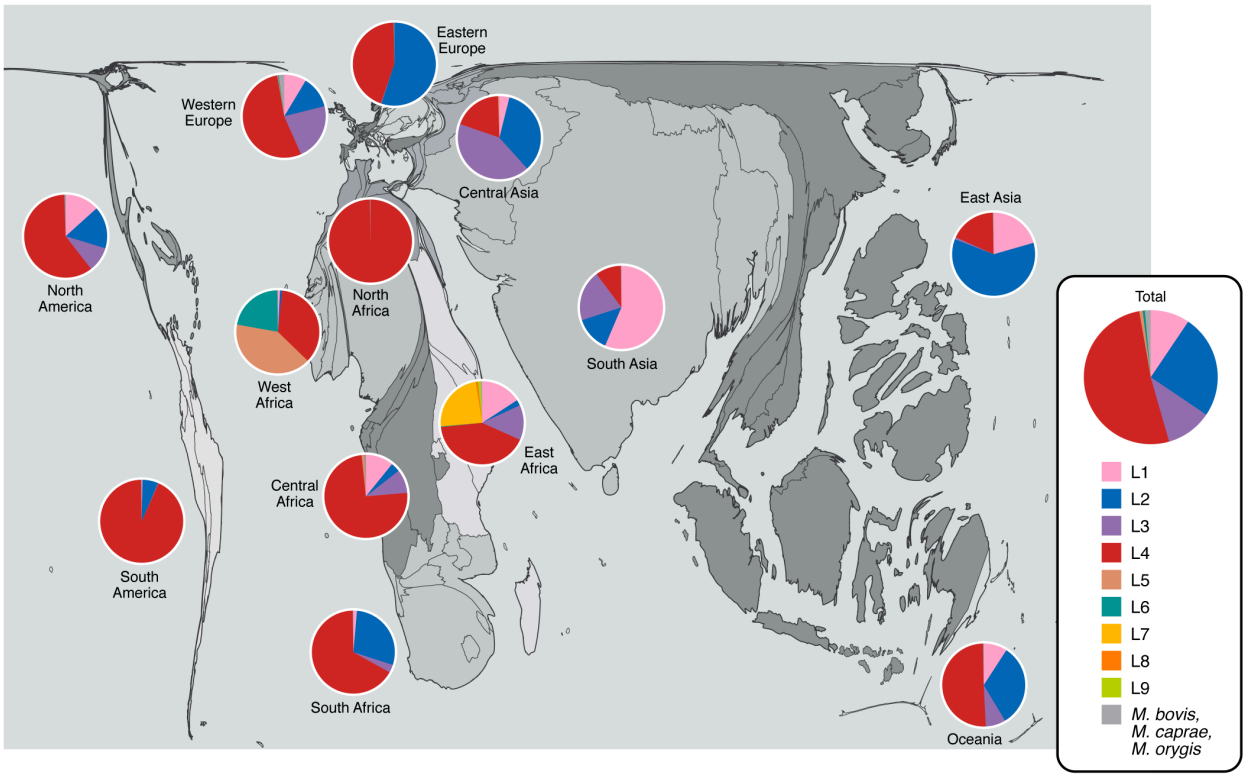


Figure 6: Cartogram showing the global TB burden by *M. tuberculosis* lineage. Countries are scaled to indicate TB incidence in 2021. Pie charts show the distributions of the *M. tuberculosis* lineage Lineage 1-Lineage 9 as well as the animal-adapted lineages *M. bovis, M. caprae* and *M. orygis* from clinical strains by geographic region according to Napier *et al.*'s scheme [61, 172]. Used from Napier *et al.*[172] under the terms of the Creative Commons Attribution Licence (CC BY) http://creativecommons.org/licenses/by/4.0/.

1.7.1 Lineage 1

Lineage 1 strains are predominantly prevalent around the rim of the Indian Ocean as shown in Figure 6 above. This is the region that encompasses East Africa, South Asia, and Southeast Asia [173]. Over the years, TB research has been mainly targeted on the laboratory strain H37Rv, which belongs to Lineage 4, as well as strains belonging to Lineage 2 and Lineage 4, due to their wide geographical range and association with global TB outbreaks as well as drug resistance. However, research on Lineage 1 global diversity has been very limited and often restricted to local populations [174, 175]. Earlier classifications of Lineage 1 strains using spoligotyping classify Lineage 1 strains as the East African Indian (EAI) group based on the presence of deletions at DVR39-42, 44 and 48. Strains that harbour the SIT19 spoligotype are collectively known as EAI2 and can be divided into two main groups, the

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Philippines-associated WAI2_MNL (Manila) and the EAI2_NTB (Nonthaburi), which is mostly found in Thailand [176]. WGS-based SNP typing methods offered a better resolution in classifying Lineage 1 strains and earlier studies classified them into two major clades Lineage 1.1 and Lineage 1.2. These two main groups were further divided into five sublineages, namely Lineage 1.1.1, Lineage 1.1.2, Lineage 1.1.3, Lineage 1.2.1, and Lineage 1.2.2 [176, 177], which were further expanded as shown in Figure 7 below. A recent scheme by Napier *et al.*[172] proposed a more refined scheme that divides Lineage 1 into 3 main groups Lineage 1.1-Lineage 1.3, the previous scheme was expanded into seven sub-lineages, Lineage 1.1.1.1-Lineage 1.1.3. In this classification scheme, the authors propose a new Lineage 1.2.1 the previous Lineage 1.2.1 was reassigned to Lineage 1.2.2, and the previous Lineage 1.2.2 became Lineage 1.3 which is further divided into Lineage 1.3.1 and Lineage 1.3.2 [176]. Sublineages of Lineage 1 also follow a phylogeographical distribution with L.1.1.1.1 mostly prevalent in Vietnam [178] while Lineage 1.1.3 strains are common in Myanmar [179], East Africa [173] and Bangladesh [180]. Lineage 1.1.2.2 and Lineage 1.2.2 (previously classified as Lineage 1.2.1) are commonly isolated in India and Southeast Asia, respectively [176, 181]. Understanding the evolutionary processes in Lineage 1 is crucial to global TB control efforts since the lineage is endemic to regions that carry the largest share of the global TB burden and is the most prevalent strain in the Philippines, a country that carries the 4th highest global TB burden [7, 173, 176].

© University of Pretoria

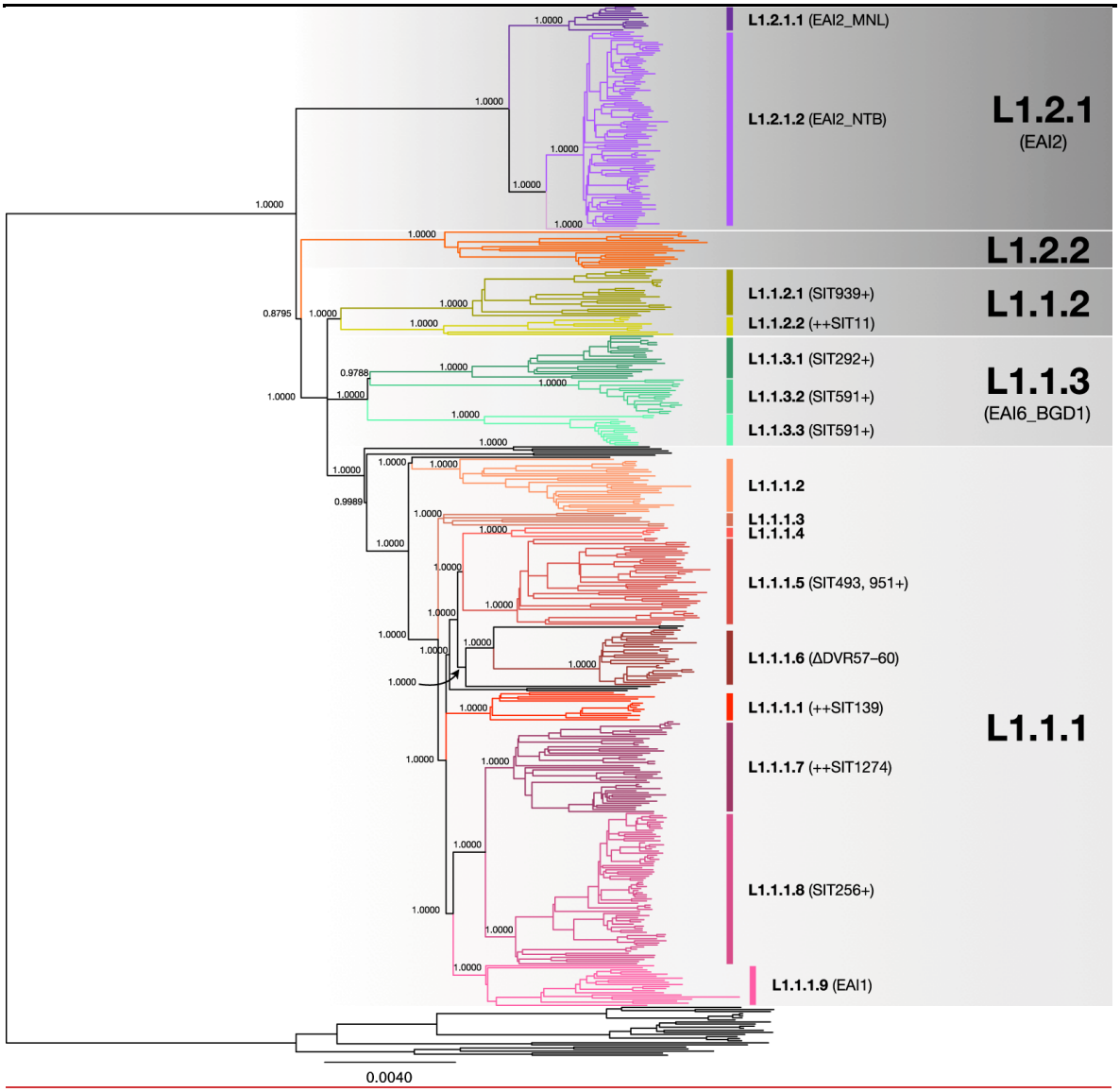


Figure 7: A Bayesian phylogenetic tree showing the sublineages of lineage 1. Used from Palittapongarnpim *et al.* [181] under the terms of the Creative Commons Attribution Licence (CC BY) http://creativecommons.org/licenses/by/4.0/.

1.7.2 Lineage 2

Lineage 2 (strains have relatively limited genetic diversity compared to other lineages [52, 178, 182]. For this reason, genotype characterisation of Lineage 2 has been poor. Several Lineage 2 classification frameworks have been proposed over the years. The LSP markers described in section 1.4 above could only resolve three major sub-lineages of Lineage 2

© University of Pretoria

(Lineage 2.1, Lineage 2.2.2, and Lineage 2.2.1) based on the presence or absence of five major deletions in the Mtb genome (RD105, RD207, RD181, RD150 and RD142) [73, 182]. Lineage 2.1 also known as proto-Beijing strains possess only the RD105 deletion, and Lineage 2.2.2 harbour the RD105 and RD207 deletions. Lastly, deletions in the RD105, RD207 and RD181 are characteristic of Lineage 2.2.1 [73]. Lineage 2.2 is also known as the Beijing strain and has gained notoriety because of its association with high virulence, heightened transmissibility, treatment failure and DR-TB [183]. Spoligotyping is often used as a budget-friendly method of genotyping Lineage 2 strains. However, its use is hampered by the low resolution of Lineage 2 sub-lineages as it can only differentiate between Lineage 2.1 (SIT523 (Manu Ancestor):777777777777771) and Lineage 2.2 (SIT1 (Beijing): 000000000003771) [182]. Earlier SNP-based approaches were used to differentiate between Ancestral and Modern Beijing strains [182, 184]. Modern Beijing strains were identified by an IS6110 insertion pattern in the NTF region [182, 185]. The MUT genes, which encode for DNA repair enzymes *ogt, mutT2,* and *mutT4,* have also been used to differentiate between modern and ancestral Lineage 2 strains with the latter harbouring the *mutT2G58C* and *mutT4R48G* mutations [44, 184]. Mestre *et al.*[186] also proposed a polymorphism-based Lineage 2 genotyping scheme based on DNA replication, recombination, and repair (3R) genes, which classified strains of the Beijing family into 26 types of Bmyc sequences.

The accumulation of publicly available WGS data over the past few years has seen the development of better genome-wide SNP-based genotyping methods for Lineage 2 with much superior resolution compared to the traditional approaches. Using a phylogenetic framework, genome-wide SNPs can be identified and extracted from WGS data to curate genotyping catalogues for Mtb. A widely used SNP-based classification scheme proposed by Coll *et al.*[177] incorporates other deletions (RD207, RD181, RD150, and RD142) in characterising Lineage 2.2 strains in addition to the classical RD105 deletion [44]. However, this scheme offers a limited resolution for Lineage 2 strains [44]. In 2015, Merker *et al.*[187] used a global dataset of ~100 Lineage 2 WGS strains to suggest a genotyping scheme that divides Lineage 2.2 (Beijing) into eight major lineages, namely: Asia Ancestral 1-3, Asia African 1-2, Pacific RD150, Europe-Russia B0/W148 and Central Asian [182]. Another study by Liu *et al.*[183] identified a set of 44 SNPS that further resolves Modern Beijing strains within Lineage 2.2.1, in their study, they refer to Modern Beijing sub-lineage as Lineage 2.3. The Lineage 2.3 nomenclature for Modern Beijing strains has been used in

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

other studies [188, 189], however, it is problematic as it renders Lineage 2.2 paraphyletic [182]. Shitikov *et al.* proposed a high-resolution scheme (HGDI 0.79), which splits Lineage 2 into two sub-lineages: Lineage 2.1 and Lineage 2.2. In this scheme, Lineage 2.2 (Beijing) is further split into 10 sublineages, namely the three "ancestral" Beijing groups: ancestral Asia 1, ancestral Asia 2, ancestral Asia 3; and the seven "modern" Beijing sub-lineages: Asiatic-African 1, Asia-Africa 2, Asia-Africa 2/RD142, Asia-Africa 3, Pacific RD150, Europe/Russia B0/W148 outbreak and Central Asia [44, 190].

Recently, two additional Lineage 2 sub-lineages, Asia Ancestral 4 [191] and Asian African 3 [190], have been proposed [182]. Using a global dataset of 781 Lineage 2 clinical strains, Rutaihwa *et al.*[192] drifted away from the previous Lineage 2.2.1 and Lineage 2.2.2 definitions and proposed a different Lineage 2 classification scheme, which divides Lineage 2.2 into ten sub-lineages namely: Lineage 2.2.1-Lineage 2.2.10 for Asia Ancestral 1-3, Asian African 1, Asian African 3, Pacific RD150, Asian African 2, RD142, Europe-Russia B0/W148 and Central Asian, respectively. The most comprehensive Lineage 2 classification scheme to date was proposed by Thawornwattana *et al.*[182]. This classification scheme, which expands the Lineage 2.2 sub-lineage, consists of 15 genotypes and three levels as seen in Figure 8 below [182]. The classification of Lineage 2 strains is complicated due to the high diversity within the lineage, which makes it challenging to define sublineages clearly. The scheme proposed by Thawornwattana *et al.*[182] not only clears controversy by acknowledging and integrating previously proposed schemes, but it's also easy to interpret and is flexible enough to accommodate other novel genotypes should they be discovered in future. As one of the most prevalent lineages in high TB burden areas, fully understanding the genetic diversity of Lineage 2 has major global implications as it has the potential to shape global TB policy and help improve global surveillance efforts.

© University of Pretoria

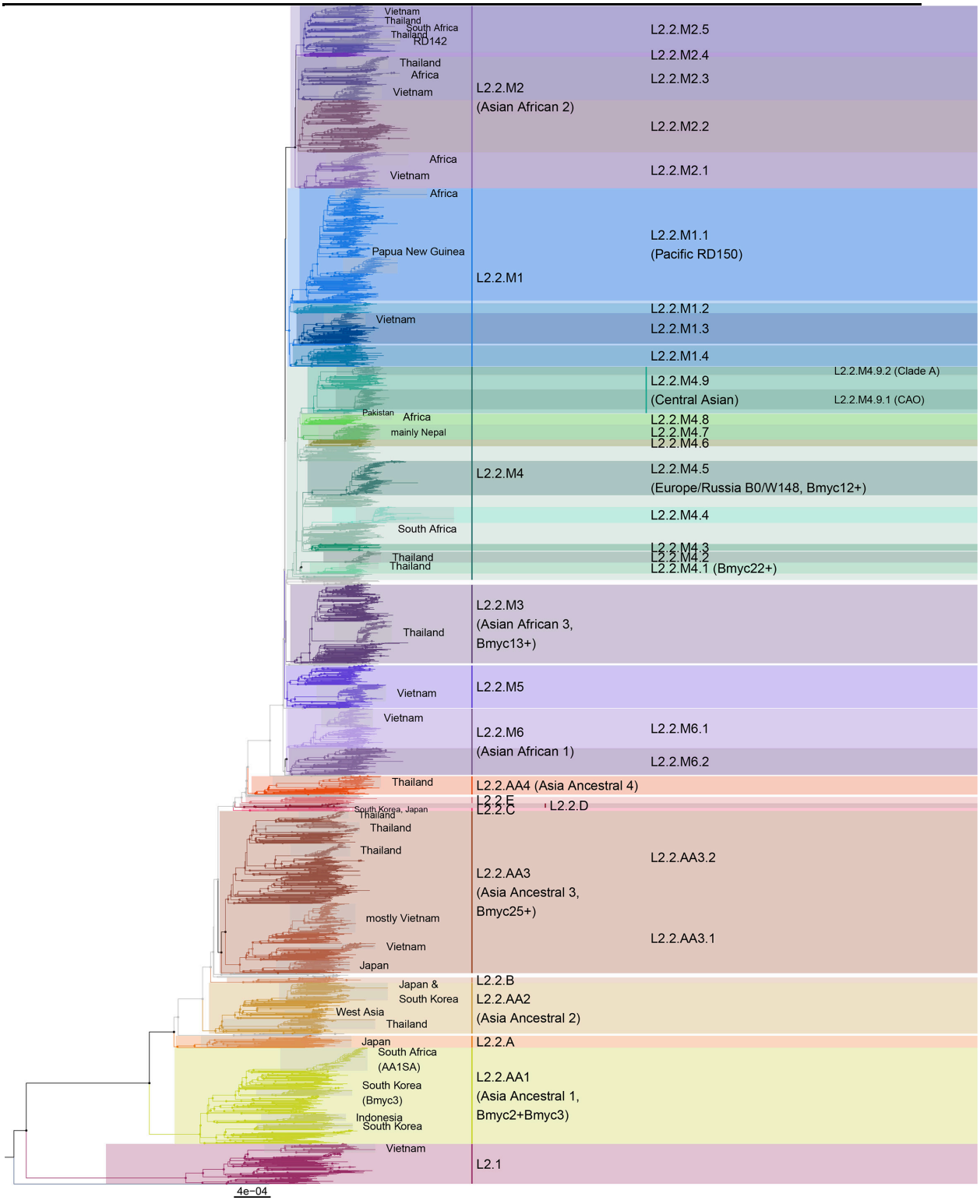


Figure 8: A maximum likelihood tree showing the sub-lineages of lineage 2. Used from Thawornwattana *et al.*[182] under the terms of the Creative Commons Attribution Licence (CC BY) http://creativecommons.org/licenses/by/4.0/.

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

1.7.3 Lineage 3

Like Lineage 1, Lineage 3, also known as Central Asian Strains (CAS), remains understudied despite being endemic to some of the global regions carrying the largest share of the world's global TB burden. Lineage 3 is predominantly found in Central Asia, Western Asia, East Africa and South Asia [193]. Lineage 3 strains are characterised by the unique RD750 deletion and include the sublineages for the CAS1-Delhi, CAS1-Kili CAS2 spoligotypes. A classification scheme by Coll *et al.*[177] grouped all non-CAS1-Delhi strains into sub-lineage Lineage 3.1 and further divided sub-lineage Lineage 3.1 into two other sublineages Lineage 3.1.1 and 3.1.2 to accommodate the CAS1-Kili and the CAS/CAS2 strains, respectively. Napier *et al.*[172] proposed an improvement to Coll's scheme that divides Lineage 3.1 into five sublineages. Sublineage Lineage 3.1 was further divided into Lineage 3.1.1 and Lineage 3.1.2. The latter is further divided into two sublineages; Lineage 3.1.2.1 and Lineage 3.1.2.2. This scheme also introduced a new lineage, Lineage 3.2 shown in Figure 9. Recently, Shuaib *et al.*[194] investigated how Lineage 3 came to the African continent and proposed an updated Lineage 3 classification scheme that complements the scheme proposed by Napier *et al.*[172] In Shuaib's scheme, Lineage 3 is divided into 6 sublineages, from Lineage 3.1 to Lineage 3.6. This new scheme also introduces new sublineages Lineage 3.3, Lineage 3.4, Lineage 3.4 and Lineage 3.5 based on previously unclassified Lineage 3 strains as shown in Figure 9. More work is required in exploring the genetic diversity of Lineage 3, given its prevalence in some of the global high TB burden regions.

© University of Pretoria

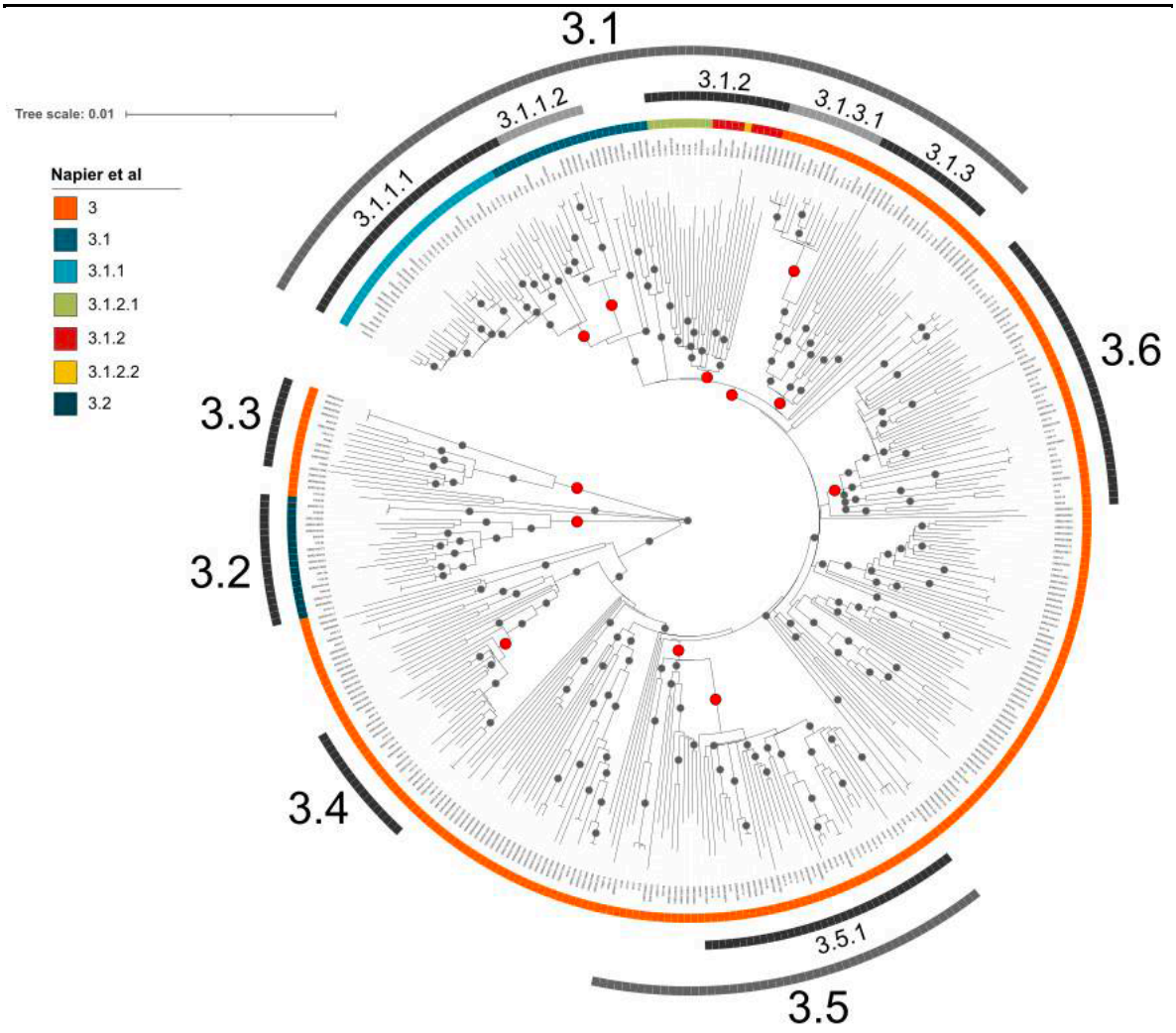


Figure 9: A Mid-point rooted maximum likelihood (ML) phylogenetic tree of Lineage 3. Used from Shuaib *et al.* [194] under the terms of the Creative Commons Attribution Licence (CC BY) http://creativecommons.org/licenses/by/4.0/.

.

1.7.4 Lineage 4

Compared to other human-adapted MTBC lineages, Lineage 4 (Euro-American lineage) is distributed at significant frequencies on all inhabited continents, making it the leading global cause of human TB [170, 195]. RD deletion analysis identified the *pks* 15/1 7-bp deletion as unique to Lineage 4 strains. Among the TbD1-deleted (modern) Mtb lineages, Lineage 4 is believed to be the most recent. Earlier studies showed that large sequence deletions can be

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

used to divide Lineage 4 into 10 distinct sublineages. These deletions are named RD115, RD122, RD174, RD182, RD183, RD193, RD219, RD724, RD726 and RD761 [196]. In addition to this, another group of Lineage 4 strains with no deletions was identified that previously was described as an "H37Rv-like" group [32]. Spoligotyping classified Lineage 4 based on the deletion of spacers 33-36 within the DR locus and divided Lineage 4 into lineages Haarlem, T-type, Latin American-Mediterranean (LAM), S-type, and X-type [196, 197]. One of the most widely used Mtb classification schemes by Coll *et al.* divided Lineage 4 into nine sub-lineages (from Lineage 4.1 to Lineage 4.9). The most comprehensive study focused on Lineage 4 was one by Stucki *et al.*[170]. Their proposed scheme guided by spoligotyping data and previous work done by Coll *et al.*[177] divided Lineage 4 into 10 sublineages (Lineage 4.1.1 to Lineage 4.10) as shown in Figure 10. Using principal component analysis (PCA), the study identified a clear distinction of seven sublineages (Lineage 4.1.1, Lineage 4.1.2, Lineage 4.1.3, Lineage 4.2, Lineage 4.3, Lineage 4.4, Lineage 4.10). However, sublineages Lineage 4.5, Lineage 4.6.1 (Uganda), and Lineage 4.6.2 (Cameroon) showed poor resolution. The study showed Lineage 4.3 (LAM) as the most frequent, accounting for ~20% of global Lineage 4 strains. Figure 11 shows the distribution of the rest of the Lineage 4 sublineages. Lineage 4 sublineages are also known to follow a phylogeographical distribution as shown in Figure 11. Lineage 4.1.2 (Haarlem), Lineage 4.3 (LAM) and Lineage 4.10 (PG33) are known to be more "generalist", as they occur globally. Lineage 4.1.3 (Ghana), Lineage 4.5, Lineage 4.6.1 (Uganda) and Lineage 4.6.2 (Cameroon) can be classified as "specialist", as they occur at high frequencies in specific regions of Africa and Asia with little or no presence in Europe and America [170]. Of the remaining three sublineages, Lineage 4.1 (X-type) is mainly prevalent in the Americas with little representation in Southern Africa, Europe and Asia; Lineage 4.2 (Ural) and Lineage 4.4 are mostly prevalent in some regions of Africa and the Russian part of Asia, but mostly absent in the Americas. Napier *et al.*[172] recently expanded Coll's scheme and proposed an updated scheme with 52 sub-lineages of Lineage 4. In contrast to Stucki's scheme, Napier's classification scheme lists Lineage 4.1 to Lineage 4.9.1.

© University of Pretoria

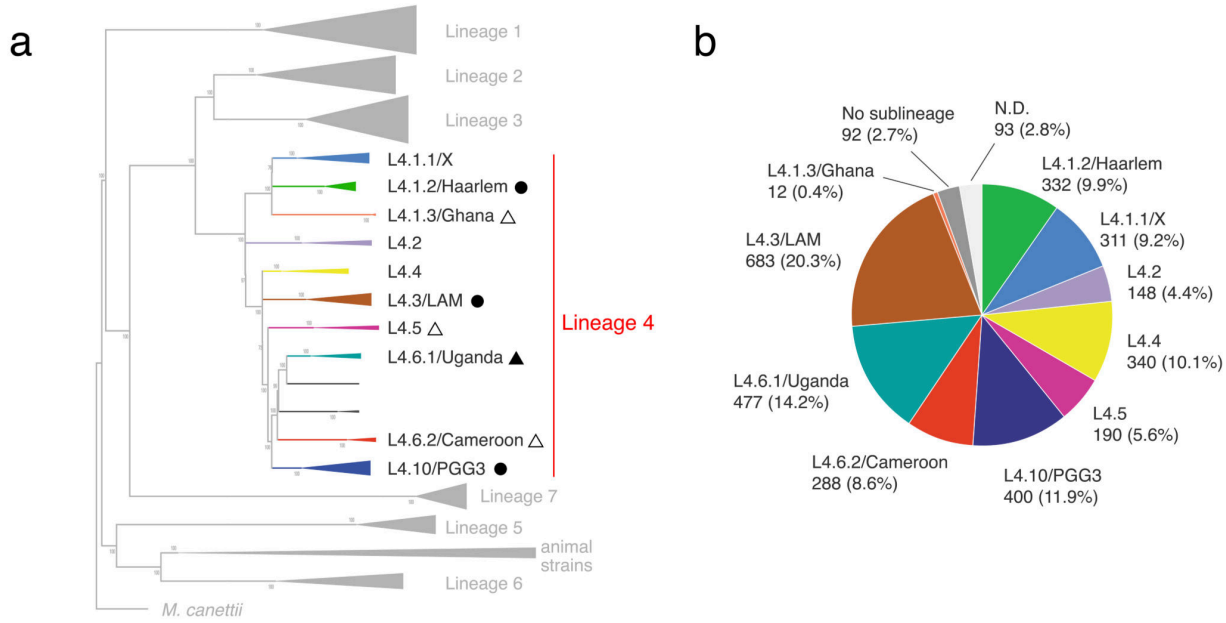


Figure 10: (a) shows ten sub-lineages of Lineage 4 according to Stucki *et al.* [170] based on the Coll *et al.*[177] scheme and supported by spoligotyping data. Specialist sublineages are represented by black triangles, while the black circles represent the generalists. (b) The global distribution of each sublineage, ND represents undefined strains. Used with permission from Springer Nature Stucki *et al.*, copyright 2016 [170]

© University of Pretoria

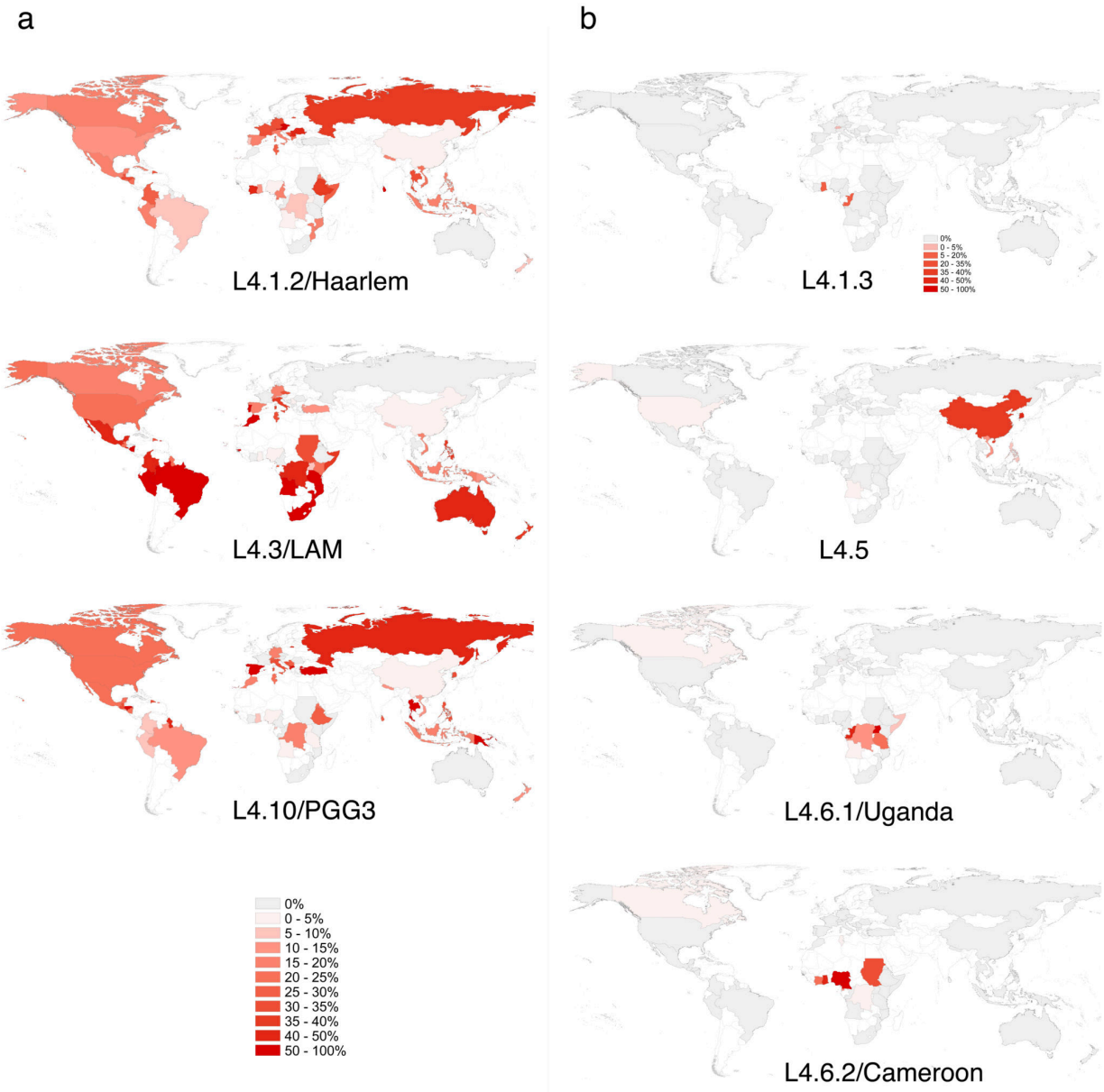


Figure 11:(a) Global distribution of the different sublineages of Lineage 4. Lineage 4.1.2(Haarlem), Lineage 4.3(LAM) and Lineage 4.10(PGG3) are found at high frequencies globally and are classified as generalists. (b)The specialists' sublineages Lineage 4.1.3 (Ghana), Lineage 4.5, Lineage 4.6.1 (Uganda) and Lineage 4.6.2 (Cameroon) were only restricted to specific geographical regions. Used with permission from Springer Nature Stucki *et al.*, [170] copyright 2016.

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

1.7.5 Lineage 5, 6, 7, 8, 9 (*M. africanum*)

The diversity of the MTBC on the African continent remains severely under-characterised despite the continent being hypothesised as the origin of the MTBC. In contrast to lineages such as Lineage 2 and Lineage 4, WGS data for *M. africanum* strains is scarce. *M. africanum* was first reported in 1968 [198] as a phonotypical intermediate between *M. bovis* and *M. tuberculosis* [199]. The primary use of phenotypic methods in characterizing *M. africanum* resulted in a rather poor understanding of the phylogenomics of these lineages [200]. The use of comparative genomics in characterizing *M. africanum* offered better resolution as it revealed that the common ancestor of *M. africanum* strains went through deletion of a specific RD9 region giving rise to the "specialist" Lineage 5 lineage (*M. africanum* 1), which is found in Ghana, Nigeria and Benin; and the Lineage 6 lineage (*M. africanum* 2), which is mostly prevalent in West Africa encompassing Senegal and Gambia [31, 32, 171]. Lineage 6 strains are also characterised by deletions RD7, RD8, and RD10, which are also missing in the animal-adapted strains of the MTBC [200]. However, these three regions are present in Lineage 5 strains suggesting that Lineage 5 branched from the same common ancestor whose RD9 region was deleted with the branching happening before RD7, RD8, and RD10 were lost [200]. It has been suggested that Lineage 5 might be locally adapted to the populations of specific regions of West Africa, a notion that has been cemented by two studies that found an association between the Ewe ethnicity in Ghana and Lineage 5 [171, 201, 202]. In 2021 Coscolla *et al.*[171] proposed a new lineage, Lineage 9, based on a set of 5 WGS genomes that had not previously been classified but showed relatedness to both Lineage 5 and Lineage 6. Phylogenetic analysis of these strains showed that they were sister clades of Lineage 6, which branched between Lineage 6 and the animal clade A1 as shown in Figure 12. This sister clade harboured the same RD702 deletions as Lineage 6 but lacked RD1 and RD5 deletions present in animal-associated lineages. The study further characterised the sublineages Lineage 5 and Lineage 6 and confirmed the previously described RDs (RD7, RD8, RD9, RD10, RD713 and RD715 [31, 32, 171, 203]. However, the study could not confirm the RD711 deletion as a marker for Lineage 5 as previously proposed by Gagneux *et al.* [32] as the deletion was only observed in a subset of Lineage 5 strains. Strains harbouring RD711 deletions were found to form a monophyletic clade within Lineage 5 [171]. The study subdivided Lineage 5 into three main sublineages, namely: Lineage 5.1, Lineage 5.2 and Lineage 5.3. Lineage 5.1 showed a high level of genomic

© University of Pretoria

diversity and was further divided into five sublineages (Lineage 5.1.1-Lineage 5.1.5) [171]. For Lineage 6, three main sublineages were proposed (Lineage 6.1-Lineage 6.3), which could also be further divided into three sublineages for each to make a total of 9 sublineages (L.6.1.1-L.6.3.3).

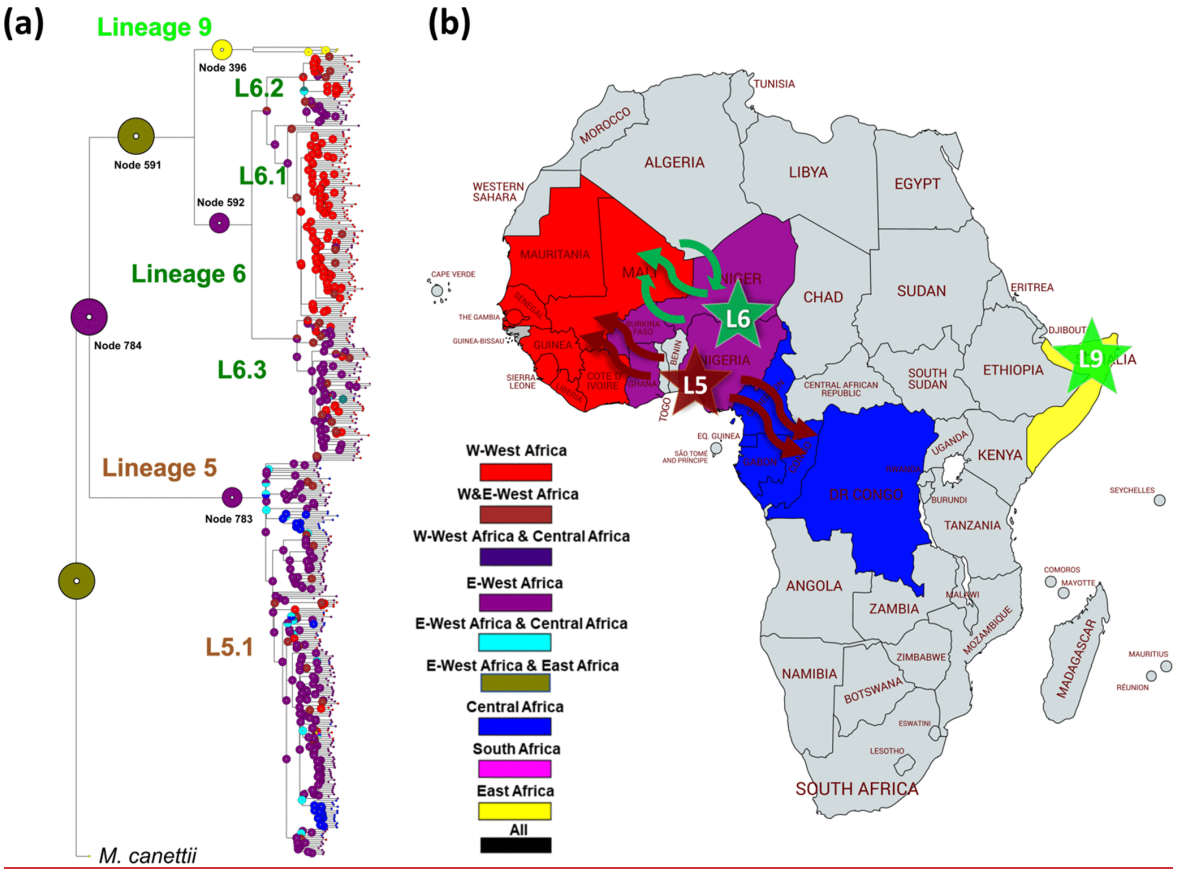


Figure 12: Geographical ancestral distribution of Lineage 5, Lineage 6 and Lineage 9. (a) Reconstruction of the ancestral region using a Bayesian approach on the ML tree with circles corresponding to the likelihood of ancestral ranges and the most probable ancestry areas are colour-coded. (b)The four geographical regions of Lineage 5, Lineage 6 and Lineage 9 with stars showing the most probable ancestry region for each lineage. The arrows indicate the movements of strains. Used from Coscolla *et al.* [171] under the terms of the Creative Commons Attribution Licence (CC BY) http://creativecommons.org/licenses/by/4.0/.

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

1.8 PHYLOGENOMICS

The unavailability of useful biological data from earlier decades necessitates the use of analytical strategies that aided in exploring events, which gave rise to the current overall Mtb population as we know it. Telford *et al.*[204] define phylogenomics as the use of large sets of sequence data of aligned genes from genomes and expressed sequence tags (EST) to infer evolutionary trees (phylogeny) of relationships between sequences and the organisms they are derived from. In simple terms, a phylogenetic tree consists of nodes that are connected by branches, where each branch represents the continuation of a genetic lineage through time, and each node represents the emergence of a new lineage as seen in Figure 13 [205]. With the availability of large amounts of WGS data and improved computational tools, phylogenetic and evolutionary analyses are now important aspects of infectious disease research. In the following sections, we describe two widely used approaches to phylogenomic inference.

1.8.1 Neighbour-joining (NJ)

The NJ algorithm [206] has been a favourite for biologists over the years mostly due to its relative speed and robustness for sequences with low divergence [207]. The algorithm first converts the sequence (DNA or protein) into a distance matrix. The distance matrix is a representation of the evolutionary distances between the sequences. For large inputs, the algorithm is fast and reasonably accurate, especially for cases where the rate of evolution is not extremely high or low [208]. For that reason, NJ is used for a range of phylogenetic applications such as constructing initializing trees for more computationally intensive algorithms and generating fast trees from large sequence alignments. Some of the popular alignment programs implement NJ in their algorithms [209-213]. However, the reliance on precomputed distance matrices instead of sequence data in NJ can also result in information loss due to a reduction in complexity. The assumption of additivity of distances (i.e. the distance between two taxa should be a sum of branch lengths) can also produce inaccurate trees in cases where the distance matrix is not truly additive (e.g. homoplasy or convergent evolution). Inaccurate distance matrices can also result in the estimation of incorrect evolutionary relationships. Unlike other methods, NJ does not provide direct statistical

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

support and hence provides less insight into the statistical confidence in the estimated phylogenetic relationships. The method is also susceptible to long-branch attraction and can perform poorly when there is rate heterogeneity.

1.8.2 Parsimony

Unlike the distance-based methods, parsimony and ML approaches map the history of genetic sequences onto a tree. The idea of parsimony was first introduced in 1963 by Edwards and Cavalli-Sforzas [214], who proposed the concept of minimum evolution where an evolutionary estimate is derived from the minimum spanning network of a set of nodes representing the quantitative characteristics of a population in some Euclidean space [215]. Therefore, a scoring criterion that ranks trees according to the number of mutations required to produce a phylogenetic tree. This led to the development of fast and accurate algorithms [207, 216-218]. However, one disadvantage of the parsimony algorithm is its focus on minimizing the number of changes that may fail to capture the complexity of evolutionary events leading to inaccurate conclusions. Another disadvantage of the parsimony algorithm lies in its susceptibility to long-branch attraction (LBA). This happens when distantly related lineages are mistakenly estimated to be closely related due to homoplasy and/or convergent evolution. This can lead to inaccurate conclusions, especially in fast-evolving species or lineages.

1.8.3 Maximum likelihood (ML)

The works of Felsenstein ushered the transition of phylogenetics into a likelihood-based framework [219, 220]. The maximum likelihood phylogenetic approach (ML) works by finding the parameter values that maximize the observed data likelihood, generating an estimated maximum likelihood of a phylogeny [221]. In molecular phylogenetics, the ML tree is the one that makes the observed sequences the most likely based on the selected model of sequence evolution. It allows for an estimation of phylogenetic trees using complex evolutionary models coupling them with the ability to estimate model parameters [222]. An advantage of the ML approach in phylogenetics is the solid foundation of powerful statistical theory and the robustness of the likelihood function in general statistical inference [207,

© University of Pretoria

223]. In contrast to the parsimony approach, ML accounts for all mutation pathways that fit the data and will always output the optimal phylogeny given an appropriate model and enough data [222, 224, 225]. The main setback to the use widespread use of ML is the computational intensity as ML-based algorithms must navigate through the multidimensional parameter space to find the ML score [207, 226].

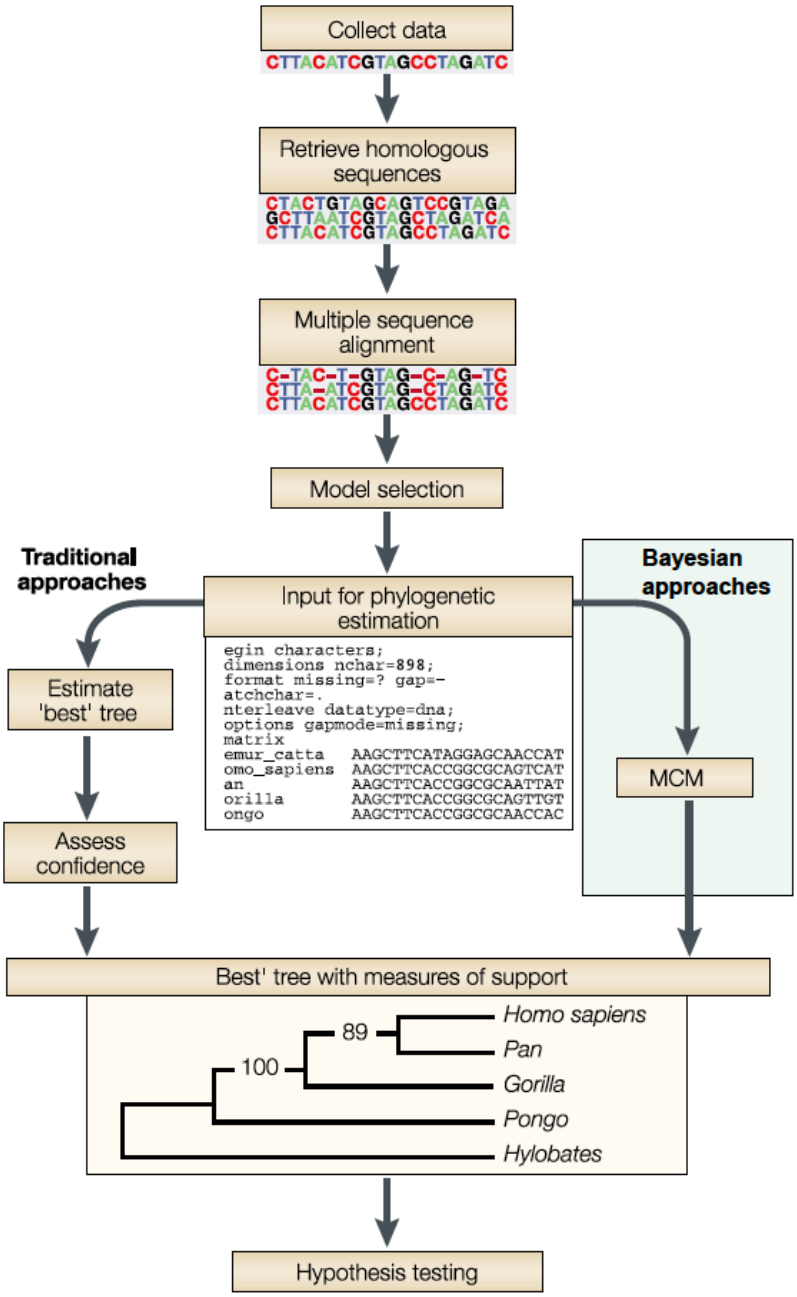


Figure 13: Flow chart summarizing a phylogenetic inference protocol (green box). The process starts with study data collection which is often combined with sequence data from

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

other relevant strains e.g. strains that can serve as an outgroup. Multiple sequence alignment is performed to make the positions in different sequences comparable. A model of sequence evolution is carefully selected, one that is appropriate to the data. Finally, different approaches to phylogenetic inference are chosen as explained in detail in the text. Used with permission from Springer Nature Holder *et al.*[227] copyright 2003.

1.8.4 Rooting and dating of phylogenies

Sampling dates of sequences can be used to infer valuable information such as molecular clock rates and divergence dates providing that enough genetic changes happened over the sampling period. Different clock models that generate such estimates and transform phylogenies into time-calibrated trees have been proposed. These include the uncorrelated relaxed clock proposed by Bromham *et al.*[228], which allows variable rates along each branch in the tree under the assumption that rates follow a pre-chosen statistical distribution. In contrast, the strict clock model assumes a homogenous substitution rate for the entire tree [221, 229]. Tree-dating can be performed in two ways, either including the sampling dates into the phylogeny reconstruction process or using a pre-generated tree to generate substitution rates. TipDate [230] is a popular program that uses the former approach. The program uses a rooted phylogeny and tip dates as input to infer substitution rates under a strict clock model and dates internal nodes using ML under a chosen evolutionary model [221].

Root-to-tip (RTT) regression [231, 232] was one of the first approaches used to infer substitution rates from fixed phylogenies. Assuming a strict molecular clock, the x-axis representing the sampling years and the y-axis representing the genetic distance from the phylogenetic tree, the points will form a linear relationship with the slope of the regression representing the substitution rates. The R2 can be used to indicate how well the molecular clock fits the data with higher values indicating strong clock-like behaviour, while lower values suggest a deviation from clock-like behaviour. While RTT regression is relatively fast, the method cannot be used to generate a dated phylogenetic tree as it does not estimate

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

dates for internal nodes. Another disadvantage of the method is that it assumes evolution following a relatively constant rate. This can lead to inaccurate results if the evolutionary rate varies across different lineages. RTT regression is also dependent on the tree topology, therefore, different tree-building methods or models can affect the distances calculated from the root. Lastly, the method is not suitable for testing statistical hypotheses as it violates the assumption of data independence [221]. The method has been proposed as a tool for exploratory data analysis [233].

Several tools use the LF model [234] to estimate substitution rates, The r8s method [235] is an ML-based dating approach that assumes a constant substitution rate and a strict molecular clock using a rooted tree as input. A Poisson distribution is used to model the number of substitutions along every tree branch. The Least-Squares-Dating (LSD) method [236] is another popular model that uses a normal approximation to the LF model to infer substitution rates. LSD estimates node times (divergence dates) in a phylogenetic tree by making the branch lengths consistent with both the sampling times and the molecular sequence data. This is done by minimizing the expected and observed branch lengths using a least squares criterion. A major advantage of LSD is that it allows for rate heterogeneity and therefore accommodates sequences that violate the strict molecular clock assumption. LSD is also computationally faster, which makes it suitable for large datasets [221]. A disadvantage of LSD is the simplicity of its model for substitution rate variations, which may fail to fully capture the complexity of evolution.

The program Treedater takes the notch up LSD by incorporating outlier detection [237]. TreeTime is another popular ML relaxed clock method that uses dynamic programming to optimize branch lengths on an input tree [238]. An advantage of TreeTime is that it also performs ancestral reconstruction and therefore, can further be used for other evolutionary estimates such as migration reconstruction.

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

1.8.5 Nucleotide Substitution models

Models of nucleotide substitution describe how DNA sequences evolve by accounting for the probabilities of changes (substitutions) between nucleotides over time. Below we describe the main nucleotide substitution models.

1.8.5.1 Jukes-Cantor (JC69) Model

The JC69 model is a simple model which assumes that all nucleotide substitutions occur at the same rate. The model also assumes equal base frequencies (25% of each of the four bases ATCG) and only takes a single parameter α representing the substitution rate [239]. Although the JC69 model can be used for closely related sequences with minimal bias in nucleotide composition, the model assumptions are highly restrictive when describing the evolution of real sequences where transitions are more common than transversions [240].

1.8.5.2 Felsenstein 1981 (F81) Model

Similar to the JC69 model, the F81 model assumes equal substitution rates for all bases while incorporating unequal base frequencies [219]. The model uses a single parameter (μ) which controls the substitution rate. The model works well with relatively similar sequences and substitution rates are assumed to be homogeneous across sites.

1.8.5.3 Kimura Two-Parameter (K80) Model

The K2P, also referred to as the K80 model offers a more realistic approach to modelling sequence evolution, unlike the JC69 as it distinguishes between transitions (A <-> G, C <-> T) and transversions (purine <-> pyrimidine) changes [241]. Transition rates (α) differ from transversion rates (β). Similar to JC69, the K80 model assumes equal base frequencies of 25%. The model is suited for moderately diverged sequences with observed bias in transition/transversion rates.

1.8.5.4 Hasegawa-Kishino-Yano (HKY85) Model

Unlike the JC69 and K80 models, the HKY85 model [242] assumes unequal base frequencies and distinguishes between transitions and transversions. The model requires two

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

parameters namely the transition/transversion rate ratio (K) and the base frequencies and is usually used to model real biological sequences with unequal nucleotide compositions.

1.8.5.5 General Time-Reversible (GTR) Model

The strength of the GTR model lies in its flexibility as it allows for different rates of substitution between all nucleotide pairs and includes unequal base frequencies. Its flexibility allows for the modelling of complex evolutionary processes and is often the preferred default model when analysing large datasets [243].

1.8.5.6 Tamura-Nei (TN93) Model

Borrowing from the HKY85 model, the TN93 model [244] can distinguish between two types of transitions (A <- > G, C <->T). The TN93 model assumes that substitution rates are equal among sites. However, the substitution rate is allowed to be different from both of the rates for transitions. The TN93 model also allows unequal base frequencies.

1.9 BAYESIAN INFERENCE IN PHYLOGENETICS

The use of Bayesian methods in inferring phylogenies was started in the 1990s and has proved to be a game changer in the way genomic sequence data is analysed with applications ranging from phylogeographical analysis of pathogen spread in humans, analysis of species diversification rates and population genetics [245]. The recent SARS-Covid-19 pandemic was a reminder that phylogenetic inference is more than a mere academic exercise, but a matter of life and death, as the pandemic situation required real-time tracking of how the virus was diverging [207]. The main idea behind Bayesian statistics (equation 1) is the use of probability distributions as representations of the uncertainty in the parameters of the model [220]. Bayesian inference can be broken down into three components (likelihood, prior and posterior), where observed data (D) and prior distributions $f\theta$ are combined through the likelihood $f(\theta|D)$ to infer posterior distribution for all model parameters. The unknown parameter is represented by θ. Thus, the prior distribution can be defined as the distribution of the parameters before the data is collected and analysed, while the posterior

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

distribution refers to the distribution that includes the information in the data. Figure 14 illustrates these three components for three different priors. Equation 2 is the normalizing constant that ensures that $f(\theta|D)$ is a proper statistical distribution that integrates to 1 [245].

$$f(\theta|D) = \frac{1}{z} f(\theta) f(D|\theta) \quad (1)$$

$$z = \int f(\theta) f(D|\theta) \, d\theta \quad (2)$$

The use of Bayesian methods in phylogenetic inference was not initially well received by frequentists, who argue that science should be based on objectivity, assuming that results of an analysis should exclusively rely on the observed data and the introduction of priors adds a level of problematic subjectivity. Another bone of contention is on the interpretation of probability. Frequentists interpret probability as the long-run frequency of an event occurring. In phylogenetics, frequentist statistics would consider evolutionary parameters such as tree topologies as fixed derivatives of substitution rates of unknown quantities, which can be analysed by quantifying uncertainties using confidence intervals and p-values based on the sampling distribution of data. In contrast, Bayesian inferences treat probability as a degree of belief or credibility of a hypothesis given the data. Parameters are therefore treated as random variables that have probability distributions [207, 246, 247]. The probability distributions of the parameters (posterior distributions) are continuously updated in Bayesian phylogenetics inference based on new data [248].

However, the superiority of the Bayesian method is in the way the parameters of random variables are treated. The delicate process of choosing priors requires careful designation of probability distributions that mirror a priori knowledge about parameter uncertainty (Figure 14) [249]. What has made Bayesian methods appealing to biologists over the years is that they allow for the implementation of more complex evolutionary models that might be unattainable using ML models. As model complexity increases and the data-points to parameters ratios become low, ML estimates of parameters can become inaccurate. However, Bayesian inference samples the whole parameter space such that the final result is not dependent on one specific parameter value [207].

© University of Pretoria

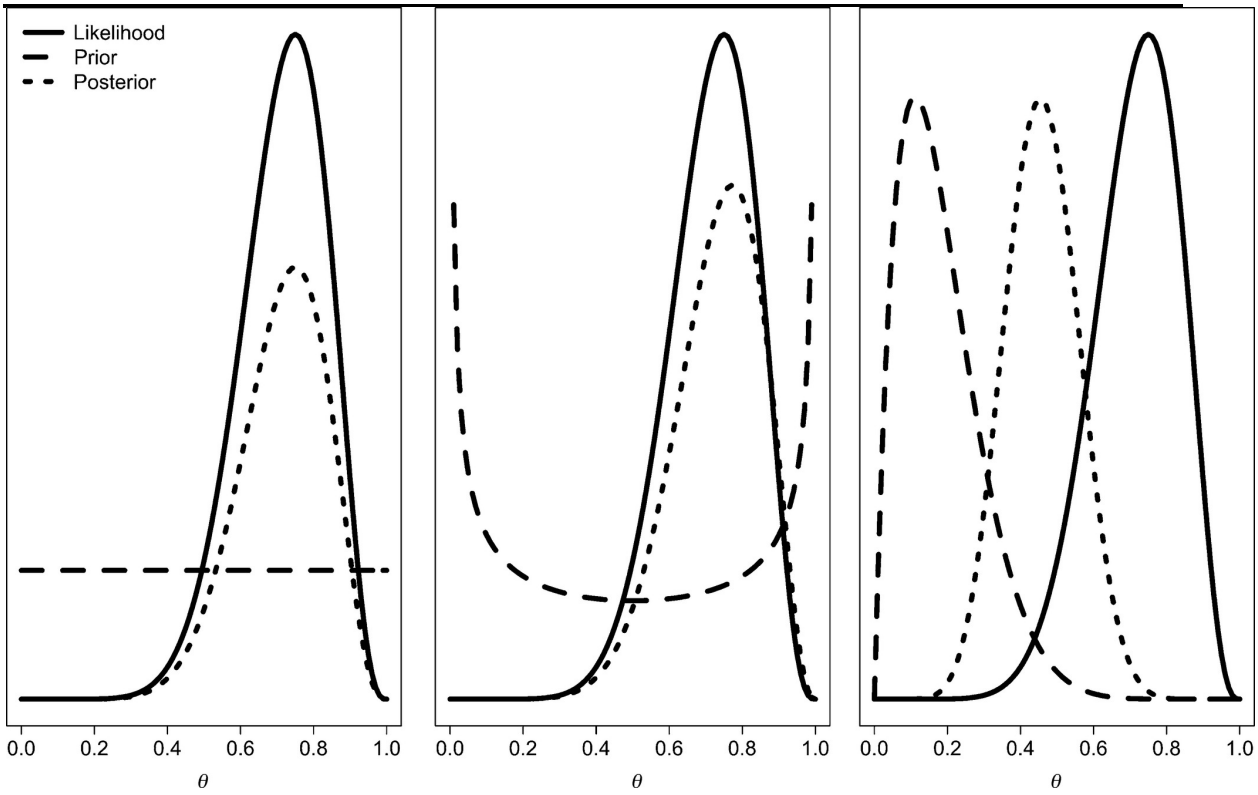


Figure 14: An example of a Beta-Binomial model showing three different choices of priors. The illustration shows a non-informative uniform prior, scale-invariant Jefferys' prior, and highly informative prior (Left to Right). Used from Banner *et al.*[249] under the terms of the Creative Commons Attribution Licence (CC BY) http://creativecommons.org/licenses/by/4.0/.

The hill-climbing algorithms that are used to calculate the ML point are also prone to speed and unreliability issues as the model complexity increases [205, 247, 250]. Figure 15 illustrates the fundamental differences between ML and Bayesian evolutionary approaches, particularly in how the parameters are treated. ML approaches commonly use the Joint estimation approach which seeks to find the highest point in the parameter space. This contrasts with the marginal estimation approach, which is used in Bayesian inferences, which integrates the whole parameter space by calculating the volume under the posterior probability surface [207]. It is important to note that some of these parameters, termed "nuisance parameters", are not of immediate interest to the sequence evolution inference, but have to be accounted for since they form part of the likelihood equation [207]. The Bayesian approach therefore "marginalizes" these nuisance parameters when inferring the marginal posterior probability of a phylogenetic tree.

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

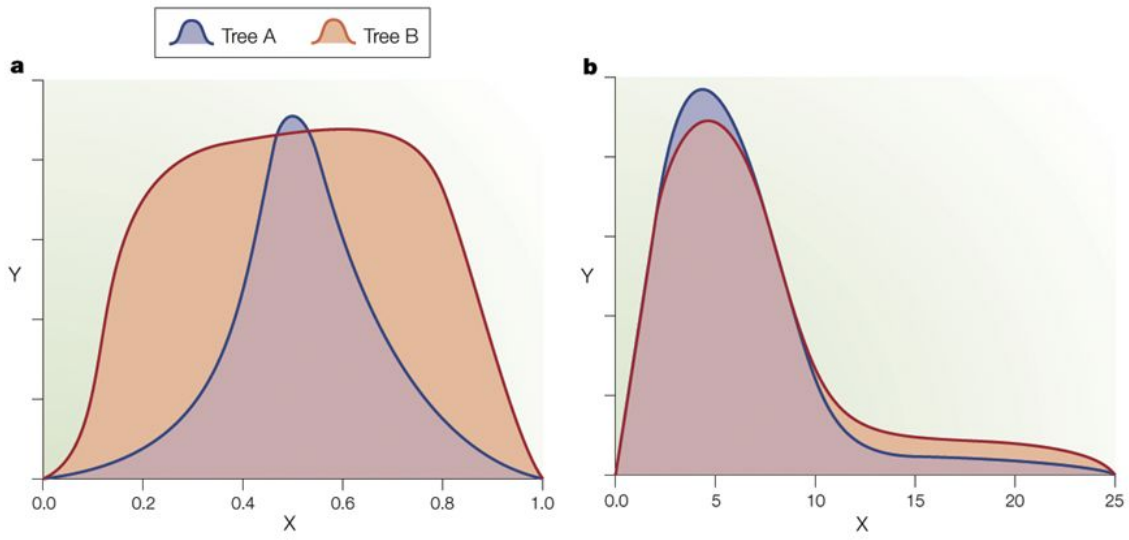


Figure 15: Difference between marginal and joint estimation. Both a and b represent the likelihood profiles for two phylogenies versus a nuisance parameter. The x-axis represents some nuisance parameter e.g. (the shape parameter of the gamma distribution). The y-axis represents the likelihood (in the context of ML) or the posterior probability density (in the context of Bayesian inference). The (AUC) area under the likelihood curve for tree A is shown in light blue. Orange represents the AUC for tree B. The AUC for both trees is coloured in mauve. A joint estimation of x and the phylogeny will choose tree A (highest peak coloured blue for both a and b). Marginal estimation over x will favour tree B (auc in the orange area is greater than in the blue area.). Used with permission from Springer Nature Holder *et al.*[227] copyright 2003.

Several Bayesian phylogenetics programs have been developed over the years. BEAST, is perhaps the most popular framework among biologists as it offers a higher number of models that can simultaneously infer tree topology as well as related evolutionary estimates such as divergence times, phylogeography and phylodynamics [245, 251]. MrBayes is often used to estimate species phylogenies and divergence as an alternative to BEAST. It also offers a large number of models for different types of sequence data (amino acid, nucleotide and morphological) [245, 252]. RevBayes, although similar to MrBayes, also comes with its programming language for setting up complex Bayesian models [253]. Other programs include MCMCTree [254], PhyloBayes [255], BPP, Migrate [256], IMa2 [257], Phycas

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

[227], BAMM [258], Structure [259] and AWTY [260]. The next section describes an innovative approach to studying evolution using Bayesian approaches.

1.9.1 Markov chain Monte Carlo (MCMC)

The posterior distribution can be considered as a multivariate function which can be illustrated by a convoluted landscape. This multidimensional parameter landscape is rather complex, however, the use of Markov Chain Monte Carlo (MCMC) provides an efficient way of navigating through the parameter space [261]. This technique allows for obtaining samples from the posterior, i.e. parameter values in proportion to their probability under the posterior distribution [251].

The MCMC algorithm works by following a series of steps starting with maintaining a location in the parameter landscape, and the instantiation of the parameters θ. These steps form a chain where each step involves the proposal of a new location in the parameter landscape. The relative posterior probability density for the new location is calculated and, if it is higher, the chain moves to the new location. If the proposed location has a lower posterior density, the move is accepted by drawing a number (p) from the unit interval, where p is the ratio of the posterior of the present location. If the move is rejected, the present location is added as the next link and the cycle is repeated [245, 248]. Repeating this process millions of times generates a chain location in the parameter space. The chain has an affinity for regions with high posterior probability, thus moves to lower posterior probability regions are rarely accepted.

1.9.1.1 Convergence, burn-in and mixing of the MCMC

Bayesian phylogenetics analyses can be complex due to the technical requirements that are needed to ensure convergence. In this section, three critical concepts (convergence, burn-in and mixing) for interpreting and assessing the results of MCMC analysis are discussed. Convergence refers to the point at which the MCMC chain has sufficiently sampled from the target posterior distribution of trees or parameters (e.g. substitution rates, divergence times, effective population sizes). This means that after running the MCMC long enough, the distribution of trees or parameters should stabilise and reflect the true posterior distribution, ideally the high probability regions of the posterior [262]. Thus ,the

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

convergence rate refers to the rate at which an MCMC chain starting from any initial region moves to the high-posterior region of the parameter landscape [245, 263].

Burn-in refers to the starting phase of the MCMC chain where the sampled trees or parameter values are yet to reflect the posterior distribution. To avoid inaccurate results, the burn-in period is often discarded. Mixing refers to how well the MCMC chain explores the parameter landscape and navigates between different possible states such as different tree topologies or parameter values. Good mixing is evidenced by a chain that traverses different parts of the posterior distribution with a smaller variance with results showing less variation among independent runs [245]. Poor mixing is usually characterised by long chains with poor estimates or chains that get stuck in certain regions or repeatedly sample from a small subset of trees or parameters [251, 262, 264]. Poor mixing can be indicative of an overly complex model, highly restrictive priors or an improperly tuned MCMC algorithm. Mixing and convergence issues can be diagnosed by plotting the log-likelihood or sampled parameter estimates against the MCMC iteration. This can be done using dedicated programs for BEAST analyses implemented in R [265] or Tracer [266]. It is also a best practice to perform multiple runs of the same analysis as a way of checking the consistency between runs. Figures 18 and 19 illustrate two trace plots for different MCMC chains resulting from a BEAST2 analysis. Figure 16 shows inefficient mixing as evidenced by low ESS values for important parameters such as the likelihood and the likelihood despite the chain running for 62,100,000 iterations with a 10% burn-in. An ESS value of at least 200 for the posterior distribution is generally acceptable to accurately infer the posterior distribution [248, 267]. Figure 17 shows a chain with efficient mixing as evidenced by the high ESS values and horizontal trace plot.

© University of Pretoria

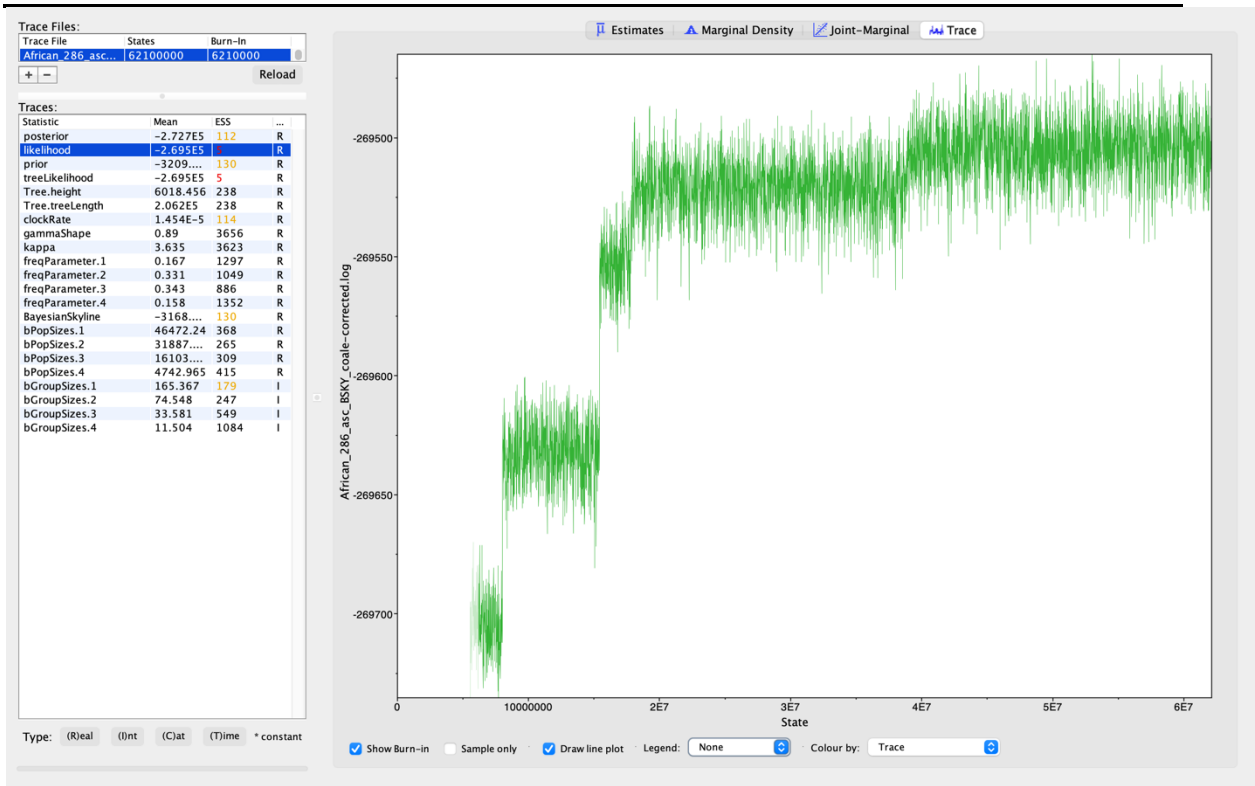


Figure 16: Trace plot from sampling a posterior distribution using an inefficient MCMC chain with poor mixing. Despite the chain running for long, some important parameters remained low.

© University of Pretoria

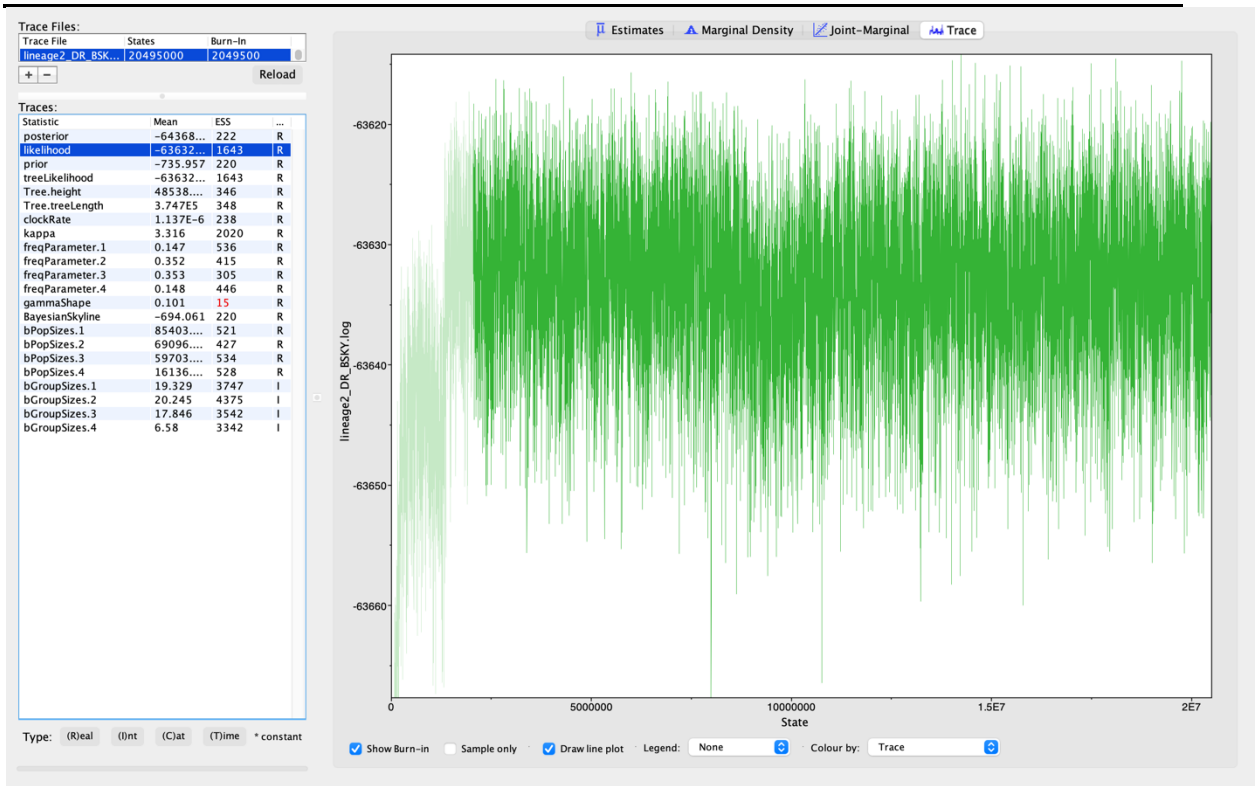


Figure 17: Trace plot from sampling a posterior distribution using an efficient MCMC chain with good mixing. The chain ran long enough and all-important parameters had ESS values > 200.

1.10 PHYLOGEOGRAPHY

The earliest definition of phylogeography was coined by Avise *et al.* [268], who described it as "the study of the geographical context for genetic variation within species – as a bridge between the process acting on populations and supra-specific evolutionary patterns". The spread of epidemics through space and time can be studied by integrating non-ordered sets of locations (countries, cities, continents, etc.) with ancestral character reconstruction (ACR) and a given phylogeny. The locations are then mapped to the tree as character states based on the known tip locations. ACR aims to extrapolate the trajectory of a character on the tree from the root to the tips through time by assigning the most likely ancestral character state to each internal node [221]. Several methods have been developed for ACR.

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Parsimony-based ACR methods aim to infer an evolutionary process that minimizes the changing of states along the phylogenetic tree. As we discussed in the previous sections, parsimony methods are fast and simple but suffer from low accuracy issues as they oversimplify evolutionary processes [221, 269, 270]. ML and Bayesian methods provide greater accuracy by using robust probabilistic evolutionary models, which incorporate standard nucleotide substitution models to s-state (discrete) trait characters [221]. The likelihood methods predict ancestral character states on selected character evolution models using approaches such as the marginal posterior probabilities of the character state [219, 221, 254] and the joint reconstruction of the most likely scenario [221, 271]. However, both approaches have limitations. For example, the state probability estimate derived from marginal reconstructions can be too complex to interpret and visualize [221]. Joint reconstructions on the other hand lack the uncertainty estimation as they provide only a single "best estimate" for each ancestral state. Other limitations of the joint reconstruction include sensitivity to errors in the tree topology and evolutionary model, and the inability to account for alternative evolutionary scenarios. Intermediate approaches have been developed to compensate for these limitations. These methods predict a unique state in the easily estimated regions of the tree (typically near the tips [221, 272]) while maintaining multiple probable states in more complex regions (closer to the root), reflecting the uncertainty of these estimates [221].

Several ACR tools have been developed over the years. Among them, Mesquite is one of the most frequently used, as it uses both parsimony and likelihood-based approaches for ACR inference. The tool also allows for interactive visualization which makes it user-friendly for people with limited programming [273]. Mesquite is also versatile as it supports both molecular and morphological data. Randomized Accelerated Maximum Likelihood (RAxML) is another popular program that offers ACR in addition to ML-based phylogenetic inference. RaxML performs ACR by identifying states that maximize the likelihood of observing the given data on the phylogeny under specific character evolution models. What makes the tool popular is that it accepts nucleotide as well as protein sequence data and is highly efficient for big datasets. One limitation of RAxML is that it's restricted to molecular data, thus it is unsuitable for continuous or morphological traits. PastML [274], a recent tool has also gained popularity for fast ACR inference, especially for larger datasets. Like Mesquite, the tool comes with several parsimony and ML ACR methods. The tool comes

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

with marginal posterior probabilities approximation (MMPA) as well as Joint and MAP (maximum a posteriori). The former approach is an intermediate method that employs decision-theory concepts and the Brier criterion to link each node in the phylogeny to a set of most probable states [221], while the latter selects the state with the highest marginal probability. A limitation of PastML is its reliance on ML limits and its flexibility in modelling complex evolutionary scenarios. Bayesian tools discussed earlier are more robust for performing ACR inferences as they handle uncertainty better than the other methods described here. However, bayesian methods are often complex to use and require substantial computation power which renders them unsuitable for large datasets.

1.11 PROJECT RATIONALE

TB remains one of the leading global causes of death due to a single infectious agent, even killing more people than HIV. The recent COVID-19 reversed some of the gains that had been achieved in the past decade to reduce global TB incidence and mortality rates [1, 8]. Resistance to anti-TB drugs remains a challenge in global, regional and local TB control programs. Recently introduced anti-TB drugs such as bedaquiline, delamanid and pretonamid have been a game changer in the treatment of multi-drug resistant TB [275]. However, resistance to these new drugs has already been reported in different parts of the world [276, 277]. Understanding the genetic basis of drug resistance is important to the development of new anti-TB drugs.

Development of a new TB vaccine, rapid diagnosis tools and new drugs are some of the major priorities in global TB control policy. For these to be achieved, critical gaps in understanding the dynamics that influence Mtb evolution need to be addressed. While valuable insights have been gained from previous studies conducted in high-income settings, exploring Mtb evolutionary dynamics in high-budget low-income regions such as Sub-Saharan Africa has the potential to improve our understanding of the pathogen.

This thesis leverages large-scale genomic datasets and state-of-the-art phylogenetic tools to explore the evolutionary history and genetic diversity of global Mtb. By focusing on a more

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Afrocentric approach, this research aims to uncover lineage-specific adaptations that have led to the success of this pathogen. This study also seeks to elucidate the phylogeographic dynamics that have shaped Mtb populations over time, offering new perspectives on how historical and ecological factors influence the pathogen's evolution.

The findings from this research will contribute to a deeper understanding of Mtb evolution and drug resistance, potentially informing health policies and global TB eradication strategies. Additionally, it highlights the importance of genomic studies in addressing regional challenges, paving the way for more targeted and effective interventions in high-burden regions.

1.12 THESIS AIM AND OBJECTIVES

The general aim of this thesis is to characterise the global diversity of *Mycobacterium tuberculosis* using public Mtb genomic datasets. Specifically:

1. To optimise the analysis of large Mtb genomic datasets, focusing on computational efficiency and accuracy to gain deeper insights into the pathogen's evolutionary dynamics.
2. To analyse available Mtb genomic data from African isolates to investigate the evolutionary dynamics that shaped the Mtb population landscape on the continent.
3. Investigate the genetic diversity of Lineages 2 and 4 in the SADC region and explore the phylogeographic dynamics within the region.
4. Elucidate the epistatic interactions that govern the evolution of drug resistance in Mtb.

© University of Pretoria

CHAPTER 2 SCALABLE APPROACHES FOR COMPUTATIONAL PHYLOGENOMICS OF LARGE MTB DATASETS

© University of Pretoria

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

2.1 ABSTRACT

Next-generation sequencing (NGS) of *Mycobacterium tuberculosis* complex (MTBC) strains is yielding increasingly large volumes of data. The availability of sequence data has transformed the tuberculosis (TB) research landscape and provided vital insights into shaping global TB control policies. MTBC diversity is an important area of TB research that has gained immensely from the accumulation of NGS datasets. However, analysing such big datasets comes at a significant computational cost especially when performing evolutionary analyses studies, not only inferring phylogenies but also analysing associated evolutionary metrics such as substitution rates, time scales and ancestral reconstruction patterns. In this study we curated a global NGS dataset of 9073 human-adapted Mtb strains and their associated metadata. We then used a computationally efficient strategy to perform an evolutionary analysis of this large dataset. Insilico drug resistance profiling identified a catalogue of drug resistance mutations that are associated with resistance to all the 14 anti-TB drugs that were tested in this study. Importantly, we also identified mutations associated with resistance to the recently adopted novel drugs bedaquiline, linezolid, delamanid and clofazimine. Lastly, we generated a high-quality global reference dataset of 8775 Mtb isolates that served as a reference for downstream analyses.

2.2 INTRODUCTION

The latest global figures on Tuberculosis (TB) show an encouraging stabilisation trend in the number of people falling ill with TB each year [1]. TB caused by *Mycobacterium tuberculosis* (Mtb) remains a global killer, reclaiming the number one spot (from COVID-19) as the leading global cause of death due to an infectious agent [1]. An estimated 10.8 million new TB cases and 1.25 million associated deaths were reported in 2023 [1]. Disruptions to TB care programs during the COVID-19 pandemic have been blamed for the reversals of some of the gains that have been made in the fight against TB over the past decade before the pandemic. The WHO End TB Strategy has a set milestone of a 50% reduction in TB incidence rates by 2025 [278]. The same strategy also aims to achieve a 75% reduction in the global number of deaths attributed to TB by 2022. However, between 2015 and 2023, a reduction in TB mortality of only 23% was achieved. The same timeframe

© University of Pretoria

only saw a net reduction in the TB incidence rate by 8.3% [1]. Recently, global commitment to ending TB has been promising, with new global targets for 2027 set at the 2023 UN high-level meeting on TB [279]. Notably, the meeting set targets that include targeting 100% of newly diagnosed TB cases (rapid testing), developing a new TB vaccine preferably within five years and increasing funding for prevention, diagnostics, treatment and TB research.

The MTBC consists of diverse lineages that follow a phylogeographical distribution. TB is a curable disease, however, global efforts to eradicate TB have been challenged by the emergence and spread of TB strains that do not respond to antibiotics. Drug-resistant TB can be classified as multi-drug-resistant TB (MDR-TB), which refers to strains that have developed resistance to the first-line drugs isoniazid and rifampicin. Once resistance to the first-line drugs develops, Second-line drugs are added to the treatment regimen. Extensive drug-resistant TB (XDR-TB) refers to TB caused by MDR-TB or rifampicin-resistant (RR-TB) strains that have developed further resistance to at least one fluoroquinolone (levofloxacin and moxifloxacin) and at least one other Group A drug (bedaquiline or linezolid). Pre-XDR-TB is caused by MDR/RR-TB that has developed resistance to at least one fluoroquinolone. The past decade has seen a concerted effort towards the development of new TB drugs and the successful adoption of bedaquiline, delamanid and pretomanid in treating MDR-TB has been a testament to that [280-282]. Concerningly, several studies have reported resistance to bedaquiline in different countries [275, 283, 284]. Drug resistance in MTBC isolates is sorely conferred through the acquisition of point mutations in specific gene targets for anti-TB drugs [94, 285]. Early diagnosis and adequate treatment are essential in preventing the development of DR-TB [286]. The ability of Whole genome sequencing (WGS) to interrogate whole genes makes the technology attractive for the rapid diagnosis of DR-TB when compared to conventional culture-based drug susceptibility testing (DST) which is slow and labour-intensive [287].

WGS is also now the standard in the tracking of disease outbreaks as recently witnessed in HIV-AIDS, H1N1, SARS-COVID-19, Zika and Monkey Pox outbreaks [288-291]. The accumulation of WGS data from global sequencing efforts provides unparalleled benefits in TB research, especially in understanding the evolutionary dynamics of global TB. The increased availability of TB genome sequences has also had a huge impact on TB epidemiology and helped researchers to better understand TB transmission dynamics that

© University of Pretoria

can be used to formulate better TB management strategies at global, regional, national and even local levels. However, the exponential increase in WGS datasets comes with challenges. The total amount of available WGS data is known to be increasing at a rate that is way faster than Moore's Law for computational processing data, as the size of biological databases increases so does redundancy [292]. Extracting new insights from WGS will not only require improved computational power but also more innovative ways of analysing large datasets.

While COVID-19 had a devastating impact on TB management programs, the response to the pandemic inadvertently boosted the TB research field. Over the years, analysing the increasing volumes of Mtb genomic data was often limited by the lack of sufficient computational power and tools for large datasets. The large volumes of data that were generated from COVID-19 genomic sequencing necessitated the development of improved computational infrastructure, and this legacy can now be pivoted towards TB genomic research. Innovative bioinformatics tools that were developed for characterising COVID-19 genetic diversity can now be leveraged to perform more intensive analyses in TB research such as complex phylogenetic analyses. One such tool was IQTREE2 which was published in May 2020, according to the developers, the software could estimate an evolutionary tree from 17,000 COVID isolates in only three minutes [293]. This was a vast improvement considering that its predecessor IQTREE 1 would require 37 hours to perform the same task [294].

In this chapter, we curate a large WGS dataset comprising 9,073 Mtb isolates, representing the global diversity of human-adapted Mtb lineages. The isolates were sourced from public repositories and were carefully selected based on their associated metadata, which included important information such as geographical origin, year of isolation, lineage and drug susceptibility testing profiles. To efficiently analyse the diversity and evolutionary dynamics within the MTBC at this scale, we developed and implemented a robust protocol tailored for large genomic datasets. This protocol ensures rigorous quality control, minimises redundancy while maintaining phylogenetic diversity, and allows for accurate phylogenetic and evolutionary characterisation of MTBC.

© University of Pretoria

2.3 MATERIALS AND METHODS

A detailed workflow outlining the data sourcing and curation approach utilized to prepare a reference MTBC dataset for other downstream analyses is summarized in Figure 18.

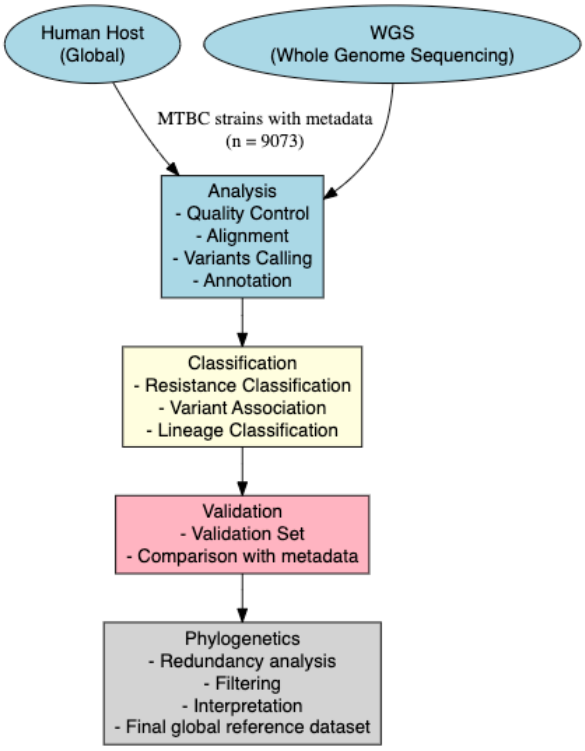


Figure 18: A workflow summarising the data collection process from sourcing data from public databases to curating a high-quality reference dataset. The workflow highlights the key steps, including data sourcing, quality control, redundancy reduction, and metadata validation, ensuring the dataset is robust and representative of global MTB diversity.

2.3.1 Data collection

In this study, we curated a large WGS dataset that was representative of the global genetic and geographical human Mtb diversity from different public repositories as summarized in Figure 18. We first sourced all the 3553 sequence (fastq) and variant data (vcf) from the Genome-wide *Mycobacterium tuberculosis* variation (GMTV) database [295] and their associated metadata. The database integrates data from epidemiology, phylogenetic lineage, year and place of isolation, drug, resistance profile, clinical outcome as well as HIV status (if available). We also sourced vcf files and associated metadata from the Afro-TB dataset

© University of Pretoria

which is the most comprehensive database of Mtb genomic data from the African continent [296]. We also downloaded raw sequencing data of MTBC isolates from the European Nucleotide Archive (ENA) and the Sequence Read Archive (SRA) using BIGSI. Associated metadata was also sourced from the Bacterial and Viral Bioinformatics Resource Centre (BVBRC) [297], associated studies and in some cases personal correspondence with the researchers affiliated with the studies.

2.3.2 Raw data processing and variant calling.

The validity of the vcf files sourced from the GMTV database was assessed using the checkVCF program [298]. Invalid VCF files were discarded and their associated raw sequencing data files were added to the variant calling pipeline. The raw sequence data were quality controlled using Trimgalore v.06.0 [299] and the quality metrics were assessed using FastQC v0.12.1 [300] and MultiQC v1.25 [301]. The raw data was filtered to keep samples with read depths of 30X or higher and a genome-wide coverage of 99% or higher. The high-quality reads were aligned to the H37Rv reference genome (Genbank: NC000962.3) using BWA mem v0.717. SAMTOOLS v1.10 [302, 303] was used to process the alignment files. FreeBayes v1.3.7 [304] was used to perform variant calling. The generated variants were quality controlled using BCFtools v1.9 [302] and positions with phred-scale quality scores < Q20 and mapping quality < 40 were discarded. All VCF files were validated for format compliance with VCF 4.2 specifications using BCFtools v1.9, identifying and rectifying any format inconsistencies. Chromosome notation was also standardised to ensure consistency, with all chromosomes labelled as “Chromosome” across all VCF datasets. Final quality control was performed using VCFtools v0.1.17 to confirm accurate variant positions, allele consistency and completeness of annotations. The code used for validation is provided in the appendices. Our combined set of high-quality VCF consisted of 9073 Mtb strains isolated from 27 countries representing the global diversity of Mtb (Supplementary file 2.1). While more data were available in public repositories, our sample size allowed us to conduct a high-quality quality computationally efficient analysis without sacrificing the reliability of our findings.

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

2.3.3 Lineage classification, drug resistance profiling

The Resistance Sniffer [24] and the FastLineageCaller v.0.3.2 [305] tools were used to predict the phylogenetic lineages for all the isolates in this study. We used the default settings for FastLineageCaller to allow for different lineage classification schemes. The results were compared to the information derived from the metadata. Strains with ambiguous lineage classifications (belonging to more than one lineage) were assumed to be mixed infections and were discarded from further analysis. The specific drug-resistance molecular markers were identified by extracting mutations from known drug target genes and results were validated using the TB-profiler v4.4.0 software [306].

2.3.4 Evolutionary analysis and redundancy assessment

Insertions and deletions (INDELS) were filtered from the high-quality VCF files using VCFtools v0.1.17 [307]. To mask low-confidence regions (LCR), we generated a BED file following the method described by Marin *et al.*[308] (Supplementary file 2.2). This BED file was applied to mask the LCRs in each VCF file, and SNPs within 50 bp of these regions were excluded. The filtered VCF files were then merged into a single multi-sample VCF file using BCFtools v1.9 where missing genotypes in each vcf file were recorded as “missing” during the merging process. Next, we used vcf2phylip v2.9 [309] to convert the merged VCF file into a multi-sample FASTA file containing SNP positions. From this FASTA file, we generated an SNP alignment of all polymorphic sites across the analysed Mtb genomes using Snp-sites v2.5.1 [310].

Using this SNP-alignment file, an initial maximum likelihood (ML) tree was constructed with IQTREE2 v2.1.3 [293]. To account for the size of our dataset, we applied the general time reversible nucleotide substitution model (GTR) combined with ultrafast bootstrap (UFBoot) analysis and fast likelihood approximation. The parameters were set as follows: 50 initial parsimony tree searches, 20 perturbation rounds, and a minimum epsilon value of 0.05 for tree optimization. The resulting unrooted phylogenetic tree was annotated with lineage metadata and manually checked for any classification discrepancies by identifying isolates clustering with incorrect lineages. To optimize the dataset for diversity and

© University of Pretoria

computational efficiency, we assessed redundancy using Treemer v3 [311], while prioritizing underrepresented lineages and geographical regions. Redundant strains were excluded and the list of preserved isolates was used to select preserved isolates which were then used to create an SNP alignment of non-redundant genomic sequences. For more robust phylogenetic reconstruction, we ran IQTREE2 v2.1.3 again with bootstrap analysis. ModelFinder [312] identified the K3 + F +ASC +R4 model as the best fit based on Akaike Information Criterion (AIC) scoring [313]. Ultrafast bootstrapping was performed with 50 replicates to confirm genome clustering fidelity. The resulting unrooted tree was refined using TreeTime v0.94 [238] , implemented via the augur-refine library in the Nextstrain Augur pipeline [314] with isolation dates from metadata used for molecular clock calibration. Finally, tree annotation was conducted using Nextstrain Augur [314] incorporating lineage and location metadata. Ancestral state reconstruction (ACR) was also performed within the Nextstrain Augur program a summary of the code used in this analysis is provided in the Supplementary file 2.3.

2.4 RESULTS

2.4.1 Lineage distribution

To curate a globally representative whole-genome sequencing (WGS) dataset of *Mycobacterium tuberculosis* diversity, we collected isolates and associated metadata from public repositories, prioritising samples with comprehensive metadata including origin, collection date, lineage diversity and drug resistance profiles. Our analysis revealed that approximately 67% of genomes in public databases contained the minimum required metadata, such as geographical location and the year of isolation. The final dataset comprised 9,073 samples collected by various laboratories in different countries over 30 years (1991-2021).

The dataset included eight lineages of the human-adapted MTBC. Lineage classifications were further validated using FastLineageCaller v0.3.2. which identified 544 strains from the GMTV database as incorrectly classified. Additionally, we resolved the lineage classification of 1,320 strains, labelled ambiguously as “New-1, Uganda, X-type, nd” in the

© University of Pretoria

GMTV database. Among these, 17 strains showed signs of a mixed infection and were excluded from further analysis.

Lineage distribution analysis revealed that Lineage 4 strains were the most prevalent representing 55% of the dataset (n=4989, Figure 19A), followed by Lineage 2 (21% n=1897), Lineage 3 (17.27%, n=1567), and Lineage 1 (6.14%, n=557). Lineages 6, 5, and 7 were the least represented, together accounting for approximately 5.47% of the dataset, with Lineages 5, and 7 contributing just 1%.

Within the Lineage 4, sublineages Lineage 4.9 and Lineage 4.1 were the most prevalent, each contributing approximately 25% of Lineage 4 genomes (Figure 19C). Other sublineages including Lineage 4.3 (18%), Lineage 4.8 (11.8%), Lineage 4.4 (6.0%) and Lineage 4.2 (6.2%) accounted for most of the remaining Lineage 4 genomes, while underrepresented sublineages contributed 7% combined.

In Lineage 2, sub-lineage 2.2.1 dominated, constituting 95% of the lineage (n=1,811), with the remaining 5% split between sublineages 2.1 (n=21) and 2.2.2 (n=65) strains. For Lineage 3, only 285 strains were resolved to the sublineage level with sublineages 3.1.1 and 3.1.2 contributing 44 % (n=124) and 56% (n=160), respectively. Among Lineage 1 strains, sublineage 1.1 represented 53% (n=293), while 1.2 comprises the remaining 47% isolates while lineage 1.2 strains made the remaining 47% (n=264). Finally, we identified 85 *M. bovis* strains that were ambiguously classified as Lineage6/Bovis in the GMTV database. These strains were reanalysed to refine their classification.

© University of Pretoria

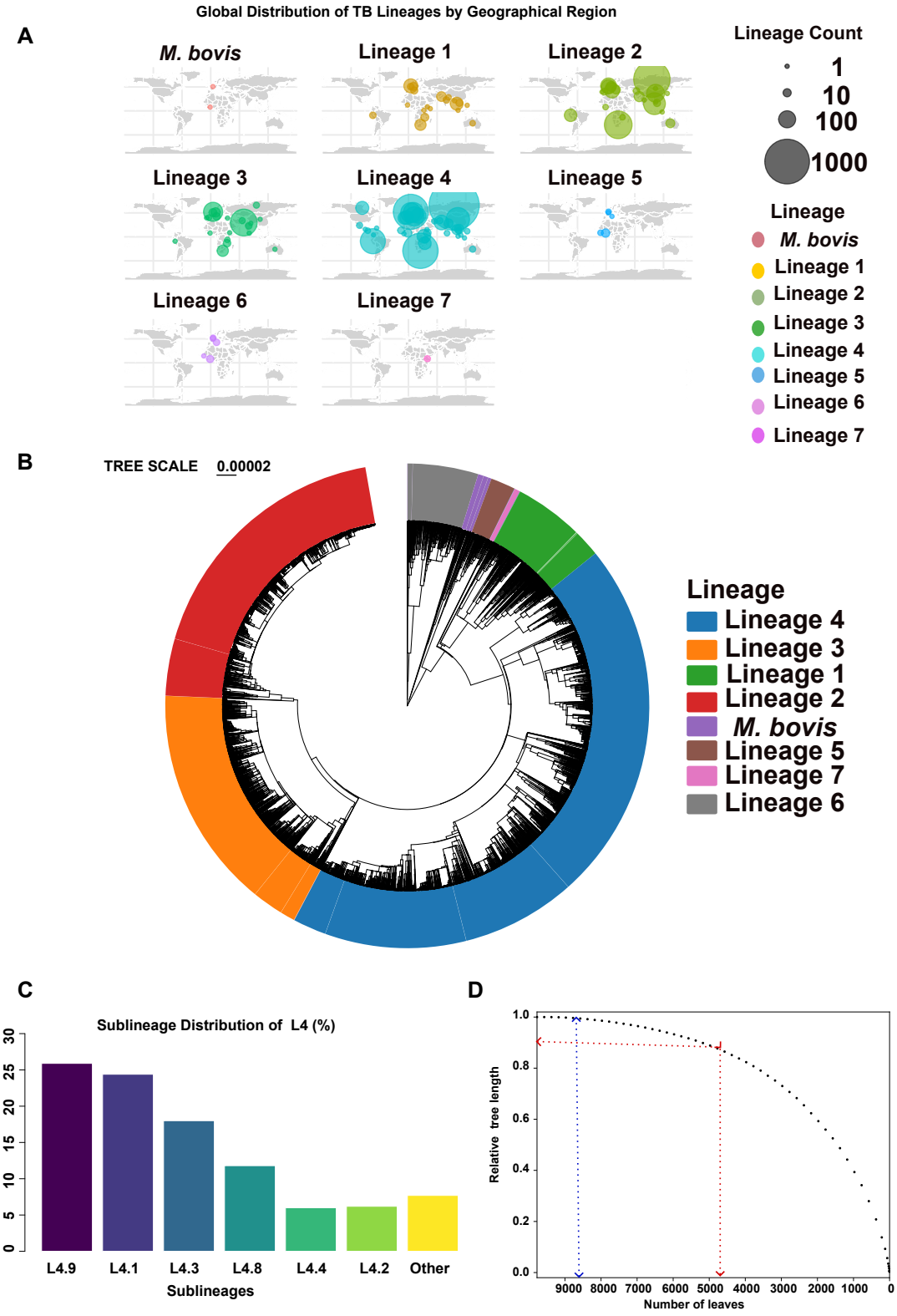


Figure 19: Global distribution of Mtb lineages based on the 9073 isolates used in this study (A)

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

The phylogeographic structure of the 9,073 global samples used in this study. (B) Maximum likelihood phylogenetic tree of the global 8301 Mtb isolates. The tree was constructed based on 148555 genome-wide SNPs. Different coloured blocks are used to represent different lineages. (C) A barplot showing the distribution of the sublineages of lineage 4 which was the most prevalent lineage in this study. (D) Relative tree length (RTL) decay plot for the 9073 Mtb isolates, trimming the original tree to 99% of the original diversity preserves 8565 isolates as represented by the blue dashed lines. The red dashed lines show that 90% of the original diversity is still maintained even if half of the leaves are trimmed indicating high levels of redundancy.

2.4.2 Evolutionary analysis

We generated a SNP alignment of 9,073 sequences spanning 450,135 polymorphic sites including 151,793 distinct patterns that representing unique allelic states across all analysed Mtb genomes. This dataset contained 149,827 parsimony-informative site positions with sufficient variation to determine the most parsimonious evolutionary tree, alongside 266,171 singleton sites and 34,137 invariant sites. The SNP alignment was used to reconstruct an unrooted ML tree.

To optimize for dataset manageability while preserving 99% of the original tree diversity, we trimmed the unrooted phylogenetic tree reducing the dataset to 8,565 strains (Figure 19D). A new SNP alignment was generated from the reduced dataset, containing 148,555 parsimony informative sites. This refined dataset was used for a more robust phylogenetic reconstruction of the 8565 strains using IQTREE 2 with 50 ultrafast bootstrap replications [315] (Figure 19B). The ModelFinder [312] program identified the GTR+F+I+R2 model as the best fit for the data.

Tree branches were corrected using a molecular clock dating approach [221] calibrated with the sampling year for each strain. The estimated substitution rate for the entire dataset was 7.45×10^{-7} substitutions per nucleotide site per year (standard deviation: 1×10^{-7}), indicating low variability in the mutation rate across the sequences analysed (Table 1). The covariance matrix for the molecular clock parameters showed minimal uncertainty, with a variance of 1.007×10^{-7} for the intercept, 4.08×10^{-8} for the rate and a negative covariance value (-2.03 x

© University of Pretoria

10^{-11}), suggesting a weak relationship between the intercept and the rate (Table 2). The intercept of the molecular clock regression was -0.00141, reflecting the baseline adjustment required for fitting the model to the data.

Based on the molecular clock analysis, the estimated TMRCA was inferred to be 193 CE. The tree dating step also provides inferred calendar dates for each tip of the tree while flagging outlier samples which were excluded from the analysis. These inferred dates will be used in downstream analyses, requiring the full sampling dates as input.

Table 1: Summary of the molecular clock parameter estimates in the phylogenetic analysis for 8301 Mtb isolates. The substitution rate is estimated at 7.45 x 10^{-7} substitutions /site/year with a standard deviation of 1 x 10^{-7} indicating the rate of evolutionary change across the sequences. The intercept of -0.00141 represents the baseline divergence at the root of the tree, used in the clock model regression.

Parameter	Estimate	Standard Deviation
Substitution Rate	7.45E-07	1.00E-07
Intercept	-1.41E-03	

Table 2: Covariance matrix of the intercept and substitution rate parameters from the molecular clock model. The diagonal elements represent the variances of the intercept (1.007 x 10^{-14}) and substitution rate (4.08x10^{-8}), while the off-diagonal elements (-2.03 x 10^{-11}) represent the covariance between these two parameters, indicating the degree of correlation between the intercept and rate estimates.

Parameter	Intercept	Rate
Intercept	1.01E-04	-2.03E-11
Rate	-2.03E-11	4.08E-08

© University of Pretoria

2.4.3 Insilico drug resistance prediction

We aimed to computationally predict antibiotic resistance for 14 widely used anti-TB drugs including rifampicin, isoniazid, streptomycin, ethambutol, ethionamide, kanamycin, capreomycin, amikacin, pyrazinamide, fluoroquinolones, delamanid, clofazimine, linezolid and bedaquiline. Drug resistance was predicted for each strain using the Resistance Sniffer program [24]. We then analysed each VCF file to identify drug-resistance molecular markers based on a list of known drug target genes and their loci. As we expected, most of the strains were predicted to be susceptible to TB antibiotics. Tables 3, 4 and 5 show the entire repertoire of drug resistance-associated mutations identified in this study. Most of the identified mutations were associated with resistance to the first-line drugs isoniazid and rifampicin as shown in Figure 20.

A total of 2,269 isolates carried mutations associated with resistance to at least one TB antibiotic [316]. Resistance was prevalent for first-line drugs and distributed as follows: isoniazid (78%), rifampicin (71%), ethambutol (45%) and pyrazinamide (29.20%). Our predictions for first-line drugs were highly concordant with pDST results for isolates that had metadata on phenotypic results rifampicin (95.2%) and isoniazid (93%). Resistance to non-first-line drugs was predicted as follows streptomycin (41%), aminoglycosides (12%), ethionamide (41%) and fluoroquinolone (15.2%). 1,112 isolates were predicted to be MDR-TB strains, while 157 isolates were classified as pre-XDR-TB status. Among the pre-XDR-TB strains, 18 exhibited resistance to ten drugs (all tested drugs except for bedaquiline, delamanid, clofazimine and linezolid). The proportions of predicted MDR-TB isolates were as follows: Lineage 2 (38.3%), Lineage 4 (29.2%), Lineage 1 (15.8%), Lineage 3 (12%), with the remaining isolates belonging to Lineage 5 and Lineage 6.

Rare resistance-associated mutations included *ddnW20** (delamanid resistance) and *rplCC154R* (linezolid resistance), each identified in a single strain. Cross-resistance to bedaquiline and clofazimine was predicted in 12 strains predominantly linked to the *Rv0678 G121R* mutation, with one isolate carrying the *Rv0678 R156** mutation.

Table 3: Drug resistance-associated mutations identified in this study across the first-line anti-

© University of Pretoria

TB drug.

Pyrazinamide	Ethambutol	Isoniazid	Rifampicin
pncA: C14R	*embB: M306V*	*fabG1: C-15T*	*rpoB: L452P*
pncA: L172P	*embA: C-16G,embB: E405D*	*fabG1: G-17T*	*rpoB: V170F*
pncA: G162D	*embB: D1024N*	*ahpC: C-57T*	*rpoB: S450L*
pncA: H57D	*embB: T1082A*	*katG: W300C*	*rpoB: L430R*
pncA: T100P	*embB: E405D*	*fabG1: C-15T,katG: W191R*	*rpoB: Q432K,rpoB: H445D*
*pncA: Y103**	*embB: G406A*	*katG: W90R*	*rpoB: D435F*
pncA: V155M	*embB: D354A*	*katG: W328L*	*rpoB: L430P*
pncA: P69L	*embB: D328Y*	*inhA: S94A*	*rpoB: I491F*
pncA: Q10P	*embB: M306L*	*katG: W438**	*rpoB: S441V*
pncA: C72R	*embB: Q497R*		*rpoB: Q429L*
pncA: G105D	*embB: G406D*	*inhA_c.-154G>A*	*rpoB: S450W*
pncA: S59P		*katG: S315T*	*rpoB: D435V*
pncA: Q141P		*katG: S315T,fabG1: C-15T*	*rpoB: Q432P*
pncA: T135P		*katG: S315T,inhA_c.-770T>A*	
pncA: D8N		*katG: S315T,inhA_c.-154G>A*	
pncA: R154G		*inhA_c.-779G>T*	
pncA_c.-11A>C		*katG: S671**	
pncA: S164P		*inhA_c.-770T>A*	
pncA: H57Y			
pncA: F94L			
pncA: K96T			
pncA: D136G			
pncA: W68C			
pncA: T76P			
pncA: T61P			
pncA: R140P			
pncA: L19P			
pncA_c.-11A>G			

Table 4: Drug resistance-associated mutations identified in this study across the second-line anti-TB drugs.

© University of Pretoria

Capreomycin	Fluoroquinolones	Ethionamide	Streptomycin
rrs: A1401G	*gyrA: A90V*	*fabG1: C-15T*	*rrs: C513T*
tlyA: N236K	*gyrA: S91P*	*ethA: N345K*	*rrs: C517T*
tylA: C-83T	*gyrA: D94G*	*ethA: S266R*	*gid: A134E*
rrs: C1402T	*gyrB: D461N*	*ethA: W256**	*rrs: A907T*
rrs: G1484T	*gyrB: E501D*	*ethA: P334A*	*gid: L79S*
	gyrB: S447F	*fabG1: C-15T,ethA: S266R*	*gid: L50R*
	gyrB: E501D,gyrB: S447F	*fabG1: G-17T*	*rrs: A906G*
		fabG1: C-15T,ethA: N345K	*gid: A138V*
		ethA: I337V	*gid: L108R*
		fabG1: C-15T,ethA: H281P	*gid: V88A*
		ethA: C137R	*gid: A200E*
		*ethA: Y143**	*gid: L94P*
		*ethA: R279**	*rrs: A514C*
		*ethA: W240**	*rpsL: K43R*
		inhA_c.-154G>A	*rpsL: K88R*
		inhA_c.-770T>A	
		inhA_c.-779G>T	

Table 5: Drug resistance-associated mutations identified in this study across the novel anti-TB drugs clofazimine, delamanid, linezolid and bedaquiline.

Clofazimine	Delamanid	Linezolid	Bedaquiline
*Rv0678: R156**	*ddn: W20**	*rplC: C154R*	*Rv0678: R156**
Rv0678: G121R			*Rv0678: G121R*

© University of Pretoria

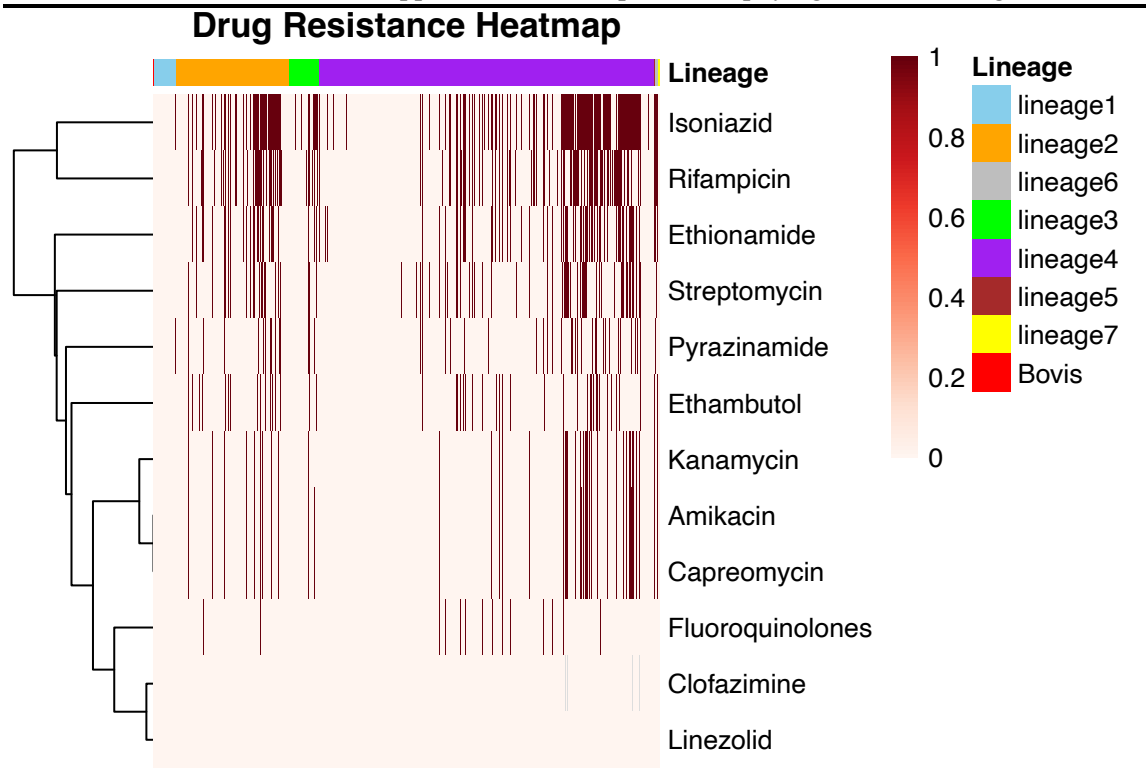


Figure 20: A heat map showing the presence (1) or absence (0) of drug resistance-associated mutations for the most used anti-TB drug across all lineages in this study. Resistance was more prevalent for the first-line drugs isoniazid and rifampicin compared to the newly introduced drugs, clofazimine and Linezolid.

Among the 1,769 isoniazid-resistant strains, the *katGS315T* mutation was the most prevalent (n=1,228), with 188 isolates also carrying the *fabG1C-15T* mutation. 41 strains harboured the *katGS315T* mutation alongside mutations in the *inhA* gene: *katGS315T* + *inhA T-770A* (n=15) and *katGS315T* + *inhA G-154A* (n=26). Other *katG* resistance mutations included *katG W300C* (n=10), *katG W90R* (n=6), *katG W328L* (n=2), *katG W438** (n=13), *katGS671** and a *katGW191R (* the later coupled with the *fabG1 C-15T*). Additional isoniazid-resistance mutations were observed in *fabG1 (C-15T* (n=160)*, G-17T* (n=121)), *ahpC C-57T* (n=10), *inhA (S94A* (n=12), *G-779T* (n=4), *G-154A* (n=5) T-*770A* (n=7)).

For the 1,610 rifampicin-resistant strains, mutations were exclusively detected in the *rpoB* gene. The *rpoBS450L* was the most prevalent (n=1071), followed by *rpoB L452P* (n=143) Less frequent mutations included *rpoB V170F* (n=63), *rpoB L430R* (n=7), *rpoB Q432K* (n=50), *rpoB D435F* (n=5), *rpoB D435V* (n=22) *rpoB L430P* (n=5), *rpoB I491F* (n=6), *rpoB*

© University of Pretoria

S441V (n=4), *rpoB Q429L* (n=23), *rpoB S450W* (n=2), *rpoB Q432P* (n=8), and a combination of *Q432K, rpoBH445D* in 8 strains.

Fluoroquinolone resistance was linked to 8 mutations (Table 4). Among the 344 fluoroquinolone-resistant strains, the most frequent was *gyrA: D94G/H/A/N/Y* mutation (n=115), followed by *gyrA A90V* (n=107) and *gyrA S91P* (n=5). Mutations in the *gyrB* included *D461N* (n=8), *S447F (n=3)* and *E501D* (n=12), with 6 strains carrying both *gyrB E501D* and *gyrB S447F*.

2.5 DISCUSSION

The availability of large volumes of whole-genome sequencing (WGS) data for *Mycobacterium tuberculosis* (Mtb) coupled with the bioinformatics innovation that came from the COVID-19 pandemic provides unprecedented opportunities to revisit traditional concepts in MTB research. In this study, we curated a dataset of 9,073 global Mtb isolates, addressing the challenges of redundancy, sampling bias, and metadata discrepancies to generate a high-quality dataset of 8,575 strains (Supplementary file 2.3).

Lineage 4 emerged as the most prevalent globally, consistent with its historical association with human migration. However, Lineage 2 showed a disproportionately high prevalence of drug-resistant strains, corroborating previous studies linking some sublineages of Lineage 2 to high virulence, transmissibility, and MDR-TB outbreaks [317]. These findings underscore the importance of monitoring Lineage 2 in TB control. Nevertheless, interpreting drug-resistance results inferred from public databases requires caution due to inherent biases, such as the overrepresentation of DR-TB strains in outbreak studies. Computational drug resistance predictions must be followed up by more robust phenotypic drug susceptibility testing.

Our work identified critical resistance-associated mutations in both established (e.g., rifampicin, isoniazid) and novel drugs (e.g., bedaquiline, linezolid). The prevalence of **rpoB S450L** (rifampicin) and **katG S315T** (isoniazid) mutations aligns with global trends and

© University of Pretoria

systematic reviews, emphasising their utility as diagnostic markers [317, 318]. Alarmingly, we detected mutations conferring resistance to newly introduced drugs, threatening recent advances in MDR-TB treatment. Our results were concordant with previous work that established that the emergence of mutations in the *mmpR5* gene (*Rv0678*) is not solely driven by the use of bedaquiline [276]. These findings highlight the need for proactive monitoring of resistance evolution as part of global TB control efforts.

Beyond the biological findings, our study demonstrates the robustness of maximum likelihood (ML) approaches in handling large genomic datasets. ML enabled accurate lineage classification and drug resistance prediction, as well as reliable phylogenetic reconstruction and molecular clock estimation. The estimated substitution rate of 7.45 E-7 substitutions /site/year (standard deviation: 1E-7) aligns with previous research [319]. However, tip-dating analyses on large datasets remain challenging, and results must be interpreted with caution.

Lastly, discrepancies between lineage classifications in metadata and our reanalysis reveal the urgent need for a standardized and regularly updated global classification framework. This will improve consistency in future studies and facilitate more accurate global comparisons of Mtb diversity.

2.6 CONCLUSION

In this study, we collected and analysed the genome sequences of 9,073 global *Mycobacterium tuberculosis* strains, generating a high-quality reference dataset of 8,575 non-redundant isolates. Using computational approaches that strike a balance between accuracy and efficiency, we characterised their phylogenetic lineages and drug resistance profiles, providing critical insights into the global distribution and evolution of Mtb.

Our findings reaffirm the dominance of Lineage 4 globally and the strong association of Lineage 2 with drug resistance and outbreaks. Moreover, the identification of mutations

© University of Pretoria

conferring resistance to newly introduced drugs underscores the urgent need for vigilance in monitoring resistance evolution.

This study demonstrates the utility of maximum likelihood (ML) approaches in analysing large genomic datasets, enabling robust phylogenetic reconstruction and molecular clock estimation. Moving forward, this work will inform efforts to understand the temporal and geographical spread of drug resistance in Mtb, guiding interventions for global TB control.

© University of Pretoria

CHAPTER 3 PREDICTING ANTIBIOTIC RESISTANCE AND POPULATION DYNAMICS OF MYCOBACTERIUM TUBERCULOSIS IN SUB-SAHARAN AFRICA: AN IN SILICO APPROACH

© University of Pretoria

This chapter has been submitted for peer review as:" D Muzondiwa, ON Reva: *Insilico historical prediction of development of antibiotic resistance and fluctuations of the effective size of Mycobacterium tuberculosis population in Sub-Saharan Africa"*

…

© University of Pretoria

3.1 ABSTRACT

Mycobacterium tuberculosis (Mtb) remains a significant threat to public health in Africa and globally. The African continent, hypothesised to be the origin of Mtb, remains a remarkable area of Mtb diversity where lineages, that were recently reintroduced to the continent from other parts of the world are now dominant. The potential for discovery through African Mtb genomic studies is unparalleled, offering invaluable insights into the genomic and molecular foundations of tuberculosis (TB).

In this study, we analysed a whole genome sequencing (WGS) dataset of 5,745 Mtb strains to characterise the evolutionary history of Mtb on the African continent. Our findings reviewed a stepwise acquisition of mutations associated with drug resistance highlighting the chronological progression of resistance mechanisms. Notably, we identified the *Rv0678 G121R* mutation in a group of Lineage 2 strains isolated before the use of bedaquiline and clofazimine. This mutation has been recently shown to be associated with resistance to the two novel drugs.

We identified fluctuations in the Mtb effective population size at the turn of the 18th century, which coincided with increased migration rates from Europe into Africa during that period. Lineage-specific Bayesian analysis also showed a rapid expansion of different Mtb lineages at the turn of the new millennium, a time when HIV/Aids was ravaging the continent. Our result highlights the role of human mobility and migration, antibiotic resistance development and HIV pandemic outbreaks on the severity of the TB pandemic in Africa.

3.2 INTRODUCTION

According to the latest World Health Organization report, TB has recently regained the top spot as the leading global cause of death from a single infectious agent. Covid-19 has overtaken TB during the past three years [8]. Africa and Asia are the worst affected regions carrying over two-thirds of the global TB burden [8, 320]. The African continent is considered the origin of *Mycobacterium tuberculosis* (Mtb) because of the highest genetic

© University of Pretoria

diversity of *M. tuberculosis* isolates on the continent [6, 34]. The continent is home to the specialist Mtb Lineages 5, 6, 7 and 9 which are found exclusively in some regions of Africa [192] . However, the modern TB pandemic in Africa is predominantly associated with Lineages 2 and 4, which are strongly believed to have been re-introduced to the continent from Europe, Asia and Latin America [169, 170]. Some strains from these lineages are characterised by a higher transmissibility rate, increased virulence and rapid development of antibiotic resistance [321, 322]. The dynamics influencing the re-introduction of Mtb into Africa are not yet well understood and studies that explore the evolutionary and population dynamics of the pandemic on the continent have the potential to offer significant insights that support global and regional TB elimination efforts.

The emergence of drug resistance (DR) in Mtb has been an obstacle to the global set goals of ending TB [323, 324]. TB strains can be classified as either drug-susceptible, multi-drug-resistant TB (MDR-TB), pre-extensive drug-resistant TB (pre-XDR-TB) or extensive drug-resistant TB (XDR-TB). Drug-susceptible TB strains are treated using a first-line regimen of isoniazid (INH), rifampicin (RIF), ethambutol (EMB), and pyrazinamide (PZA). MDR-TB refers to strains that have acquired resistance to the first-line drugs INH and RIF. Pre-XDR-TB refers to MDR-TB and RIF-resistant (RR-TB) strains that have additionally gained resistance to any fluoroquinolone [8]. XDR-TB is TB caused by MDR/RR-TB strains that have gained additional resistance to any fluoroquinolone and at least one additional Group A drug [82, 323].

Due to the serious implications of tuberculosis on public health, *M. tuberculosis* is one of the most studied bacterial species. Tuberculosis in humans is mainly caused by *M. tuberculosis sensu stricto* (Lineages 1, 2, 3, 4, 7 and 8) and *M. africanum* (Lineages 5, 6 and 9) strains forming nine human-adapted lineages [171, 172, 325]. These lineages form part of the *M. tuberculosis* complex (MTBC), which consists of several strains affecting different mammalian hosts as well as the ancestral smooth tubercle bacilli, *M. canettii* [326]. Human-adapted MTBC strains follow a phylo-geographical population structure with some lineages more widely distributed around the globe than others [52]. MTBC lineages can be classified as either generalist or specialist with the former referring to lineages that are widely distributed worldwide and the latter referring to lineages that are only restricted to a specific

© University of Pretoria

geographical region [327]. Lineage 2 and Lineage 4 are more widespread compared to the other lineages with the former (also known as the East Asian strain) predominantly present in East Asia and the latter (also known as the Euro-American lineage) frequenting Asia, Europe, Africa and America. It is also important to note that some Lineage 4 sub-lineages are exclusively restricted to certain geographical niches. Lineage 1 and Lineage 3 have a less widespread distribution that is restricted to East Africa, Central, South and Southeast Asia[52]. Lineages 5, 6 and 7 are much more geographically restricted to some specific regions in Africa [52]. Lineages 5 and 6 also known as *M. africanum* West Africa 1 and West Africa 2, respectively are mostly found exclusively in the Western Africa region or from recent immigrants of West African origin [52, 167]. The Lineage 5 strains are predominantly found further to the East in regions bordering the Gulf of Guinea [52, 168]. Lineage 7 strains are mostly restricted to Ethiopians and recent immigrants from the Horn of Africa region [328]. Lastly, the recently discovered lineage 9 has only been isolated in Somalia [329]. The reason why these lineages are exclusive to Africa is still elusive. Interestingly, Africa is the only continent that harbours all the human-adapted lineages [327, 330]. The current body of evidence also places the origins of the MTBC in Africa and attributes the global spread of the Mtb strains to the out-of-Africa migrations of modern humans [45, 50, 331, 332]. Although Africa has taken significant strides in the fight against TB over the past decade, TB incidence in the continent remains high mostly fuelled by high HIV prevalence rates and development of resistance to anti-TB drugs [333].

The past decade has seen a concerted effort towards the development of new TB drugs, and the successful adoption of bedaquiline, delamanid and pretomanid in treating MDR-TB has been a testament to that [280, 282, 284]. South Africa was one of the first countries to officially avail the drug for clinical use [281]. Concerningly, several studies have reported on bedaquiline resistance in different countries [275, 283, 284]. Recent studies have implicated mutations in ATP synthesis genes in causing cross-resistance to bedaquiline and clofazimine [277]. The widespread presence of preexisting *Rv0678* mutations in isolates that have never been exposed to bedaquiline raises significant concerns [276]. Resistance to the nitroimidazoles pretomanid and delamanid has been attributed to mutations in the *fdg1*, *ddn* and *fbiABCD* genes that are involved in ketomycolic acid synthesis pathways [61, 69, 334].

© University of Pretoria

Characterizing resistance in these recently introduced anti-TB drugs is further complicated by our limited understanding of the variants involved.

Understanding the processes that drive the evolution of Mtb is imperative in eradicating the global scourge of TB and the characterisation of the genetic diversity of strains isolated in Africa holds significant potential for advancing TB research and adding valuable insights on treatment of the disease on the continent and beyond. Due to the low information content of the MTBC genome, researchers in the past showed little interest in studying the evolutionary processes in the MTBC [3]. This has resulted in a gap in theoretical and conceptual research on MTBC evolution [3]. Next-generation sequencing (NGS) technology has not only revolutionised how TB is diagnosed and treated but has impacted every facet of the TB field since the first *M. tuberculosis* genome sequence was completed 26 years ago [51]. However, very few studies have been undertaken to explore the evolutionary dynamics of Mtb on the African continent. Leveraging the extensive genomic data of thousands of Mtb strains that have been sequenced across the globe, we curated a dataset of WGS sequences and investigated the evolutionary history of Mtb with a specific focus on the African continent.

3.3 MATERIALS AND METHODS

3.3.1 Data collection, processing and variant calling

A WGS dataset of 6,058 MTB strains (Supplementary 3.1) isolated from 23 countries covering the 5 United Nations sub-regions of the African continent between 1991 and 2021 was curated from the Bacterial and Viral Bioinformatics Resource Centre (BV-BRC) [297], the Genome-wide *Mycobacterium tuberculosis* variation (GMTV) [295], European Nucleotide Archive (ENA) [335] and the AFRO-TB dataset [296]. Additionally, we compiled metadata such as geographical location, lineage classification, drug-resistance profiles and date of isolation for each sample. In instances where metadata was missing, we sought to manually add the information by going through the associated literature and in some cases, personal correspondence with stakeholders involved in some of the studies.

© University of Pretoria

Raw sequences were quality-controlled and trimmed using Trimgalore v0.610 [299]. The raw data was filtered to keep samples that had sequences read depths of 30X or higher and a genome-wide coverage of 99% or higher. The high-quality reads were mapped to the H37Rv reference genome (Genbank: NC000962.3) using BWA mem v0.7.17. SAMTOOLS v1.10 [302, 303]. Variant calling was performed using FreeBayes v1.3.7 [304]. Quality control on the variants was performed using BCFtools v1.9 [302], where positions with phred-scale quality scores < Q20 and mapping quality < 40 were discarded. Lineage classification and drug resistance prediction were performed using Resistance Sniffer [24] and the FastLineageCaller v.0.3.2 [305] was used to validate the lineage classifications. Treemer v3 [311] was used to identify and filter out redundant strains while preserving underrepresented strains based on phylogenetic lineage and geographical location. Strains associated with more than one lineage were excluded from further downstream analysis.

3.3.2 Evolutionary analysis

INDELS were filtered from the VCF files using VCFtools v0.1.17 [307]. Low confidence regions (LCR) as described by Marin *et al.*[308] were masked from the VCF files. SNPs that were in proximity to LCRs (within 50bp) were also filtered out. The resulting vcf files were then merged using BCFtools v1.9 [302] and missing genotypes in individual vcf files were set to missing in the BCFtools merging process. The vcf2phylip v2.9 tool was used to convert the multi-sample vcf file to fasta format [309] and the SNP-sites program was used to generate a multiple sequence alignment of variable positions [310]. To minimise bias in the tree topology, we also excluded known drug resistance-associated positions (from the literature) from the alignment. An initial maximum likelihood (ML) phylogeny was created with IQTREE v2.1.3 [293] using the GTR model for nucleotide substitutions. For all the other ML-based analyses, IQTREE2 v2.1.3 was used to construct the phylogenetic tree with the best model selected by ModelFinder [312] with 1000 ultrafast bootstrap replications [315]. To resolve branch lengths and obtain estimates for clock rates and Time to the Most Recent Common Ancestor (TMRCA), the inferred trees were further processed using TreeTime v0.94 [238] as implemented in Augur [314]. Tree annotation and ancestral state

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

reconstruction were performed with Augur [314] and the results were visualized using Auspice and Chiplot v2.6.1[336].

3.3.3 Bayesian evolutionary analysis

Lineage-specific dated phylogenetic trees were reconstructed using Markov chain Monte Carlo (MCMC) Bayesian phylogenetic inference as implemented in BEAST v2.7.7 [251]. This also allowed obtaining absolute TMRCA values for variants of interest, specifically those associated with resistance to anti-TB drugs. The sampling dates of sequences can be used to calibrate a phylogeny into a dated tree whose branches are scaled in units of time. However, this is only possible when adequate genetic change has occurred throughout sampling such that there is a correlation between the divergence in each sequence and its sampling date [221], this is known as a temporal signal. To assess the strength of the temporal signal, we performed root-dating on the updated ML tree using Tempest v1.5.3 and TreeTime v0.11.2 [233, 238]. We then performed randomised tip dating in BEAST to compare substitution rates derived from the actual tip dates and those obtained from cluster-randomised tip dates [337, 338]. Based on the insufficient temporal signal in the data (Figure 21D and Supplementary Figure 3.1), we decided to use a strict molecular clock rate of 4.6E-8 suggested in another publication [16] with a lognormal prior distribution on the clock rate for all our Bayesian phylogenetic inferences. The general time reversible (GTR) and Hasegawa-Kishino-Yano [242] nucleotide-substitution models were chosen, and a gamma distribution with four rate categories was selected to model among-site heterogeneity. These two complex nucleotide models were chosen because of their robustness in reconstructing deep phylogenies [245]. Since alignments of only variable sites were being used for Bayesian inference, constant site weights for each of the four nucleotides were calculated and used to modify the BEAST XML files generated by BEAUTI v2.7.7 [339]. This was done to correct for ascertainment bias as recommended by the developers of BEAST 2 [339]. A chain length of 300,000,000 was chosen with a burn-in of 100,000 sampling every 10,000 steps of the chain. Tracer v1.7.2 [266] was used to assess the runs with effective sample sizes > 200 chosen as sufficient convergence. Maximum clade credibility trees with median node

© University of Pretoria

heights were reconstructed using TreeAnnonator v2.7.6 after discarding 10% burn-in. FigTree v1.4.4 was used to visualize the trees.

3.3.4 Dynamics of effective population size

For each lineage, we extracted sub-datasets of strains showing the strongest temporal signal. For computational efficiency, we limited each dataset to a maximum of 250 isolates (Appendices). We further split Lineage 4 into datasets representing the sublineages. The effective size of the Mtb population in Africa over time was inferred using Bayesian skyline analyses as implemented in Tracer using the log files of the MCMC sampling process as described above [248]. To minimize the confounding effect of population structure on Bayesian skyline analyses, we carefully selected strains belonging to the same monophyletic clade for each lineage. To account for potential differences in clock rates among different lineages, we adopted lineage-specific rates reported in previous studies [319]. Lineages 5,6, and 7, were excluded from the analysis due to the datasets being homochronous (assumed to be sampled at the same time). Skyline plots were used to visualize the changes in genetic diversity over time. The XML files can be found in the supplementary material.

3.4 RESULTS

3.4.1 Initial data curation and lineage classification

In this study, we curated an initial WGS dataset from public databases consisting of 6,058 Mtb strains isolated from 23 countries on the African continent (Supplementary file 3.1). Our dataset encompassed all 5 subregions of the African continent as defined by the UN as shown in Figure 21A.

© University of Pretoria

In total, eight lineages of the human-adapted MTBC were identified among African Mtb isolates as shown in Figure 21B. Lineage classifications were further validated using the FastLineage caller tool [305] and the program TB-profiler v4.4.0 [306]. Ten isolates from the AFRO-TB database had missing lineage information and our classification identified seven isolates from Djibouti as *M. canettii* while two isolates were ambiguously assigned to both Lineage 4 and Lineage 2. We classified the remaining sample with ID SRR1082883 from Rwanda as belonging to Lineage 8. From the collected metadata, 106 isolates were classified as "Africanum_Bovis", our lineage classification identified 59 of the samples as *M. bovis.* Twenty-two samples had ambiguous lineage classification possibly due to mixed infection. The 88 identified *M. canetti* , *M. bovis* and ambiguously classified isolates were excluded from further analysis. Scrutiny of the sampling location metadata also identified 185 samples from Papua New Guinea that were mistakenly recorded as isolated from the Republic of Guinea in West Africa. As a quality control measure, we excluded all samples reported to have been isolated from Guinea. The remaining high-quality dataset consisted of 5,745 isolates (Supplementary 3.2).

3.4.2 Phylogenetic analysis

Allelic states of predicted polymorphic sites in the 5,745 genomes of African Mtb isolates were retrieved from VCF files using vcf2phylip [309]. The resulting SNP alignment (Supplementary Materials) comprised of 111,433 sites, of which 42,775 were parsimony-informative and 19,776 were non-informative invariants. Phylogenetic inference was performed using IQTREE2 with a maximum likelihood approach focusing on parsimony informative sites. Model GTR+I+G was predicted by ModelFinder to be the best fit to the generated SNP alignment. Figure 24 shows the inferred phylogenetic tree.

Redundancy in the dataset was assessed with Treemer, showing that retaining 99% diversity preserved 4,003 isolates (Figure 21C). A weak but significant temporal signal was detected via root-to-tip regression ($R^2 = 0.05$, $p < 0.05$; Figure 21D). Molecular clock estimation using TreeTime identified 1643 outliers and estimated a substitution rate of 1.19×10^{-7}. The final time-resolved phylogeny is shown in Figure 22. After excluding the identified redundant

© University of Pretoria

isolates and the outliers from the molecular clock estimates, a dataset of 2168 samples was retained as a "reservoir" (Supplementary file 3.3) for subsampling isolates for downstream Bayesian analyses which require smaller datasets due to high computational demands.

© University of Pretoria

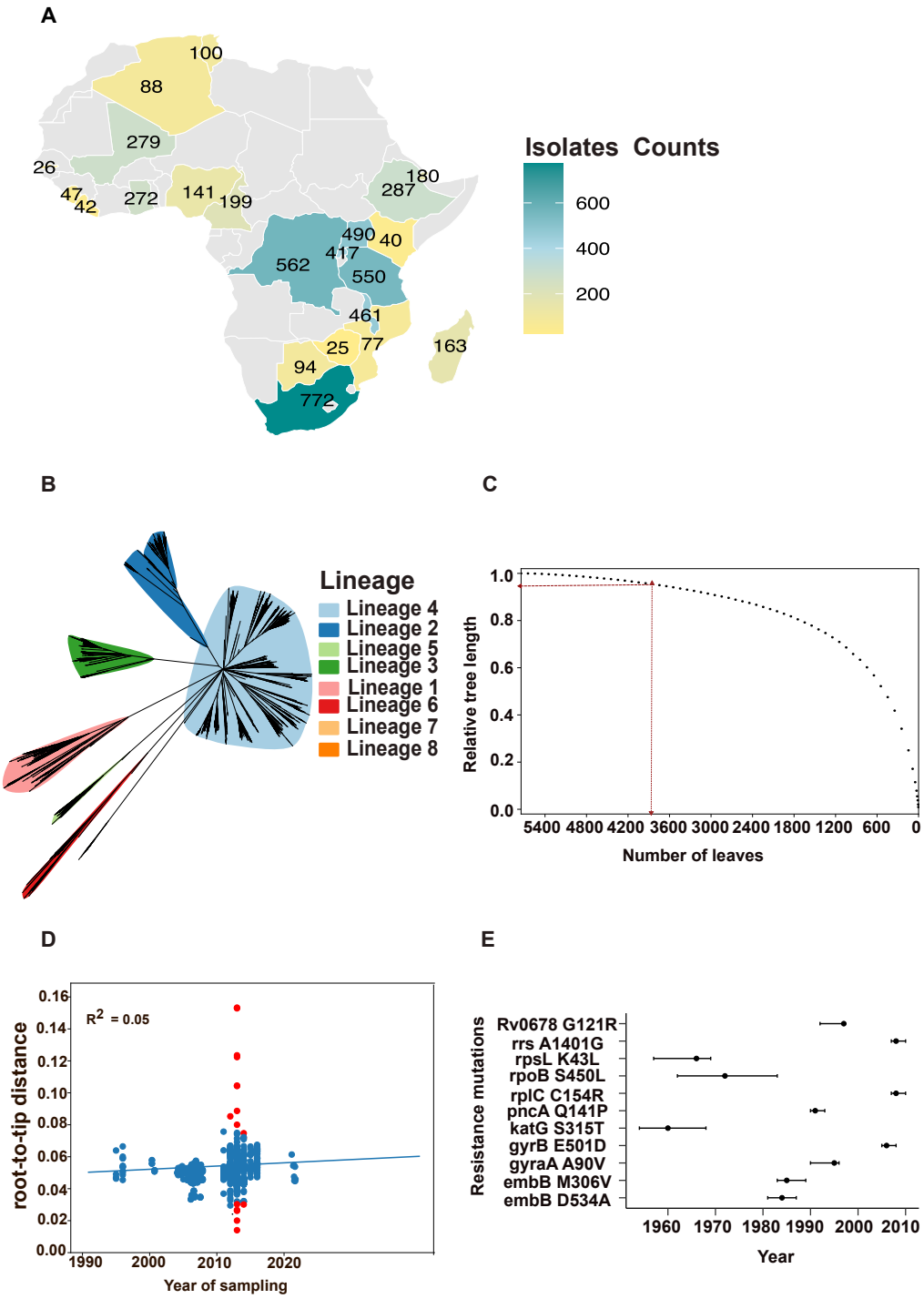


Figure 21 Sampling scheme, lineage distribution redundancy analysis, temporal signal analysis and antibiotic resistance emergence timing of 5,745 African Mtb strains. (A) Map of

© University of Pretoria

Africa showing the countries represented in our original dataset as well as the number of isolates for each country. The map is coloured according to several sequences isolated per country. (B) Unrooted phylogenetic showing the main lineages of Mtb identified in our African dataset. (C)The decay curve shows the high redundancy levels in our starting dataset, the red arrows show that 99% diversity of the dataset can still be preserved after pruning more than 2000 leaves from the phylogenetic tree. (D) Root-to-tip regression plot showing a weak but positive temporal signal. The red dots represent identified outliers. (E) the estimated time of emergence in Africa for some of the known drug resistance mutations.

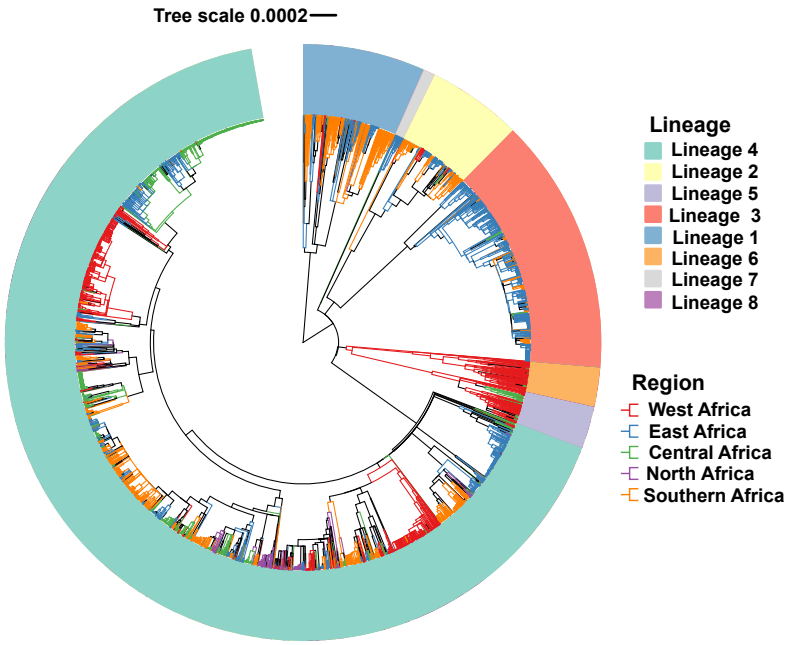


Figure 22: Maximum likelihood tree of the 5745 African Mtb strains. The outer bands show the distribution of the main lineages, with Lineage 4 being the most dominant. The branches are coloured by regional location. The phylogeographical structure of the Mtb is visible with Lineage 3 (peach band) prevalent in East Africa (blue branches), and Lineage 5 and 6 (orange and grey bands) mostly dominate West Africa (red branches).

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

3.4.3 Lineage distribution and antibiotic resistance profiling

Our results suggest that Lineage 4 strains were the most dominant in Africa (59.94%) followed by Lineage 2 (17%), Lineage 3 (12.53%) and Lineage 1 (5.97%). Lineages 5, 6 and 7 combined were the least represented (4.56%). Within Lineage 4 strains, sub-lineage 4.3 was the most prevalent (42%), followed by sub-lineages 4.1 (16%), 4.6 (14%), 4.4 (10%), 4.8 (6%), 4.2 (3%), 4.7 (2%), 4.9 (2%) in that order. The rest of the sub-lineages of Lineage 4 constituted the remaining 3%. We identified 2 sub-lineages of Lineage 2 with the sub-Lineage 2.2.1 representing 74% and 2.2.2 representing 26%, respectively.

We sought to infer antimicrobial resistance (AMR) in silico for 14 anti-TB drugs (rifampicin, isoniazid, streptomycin, ethambutol, ethionamide, kanamycin, capreomycin, amikacin, pyrazinamide, fluoroquinolones, delamanid, clofazimine, linezolid, bedaquiline) using the program Resistance Sniffer [24]. Most of the samples were predicted to be susceptible to anti-TB drugs (n=3986), and 1759 strains were resistant to at least one anti-TB drug. Among the predicted DR-TB strains, 7 % of the strains were resistant to only one anti-TB drug (n=398, including 363 strains resistant to one first-line anti-TB drug). Our Insilico drug resistance profiling identified resistance-associated variants for all drugs tested. For the mono-resistant strains showing resistance to at least one first-line anti-TB drug [340], the resistance was predicted against Isoniazid (n=218), Rifampicin (n=138), Pyrazinamide (n=14), and Ethambutol (n=9). Among the 1374 strains, which were resistant to more than one drug, 958 belonged to the MDR-TB category, and 96 were pre-XDR-TB strains. The classification of the multidrug-resistant corresponds to DR-TB classification criteria as defined by WHO [82]. No strains were classified as XDR-TB. Among the predicted pre-XDR-TB isolates, we identified five strains resistant to nine drugs, with susceptibility retained only to bedaquiline, linezolid, delamanid, bedaquiline and clofazimine.

A notable pattern was observed in the distribution of known AMR mutations among resistant Mtb strains. Notably, of the 1513 isoniazid-resistant strains, 984 (65%) strains carried the solo *katGS315T* mutation, 197 strains carried the solo *fabG1C-15T* mutation; 167 strains carried the *fabG1G-17T* mutation in tandem with the *katGS315T*; and 326 strains carried the

© University of Pretoria

fabG1C-15T mutation in tandem with the mutation in *katGS315T*. The rest of the isoniazid-resistant strains (n=12) harboured the *ahpCC-57T* mutation in tandem with *katGS315T*. For all the 1151 rifampicin-resistant strains, mutations were identified only in *the rpoB* gene with *rpoB S450L* (n=947) mutation accounting for 86.73%. The rest of the *rpoB* mutations were distributed as follows: *rpoB L452P* (n=101); *rpoB Q429L* (n=17,) *rpoB S450W* (n=16), *rpoB L430P* (n=11); *rpoB V170F* (n=10); *rpoB I491F* (n=8); *rpoB D435F* (n=6), *rpoB S441V* (n=5). The *rpoB Q429L* mutation was only detected in Lineage 2.2.2 strains, while *rpoB S450W* was detected in strains belonging to Lineage 4.3. In total, 7 % (n=124) of all analysed African Mtb isolates carried mutations associated with resistance to fluoroquinolones. Among them, the substitution *gyrA A90V* was the most frequent (88%). The rest of the mutations associated with resistance to fluoroquinolones were as follows: *gyrA S91P* (n=1), *gyrB D461N* (n=2), *gyrB E501D* (n=3), and *gyrB S447F* (n=2). Twelve strains classified as Lineage 2.2.2 carried the *Rv0678 G121R* mutation, which is associated with bedaquiline and clofazimine resistance [277, 341]. Among the fluoroquinolone-resistant strains of Lineage 4.3, the most dominant mutation was *rpoB L452P* (82%) associated with rifampicin resistance. In all other lineages, fluoroquinolone-resistant strains carried the *rpoBS450L* mutation. In 28% of Lineage 4.4.1 isolates, the MDR phenotype was conferred by a combination of *fabG1 C-15T* and *rpoBS450L* mutations, while 71 % carried a combination of *katGS315T* and *rpoBS450L* mutations. A single Lineage 2.2.1 strain was harbouring *rplC C154R* mutation associated with linezolid resistance.

To understand the evolution of drug-resistant tuberculosis (DR-TB) in Africa, we estimated the timescales over which resistance-associated mutations were acquired in *M. tuberculosis*. First, we mapped known resistance-associated variants onto a time-scaled phylogenetic tree generated with TreeTime (Supplementary Figure 3.2), identifying the nodes where these mutations emerged and the estimated timeframes associated with these nodes. To enhance robustness, we performed a complementary analysis using a Bayesian phylogenetic framework (BEAST2), focusing on a subset of 245 samples representing all major lineages except the ancient Lineages 5, 6, and 7, while ensuring the inclusion of diverse drug resistance profiles (susceptible, mono-resistant, MDR, pre-XDR). Additionally, lineage-specific datasets were analysed to capture finer evolutionary patterns.

© University of Pretoria

Bayesian inference placed the time to the most recent ancestor (TMRCA) of the 245 Mtb strains in the year 1781 (95% highest posterior density (HPD) interval: 1760 -1796) while the first emergence of drug resistance-conferring mutations (*katGS315T* associated with isoniazid resistance) was inferred to the year 1959 (95% HPD interval from 1945 - 1961) (Figure 21E and Table 6). Rifampicin resistance-associated mutations (*rpoB S450L*) emerged 14 years later in 1983 (95% HPD interval: 1979 - 1984). Resistance to streptomycin (*rpsL K43R*), ethambutol, (*embBM306V*), and pyrazinamide emerged in 1960-1990 with 95% HPD intervals of 1957 - 1969, 1983 - 1989, and 1990 – 1993, respectively. Fluoroquinolone resistance mutations appeared later around 1995 (95% HPD interval: 1990 – 1996) and 2006 (95% HPD interval: 2005 – 2008) for *gyrA A90V* and *gyrB E501D*, respectively (Supplementary figure 3.3). Lastly, mutations associated with amikacin and capreomycin resistance (*rrs A1401G* and *rrs G1484T*) were the last to emerge in 2008 (95% HPD interval: 2007 – 2010).

Table 6: Comparison of estimated time of emergence of resistance-associated variants in different lineages of Mtb strains isolated on the African continent. The results show that resistance-associated mutations were acquired chronologically across all lineages with isoniazid resistance-associated mutations emerging before rifampicin resistance-associated mutations

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Gene	Lineage 1	Lineage 2	Lineage 3	Lineage 4.1	Lineage 4.3	Lineage 4.4.1
TMRCA	1916 (1891 - 1937)	1913 (1895 - 1928)	1932 (1912 - 1950)	1880 (1847 - 1910)	1875 (1855 - 1892)	1887 (1847 - 1858)
fabG1 C-15T	1974 (1964 - 1984)	1987 (1983 - 1992)	NA	1996 (1991 - 2001)	NA	1997 (1994 - 2000)
inhA G-154A	2011 (2011 - 2012)	NA	NA	NA	1983 (1979 - 1990)	2006 (2003 - 2008)
inhA S94A	2008 (2006 - 2009)	NA	NA	NA	NA	NA
katG S315T	1974 (1964 - 1984)	1965 (1958 - 1973)	1996 (1994 - 1999)	1963 (1951 - 1974)	1960 (1954 - 1968)	1953 (1940 - 1965)
*katG W412**	NA	NA	NA	NA	1980 (1975 - 1985)	NA
rpoB Q429L	NA	1999 (1997 - 2001)	NA	NA	NA	NA
rpoB L430P	2005 (2003 - 2007)	1973 (1967 - 1979)	NA	NA	NA	2010 (2008 - 2012)
rpoB S450L	2011 (2011 - 2012)	1980 (1974 - 1986)	2005 (2002 - 2008)	1972 (1962 - 1983)	1973 (1967 - 1978)	1981 (1974 - 1987)
rpoB S450W	NA	NA	NA	NA	1992 (1990 - 1994)	NA
rpoB V170F	NA	NA	NA	NA	NA	2004 (2000 - 2007)
rpoB S441V	NA	NA	NA	1992 (1986 - 2000)	NA	NA
rpoB L52P	NA	2008 (2008 - 2009)	NA	1980 (1973 - 1987)	1986 (1982 - 1989)	2005 (2004 - 2006)
embB M306V	2013 (2012 - 2013)	1987 (1983 - 1992)	NA	2000 (1995 - 2004)	1987 (1984 - 1990)	1982 (1977 - 1989)
embB M306L	NA	NA	NA	NA	NA	2006 (2005 - 2008)
pncA V9G	NA	NA	1996 (1994 - 1999)	2012 (2011 - 2012)	NA	NA
pncA W68C	NA	NA	1988 (1983 - 1993)	NA	1973 (1967 - 1979)	NA
gyrAA90V	2013 (2013 - 2014)	2006 (2011 - 2014)	NA	2008 (1995 - 2010)	2001 (1998 - 2003)	2007 (2004 - 2009)
rrs A1401G	NA	1997 (1992 - 1997)	NA	2001 (1998 - 2003)	1987 (1984 - 1990)	NA
Rv0678 G121R	NA	1997 (1992 - 1997)	NA	NA	NA	NA
rplC C154R)	NA	2008 (2007 - 2009)	NA	NA	NA	NA
*gidS136**	NA	1965 (1958 - 1973)	1995 (1990 - 2000)	NA	1965 (1958 - 1973)	NA

© University of Pretoria

rpsL K43L	NA	1972 (1962 - 1983)	NA	1973 (1966 - 1984)	1972 (1967 - 1979)	1981 (1974 - 1987)
gidA134E	NA	NA	NA	1967 (1959 - 1973)	1966 (1959 - 1973)	NA
*gidR102**	NA	NA	NA	NA	1973 (1967 - 1979)	NA
rrsA514C	NA	1996 (1993 - 1997)	NA	NA	NA	NA
rrsC513T	NA	NA	NA	NA	1971 (1969 - 1979)	NA
rrsC517T	NA	1980 (1976 - 1986)	NA	NA	NA	NA

© University of Pretoria

3.4.4 Dynamics of Mtb lineage expansion in Africa

To estimate population changes over time in Mtb, we subsampled datasets for Lineages 1,2,3 and sublineages 4.1, 4.3, 4.8 and 4.4. We then applied the Bayesian skyline population model to each dataset. This model estimates a posterior distribution of effective population size *(N_e)* through time under a specified nucleotide substitution model [342, 343].

Bayesian coalescence-based demographic reconstruction using a subsample of 245 Mtb strains representing all lineages showed a period of relative stability in the *(N_e)* of MTBC strains isolated in Africa between 1400 to ~1650 as shown in Supplementary Figure 3.4A. Please note that the y-axis scales in Supplementary Figure 3.4 are adjusted dynamically for each panel to fit the range of population size changes in that specific dataset. We avoided uniform scaling across all panels as that would have obscured details for datasets with smaller population sizes.

There was a steady increase in *(N_e)* between ~1700 and 1800 followed by a more pronounced growth in the 19th and 20th centuries. There was also a sharp rise close to the year 2000 followed by a period of stabilization from 2010 to 2021. Figure 23 shows Bayesian skyline plots for Lineages 2 and 4. The figure shows a relative stability of *(N_e)* in Lineage 4 (Figure 23A) between ~1500-1700 followed by a gradual increase starting from ~1700-1800. This was then followed by a decline around ~1900 , the *(N_e)* rapidly increases around 2000, with high certainty. Figure 23B also shows relative stability in Lineage 2 *(N_e)* between ~1600 and 1800. Around 1800 -1900, a slight increase is observed. Interestingly, there is more fluctuation towards 2000, with a final drop towards the present day.

Supplementary Figure 3.4 shows Bayesian skyline plots for Lineage 1, Lineage 3 and the sublineages of Lineage 4. Bayesian skyline analysis for Lineage 1 strains showed a steep decline around 2005 which was followed by a slight expansion after 2010 followed by a period of stabilisation (Supplementary Figure 3.4B). In Lineage 3, the plot shows a relatively stable *(N_e)* pre-1950 to ~2000 with a gradual increase over time. (Supplementary Figure 3.4E) Around ~2000, there was a noticeable increase in *(N_e)* followed by a period of relative stability. The same trend was noticed in Lineage 4.3 (Supplementary Figure 3.4G)

© University of Pretoria

and Lineage 4.4.1 (Supplementary Figure 3.4F), where the population increases in the early 2000s and then plateaus to the present time. The Lineage 4.8 plot (Supplementary Figure 3.4H) shows relative population stability through time with high 95% HPD intervals for the population estimates.

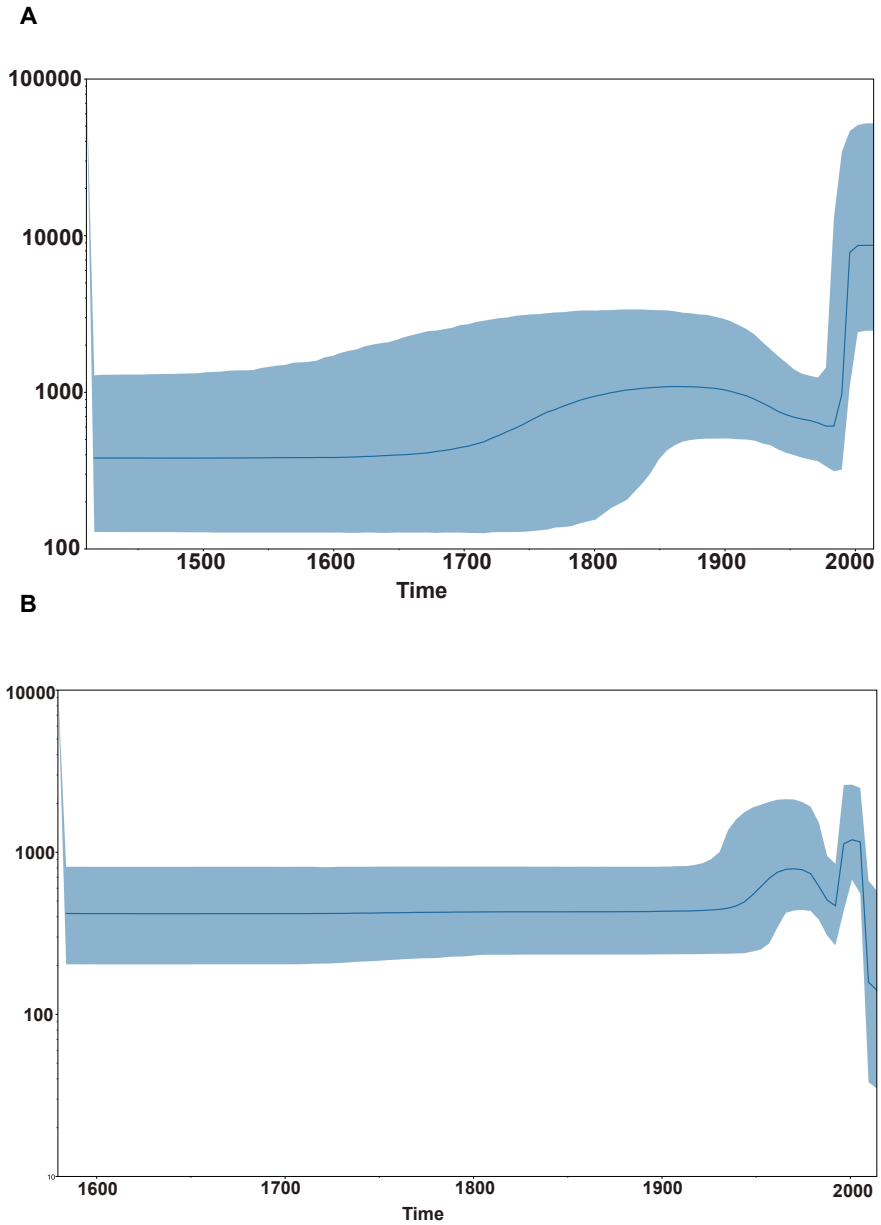


Figure 23: Bayesian skyline plots (BSPs) showing the inferred effective population size over time for Lineage 2 and 4. The y-axis scales vary between panels to accommodate differences in

© University of Pretoria

population size estimates across lineages or datasets. (A) Bayesian skyline plot inferred from 245 Mtb isolates sub-sampled from Lineage 4 strains showing a relative stability between ~1500-1700, there was a graduation population growth starting around 1700-1800. This was followed by a decline in ~1900. Around 2000, the population size rapidly increases. (B) Lineage 2 showed stability between ~1600-1800 followed by a slight expansion between 1800-1900. Towards the year 2000, there was a fluctuation which was followed by a drop.

3.5 DISCUSSION

In this study, we curated a WGS dataset of 5,745 MTBC strains isolated in Africa to characterise the evolution of MTBC lineages on the continent. Our results suggested that Lineage 4 was the dominant strain in Africa, followed by Lineage 2. However, lineage 2 strains had the highest proportion of strains that were predicted to be resistant to at least one anti-TB drug. Our results are consistent with findings from several studies that have associated Lineage 2 strains with high virulence, transmissibility and MDR-TB outbreaks [344-346]. Although several of our strains were isolated before the use of bedaquiline and other novel drugs, we sought to detect variants in genes that have been associated with pre-resistance to the novel anti-TB drugs. Our study identified a mutation in the *mmpR5* gene (*Rv0678 G121R)*, which has been linked to bedaquiline resistance. Studies on clinical isolates have shown that mutations in the *mmpR5,* which represses the expression of the *mmpL5* efflux pump, lead to overexpression of the pump, resulting in elevated minimum inhibitory concentrations (MIC) to clofazimine, bedaquiline, fusidic acid and other novel drugs [276, 277].

Our results are in agreement with a prior study that reported on the emergence of bedaquiline resistance-associated mutations before the use of the drug [276]. We also identified a single Lineage 2 strain which carried the *rplC C154R* mutation which is associated with linezolid resistance. Linezolid was recommended as the most important second-line drug in MDR-TB treatment in 2016; therefore understanding the genetic basis of Linezolid resistance is of crucial importance [347, 348]. Characterising the age and diversity of DR-associated mutations provides important insights into our understanding of widespread resistance as a current challenge. Furthermore, we sought to explore the evolutionary trajectory of the

© University of Pretoria

XDR-TB phenotype. Although resistance to anti-TB medication is often reported to have arisen in the 1980s, our results place the first emergence of isoniazid resistance-conferring mutations in our dataset much earlier to the year 1959 (95% HPD interval: 1945 - 1961). Our results were consistent with past and present clinical guidelines for TB treatment as well as previous study findings. Isoniazid resistance develops before rifampicin resistance, and this was consistent in all lineages and geographical locations. Similar to previous findings, the most frequent mutation among the rifampicin-resistant strains was *rpoB S450L* [349, 350]. Intriguingly, 65% of strains predicted to be resistant to isoniazid carried the *katG S315T*. A more comprehensive systematic review on gene mutations associated with isoniazid resistance using 11,411 Mtb strains from 49 countries concluded that globally 64% of all observed phenotypic isoniazid resistance was linked to *katG 315* mutations [351]. Fluoroquinolone resistance is a known risk factor in MDR-TB treatment failure [349, 352, 353]. Our results predicted that mutations associated with fluoroquinolone resistance were recently acquired with the latest predicted to have emerged in 2013. Although this is in line with common treatment guidelines where fluoroquinolones are only added to the treatment regimen after the failure of first-line therapy, this recent emergence of resistance poses a risk of the transmission of highly resistant MTBC strains.

Bayesian skyline inferences on the evolution of MTBC strains in Africa suggested that the MTBC population went through a steady expansion at the beginning of the 18th century, followed by a more significant growth in the 20th century. The expansion might be due to the introduction of MTBC strains into Africa by infected European immigrants. The industrialisation of Africa during that time was associated with unsanitary working conditions, which might have fuelled the expansion [354]. Furthermore, lineage-specific Bayesian skyline analysis on strains that were predicted to be resistant to at least one TB drug captured a gradual expansion in the effective population sizes of DR-TB strains in the late 1990s and rapid expansions between 2000 and 2010. This period marks the beginning of the widespread HIV epidemic and the subsequent rise in TB incidence that followed [355-357]. There is also a notable stabilisation and decrease in the effective population size across lineages from 2010 to the present. This might be attributed to the period when antiretroviral therapy was rolled out across the continent, and several African countries reported decreases

© University of Pretoria

in HIV prevalence [358]. Our results are in agreement with what was reported in a study on LAM4/KZN XDR-TB strains where researchers used both TB incidence and HIV prevalence data to show a link between the expansion of the effective population sizes of XDR-TB strains and HIV prevalence [337]. Future work will also explore the role of the COVID-19 global pandemic in the population dynamics of MTBC on the African continent. A major limitation of our study was that we could not obtain data from all countries on the continent; hence, our dataset suffered from sampling bias. The range of the Bayesian models we could explore was also limited by our restricted access to more powerful computational resources. However, in future, we will revisit this study armed with more data and computational power.

3.6 CONCLUSION

This study was aimed at characterising the evolution of MTBC in Africa. This plays an important role in the development of TB, guiding treatment outcomes and Mtb control on the continent. Lineage 4.3 strains were the most prevalent on the continent, while Lineage 2 strains were highly associated with drug resistance. Through in silico drug resistance inference, we also identified several lineage 2 isolates that carried the *mmpR5* gene (*Rv0678 G121R)* that has been associated with bedaquiline/clofazimine resistance. With the implementation of initiatives such as the TB Drug Accelerator program [359], we expect to see the development of improved TB drugs shortly. While the use of novel drugs such as delamanid, bedaquiline and clofazimine offer substantial benefits in improving treatment outcomes, this study implores stakeholders on the risks associated with the introduction of novel game-changing drugs. Drug susceptibility testing remains a key priority in guiding the selection of treatment regimens. Our Afrocentric approach to this study also echoes the sentiments of researchers on the continent on the need to build research capacity on the continent, especially after the SARS-COV2 global pandemic [360]. As sequencing technology improves, we expect to see a decrease in the cost and turnaround times of WGS clinical applications. There is also great potential in undertaking massive continent-wise retrospective WGS projects using specimens collected from different TB referral centres across the continent over the years. In this study, we also touched on the link between HIV

© University of Pretoria

and TB as a reminder of the need to carefully treat HIV coinfection and MDR-TB. Future work will explore the role of the SARS-COV2 global pandemic on the evolution of the MTBC.

© University of Pretoria

CHAPTER 4 PHYLOGEOGRAPHY AND GENETIC DIVERSITY OF MTB LINEAGE 4 AND 2 IN THE SADC REGION

© University of Pretoria

This chapter forms part of a manuscript in preparation.

…

© University of Pretoria

4.1 ABSTRACT

Tuberculosis (TB) caused by members of the *Mycobacterium tuberculosis* complex (MTBC), remains a massive public health challenge in the Southern Africa Development Community (SADC), a region bearing some of the highest global TB burdens. Among the MTBC, Lineage 4 and Lineage 2 dominate the SADC region, characterised by unique transmission dynamics, antibiotic resistance patterns and evolutionary histories. Lineage 2 has gained global notoriety due to its association with highly transmissible and virulent outbreaks. Lineage 2 and 4 strains have also been linked to anti-TB drug resistance and poor treatment outcomes. The recent emergence of Lineage 2 in Africa poses a great challenge to TB control efforts in the Southern Africa Development Community (SADC). Concerningly, very few studies have explored the evolutionary dynamics of the two lineages on this part of the continent. In this study, we curated a Whole Genome Sequencing (WGS) dataset of 1214 SADC and 261 non-SADC global representative Lineage 2 and Lineage 4 clinical strains to characterise the evolution and diversity of the two lineages in the SADC region. Our study identified 13 sublineages of Lineage 2 in the SADC region out of 16 identified sublineages. We identified 8 sublineages of Lineage 4 in the region with Lineage 4.3 being the most dominant. Through Bayesian and Maximum Likelihood (ML) phylogeographical inference, we identified East and Southeast Asia as the most likely source regions of Lineage 2 strains that were introduced multiple times into Southern Africa in the late 17th century while excluding Europe as a potential region of origin. Europe was identified as the exclusive region of origin for Lineage 4.3 isolates in the SADC region. Our study also highlights how labour and colonial migration might have contributed to the diversity of Lineage 2 and Lineage 4 MTBC in SADC.

4.2 INTRODUCTION

Tuberculosis (TB) remains a global killer, responsible for over a million deaths per year [8]. According to the World Health Organization's latest report, TB has reclaimed the number one spot as the leading cause of death due to an infectious agent from the SARS Coronavirus even surpassing HIV[8]. TB is caused by the *Mycobacterium tuberculosis* complex (MTBC) which is a group of highly clonal bacteria which affect humans and animals. *M. tuberculosis*

© University of Pretoria

has been classified into nine lineages which follow a phylogeographic distribution [61]. Lineages 2, 3, and 4 also known as the "modern lineages are responsible for most of the global TB burden [361]. Among the modern lineages, Lineage 4 is the most prevalent lineage with a range that spans Africa, the Americas, Asia and Europe [170]. Lineage 2 strains are more prevalent in the eastern parts of Asia and some parts of Central Asia [361]. Lineage 3 is more dominant in India with some cases in East Africa [173]. Lineage 1 forms what have been coined as the ancestral lineages together with lineages 5-9 [61, 361]. Lineage 1 strains are more prevalent in Southeast Asia and India while lineages 5-9 (*M. africanum*) are exclusively found in Africa [171]. In 1995, van Soolingen *et al.*[362] reported on Lineage 2 also known as Beijing lineage, which has gained notoriety as the cause of some of the most devastating TB outbreaks globally. Lineage 2 strains have been associated with high transmissibility, an increase in febrile response in patients during treatment [363], multi-drug-resistance (MDR) [321, 322], high virulence [364], frequent co-infection with Human Immunodeficiency Virus (HIV) [365, 366] as well as the ability to evade protective immunity conferred by Bacillus Calmette Guerin (BCG) [362].

Characterising the evolution of Lineage 2 and Lineage 4 strains is imperative in understanding the dynamics of TB transmission and implementing effective TB control measures. However, the global genomic diversity and population structure of Lineage 2 Mtb remains poorly understood [182]. Southeast Asia has been proposed as the most likely geographic origin of Lineage 2 before it expanded to other parts of the globe [184, 367]. Lineage 4.3 strains have been implicated in causing some of the devastating XDR-TB outbreaks in the region, specifically the LAM4/KZN family of strains that were identified among hospitalised HIV patients in Tugela Ferry KwaZulu Natal in 2005 [83, 337]. Although Lineage 4.3 (LAM) has been the most prevalent in sub-Saharan Africa, recent studies have reported an increase in Lineage 2 prevalence in the African continent. Van Helden *et al.*[368] hypothesised that Lineage 2 strains were disseminated into Southern Africa through the importation of slaves from Southeast Asia as well as increases in the number of Chinese migrant labourers in the 1900s [368]. Sub-Saharan Africa is the worst affected region in terms of HIV/AIDS with the Southern African Development Community (SADC) region having the highest global HIV prevalence rates [369]. There was a notable two-fold increase in TB notification rates in eight SADC member states from the end of 2021 to mid-2023 [370]. HIV/TB coinfection remains a challenge in the region with ~ 26%

© University of Pretoria

of HIV-related deaths in the SADC region due to TB [8, 370]. The global expansion of Lineage 2 Beijing strains is a cause for concern in the SADC region. The coexistence of Lineage 2 and Lineage 4 in the SADC region offers a unique opportunity to study their evolutionary dynamics. However, although previous studies have focused on understanding the phylogeography and population structure of Mtb globally [6, 34], the evolutionary dynamics influencing the success of Lineages 2 and 4 in the SADC region remain poorly understood. It is also concerning that very few studies have formally investigated the evolution of Lineage2 in this part of the continent. In this study, we leveraged publicly available whole genome sequencing data (WGS) to characterise the genetic diversity of Lineage 2 and Lineage 4 Mtb strains in the SADC region through phylogenetic reconstruction. Through WGS and phylogenetic analysis, we aim to unravel the temporal and geographic patterns underlying the adaptation of these lineages in the region and provide insights into their implications for TB control strategies.

4.3 MATERIALS AND METHODS

4.3.1 Collection of genome sequences of MTBC

We filtered the global reference dataset as described in Chapter 2 by geographical location and lineage to extract isolates from the SADC region, resulting in 224 Lineage 2 and 990 Lineage 4 isolates. Additionally, we created a global reference for Lineage 2 by selecting 189 isolates from other regions, including the endemic regions of East Asia and Southeast Asia, Europe and other non-SADC regions of Africa. For lineage 4 we sub-sampled 301 additional isolates from Europe, Asia, the Americas, and non-SADC regions of Africa to serve as a global reference. It is important to note that for this study, the Americas refers to territories that fall under North America, Central America, South America and the associated islands Detailed sampled metadata for these analyses is provided in the Supplementary file 4.1.

© University of Pretoria

4.3.2 Sub-lineage classification and Insilico drug resistance profiling

Lineage classification was performed using FastLineageCaller v 0.3.2 [305] as it allows the user to also incorporate custom lineage classification schemes. We adopted different classification schemes to assign and compare Lineage 2 and Lineage 4 sublineages namely the widely used Coll *et al.*[177] scheme which was curated from a global collection of 1601 MTBC strains, Shitikov *et al.*[190] scheme, which was derived from other existing schemes and the Freschi *et al.*[305] scheme, which was generated from 9584 Mtb isolates, the Bmyc scheme by Mestre *et al.*[186] which utilises genetic markers in a group of DNA repair, replication and recombination (3R) genes. For the Lineage 2 isolates that could not be resolved to a sublineage level, we verified the authenticity of Lineage 2 classification by first assessing the associated BAM files (see Chapter 2 methods) for the presence or absence of three regions of difference (RDs) that are specific to Lineage 2 namely RD105 (79,567 – 83,034), RD181(2,535,429 – 2,536,140) and RD207 (3,120,521 – 3,127,920) using the RD positions described by Faksri *et al.*[371] using the Integrated Genome Viewer IGV v3.0.9. For a more comprehensive classification, we adopted a phylogenetic approach where we performed Bayesian phylogenetic inferences. We then performed a rigorous manual lineage classification using the more comprehensive scheme by Thaworttana *et al.*[182].

4.3.3 Phylogenetic reconstruction

As the first quality control step, insertions and deletions (INDELS) were excluded from the VCF files using VCFtools v0.1.17 [307]. 10% of the Mtb reference genome (H37RV) encompasses regions that are known to be error-prone in genomic analyses [308]. These regions including PE/PPE genes, mobile genetic elements and other large repetitive regions, which are classified as low confidence regions (LCR) are routinely excluded during analysis [52, 372]. In this study we masked LCR in all VCF files as described in a paper by Marin *et al.*[308] . SNPs that were in proximity to LCRs (within 50 bp) were also filtered out. The resulting filtered VCF files were then merged using BCFtools v1.9 [302] and missing genotypes were set to missing, these missing genotypes occur when a sample does not have information for a particular variant in their original dataset. The vcf2phylip v2.9 tool was

© University of Pretoria

used to convert the multi-sample VCF file to fasta format [309]. Program SNP-sites v2.5.1 was used to generate a pseudo-alignment of allelic states of the SNPs identified in multiple genomes [310]. Additionally, SNPs described in the literature as possible antibiotic resistance mutations were automatically removed from this alignment using a precompiled list as they tend to distort phylogenetic inferences due to their susceptibility to homoplasy and the selection pressures imposed by antibiotics which can result in convergent evolution. A maximum likelihood phylogeny was inferred with IQTREE v2.1.3 [293] applying the best nucleotide substitution model proposed by the program ModelFinder [312]. An ultrafast bootstrap algorithm with 500 dataset replications was utilized to assess the robustness of the inferred evolutionary relationships [315]. Branch length estimation and molecular clock simulation were performed using TreeTime v0.94, [238] implemented in the Nextstrain Augur package [314]. This analysis allowed estimation of the time to the most recent ancestor (TMRCA). The Nextstrain Augur package was also used for tree annotation as well as ancestral state reconstructions (ASR) [314]. The results were visualised using the interactive Nextstrain visualization platform Auspice (https://github.com/nextstrain/auspice).

4.3.4 Bayesian Molecular Clock Dating

Markov Chain Monte Carlo (MCMC) Bayesian phylogenetic inference implemented in BEAST v2.7.7 [251] was used to estimate the timeframes at which Lineage 2 and Lineage 4 were introduced into the SADC region on multiple occasions. For tip calibration, we used date estimations generated by TreeTime v0.94. We used the j-model test [373] to determine the most appropriate nucleotide substitution model for the data based on Akaike information criteria scores [313]. We used a relaxed log-normal clock [374] (determined by jModelTest) v2.1.10 with a lognormal prior distribution on the clock rates for the Bayesian phylogenetic estimation. Since the TVM model of nucleotide substitution was not available in BEAST2, we used the general time reversible (GTR) model of nucleotide substitutions [375] with a gamma distribution and four rate categories were selected to model among-site heterogeneity. Since we were using a matrix of only variable sites, constant site weights for each of the four nucleotides were calculated and manually added to the BEAST2 XML files

© University of Pretoria

generated by BEAUTI v2.7.7 [339]. This was done to correct for ascertainment bias. A coalescent constant population prior was chosen under a log-normal distribution with the upper limit set to 250. For the substitution model, we modelled the prior rates with a gamma distribution set from 0 to infinity. A chain length of 300,000,000 was chosen with a burn-in of 100,000 sampling every 10,000 steps of the chain. Tracer v1.7.2 [266] was used to assess the runs with effective sample sizes >200 chosen as sufficient convergence. Maximum clade credibility trees with median node heights were reconstructed using TreeAnnonator v2.7.6 after excluding 10% burn-in. The 95% highest posterior density (HPD) metric was used to assess uncertainty in all Bayesian estimates. FigTree v1.4.4 and Chiplot [336] were used to visualize the trees. All BEAST2 XML files are available in the supplementary material.

4.3.5 Phylogeography

To analyse the spread of the MTBC through space and time, we used the inferred ML phylogenetic trees to perform ancestral character reconstruction using the geographical location information from the provided metadata as possible character states using TreeTime [238] and Nextstrain Augur [314]. For Lineage 4 analyses, we grouped samples from different regions of America as “Americas”. TreeTime inference yielded equilibrium frequencies (πi), representing the likelihood of an isolate originating from a specific region, symmetrised rates (W_{ij}) indicating the bidirectional movement between regions and directional rates (Q_{ij}) estimating region-to-region migration events. Interactive visualisation was done using Auspice, while maps and tables were generated using R [265]. To complement the ML-based phylogeographical analysis and estimate the regions of origin for Lineage 2 and Lineage 4, we employed a coalescent theory-based Bayesian approach as implemented in the Marginal Approximation of the Structured Coalescent (Mascot) tool [376] to infer migration rates between Southern Africa, East Asia, Southeast Asia and Europe for Lineage 2. For Lineage 4, we limited our analysis to dominant Lineage 4.3 and compared the SADC region to Europe and the Americas. We limited our analysis to these regions to ensure that we obtain adequate MCC chain convergence and adequate mixing (see the Bayesian phylogeny section in Chapter 1). MASCOT analysis involves the use of highly complex models and algorithms, thus, an increase in the number of taxa in the analysis also

© University of Pretoria

increases the complexity of the coalescent calculations exponentially. To make our analysis computationally feasible, we limited our datasets to <120 isolates.

4.4 RESULTS

4.4.1 Classification and Phylogenetic Inference

4.4.1.1 Lineage 2

To explore the genetic diversity of Lineage 2 strains in Southern Africa, we compiled a WGS dataset of 224 Lineage 2 Mtb isolates from 7 SADC countries (Figure 27). A general lineage classification of the isolates using the scheme by Coll *et al.*[177] showed that all isolates from the SADC dataset belonged to Lineage 2.2, with 80.56% of strains classified as Lineage 2.2.1. Lineage 2.2.2 made up the remaining 19.44% of the dataset. Isolates from South Africa made up 46 % of the dataset, and Malawi was the second most represented country, contributing 38% of the isolates. The remaining 16% was made up of samples from Mozambique, Tanzania, Botswana and Zimbabwe, as shown in Figure 25. A single strain from South Africa could not be classified beyond the Ancestral or Modern Lineage 2.2.1. We curated an additional dataset of 189 Lineage 2 isolates from East Asia, Southeast Asia, Europe and the rest of Africa to serve as a global reference. Since the lineage classification using the above scheme showed low resolution in differentiating members of Lineage2, we sought to further characterise the genetic diversity of the datasets using the phylogenetic approach.

Based on Bayesian Information Criterion (BIC) scores from ModelFinder, the best evolutionary model for the SADC dataset was the Transversional Model with Variable base Frequencies, ascertainment bias correction and three gamma rate categories (R3) modelling rate heterogeneity (TVM+F+ASC+R3). To obtain a more robust tree, we estimated the phylogeny under a Bayesian constant coalescent population model. We inferred a mean substitution rate of 1.43E-7 (95% HPD, 6.75E-8 – 2.32E-7). After discarding 10% of the chain as burn-in, BEAST2 generated a posterior collection of 8023 trees. Figure 28A shows the maximum clade credibility tree obtained from the 8023 trees with the most recent

© University of Pretoria

ancestor inferred to the year 1857 (HPD 1845 – 1869). From the tree, we identified 13 distinct sub-lineages of Lineage 2. Using the Thaworttana *et al.*[182] scheme, these sublineages were identified as Lineage 2.2.M2.2, Lineage 2.2.M4.8, Lineage 2.2.AA1, Lineage 2.2.M3, Lineage 2.2.M4.4, Lineage 2.2M6.2, Lineage 2.2M2.5, Lineage 2.2.M6.1, Lineage 2.2.M1.1, Lineage 2.2.M2.1, Lineage 2.2.AA3.2, Lineage 2.2.M2.3, Lineage 2.2.M5 as shown in Figure 25. The figure also shows the distribution of the different sublineages per country. South Africa had the highest diversity of Lineage 2 sublineages (n=10) distributed as follows: Lineage 2.2AA1 (16.26 %), Lineage 2.2.M4.4 (8.87%), Lineage 2.2.M2.2 (3.45%), Lineage 2.2.M6.1 (2.96%), Lineage 2.2.M2.1 (1.97%), Lineage 2.2.M4.8 (1.97%), Lineage 2.2M6.2 (1.48%), Lineage 2.2.M2.5 (0.99%), Lineage 2.2.M5 (0.49%) and Lineage 2.2.M1.1 (0.99%). It must be noted that the percentages are calculated from the group total of the entire dataset. Malawi and Eswatini had the second highest diversity of Lineage 2 with each country having 6 different sublineages, however, the proportional representations of these sublineages differed with Lineage 2.2.M2.2 dominating in Eswatini compared to Lineage 2.2.M4.8 which was the most prevalent in Malawi. Interestingly, the Lineage 2.2.M2.2 sublineage was also dominant in both Botswana and Zimbabwe with all the strains isolated in Botswana belonging exclusively to the sublineage, while the remainder of Zimbabwean isolates were classified as Lineage 2.2.M5 which was the most dominant in their eastern neighbour Mozambique. Other two sublineages were identified in Mozambique namely Lineage 2.2.M1.1 and Lineage 2.2.AA3.2 with the latter only identified in Malawi and Mozambique.

Additionally, we sought to analyse the diversity of Lineage 2 in the SADC region in a global context. Bayesian phylogenetic inference on a combined dataset of SADC and global Lineage 2 samples identified an additional three sublineages, namely Lineage 2.1, Lineage 2.2.M4.7 and Lineage 2.2.M4. Figure 24A below shows the entire diversity of Lineage 2 strains identified in this study, while Figure 24B shows how the different Lineage 2 sublineages were distributed globally. Lineage 2.1 (proto-Beijing) strains identified in this study were exclusively from China and Thailand. Interestingly, the two countries were represented in all the sublineages identified in this study, supporting the Asian origin hypothesis of Lineage 2. Strains from South Africa and Malawi were also interspersed with strains from different global regions, a result that suggests multiple introductions of Lineage 2 in the SADC region.

© University of Pretoria

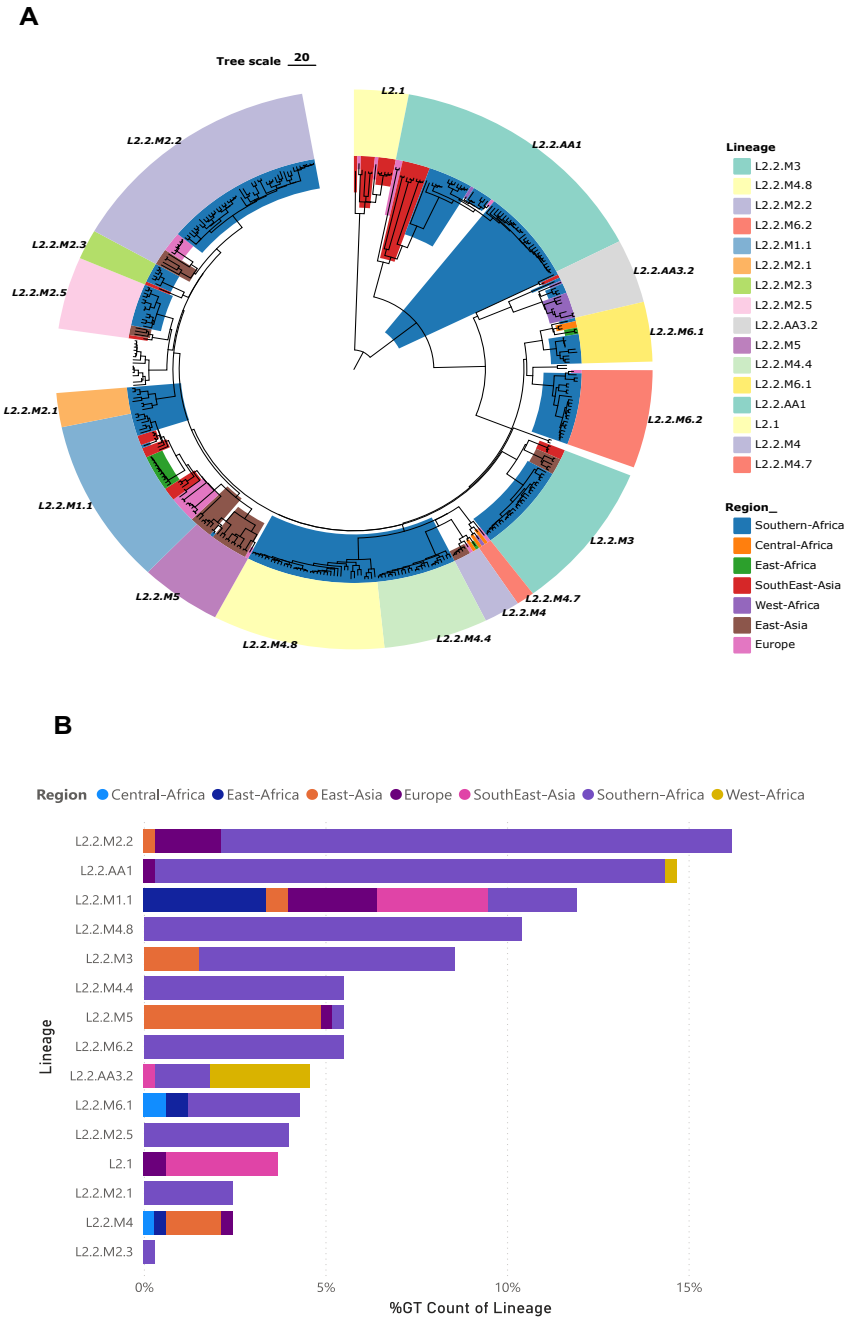
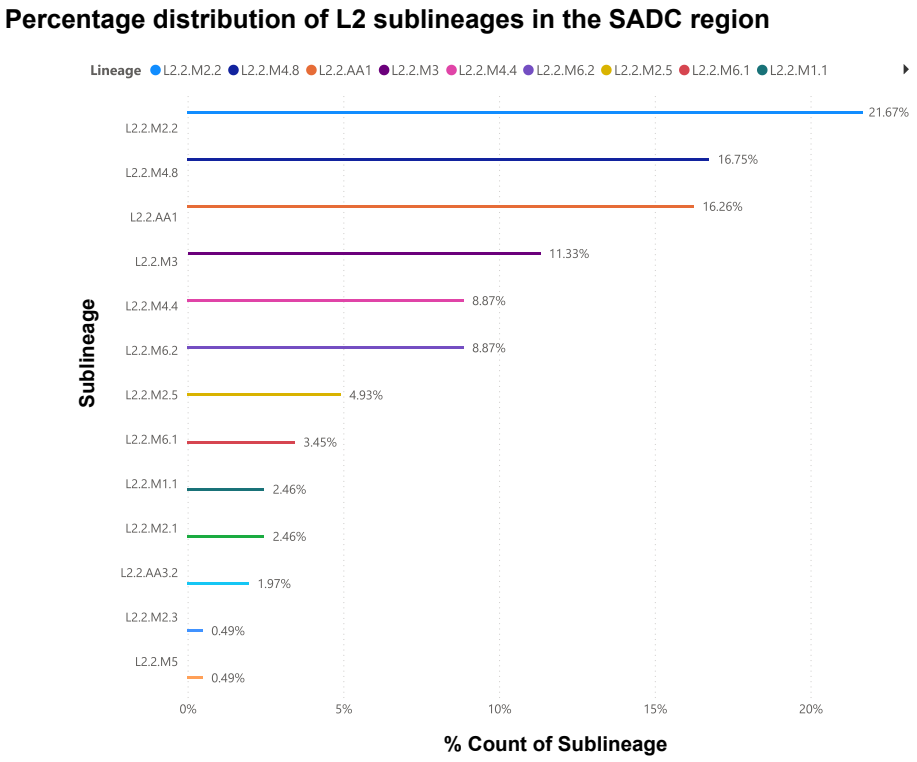


Figure 24: Global diversity of the Lineage 2 strains in this study. (A) Maximum Clade Credibility tree of the combined 413 clinical isolates analysed in this study. The outer bands show the 16 sublineages of Lineage 2 identified in this study. The inner bands shown are coloured according to the region of isolation. (B) The distribution of the 16 sublineages per region.

© University of Pretoria



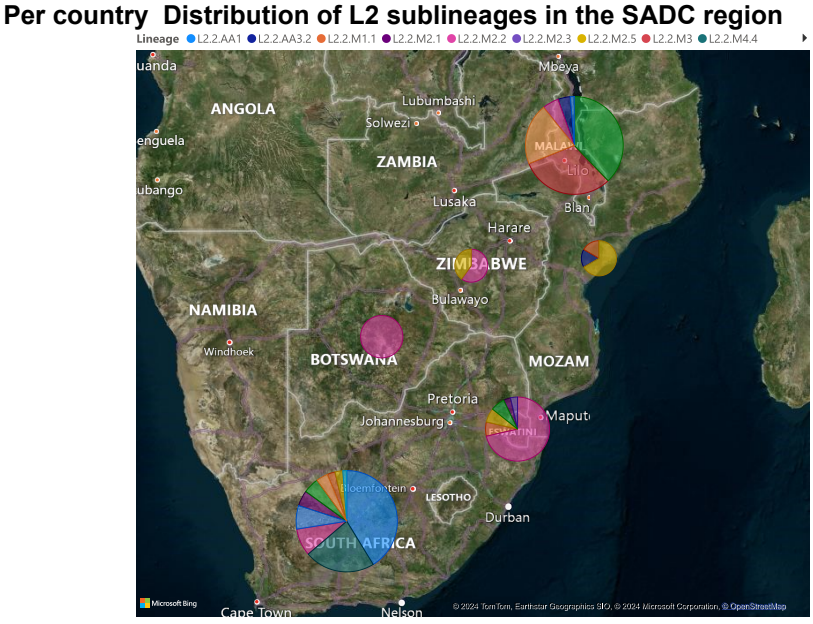


Figure 25: Distribution of sublineages of Lineage 2 Mtb in the SADC region. Sublineage Lineage 2.2.M2.2 was the most prevalent (Top bar plot). South Africa had the highest diversity of Lineage 2 in the region (bottom map).

© University of Pretoria

Strain SRR1167282 from South Africa could not be classified to sublineage level so we sought to identify three RDs which are unique to Lineage 2. Figure 26 shows the alignment views for three regions encompassing these three RD regions. RD105 and RD207 were deleted suggesting that the sample came from a patient who was infected by both Lineage 2 and Lineage 4 strains.

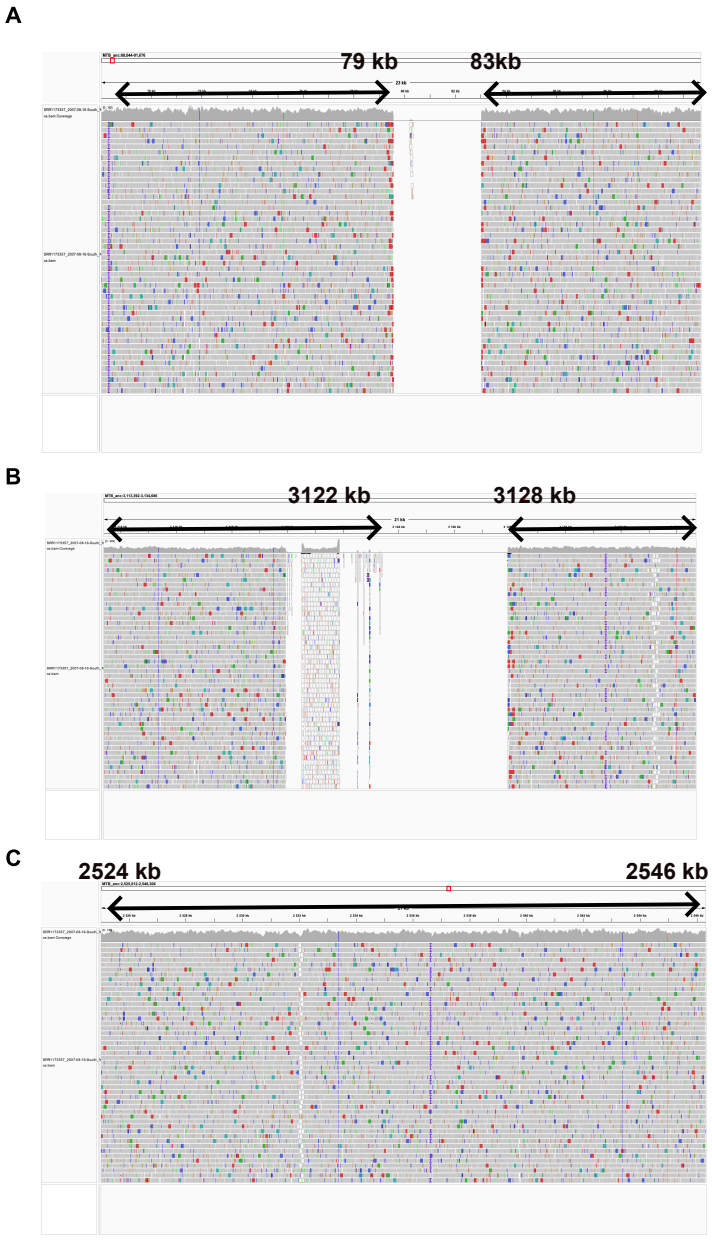


Figure 26: Alignment views showing the three different RD regions in strain SRR1167282. RD105 (A) and RD207(B) were present, while RD181(C) was absent.

© University of Pretoria

4.4.1.2 Lineage 4

The Coll *et al.* [177] classification scheme had a better resolution for Lineage 4 strains compared to Lineage 2. We identified multiple sublineages of Lineage 4 in the SADC region, including Lineage 4.1 to Lineage 4.9, except for Lineage 4.5, as shown in Figure 27A. Sublineage 4.3 was the most prevalent, accounting for 51% of SADC Lineage 4 isolates, making it the dominant sublineage in the region. The remaining sublineages were distributed as follows: Lineage 4.1 (14.64%), Lineage 4.4 (14.34%), Lineage 4.8 (8.67%), Lineage 4.6 (3.59%), Lineage 4.9 (3.19%), Lineage 4.2 (2.6%), and Lineage 4.7 (0.3%). Two samples could not be classified to the sublineage level. Figure 27B shows the sublineage distribution of Lineage 4 Mtb in the SADC countries included in this study. Tanzania exhibited the highest Lineage 4 diversity, with eight of the nine sublineages detected, excluding Lineage 4.7. Malawi showed the second highest diversity, with all identified sublineages present except Lineage 4.2, Lineage 4.7, and Lineage 4.9. Notably, sublineage 4.3 was the most dominant in all studied countries except eSwatini, where Lineage 4.4 prevailed. In South Africa, Lineage 4.4 was the second most common sublineage. In Botswana and Zimbabwe, the proportions of Lineage 4.3 and Lineage 4.1 were nearly equal; however, small sample sizes from these countries may affect the accuracy of these estimates.

© University of Pretoria

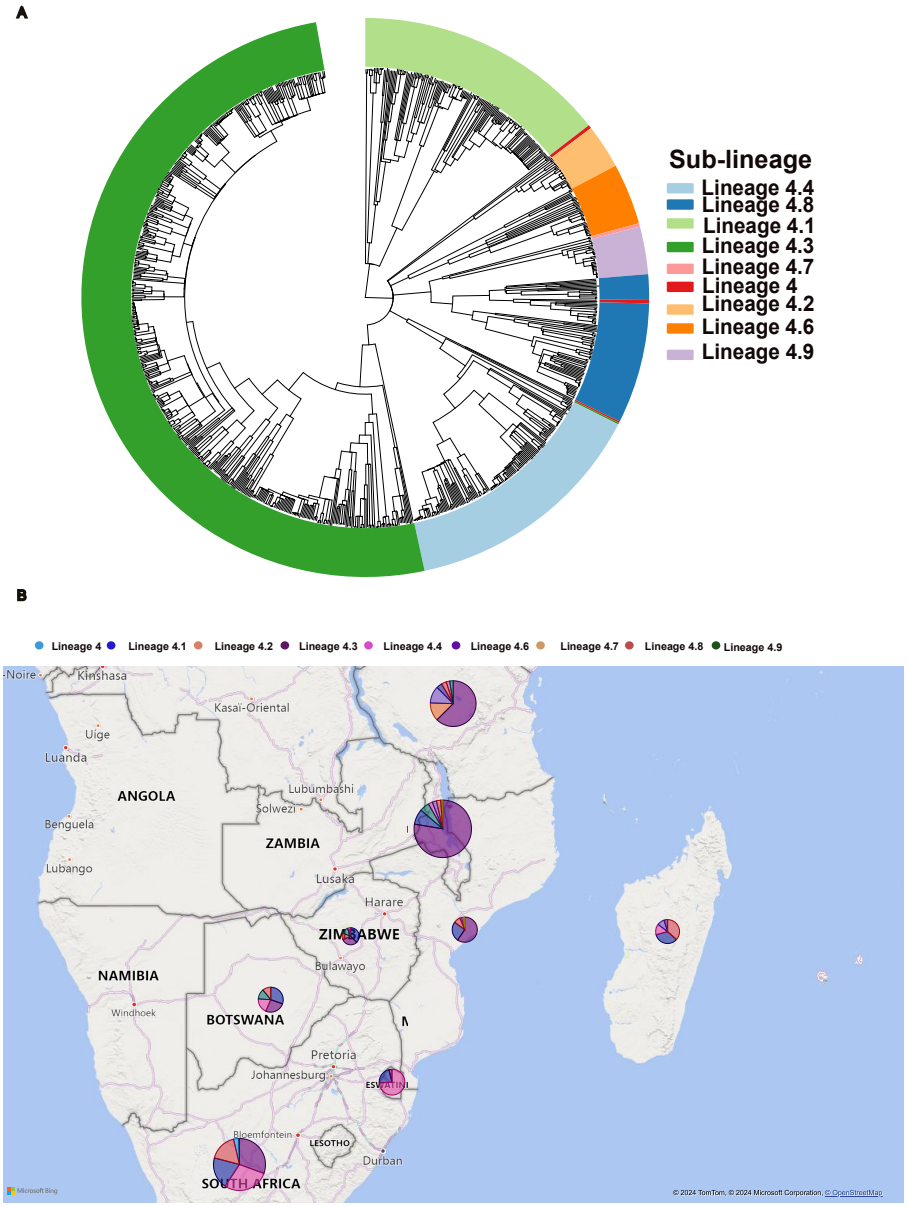


Figure 27: Genetic diversity of Lineage 4 Mtb in the Southern Africa region. (A) Maximum likelihood tree inferred from 990 Lineage 4 isolates sourced from 8 countries in the SADC region of Africa the outer ring is coloured according to sublineage. (B) map showing the sublineage distribution of Lineage 4 samples in each country included in this analysis. Lineage 4.3 was predominant across the region.

© University of Pretoria

4.4.2 Phylogeography

4.4.2.1 Lineage 2

We employed two different approaches (Augur and Mascot) to phylogeographic analysis. Our analyses show the emergence of Lineage 2 strains in South Africa and Mozambique at the beginning of the 17th century, with South Africa being a central hub for the dispersion of Lineage 2 strains to different SADC countries. Interestingly, our phylogeographic inference using Augur showed direct introductions of Lineage 2 from South Africa to all the SADC countries in this study except for Zimbabwe. The two independent introductions into Zimbabwe came through Mozambique and Botswana. However, caution must be exercised when interpreting this result due to the limited sample sizes from the three countries. Next, we sort to estimate the most likely region of origin for Lineage 2 strains isolated in the SADC region. As we expected, estimates from Augur predicted Asia as the place of origin for the Most Recent Common ancestor (MRCA) of Lineage 2 strains isolated in the SADC region with multiple independent introduction events. The result was also confirmed by Mascot which inferred East Asia as the most likely region of origin (posterior probability 86%) as shown in Figure 28B. The time to the most recent ancestor (TMRCA) was estimated to be the year 1694 (HPD 1634 - 1801). Our results (Table 7) also suggest that there were multiple independent introductions of Lineage 2 into the SADC region from Asia with East Asia contributing to the highest number of introductions (n=13 HPD 95% 9 - 18) and Southeast Asia contributing (n=3 HPD 95% 2 - 7) introductions. A study by Rutaihwa *et al.*[192] using a different approach of stochastic mapping identified 13 introduction events of Lineage 2-Beijing into Africa from Asia. Introductions from Europe to Southern Africa were estimated with a median of 0.76 (95% HPD 0 – 3), indicating that Europe did not play any role in the introduction of Lineage 2 into the SADC region.

Table 7 shows the estimated backwards migration rates for the four regions that were analysed using Mascot with all parameters effective sample sizes (ESS) values significantly above 200, hence showing high reliability with our estimates. It is important to note that the estimates are inferred as backwards in time migration rates; therefore, a backward migration rate for Southern Africa to East Asia will be an estimate of the likelihood of East Asia being

© University of Pretoria

the origin of the ancestral strains in Southern Africa. The estimated backward migration rates from Southern Africa to East Asia were the highest with a median of 2.16E-03 (HPD 4.94E-5 - 1.90E-4), this rate was 2.5 times larger than the estimates for Southern Africa to Southeast Asia with a median of 8.59E-4 (HPD 8.85 E-4 – 3.67E-3). Migration from Southern Africa to Europe showed the lowest rates with a median of 2.77E-3 (HPD 1.03-6 – 1.13E-3). The estimates in the opposite direction were smaller with East Asia to Southern Africa estimated at 2.17E-4 (HPD 7.81E-7 – 1.09E-3), Southeast Asia to Southern Africa and from Europe to Southern Africa had rates of 3.75E-4 (HPD 1.41 E-8 – 1.59 E-3) and 1.11E-3 (HPD 2.81E-4 – 3.15E-3), respectively. The consolidated Lineage 2 phylogeographical analysis results are summarized in Figure 29.

© University of Pretoria

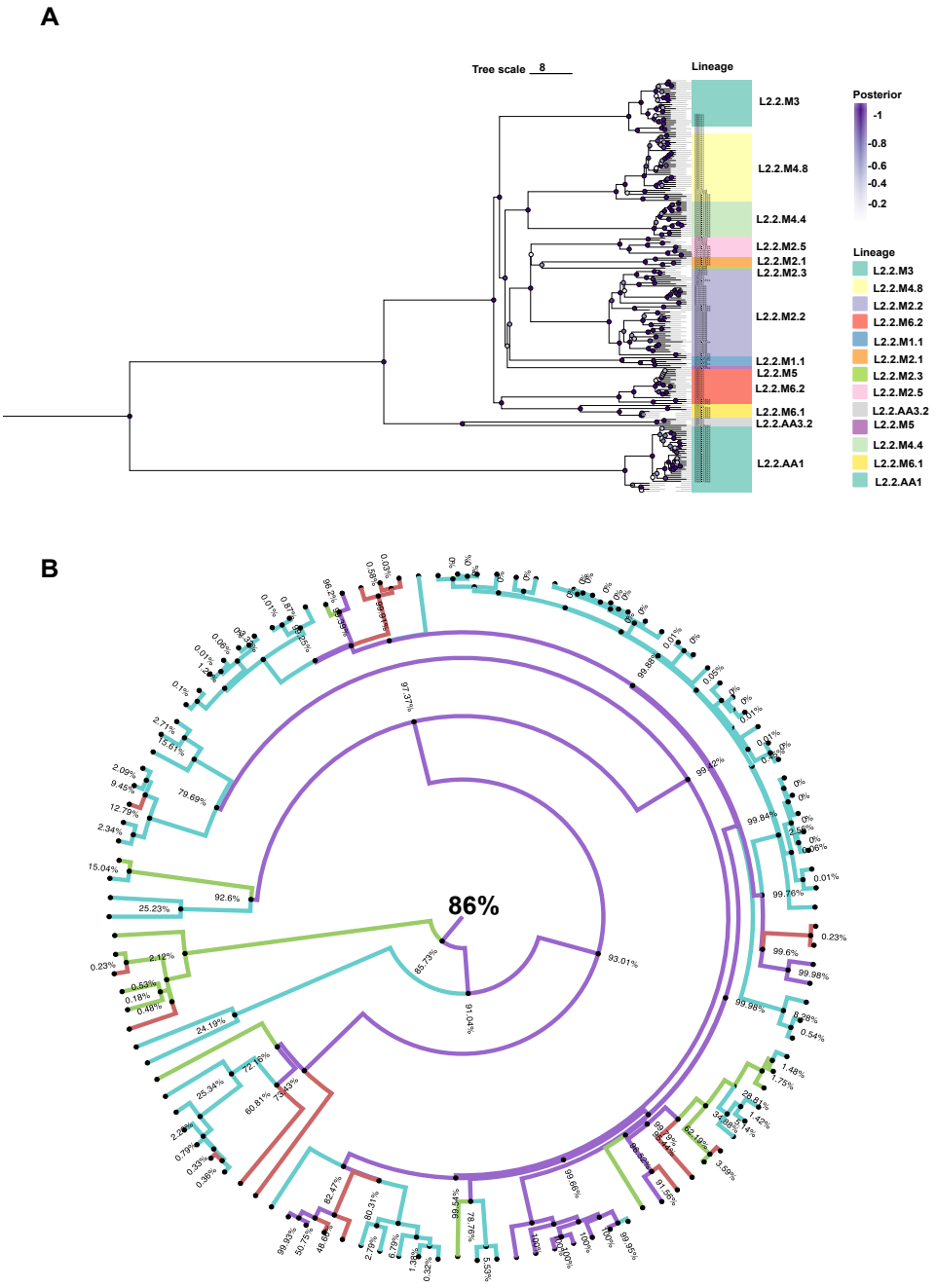


Figure 28: Bayesian phylogenetics and phylogeography of Lineage 2 strains isolated in the SADC region. (A) Maximum clade credibility tree inferred from Lineage 2 strains isolated in the SADC region. (B) A phylogenetic tree inferred from the Mascot package in Beast 2 places the root in East Asia (purple) with an 86% probability. Different regions are colour-coded as follows: Europe (red), Southern Africa (turquoise), and East Asia (green).

© University of Pretoria

Table 7: Mascot analysis results for SADC Lineage 2. Effective sample size (ESS) values for all parameters showed that the analysis produced a large number of independent samples from the posterior (high reliability)

Effective population size (Ne)	Median	95% HPD	ESS
East Asia	89.2	46.20 - 158.40	733
South-East Asia	222.64	92.96 - 418.79	1135
Southern Africa	591.58	363.67 - 898.48	401
Europe	190.57	71.82 - 353.37	1150
Migration rates (backward migration Constant)			
East Asia to Europe	6.64E-04	1.70E-5 -1.76E-3	985.8
East Asia to South-East Asia	2.69E-04	7.81E-7 - 1.09E-3	569.8
East Asia to Southern Africa	2.17E-04	3.36E-7 - 9.28E-3	707.7
Europe to East Asia	1.37E-03	2.03E-4 - 3.09E-3	967.9
Europe to South-East Asia	1.43E-03	1.60E-4 - 2.60E-3	1008.2
Europe to Southern Africa	1.11E-03	2.81E-4 - 3.15E-3	1151.7
South-East Asia to Europe	4.26E-04	6.79E-7 - 1.65E-3	599.1
South-East Asia to Southern Africa	3.75E-04	1.41E-8 - 1.59E-3	566.5
South-East Asia to East Asia	1.43E-03	2.11E-4 - 3.15E-3	1010.9
Southern Africa to Europe	2.77E-04	1.03E-6 - 1.13E-3	644
Southern Africa to East Asia	2.16E-03	4.94E-5 - 1.90E-3	991.7
Southern Africa to South-East Asia	8.59E-04	8.85E-4 - 3.67E-3	1178.2
Migration events			
East Asia to Europe	6	[1 - 9]	1392.6
East Asia to South-East Asia	7	[2 - 10]	923.9
East Asia to Southern Africa	13	[9 - 18]	1349.8
Europe to East Asia	1	[0 - 3]	1559.6
Europe to South-East Asia	1	[0 - 4]	925.6
Europe to Southern Africa	1	[0 - 3]	1088
South-East Asia to Europe	4	[2 - 7]	1498
South-East Asia to Southern Africa	3	[1 - 9]	1531
South-East Asia to East Asia	0	[0 - 6]	1316
Southern Africa to Europe	3	[1 - 5]	1482.5
Southern Africa to East Asia	0	[0 - 1]	1841
Southern Africa to South-East Asia	1	[0 - 4]	887.1
nr Migration Events	40	[36 – 46]	1478

© University of Pretoria

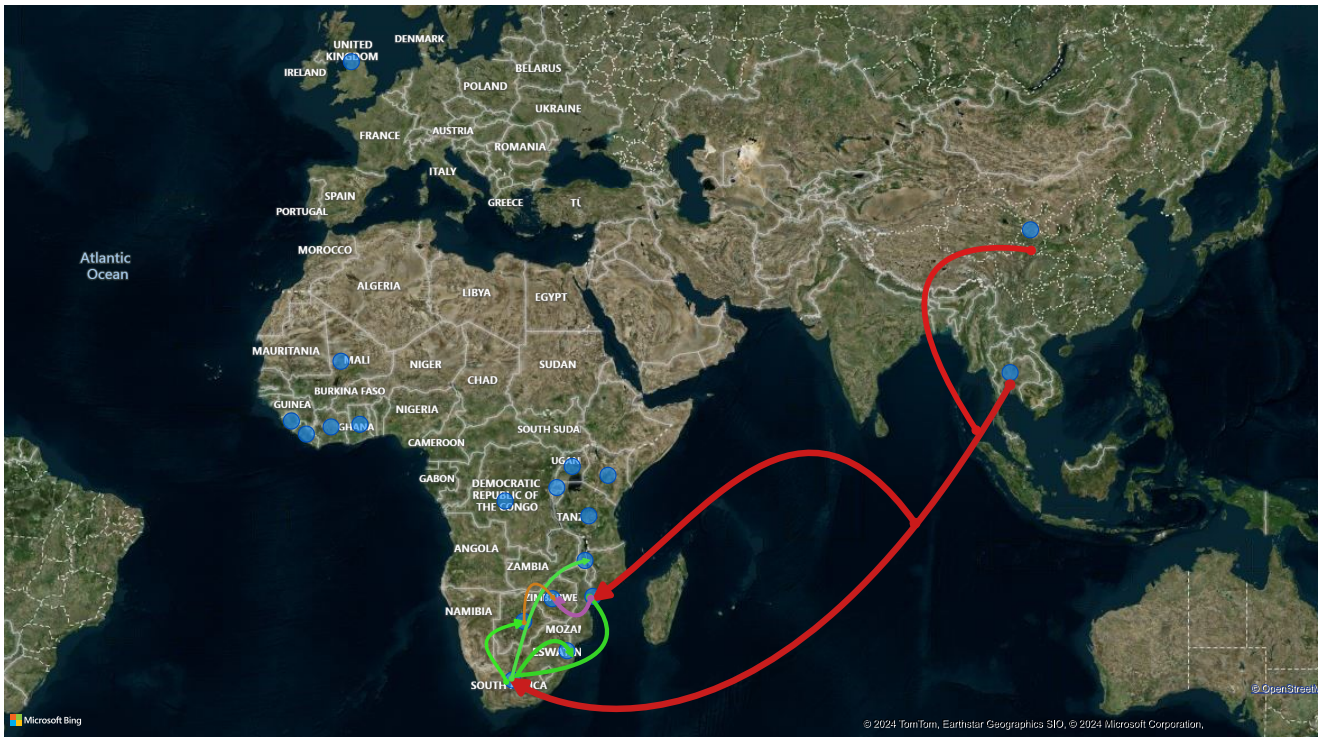


Figure 29: A simplified graphical representation of the possible introduction of Lineage 2 in the SADC region. The red arrows show Southeast Asia as the most likely origin of the MRCA of Lineage 2 strains in the region. The green arrows show the spread of Lineage 2 from South Africa upwards. The orange and purple lines represent the two introductions of Lineage 2 into Zimbabwe one from Botswana (Orange) and the second one from Mozambique (Purple).

4.4.2.2 Lineage 4

Using a time-scaled phylogenetic tree inferred with TreeTime, we analysed the geographical transitions of Lineage 4 Mtb across SADC, focusing on eight countries as shown in Table 8 below. The equilibrium frequencies (πi) indicated the long-term probability of isolates being assigned to each region. Zimbabwe (πH= 0.1483) and South Africa (πF=0.146) had the highest probabilities, while Tanzania (πH= 0.08695) had the lowest. Symmetrized migration rates (W_{ij}) showed strong connectivity between Malawi and Tanzania (W_{DG}=8.6611), moderate movement between Malawi and Mozambique (W_{DE}= 2.897) and limited movement between Tanzania and Botswana (W_{GA} =0.4728). Directional migration rates

© University of Pretoria

(Q_{ij}) revealed significant movement from Malawi to Tanzania (Q_{GD}=0.7751), Movement from South Africa to Botswana was moderate (Q_{FA}=0.1455) with low rates of movement between other regions. Our results highlight Malawi and Tanzania as historically highly connected regions. Zimbabwe and South Africa were identified as key hubs of Lineage 4 Mtb diversity in the region. Further investigation into the regional dynamics may clarify the role of these countries in the evolution of Lineage 4 in the SADC region.

Table 8: TreeTime inferred phylogeographic parameters for Lineage 4 Mtb across Southern African regions. The equilibrium frequencies (πi) represent the long-term likelihood of isolates being assigned to specific regions. Symmetrized rates (W_{ij}) show bidirectional movement between regions, while directional rates (Qij) indicate the rates of migration from one region to another. Regions correspond to the following countries: A = Botswana, B = Eswatini, C = Madagascar, D = Malawi, E = Mozambique, F = South Africa, G = Tanzania, H = Zimbabwe.

	A	B	C	D	E	F	G	H
Equilibrium frequencies (pi_i)	0.1025	0.1663	0.1101	0.0958	0.1415	0.146	0.0895	0.1483
Symmetrized rates from j->i (W_ij)								
A	0	1.4586	0.6178	4.3553	1.003	0.9966	0.4728	0.7381
B	1.4586	0	0.4885	0.7377	0.6413	1.0066	0.4112	0.5166
C	0.6178	0.4885	0	2.8725	0.9761	0.6994	0.8252	0.4216
D	4.3553	0.7377	2.8725	0	2.8978	3.0263	8.6611	0.9717
E	1.003	0.6413	0.9761	2.8978	0	0.7915	0.5805	0.5134
F	0.9966	1.0066	0.6994	3.0263	0.7915	0	0.4868	0.6275
G	0.4728	0.4112	0.8252	8.6611	0.5805	0.4868	0	0.4132
H	0.7381	0.5166	0.4216	0.9717	0.5134	0.6275	0.4132	0
Actual rates from j->i (Q_ij)								

© University of Pretoria

A	0	0.1495	0.0633	0.4463	0.1028	0.1021	0.0485	0.0756
B	0.2425	0	0.0812	0.1227	0.1066	0.1674	0.0684	0.0859
C	0.0684	0.0538	0	0.3164	0.1075	0.0775	0.0909	0.0464
D	0.4171	0.0707	0.2751	0	0.2775	0.2899	0.8295	0.0931
E	0.1419	0.0907	0.1381	0.410	0	0.112	0.0821	0.0727
F	0.1455	0.147	0.1021	0.4419	0.1156	0	0.0711	0.0916
G	0.0423	0.0368	0.0738	0.7751	0.0519	0.0436	0	0.037
H	0.1095	0.0766	0.0625	0.1441	0.0761	0.0931	0.0613	0

Our results as visualized by Auspice (Supplementary file 4.2) show multiple introductions of Lineage 4 into SADC predominantly from Central Europe. The earliest of these introductions was inferred to have occurred in the mid-16th century when lineage 4.3.4 strains from Europe were introduced into the Great Lakes part of the SADC region. Historically, this could be attributed to the early contact with European traders who actively established trading posts and forts in Africa in search of ivory and gold. Historical records show that in the mid-1400s, European traders, particularly from Portugal were active in Sub-Saharan Africa [377]. Another wave of lineage 4 introductions into the SADC region followed in the 18th century with South Africa dominating as a key point for seeding these strains up north into other SADC countries and beyond. This timeline coincides with the migration of European settlers (mainly from Britain and Holland into South Africa). By the 18th and 19th centuries, TB had become an epidemic in Europe and presumably, some of the immigrants carried Mtb infection [378-380]. A closer look into the phylogeographic dynamics of strains isolated from Malawi and South Africa showed an early introduction of lineage 4.34 from Malawi into South Africa. Later, there were multiple introductions of different sub-lineages of Lineage 4 from South Africa to Malawi and other parts of the SADC region in the 19th and 20th centuries.

© University of Pretoria

Our MASCOT analysis using Bayesian methods focusing on the dominant Lineage 4.3 strongly supported the European origin for Lineage 4.3 with a posterior probability of 90 %, as shown in Figure 30A below. Table 9 below also shows that the backward migration rates from Southern Africa to Europe were estimated at a median of 9.77E-3 (HPD 1.03E-6 – 1.13E-3), this was the highest rate of the regions compared using this analysis and three times higher than the rates in the opposite direction (Europe to Southern Africa) with an estimated median of 3.35E-3 (HPD 2.81E-4 - 3.15E-3). Backward migration rates from Southern Africa to the Americas had an estimated median of 1.21E-3 (HPD 3.36E-7 - 9.28E-3). There was an 8-fold difference in estimated backwards migration rates when comparing Southern Africa to Europe and Southern Africa to the Americas, highlighting Europe as a major source of Lineage 4.3 in the SADC region. Table 9 also shows that the estimated migration events were highest in the Europe to Southern Africa direction with a median of 20 (HPD 7 – 34). The estimated events in the opposite direction were estimated at a median of 11 (HPD 6 – 15). Southern Africa to the Americas had the lowest estimated migration events with an estimated median of 3 (HPD 1 - 5). The consolidated Lineage 4 phylogeographical analysis results are summarized in Figure 30B.

© University of Pretoria

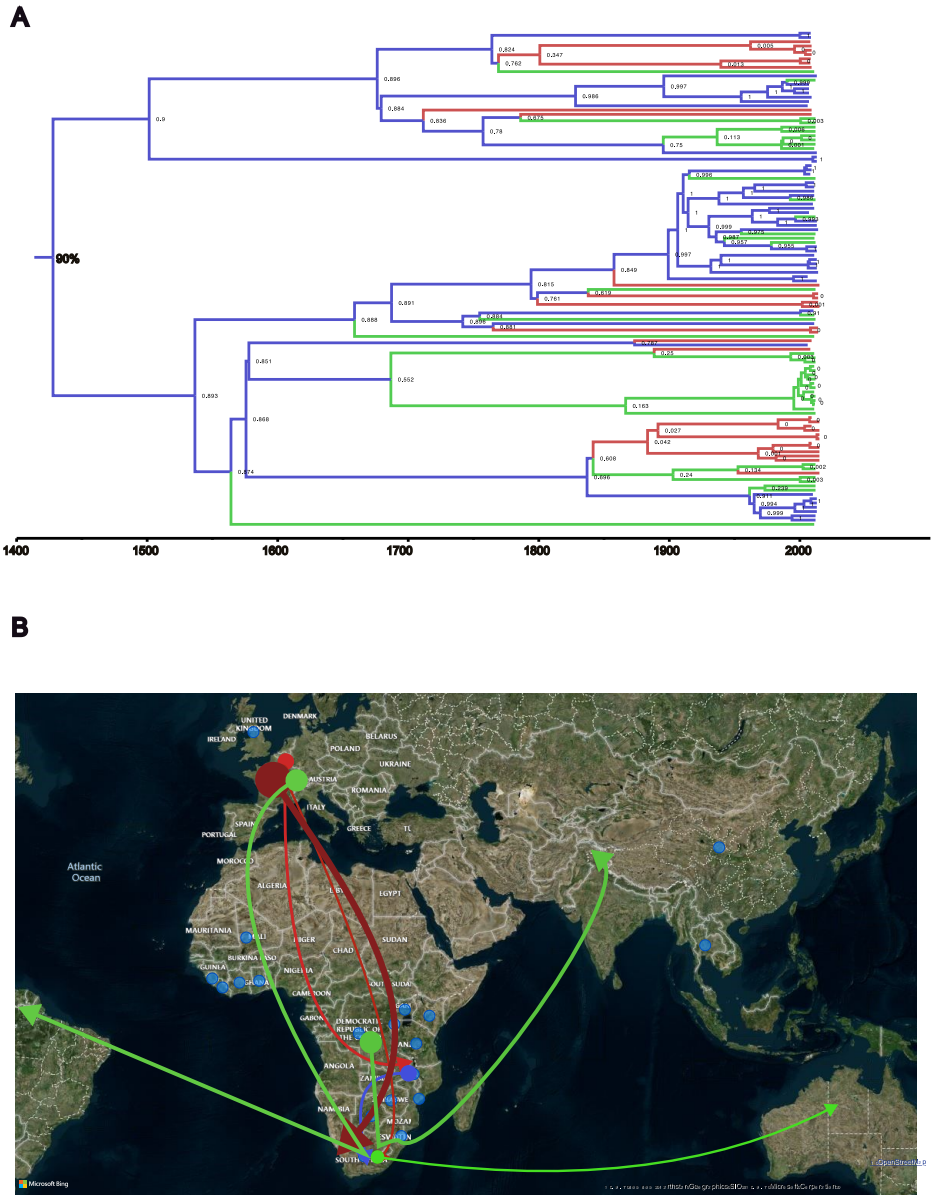


Figure 30: Spatial dynamics of Lineage 4.3 strains sampled in the Southern Africa region. (A) A phylogenetic tree inferred from the Mascot package in Beast 2 placing the root of Lineage 4.3 isolates sampled in the SADC region in Europe (blue) with a 90% posterior probability. Different regions are colour-coded as follows: Europe (blue), Southern Africa (green), and the Americas (red). (B)A simplified graphical representation of the possible introduction of Lineage 4 in the SADC region. The red arrows show Europe as the most likely origin of the MRCA of Lineage 4 strains in the region. The green arrows show the spread of Lineage 4 from South Africa to different parts of the world in the last century).

© University of Pretoria

Table 9: Mascot analysis results for SADC Lineage 4.3. Effective sample size (ESS) values for all parameters showed that the analysis produced many independent samples from the posterior indicating higher reliability of our analysis.

Effective population size (Ne)	Median	95% HPD	ESS
America	598	46.20 - 158.40	468
Southern Africa	713	363.67 - 898.48	414
Europe	694	71.82 - 353.37	656
Migration rates (backward migration Constant)			
America to Europe	9.60E-04	4.8E-5 -2.48E-3	985.8
America to Southern Africa	1.21E-03	3.36E-7 - 9.28E-3	707.7
Europe to America	4.11E-04	1.43E-5 - 3.09E-3	967.9
Europe to Southern Africa	3.35E-03	2.81E-4 - 3.15E-3	1151.7
Southern Africa to Europe	9.77E-03	1.03E-6 - 1.13E-3	644
Migration events			
America to Europe	8.00	[1 - 9]	995
America to Southern Africa	6	[4 - 18]	1349.8
Europe to America	9	[5 - 12]	1076
Europe to Southern Africa	20	[7 - 34]	1264
Southern Africa to Europe	11	[6 - 15]	1092
Southern Africa to America	3	[1 - 5]	1198
nr Migration Events	52	39 - 64	1396

© University of Pretoria

4.5 DISCUSSION

Here, we characterise the genetic diversity of Mtb Lineage 2 and Lineage 4 in the SADC region based on genomic and phylogenetic analyses of 1313 (n=413 for Lineage 2 and n=990 for Lineage 4) clinical isolates representative of the SADC region and the endemic regions in East Asia and Southeast Asia. We also included strains from Europe in this study as some studies have attributed the widespread TB disease in Africa to European colonization [34, 169]. These results show that the genetic diversity of Lineage 2 strains in the Southern African region has been previously underestimated. Importantly, for Lineage 2, we adopted a recently proposed classification scheme that adequately captures the hierarchical relationships between isolates while preserving valuable insights from other widely used classification schemes. Our findings on Lineage 2 complement and extend previous work by Rutaihwa *et al.*[192] , who used 781 Lineage 2 clinical strains from 14 geographical regions around the globe to elucidate the origins and onward spread of Lineage 2 in Africa. We placed the origins of global Lineage 2 in Asia and inferred multiple introductions of Lineage 2 strains into the SADC region. While European contact has been hypothesised to have introduced and exacerbated the spread of TB in Africa, our results exclude Europe as a source of origin of Lineage 2 TB strains in the SADC region. We also inferred that the earlier introduction in the SADC region came through South Africa and Mozambique, a finding consistent with the role of both countries in the sea routes servicing the region. Our inferred TMRCA in the late 19th century also coincides with the importation of Chinese labourers to work in the Transvaal mines of South Africa [381]. In this study, we also identified South Africa as a hub in the dissemination of Lineage 2 and Lineage 4 strains northwards in the continent, this supports previous studies' findings [368].

Our Lineage 4 phylogeographic analysis is concordant with historical records of migration within the region. For example, the early introduction of Lineage 4.3 in the Greater Lakes region can be attributed to the well-documented trade between local populations and Portuguese traders in the 15th century. The phylogeographic dynamics between South Africa and Malawi also show a pattern that is consistent with the migratory patterns of the Southern Africa region. We estimated an early introduction of Lineage 4.3.4 from Malawi into South Africa in the early 17th century. Two centuries later, there were multiple introductions of different Lineage 4 strains from South Africa into Malawi. Cross-border migration for

© University of Pretoria

employment within the SADC dates to the mid-19th century [382]. The multiple introductions of Lineage 4 from South Africa to Malawi can be linked to the recruitment drive of "Tropical Africans" to work in the Witwatersrand gold mines which were often associated with unsanitary conditions for workers, thus contributing to the spread of TB [383]. Two centuries later, the mining sector still suffers a disproportionately high TB burden in the SADC region such that SADC countries have adopted a regional strategy to combat TB in the mining industry as a way to achieve TB eradication goals [384].

Our in-depth characterisation of Lineage 2 and Lineage 4 diversity in a region that has been severely affected by both HIV/Aids and TB provides important insights into regional and global TB management efforts. A recent study by Zhu *et al.*[385] implicated a significant contribution of a few recently emerging strains to the global success of *M. tuberculosis* Modern Beijing family. In their study, they identified six phylogenetic clades of modern Beijing Mtb, namely Lineage 2.3.1 to Lineage 2.3.6 and concluded that Lineage 2.3.4 to Lineage 2.3.6 strains were associated with more cross-border transmission and virulence. These strains fall under Lineage 2.2.M2, Lineage 2.2.M5 and Lineage 2.2.M4 in our analysis, thus our study highlights the importance of future work in further exploring the characteristics of these sublineages regionally and globally. As the field of TB research has established the heterogenicity of Lineage 2, there is a growing need for harnessing the full potential of the accumulating global MTB WGS data collection to fully understand the genetic diversity of Mtb, especially strains belonging to Lineage 2.

Our characterisation of the genetic diversity of Lineage 2 Mtb in the SADC region identified 13 Lineage 2 sublineages in SADC out of the 16 that were identified by combining the dataset with a global reference. Our study represents one of the first few studies that aim to fully understand the evolution of Mtb on the African continent. By using different phylogenetic approaches, we managed to strike a balance between computational efficiency and accuracy. We acknowledge a major limitation in our study in that most of our Lineage 2 and Lineage 4 isolates were sampled in South Africa and Malawi while there were limited samples from Mozambique, Tanzania, Zimbabwe, Eswatini and Botswana and this made it challenging to adequately explore the full genetic diversity of these lineages in the region. The sampling window (time differences) for most of our samples was relatively short which also makes it difficult to accurately estimate the evolutionary timeframes in our analysis.

© University of Pretoria

We did not find any sufficient Mtb WGS data for the remaining SADC countries. Hopefully, WGS will soon be standard in the clinical management of TB in the region. The accumulation of genomes from underrepresented countries has the potential to provide better insights into phylogenetic structure resolution. The SADC region has been traditionally a hive of economic activity on the continent, and this has been associated with high rates of migration for centuries and this has always had an impact on TB on the continent. Studying the genetic diversity of Mtb in this region of Africa offers unparalleled opportunities in the fight against TB; hence there is a need for a more concerted regional approach towards TB research within the region.

© University of Pretoria

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

CHAPTER 5 THE EPISTATIC LANDSCAPE OF ANTIBIOTIC RESISTANCE OF DIFFERENT LINEAGES OF MTB

The work in this chapter has been published as: Muzondiwa, D., H. Hlanze, and O.N. Reva, *The epistatic landscape of antibiotic resistance of different clades of Mycobacterium tuberculosis.* Antibiotics, 2021. **10**(7): p. 857.

…

© University of Pretoria

5.1 ABSTRACT

Drug resistance (DR) remains a global challenge in tuberculosis (TB) control. To develop molecular-based diagnostic methods to replace the traditional culture-based diagnostics, there is a need for a thorough understanding of the processes that govern TB drug resistance. The use of whole genome sequencing coupled with statistical and computational methods has shown great potential in unravelling the complexity of the evolution of drug-resistant TB (DR-TB). In this study, we took an innovative approach that sought to determine non-random associations between polymorphic sites in *Mycobacterium tuberculosis* (Mtb) genomes. Attributable risk statistics were applied to identify epistatic determinants of DR in different clades of Mtb and possible evolutionary pathways of DR development. It was found that different lineages of Mtb exploited different evolutionary trajectories towards multidrug resistance and compensatory evolution to reduce the DR-associated fitness cost. Epistasis of DR acquisition is a new area of research that will aid in better understanding evolutionary biological processes and allow for the prediction of upcoming multidrug-resistant pathogens before a new outbreak strikes humanity.

5.2 INTRODUCTION

Despite being a disease of antiquity, TB caused by *Mycobacterium tuberculosis* (Mtb) remains one of the leading killers globally, causing twice as many fatalities as HIV/AIDS. According to the latest report by the World Health Organization WHO, TB reclaimed its spot as the global leading cause of death from a single infectious agent following three years in which coronavirus disease (COVID-19) had taken the top spot [386]. More than 10 million new cases have been reported annually and the figures have been rising since 2021. Drug susceptible TB is treatable however, drug-resistant TB (DR-TB) has been identified as a cause for concern in global efforts to eradicate TB and has been listed as one of the 24 pathogens on the WHO's Bacterial Priority Pathogens List [386, 387]. In 2023, an estimated 400,000 people (95% UI 360 000 - 440 000) developed MDR/RR-TB, with an estimated 175

© University of Pretoria

923 people diagnosed and treated for multidrug-resistant or rifampicin-resistant TB (MDR/RR-TB). MDR-TB is TB caused by strains that have become resistant to both rifampicin (RIF) and isoniazid, two of the most potent first-line anti-TB drugs. RR-TB refers to TB that is caused by strains that have developed resistance to rifampicin [11]. Extensive drug-resistant TB XDR-TB can also develop when MDR/RR-TB strains develop additional resistance to rifampicin, any fluroquinolone(a class of second-line anti-TB drug) plus at least one of either bedaquiline or linezolid [386]. Pre-XDR-TB is a term that is also used to describe strains resistant to rifampicin and any other fluoroquinolone.

Treatment for patients with MDR-TB and XDR-TB often involves prolonged and costly antibiotic regimens and poor outcomes that result in high rates of mortality and treatment failure [388, 389]. Countries that have a high burden of total cases of TB, including DR-TB, are often resource-limited. Consequently, the inadequate administration of treatment risks the amplification of resistance to further drugs and may prolong opportunities for infection transmission [390]. Timely diagnosis and correct treatment are of the uttermost importance in turning the tide against TB. DR-TB can be acquired in the course of treatment through the acquisition of chromosomal mutations in anti-TB drug target genes (secondary resistance) or can be transmitted from one infected individual to another (primary resistance) [389, 391]. The development of acquired resistance is common as a result of single nucleotide polymorphisms (SNPs), insertions and deletions [388, 390]. Apart from chromosomal mutations, the intrinsic resistance due to the bacterium's lipid-rich cell wall and efflux pumps has limited the number of drugs that can treat TB [388, 392].

Rapid molecular diagnostic assays have been developed to tackle the challenge of early diagnosis and treatment. However, the technology behind these trials is often limited to a small number of loci they can interrogate, and thus, the number of drugs that can be tested is usually restricted to the first-line drugs. The genetic basis of drug resistance is still poorly understood, with recent research suggesting that it's more complex than previously thought [189]. Whole Genome Sequencing (WGS) technology has proven to be an effective tool in shedding more insights into our understanding of TB [27]. Unlike in most bacterial pathogens, horizontal gene transfer does not play a role in the acquisition of DR in Mtb. For this reason, research has mostly focused on identifying multivariate associations between

© University of Pretoria

SNPs in the Mtb genome and the resistant phenotype. The use of mathematical methods together with WGS and other related TB data has also proved to be of importance in exploring associations between variants in the Mtb genome [133, 388, 393, 394].

Focusing exclusively on DR-associated variants does unravel the full complexity of the evolutionary landscape driving drug resistance in Mtb. Studies have also suggested that the distribution patterns of DR-TB-associated mutations in Mtb genomes appear to be lineage-specific due to epistasis limitations [52, 395]. The acquisition of a DR-associated mutation is expected to impose a bacterial fitness cost. However, several studies have shown that MD-TB strains have minimal or no fitness cost, yet they show high levels of resistance [84, 88, 127, 395]. Several DR-associated mutations in the following genes: *katG*, *gyrA*, *rpoB* and *rpsL* render resistance to different antibiotics and are frequently observed in successful DR clinical isolates [395]. On the other hand, a household-based case study done by Salvatore *et al*. [396] demonstrated that isolates carrying both *katG* S315T and *rpsL* R43K mutations were less likely to be found in households with multiple cases, suggesting decreased transmission even though both mutations conferred high-level resistance. Several studies have also identified putative compensatory mutations in *rpoA*, *rpoC* and *rpoB* genes that ameliorate the fitness cost of rifampicin resistance [397, 398]. Work done by Vargas *et al*. [399] identified 35 additional putatively novel compensatory mutations in *rpoC* genes from isolates that were rifampicin resistant in Peru. Another study done by Coll *et al*. [388] identified a *pnaB2* mutation that may compensate for pyrazinamide resistance conferred by *pncA* and mutations in the *thyX–hsdS.1* promoter that compensates for para-aminosalicylic acid resistance conferred by mutations in *thyA*. The stated examples suggest that these strains may have undergone compensatory evolution modulating the fitness of the strains, which either precedes or follows drug resistance acquisition by these strains [187, 400].

Another class of stepping-stone mutations that appear to emerge before DR status has been acquired was reported in a study by Cohen *et al*. [401]. The authors found that mutations in *ubiA* emerge before ethambutol resistance conferred by mutations in *embB*. Safi *et al*. [111] ‘s *in vitro* studies showed a multistep selection of *ubiA*, *aftA*, *embB*, and *embC* mutations required to achieve the highest level of resistance to ethambutol. Another study done by Shea *et al*. [402] identified low-level resistance conferring mutations in the *rpoB* gene.

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Associations between the genetic background and DR-TB have been investigated by several groups [111, 400, 403]. For example, Li *et al*. [404] reported that substitutions at *katG*315 and *rpoB*450 loci conferring isoniazid and rifampicin resistance were more frequent in more transmissible strains of the "modern" Beijing sub-lineages compared to the "ancient" Beijing strains. Additionally, Fenner et al. [405] found that mutations in *katG* and *inhA* conferred different levels of INH resistance depending on the lineage of the mutant TB strains. In Lineage 2 mutant strains, substitutions in *katG*315 were more prevalent and conferred high-level resistance to isoniazid, whereas in the Lineage 1 mutant strains, substitutions in *inhA* conferring low-level resistance to isoniazid were more prevalent.

5.3 MATERIALS AND METHODS

5.3.1 Data sourcing

In our initial study, we obtained data from the GMTV database [295] , a public repository which provides a comprehensive catalogue of variant call format (VCF) files as well as metadata information on phenotypic drug susceptibility testing (pDST) and lineage classification for each sample. As a follow-up study, we added 2500 isolates from the "Comprehensive Resistance Prediction for Tuberculosis: an International Consortium" (CRyPTIC) [406] . The new dataset of 5001 Mtb strains was then used to calculate the mutation frequencies for the selected Mtb lineage.

5.3.2 Functional associations between Mtb mutations

We applied the approach described by van Nierkerk *et al*.[102] , which leverages Levin's attributable risk statistic (R_a-statistics) for classifying Mtb strains [407]. The method uses R_a-statistics to infer directional epistatic interactions while distinguishing them from genetic drift associations among mutations in Mtb genomes. Statistical inference of attributable risks was performed using Equation 1, while Equation 2 was employed for Fleiss' standard error

© University of Pretoria

estimation [74]. Allele combinations frequencies were represented as P_{AB}, P_{Ab}, P_{aB} and P_{ab}, with N denoting the total number of analysed Mtb strains.

$$R_{A\to a|b} = \frac{P_{AB}P_{ab} - P_{aB}P_{Ab}}{(P_{AB}+P_{aB})(P_{aB}+P_{ab})} \quad (1)$$

$$StdErr = \sqrt{\frac{P_{Ab}+R_{aA\to a|b}(P_{AB}+P_{ab})}{N\times P_{aB}}} \quad (2)$$

The likelihood and the standard error of the mutation $B\to b$ representing the reverse direction $A\to a$ were computed using equations 3 and 4, respectively.

$$R_{B\to b|a} = \frac{P_{AB}P_{ab} - P_{aB}P_{Ab}}{(P_{AB}+P_{Ab})(P_{Ab}+P_{ab})} \quad (3)$$

$$StdErr = \sqrt{\frac{P_{aB}+R_{B\to b|a}(P_{AB}+P_{ab})}{N\times P_{Ab}}} \quad (4)$$

Finally, equation 5 was used to compute the range of confidence values for the derived attributable risks estimate [74].

$$[1 - EXP(ln(1-R_a) - 1.96\times StdErr)]\ to\ [1 - EXP(ln(1-R_a) + 1.96\times StdErr)] \quad (5)$$

The co-dependence estimates were then integrated with the pDST metadata collected from the GMTV database together with annotation from ReSeqTB [408].

5.3.3 Phylogenetic drug resistance inference

To validate the results for each lineage, we adopted a phylogeny-based approach to tracking the directions of acquisition of DR-associated mutations. Based on the metadata, we created a sub-dataset of 2486 strains, enriched with MDR-TB and pre-XDR-TB isolates. For a genome to be included in this analysis, it had to be classified under one lineage (no mixed infections) and have an average coverage of >=20 and <10% missing SNPs across all isolates [409]. We generated SNP alignment files using BCFTOOLs v1.9 [302] while choosing the

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

missing parameter for missing variants. From the merged vcf file, we generated an alignment file in phylip format and used SNPsites to extract an alignment of only variable sites. We inferred a maximum likelihood phylogenetic tree using IQTREE2 v2.1.3 [293] using the following **parameters -m GTR+G+ASC -bb 1000 -nt AUTO** where **GTR** is the general time reversible model, **G** is the rate heterogeneity among sites modelled by the gamma distribution, **ASC** is ascertainment bias correction since we used only variable sites, **-bb** is the ultrafast bootstrapping option, and **nt** allows for multicore processing.

Utilizing the Nextstrain Augur software suite [314], which enables phylogenetic trait mapping and ancestral state reconstruction. We used the tree topology from IQTREE2 to fix the branch lengths by positioning the tips by year of the date of isolation using the *augur refine* command. Using TreeTime [238], the commands output a time-resolved tree in Newick format and a JSON file with the estimated dates and mutations on each node or branch. We then used the *augur ancestral* command for ancestral sequence reconstruction for each internal node on the tree. The program outputs a VCF file with the reconstructed sequence of all the internal nodes and a JSON file with all the nucleotide mutation data for each node in the tree. We then compiled a tab-delimited file of genes that harboured mutations with the highest likelihoods from our functional association analysis as well as known drug target genes. We then downloaded a GFF of Mtb gene annotations from Mycobrowser [410] to use as input for the *augur translate* command. With this command, the program translates the nucleotide mutations generated by *Augur ancestral* using the gene coordinates in the GFF file as reference points. The output of the *Augur translate* is a JSON file with identified amino acid mutations for each tip (strain) or internal node. Using custom Python scripts, we extracted the identified mutations from the generated JSON files while discarding genes with no mutations present.

Using the *augur sequence traits* command, we sought to identify mutations that were associated with drug resistance. To do this, we created a tab-delimited file in the format specified by the augur documentation. For our analysis, we populated this file with information on well-characterised drug resistance-associated mutations as well as possible compensatory mutations inferred from our functional association inference modelling. The

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

file contains the gene name, the amino acid site, and the display information for the mutation. Table 10 below shows an example of how we compiled our file.

Table 10: Structure of input file required by *augur sequence traits.*

gene	site	ALT	Display name	Feature
atpE	63	P		Bedaquiline
nuc	1673425	T	*fabG1: C-15T*	Isoniazid
rpoA	332	S	Possible compensatory	Rifampicin
drrA	262	G	Possible compensatory	Rifampicin

We used the *Augur export* program, which consolidates all the metadata information, the inferred time-resolved phylogeny and the JSON results file to generate a visualization file. A comprehensive list of all the commands ran in this analysis is provided in the appendices. Finally, we integrated our identified mutations with their co-dependency estimates (section 5.3.2) and created drug resistance functional association networks using Cytoscape v3.7.2 [411].

5.4 RESULTS

5.4.1 Lineage 1

Lineage 1 strains are generally known to be drug-susceptible except for those belonging to sub-lineage 1.2 [295]. Figure 31 shows the inferred paths connecting the DR mutations in strains of this lineage. We identified the *pyrG* gene as an integral mutation in the early development of drug resistance in our Lineage 1 network. The gene codes for the *CTP* synthetase, *pyrG* and has been proposed as a possible drug target for new antitubercular agents. We also identified a mutation at codon 40 of the *frdC* gene as a possible prerequisite

© University of Pretoria

of the acquisition of fluoroquinolone resistance-associated *gyrA* mutations, *gyrA A90V* and *gyrA D94G*. The gene *frdC* is known to affect quinol binding [412].

We also identified the *Rv3651* gene in two pathways that lead to the development of MDR-TB. Our co-dependency analysis estimates suggest that mutations in the gene develop early on after the development of low isoniazid resistance as a result of the acquisition of the *inhA.G-154A* mutation in the promoter region of the *inhA* gene. Mutations in the *Rv3651* gene were also linked to pathways that lead to the acquisition of *rpoB L430P* and *embB D354A* mutations that are known to be associated with rifampicin and ethambutol resistance, respectively. The exact function of *Rv3651* is currently unknown, however, structural studies have reviewed that the protein possesses a unique structure that differs from other sensor-effector proteins [413].

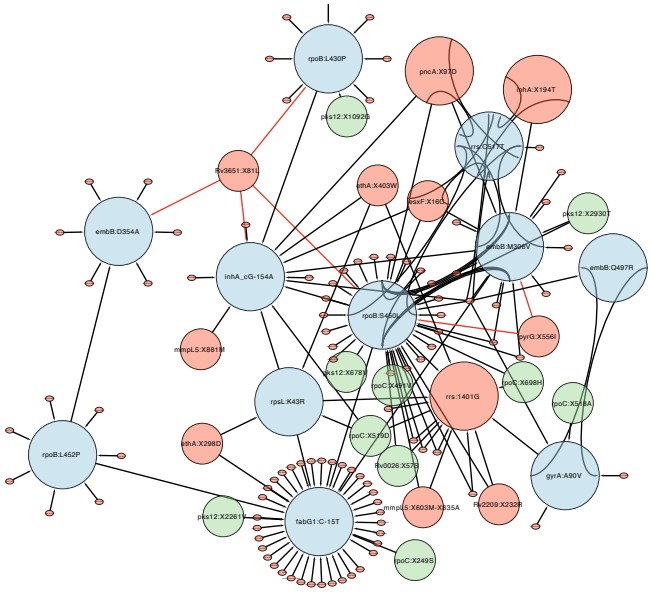


Figure 31: Network of DR-associated mutations inferred from Lineage 1. The well-characterised DR mutations are coloured blue, while the prerequisite inferred compensatory mutations are highlighted in green. Arrows illustrate relations between prerequisite mutations and the well-characterised DR mutations.

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

5.4.2 Lineage 2

Certain sublineages of Lineage 2 have been associated with hypervirulence and a propensity to develop MDR-TB. Our analysis for this lineage showed a complex network of mutations co-occurring with known antibiotic resistance mutations (Figure 32). Our analysis identified the acquisition of mutations in the *dhaA* gene as key in the progression to MDR-TB status. In the development of MDR-TB, we identified different pathways to isoniazid resistance development, mainly mutations in the *katG* gene, *inhA* and the *fabG1* promoter region. The *dhaA* gene was identified in both different pathways. Previous transcriptomic studies have shown that this gene is upregulated in bacteria living inside macrophages suggesting that it might play a role in mechanisms that allow Mtb to evade immune defenses [412].

We also identified the *ponA1* gene as a possible link to the complex networks that lead to the development of MDR-TB. In our analyses using Lineage 2 strains, we identified *ponA1* mutations in three different MDR-TB networks. The first one was the *ponA1:X213L* mutation in a pathway linked to low-level isoniazid resistance mutations *fabG1G-17T* as well as *inhA.G-154A* mutation in the *inhA* promoter region. The second one, *ponA1:X365H,* was estimated to have preceded the acquisition of *rpoB S450L* in an MDR pathway that started with the emergence of *fabG1:G-15T* mutations conferring low-level isoniazid resistance. Lastly, the third mutation, *ponA1: X402F,* was found to have been a prerequisite in the acquisition of *rpoB: V170F.* Our findings agreed with previous studies that link the *ponA1* gene to cell wall synthesis mechanisms that alter Mtb fitness during rifampicin therapy [393, 414, 415].

Another key gene identified in the complex network of drug resistance-associated mutations in this study was the *aspC* gene. We found two mutations (*aspC:X298S* and *aspC: X3S*) arose before the acquisition of *rpoB S450L* and *rpoB D435V*. The gene is known to mediate the functioning of the transcriptional regulator *Whib7,* which plays a role in regulating intrinsic multidrug resistance [416]. In Lineage 2, we also identified the *lysX* gene as an integral part of the trajectory to multi-drug resistance in TB.

© University of Pretoria

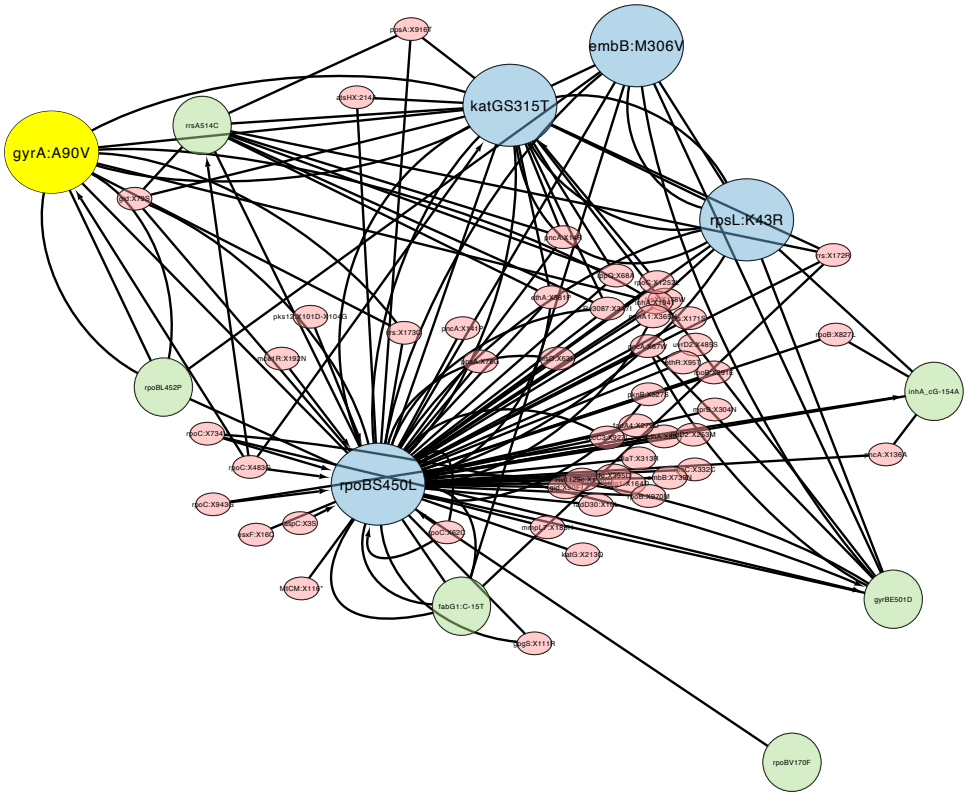


Figure 32: Network of DR-associated mutations inferred from Lineage 2. The well-characterised DR mutations are coloured blue, while the prerequisite inferred compensatory mutations are highlighted in green. Arrows illustrate relations between prerequisite mutations and the well-characterised DR mutations. *The gyrA A90V* mutation was identified as a key mutation in the trajectory towards pre-XDR-TB.

5.4.3 Lineage 4

In Lineage 4, we inferred several mutations that were prerequisites for MDR-TB emergence in different resistance trajectories (Figure 33). In Lineage 4.3, we identified one interesting possible network from drug susceptibility leading to the pre-XDR-TB phenotype. In this pathway, the *fabG1:X-15C* mutation emerges first together with several mutations such as *drrA: X262*. The *fabG1:X-15C* mutation is associated with low levels of isoniazid resistance [316], while the *drrA* mutations have been linked to resistance against rifampicin. Several

© University of Pretoria

mutations in the *rpoB* gene then emerge, notably the *L452P* mutation, which is supported by the *D435V* mutation. Other mutations outside the RRDR of the *rpoB* gene, such as *X286V*, *X399A*, *X827L* and *I1106T,* also emerge together with other previously characterised compensatory mutations such as *rpo*C (*G332C*, *F452C*, *F452S, D485Y*, *V883G*, *V483A*, *N698S*, *N698H* and *Q1125H*) [417]. We hypothesize that the acquisition of these mutations confers low-level to moderate-level rifampicin resistance. As more mutations accumulate, *katG S315T* and *rpoB S450L* emerge leading to full-blown MDR-TB and eventually pre-XDR-TB after the acquisition of *gyrA A90V*. In Lineage 4, we also identified other putative *rpoC* compensatory mutations such as *X1017Q*, *X746V*, *X734V, X62C*, and *X861T* as possible prerequisites for rifampicin resistance.

We also inferred mutations in the *Rv0147*, *Rv0893* and *Rv3087* genes as possible prerequisites for rifampicin resistance. The *Rv1047* gene has been suggested as a possible drug target for MDR-TB treatment, although no differences were reported when RNA expression profiles of the protein were compared between drug-susceptible and MDR-TB isolates [418]. *Rv0893* was confirmed to be a sam-dependent MTase, however, very few studies have characterised the enzyme [419].

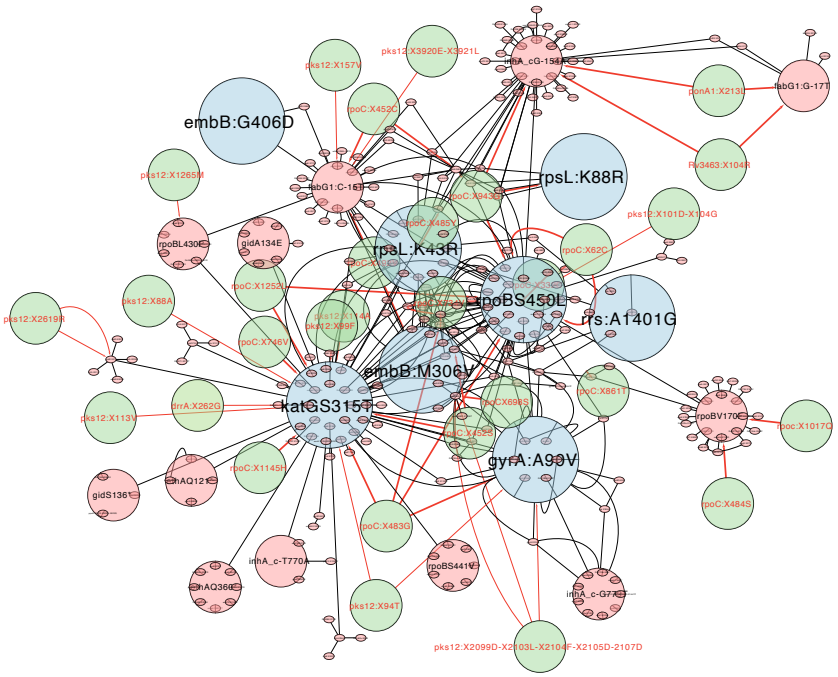


Figure 33: Network of DR-associated mutations inferred from Lineage 4. The well-

© University of Pretoria

characterised DR mutations are coloured blue, while the prerequisite inferred compensatory mutations are highlighted in green. Arrows illustrate relations between prerequisite mutations and the well-characterised DR mutations.

5.5 DISCUSSION

The evolution of drug resistance (DR) in *Mycobacterium tuberculosis* (Mtb) is influenced by factors such as fitness cost, compensatory mutations, and phylogenetic background. Since some of these factors cannot be directly measured experimentally, mathematical modelling and statistical approaches are essential. These tools challenge earlier assumptions, such as the notion that all DR Mtb strains experience fitness disadvantages compared to susceptible strains. Instead, bacterial fitness is now understood to be heterogeneous.

This study combined whole-genome sequencing (WGS) and drug susceptibility testing (DST) data with statistical modelling to investigate genetic epistasis and its role in DR evolution. We identified epistatic mutations in genes associated with stress response, efflux pumps, lipid metabolism, DNA repair, and cell wall synthesis. For example, mutations in efflux pump genes may enhance tolerance to isoniazid, while loss of *katG* function leaves Mtb vulnerable to reactive oxygen species (ROS). This vulnerability, coupled with mechanisms like SOS-induced mutagenesis, accelerates the accumulation of resistance mutations. Rifampicin and aminoglycosides (e.g., kanamycin, streptomycin) directly induce this SOS response, potentially explaining the role of *rpsL* mutations in DR evolution. Moreover, mutations in lipid metabolism genes likely compensate for fitness losses associated with DR.

Using Levin's attributable risk statistic, we mapped networks of DR-associated mutations and identified evolutionary trajectories of resistance development. Isoniazid resistance often precedes rifampicin resistance, while fluoroquinolone resistance emerges last, consistent with global treatment practices. While these results align with known resistance mechanisms, biases in public genomic datasets, such as the overrepresentation of DR strains and incomplete phenotype data, pose limitations. Additionally, the evolutionary pathways

© University of Pretoria

of DR can vary by geographic region and treatment practices, further complicating the analysis.

Advances in WGS technologies, including long-read sequencing, are expected to overcome current challenges, particularly in characterizing repetitive regions of the Mtb genome. Such developments will enhance future studies on DR evolution and epistatic networks, enabling earlier prediction of potential DR outbreaks. A major limitation of our study is that epistasis and drug-resistance evolution in Mtb is highly complex, and the application of mathematical models to infer these processes may not be enough. However, these models, coupled with experimental validation, have the potential to improve our understanding of the processes that drive the evolution of drug resistance in Mtb.

© University of Pretoria

CHAPTER 6 GENERAL DISCUSSION AND CONCLUSIONS

6.1 INTRODUCTION

The devasting effects of the recent COVID-19 pandemic have had a major impact on the fight against TB as a result of disruptions to TB care services, such as access to treatment and diagnosis. The United Nations has set new global targets for 2027 in the war against TB [1, 279]. These targets include 100% rapid TB testing for newly diagnosed cases, 90% coverage of TB preventative treatment among high-risk populations, and the availability of a new TB vaccine within the next 5 years. Whole genome sequencing has proven to be a formidable force in the fight against TB.

Leveraging on the rich repository of publicly available Mtb genomic datasets generated from across the globe, along with advancements in bioinformatics and analytical methods that arose from the COVID-19 pandemic, this study aimed to characterise the global diversity of Mtb. The goal was to advance our understanding of how this pathogen has evolved into one of the major global health challenges.

The African continent has been proposed as the origin of Mtb, yet very few studies have been done to explore Mtb evolution on the continent. Moreover, very few Mtb genomes have been sequenced in Africa compared to high-income countries with low TB incidence rates. This thesis aimed to analyse available Mtb genomic data from African isolates to investigate the evolutionary dynamics that shaped the Mtb population landscape on the continent,

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

including the factors driving the expansion of modern Mtb lineages and the evolution of drug-resistant TB in Africa.

The Southern African Development Community (SADC) region has faced decades of high HIV and TB incidence rates, yet the genomic diversity of the MTBC in this region remains under-characterised. This project aimed to investigate the genetic diversity of *M. tuberculosis* Lineages 2 and 4 in the SADC region and to explore the phylogeographic dynamics of MTBC within the region. Furthermore, drug-resistant TB continues to threaten global TB eradication efforts, with resistance reported even in newly introduced anti-TB drugs. Although the evolution of drug resistance in TB has been a major focus of research, the genetic basis of resistance in *M. tuberculosis* remains poorly understood. This study aimed to elucidate the epistatic interactions that govern the evolution of drug resistance in *M. tuberculosis*.

6.2 KEY CONCLUSIONS AND RECOMMENDATIONS

6.2.1 Data curation

While COVID-19 disruptions might have hurt global TB control efforts, the fight against COVID-19 unlocked enormous benefits for TB research, especially in the area of WGS. The massive volumes of genomic data that were being constantly generated necessitated innovative approaches to tackle large volumes of genomic data. Here, we adopted an analysis protocol that allowed us to analyse large volumes of Mtb WGS data efficiently. Because of its importance to public health, Mtb is one of the most studied and sequenced pathogens. Over the last decades, an extensive global dataset of Mtb genomes has accumulated. Using Maximum Likelihood phylogenomic tools on a dataset of more than 9000 genome isolates, we were able to get insights into drug resistance mechanisms, divergence time estimations and phylogeographic dynamics of the Mtb population. Through our analysis, we also highlighted the importance of rigorous quality control of publicly available data due to multiple erroneous records and redundancy in the public sequence repositories. For example, 185 Papua New Guinea samples sourced from the AFROTB dataset, which only

© University of Pretoria

contains strains from Africa, were erroneously recorded as sampled from the Guinea Republic in West Africa. The redundancy and bias are usually caused by oversampling from high-income countries with better capacity to routinely perform WGS on Mtb isolates. Sampling bias can also be a result of the accumulation of WGS data from large outbreak samples. For example, if a particular strain has infected 500 people, sampling from each person will result in 500 WGS sequences of the same strain being deposited in public databases. For computational efficiency reasons, it is important to analyse the redundancy levels of large MTB datasets before proceeding into more complex downstream analyses.

We do acknowledge that our dataset was a product of convenience sampling, which may not reflect the real-world Mtb allele frequencies; however, valuable insights were gained into evolutionary Mtb processes. Phylogenomic characterisation of the large dataset of Mtb sequences we compiled from public sources revealed a broad representation of the pathogen's global genetic diversity. Using advanced lineage classification tools, that leverage SNP data, we identified Lineage 4 and Lineage 2 as the most dominant lineages globally and in African countries. These two generalist lineages were, as expected, responsible for MDR-TB outbreaks due to their enhanced capacity to acquire new drug resistance mutations and associated compensatory mutations that mitigate the fitness costs of DR mutations. However, the patterns and frequencies of these acquired mutations varied significantly between lineages.

Of particular concern was the identification of mutations associated with resistance to recently introduced anti-TB drugs, including bedaquiline, delamanid and linezolid. This finding underscores the urgent need for rapid diagnostics and appropriate treatment regimens that preserve the efficacy of these game-changing drugs.

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

6.2.2 Evolution of Mtb in Africa

The African continent is known to harbour the highest genetic diversity of Mtb, yet very few studies have explored evolutionary dynamics on the continent. As we observed in the global dataset, Lineage 4 was the most prevalent strain on the continent, specifically the sublineage 4.3. Lineage 2 was the second most prevalent in Africa, in agreement with previous studies that have also reported on the increasing prevalence of Lineage 2 on the continent [346, 420, 421].

The prevalence of Lineage 2 and Lineage 4.3 on the continent highlights the need for robust surveillance of these two lineages as, in some instances, they have been associated with faster progression to cause active TB and DR-TB outbreaks [83, 422-424]. Our drug resistance prediction results were concordant with the pDST results provided in the metadata, especially for the first-line drugs. By placing the identified drug-resistant mutations in a temporal phylogenetic context, we found that TB drug resistance is a recent phenomenon, with the first drug-resistance mutation mapped to the year 1959. Although caution must be exercised when interpreting time-scaled phylogenies [319], our results are in agreement with previous studies [192, 425]. In exploring the evolutionary trajectory of DR-TB, we identified that drug resistance in Africa emerged as a series of multiple independent events. However, one consistency in the trajectory of the development of resistance from susceptibility to pre-XDR-TB was that the isoniazid resistance developed first before rifampicin resistance emerged, while fluoroquinolone resistance was the last to emerge. This order of events was consistent in all lineages and geographical locations further emphasizing the importance of early diagnosis of rifampicin-resistant TB, which can serve as a marker of MDR-TB development.

Concerningly, we identified several recently acquired fluoroquinolone resistance-associated mutations with the latest predicted to have emerged 10 years ago. Although the late emergence of fluoroquinolone resistance is consistent with common treatment guidelines where the drugs are only added to the treatment regimen after the failure of the first-line therapy, fluoroquinolone resistance is a known risk factor in MDR-TB treatment failure [349, 351-353] , and this recent emergence of resistance poses a risk of transmission of highly resistant strains. The presence of variants associated with newly introduced TB drugs

© University of Pretoria

raises a concern, especially given the successful adoption of bedaquiline, delamanid and pretomanid in treating MDR-TB. We implore various stakeholders in the TB field to properly manage the risk of drug resistance developing against new TB antibiotics. Drug susceptibility testing remains a key priority in guiding the selection of treatment regimens.

Population genetic signatures showed an expansion of modern MTBC lineages population at the turn of the 19th century. This supports the hypothesis that the recent TB epidemic on the continent was fuelled by increased migration from Europe and Asia towards the continent during the colonial era. However, this should not be interpreted as support of the "Virgin Soil" hypothesis, which assumes that Africa was TB-free before European contact [426-428], as the current evidence supports the "Out of Africa" hypothesis that places the continent as the origin of MTBC. We hypothesize the presence of a different MTBC landscape in Africa before colonialism, where strains in circulation were less virulent and less transmissible compared to the highly successful modern lineages that had been refined by years of selection towards highly virulent and transmissible pathogens [50].

As the wave of colonialism swept throughout the continent, it drastically changed African societies as shifts in economic systems led to the displacement of African populations to more crowded economic hubs. This, coupled with poor working environments and unsanitary living conditions, might have provided the perfect opportunities for the modern lineages to take over and reshape the MTBC diversity landscape on the continent. The stabilization of MTBC populations in the mid-20th century could be attributed to changes in attitudes towards TB control about industrial policy as well as the introduction of antibiotics. There was also another expansion of Mtb populations in the late 1990s to the mid-2000s. This period coincides with the beginning of the widespread HIV epidemic and the subsequent rise in TB incidence that followed [355-357]. A study focusing on LAM4/KZN XDR-TB strains and HIV prevalence data also showed a link between the expansion of the effective population sizes of XDR-TB strains and HIV prevalence [337].

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

6.2.3 The genetic diversity of Lineage 2 and Lineage 4 Mtb in the SADC region

Our study shows that the genetic diversity of Mtb in the SADC region has not been fully explored, especially for Lineage 2, with most of the available isolates not classified beyond the Ancestral or modern Beijing strains. This has a major impact on the accurate reporting and monitoring of transmission and outbreaks caused by Lineage 2 strains. Our study supports an earlier study that reported the heterogeneity of Lineage 2 in Africa [192]. While most of our findings support the earlier work done on Lineage 2 in Africa by Rutaihwa *et al*.[192], we did not use their proposed classification scheme to characterise Lineage 2 in SADC, but instead opted for a more comprehensive classification scheme characterised by backward compatibility with previously proposed schemes, while adequately capturing Lineage 2 diversity [44, 182]. We identified 13 sublineages of Lineage 2 in SADC, reflecting higher diversity compared to Central Asia and Eastern parts of Europe, where a few highly similar Lineage 2-Beijings strains predominate [192, 391, 429]. While Lineage 2 has gained notoriety among other Mtb lineages, studies have shown that only sublineages 2.2M2, 2.2M5 and 2.2.M4 drive the global success of Lineage 2 in causing TB outbreaks [385]. These three lineages were the most prevalent in four SADC countries as demonstrated by this study, namely: Eswatini (Lineage 2.2.M2), Botswana (Lineage 2.2.M2), Zimbabwe (Lineage 2.2.M2 and Lineage 2.2.M5) and Mozambique (Lineage 2.2.M5). While this observation might be a result of biased sampling, the presence of these sublineages in the region requires interventions of the regional TB surveillance programs that actively track the transmission of these strains through early diagnosis and data sharing.

We inferred East Asia as the origin of Lineage 2 strains in the SADC region, in concordance with various studies [182, 385]. We also placed Europe as the origin of Lineage 4 in the region. Our results indicate that both lineages were introduced to the region multiple times independently as a result of migration from Europe and Asia. This finding agrees with the lineage expansion schemes we reported in Chapter 3. Our results presented in Chapter 3 place the first emergence of drug resistance in Africa in the mid of 20th century, while lineage expansion and phylogeographic predictions of this work suggest an earlier introduction of both Lineage 2 and Lineage 4. From these findings, we hypothesised that drug resistance was not a driving factor in the early emergence of both Lineage 2 and Lineage 4 in SADC that is a different narrative compared to Eastern Asia and Europe evolutionary hypothesis of

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

TB expansion, where the Mtb landscape, particularly that of Lineage 2, has been predominantly shaped by a few rapidly emerged drug-resistant strains [192, 391, 429].

Although the low number of sequences from other regions might have limited our study, we draw some interesting insights from our phylogeographic analyses on a regional scale. Firstly, our results agree with both historical and current migration patterns in the region. From the inferred symmetrized migration rates, we identified a strong connectivity between Malawi and Tanzania, which makes sense since the two countries are neighbours. While the migration rates between Malawi and Botswana were the lowest. We inferred a moderately higher rate of Mtb exchange between Malawi and South Africa. Historically, Malawi has been a source of labour for South African mines [383, 430]. We identified South Africa as a historical hub for the dissemination of Lineage 2 to other parts of the region and beyond. We also identified the same trend for Lineage 4, where both South Africa and Zimbabwe were identified as transmission hubs. In the modern day, the SADC region still experiences high rates of labour migration. The mining industry, which is one of the major beneficiaries of migrant labour, carries some of the highest TB burdens in the region [384, 431, 432]. A study reported on an active cross-border transmission of RR/MDR-TB within the region [433].

There is a need for an integrated regional approach to managing TB within the SADC region. Policymakers can adopt a cross-border TB control system that mirrors the one in the European Union [434]. The limited availability of MTB WGS data from most countries in the region is also a cause for concern, and we recommend urgent capacitation of TB laboratories in the region.

6.2.4 Epistatic landscape of drug-resistant TB

Mathematical models play a vital role in the characterisation of the complex factors that drive the evolution of drug resistance in Mtb. These approaches have been instrumental in supporting experimental TB research that has translated into the discovery of new TB drugs, as we explained in Chapter 1. Coupling statistical modelling with WGS data and phenotypic drug susceptibility profiles, we inferred functional associations between polymorphisms to model the epistatic dynamics of drug resistance in Mtb. All the well-characterised drug resistance mutations (from the literature) had high functional association estimates. We also

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

identified some novel mutations in genes that code for efflux pumps, DNA repair, and lipid metabolism pathways. We identified a key role of the polyketide synthase gene *pks12* in the development of intrinsic multi-drug resistance. The same gene has been implicated as a prerequisite for multi-drug resistance development in *M. avium* [435]. Although we identified some strains that carried mutations implicated in the development of drug resistance to the newly added drugs bedaquiline, clofazimine and linezolid, attributable risk statistics conducted in this study did not predict any additional mutations significantly associated with resistance to any of these drugs. A reason for this could be that the attributable risk approach requires a sufficiently large training dataset, whereas the available data on resistance strains to these novel drugs in Africa is limited, hence the low statistical power. From the performed network analysis, Lineage 2 showed the most complex epistatic network, and this could be attributed to the recent global expansion of the lineage, in some cases driven by drug resistance. We do acknowledge that our inferences are highly theoretical, and further experimental validation is thus required.

6.3 FUTURE DIRECTIONS AND RESEARCH NEEDS

This work is particularly relevant and well-aligned with the current research priorities. The recent COVID-19 pandemic has rejuvenated global public health research, as evidenced by recent commitments to strategies and goals that aim to bring an end to TB. As sequencing costs continue to drop and technologies advance, we expect WGS to become routine in clinical laboratories. Hopefully, this will be coupled with better data management systems that will be robust enough to capture all required metadata. The availability of WGS data from previously underrepresented countries will allow us to get a better understanding of the global Mtb genetic diversity. Future work will also involve exploring how COVID-19 has impacted the Mtb population on its genetic level.

The COVID-19 pandemic also showed the capabilities of African scientists in tackling global health challenges. We expect to see an enhanced WGS capacity on the continent and better regional and international collaboration in TB research. As more data and better computational technologies become available, our future work will dive deeper into understanding the dynamics of HIV/TB coinfection. We also hypothesize that the full

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

genetic diversity of the Mtb on the continent has not been elucidated. Future work will also focus on understanding the global origins of the Mtb. Our work on epistasis will also be expanded to refine and develop drug resistance and lineage classification computational tools that can be easily adopted in clinical laboratories.

A major limitation of this study was that the Bayesian analyses required substantial amounts of computational power, even on graphical processing units (GPU). We had very limited access to GPU such that we had to limit our dataset sizes and models. With improved access to computational power, we aim to incorporate more parameter-rich models to get better insights into Mtb evolution.

6.4 CLOSING REMARKS

The *M. tuberculosis* genome sequence ushered unprecedented opportunities that have impacted nearly every aspect of TB research. We expect to see a boom in innovations stemming out of Mtb genome sequencing. The study showed the feasibility of analysing large Mtb datasets by leveraging the innovations that stemmed out of the COVID-19 pandemic. Future research will also focus on more efficient strategies to utilise Mtb genomes at scale.

The work in this thesis not only adds to the body of knowledge on Mtb diversity but also raises awareness of the need for capacity building in Mtb WGS applications in low-income high-budget settings. This study is one of the few that have sought to exploit the rich diversity on the African continent to understand the evolution of Mtb. Furthermore, this study confirmed the heterogeneity of Lineage 2 Mtb in the SADC region and highlighted the distribution of Lineage 4 in the region. Insights gained here can also help public health stakeholders make more informed decisions when managing outbreaks or tracking TB disease.

Using different phylogeographical approaches, this study also inferred the origins of Lineage 4 and Lineage 2 Mtb in the SADC region and also inferred some of the migratory dynamics in the region in the context of Mtb evolution. Insights gained in this study can help in crafting

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

labour and migration policies that impact TB control. Our work on epistatic dynamics in TB allowed us to get better insights into the genetic basis of phenotypic drug resistance. Our findings helped us in refining our in-house computational tools for Mtb research.

In closing, the full potential of Mtb genomics can only be harnessed by complementing it with experimental wet lab work. Thus, there is a need for more collaboration between computational researchers and their wet lab counterparts.

© University of Pretoria

CHAPTER 7 REFERENCES

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Bibliography

1. Organization, G.W.H., *Global tuberculosis report 2024.* 2024.
2. Kanabalan, R.D., et al., *Human tuberculosis and Mycobacterium tuberculosis complex: A review on genetic diversity, pathogenesis and omics approaches in host biomarkers discovery.* Microbiological research, 2021. **246**: p. 126674.
3. Stritt, C. and S. Gagneux, *How do monomorphic bacteria evolve? The Mycobacterium tuberculosis complex and the awkward population genetics of extreme clonality.* Peer Community Journal, 2023. **3**.
4. Achtman, M., *Evolution, population structure, and phylogeography of genetically monomorphic bacterial pathogens.* Annu. Rev. Microbiol., 2008. **62**(1): p. 53-70.
5. Madacki, J., et al., *Esx-1-independent horizontal gene transfer by Mycobacterium tuberculosis complex strains.* MBio, 2021. **12**(3): p. 10.1128/mbio. 00965-21.
6. Gagneux, S., *Ecology and evolution of Mycobacterium tuberculosis.* Nature Reviews Microbiology, 2018. **16**(4): p. 202-213.
7. Wiens, K.E., et al., *Global variation in bacterial strains that cause tuberculosis disease: a systematic review and meta-analysis.* BMC medicine, 2018. **16**: p. 1-13.
8. Organization, G.W.H., *Global tuberculosis report 2023.* 2023.
9. Raviglione, M. and G. Director, *Global strategy and targets for tuberculosis prevention, care and control after 2015.* World Health Organization, Geneva, 2013.
10. Rao, S.N., et al., *Errors in the treatment of tuberculosis in Baltimore.* Chest, 2000. **117**(3): p. 734-737.
11. Organization, W.H., *WHO consolidated guidelines on tuberculosis. Module 4: treatment-drug-resistant tuberculosis treatment, 2022 update*. 2022: World Health Organization.
12. Pooran, A., et al., *What is the cost of diagnosis and management of drug resistant tuberculosis in South Africa?* PloS one, 2013. **8**(1): p. e54587.
13. Sudre, P., G. Ten Dam, and A. Kochi, *Tuberculosis: a global overview of the situation today.* Bulletin of the World Health Organization, 1992. **70**(2): p. 149.
14. WHO, G., *Global tuberculosis report 2020.* Glob. Tuberc. Rep, 2020. **2020**.
15. info, A., *AIDS info.* 2023.
16. Bos, K.I., et al., *Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis.* Nature, 2014. **514**(7523): p. 494-497.
17. Sharma, A., et al., *Tuberculosis caused by Mycobacterium africanum, United States, 2004–2013.* Emerging infectious diseases, 2016. **22**(3): p. 396.
18. Borrell, S. and S. Gagneux, *Strain diversity, epistasis and the evolution of drug resistance in Mycobacterium tuberculosis.* Clinical Microbiology and Infection, 2011. **17**(6): p. 815-820.
19. Comas, I. and S. Gagneux, *The past and future of tuberculosis research.* PLoS pathog, 2009. **5**(10): p. e1000600.
20. Kaufmann, S.H. and S.K. Parida, *Changing funding patterns in tuberculosis.* Nature Medicine, 2007. **13**(3): p. 299-303.
21. Frieden, T., *Tuberculosis in New York City–turning the tide.* New England Journal of Medicine, 1996. **333**: p. 5.
22. Theron, G., et al., *Accuracy and impact of Xpert MTB/RIF for the diagnosis of smear-negative or sputum-scarce tuberculosis using bronchoalveolar lavage fluid.* Thorax, 2013. **68**(11): p. 1043-51.

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

23. Rousseau, P. and M. Dupuis, *Antituberculous drug susceptibility testing of Mycobacterium bovis BCG strain Montreal.* Canadian journal of microbiology, 1990. **36**(10): p. 735-737.
24. Muzondiwa, D., et al., *Resistance Sniffer: An online tool for prediction of drug resistance patterns of Mycobacterium tuberculosis isolates using next generation sequencing data.* International Journal of Medical Microbiology, 2020. **310**(2): p. 151399.
25. Van Embden, J., et al., *Strain identification of Mycobacterium tuberculosis by DNA fingerprinting: recommendations for a standardized methodology.* Journal of clinical microbiology, 1993. **31**(2): p. 406-409.
26. Plikaytis, B.B., et al., *Multiplex PCR assay specific for the multidrug-resistant strain W of Mycobacterium tuberculosis.* Journal of Clinical Microbiology, 1994. **32**(6): p. 1542-1546.
27. Comas, I. and S. Gagneux, *The past and future of tuberculosis research.* PLoS pathogens, 2009. **5**(10): p. e1000600.
28. Gagneux, S., *Host-pathogen coevolution in human tuberculosis.* Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 2012. **367**: p. 10.
29. Cole, S.T., et al., *Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence.* Nature, 1998. **393**(6685): p. 537-44.
30. Tsolaki, A.G., et al., *Functional and evolutionary genomics of Mycobacterium tuberculosis: insights from genomic deletions in 100 strains.* Proceedings of the National Academy of Sciences, 2004. **101**(14): p. 4865-4870.
31. Brosch, R., et al., *A new evolutionary scenario for the Mycobacterium tuberculosis complex.* Proceedings of the national academy of Sciences, 2002. **99**(6): p. 3684-3689.
32. Gagneux, S., et al., *Variable host–pathogen compatibility in Mycobacterium tuberculosis.* Proceedings of the National Academy of Sciences, 2006. **103**(8): p. 2869-2873.
33. Gagneux, S., *Ecology and evolution of Mycobacterium tuberculosis.* Nature Reviews Microbiology, 2018. **16**: p. 202.
34. Comas, I., et al., *Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans.* Nature genetics, 2013. **45**(10): p. 1176-1182.
35. Gagneux, S. and P.M. Small, *Global phylogeography of Mycobacterium tuberculosis and implications for tuberculosis product development.* The Lancet infectious diseases, 2007. **7**(5): p. 328-337.
36. Wirth, T., *Origin, spread and demography of the Mycobacterium tuberculosis complex.* PLoS Pathogens, 2008. **4**.
37. Hershberg, R., et al., *High functional diversity in Mycobacterium tuberculosis driven by genetic drift and human demography.* PLoS Biol, 2008. **6**(12): p. e311.
38. Brudey, K., et al., *Mycobacterium tuberculosis complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology.* BMC microbiology, 2006. **6**(1): p. 1-17.
39. Baker, L., et al., *Silent nucleotide polymorphisms and a phylogeny for Mycobacterium tuberculosis.* Emerging infectious diseases, 2004. **10**(9): p. 1568.

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

40. Filliol, I., et al., *Global phylogeny of Mycobacterium tuberculosis based on single nucleotide polymorphism (SNP) analysis: insights into tuberculosis evolution, phylogenetic accuracy of other DNA fingerprinting systems, and recommendations for a minimal standard SNP set.* J Bacteriol, 2006. **188**(2): p. 759-72.
41. Gutacker, M.M., et al., *Single-nucleotide polymorphism–based population genetic analysis of Mycobacterium tuberculosis strains from 4 geographic sites.* The Journal of infectious diseases, 2006. **193**(1): p. 121-128.
42. Warren RM, S.E., Sampson SL, van der Spuy GD, Richardson M, et and al., *Microevolution of the direct repeat region of Mycobacterium tuberculosis: implications for interpretation of spoligotyping data.* Journal of Clinical Microbiology, 2002. **40**: p. 9.
43. Alland D, W.T., Murray MB, Cave MD, Hazbon MH, et al., *Modeling bacterial evolution with comparative-genome-based marker systems: application to Mycobacterium tuberculosis evolution and pathogenesis.* Journal Bacteriology, 2003. **185**: p. 8.
44. Atavliyeva, S., D. Auganova, and P. Tarlykov, *Genetic diversity, evolution and drug resistance of Mycobacterium tuberculosis lineage 2.* Frontiers in Microbiology, 2024. **15**: p. 1384791.
45. Gagneux, S., *Host–pathogen coevolution in human tuberculosis.* Philosophical Transactions of the Royal Society B: Biological Sciences, 2012. **367**(1590): p. 850-859.
46. Paulsen, H.J., *Tuberculosis in the native American: indigenous or introduced?* Infectious Diseases, 1987. **9**: p. 7.
47. McMillen, C.W., *The red man and the white plague': rethinking race, tuberculosis, and American Indians, ca.* Bulletin - History Medicine, 2008. **82**: p. 38.
48. Stone, A.C., Wilbur,A.K.,Buikstra, J. E.&Roberts,C.A., *Tuberculosis and leprosy in perspective.* american Journal of Physical Anthropology, 2009. **140**: p. 29.
49. Salo, W.L., Aufderheide, A. C., Buikstra, J. & Holcomb,T. A., *Identification of Mycobacterium tuberculosis DNA in a pre-Columbian Peruvian mummy.* Proc. Natl Acad. Sci. USA, 1994. **91**: p. 4.
50. Hershberg, R., et al., *High functional diversity in Mycobacterium tuberculosis driven by genetic drift and human demography.* PLoS biology, 2008. **6**(12): p. e311.
51. Cole, S., et al., *Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence.* Nature, 1998. **396**(6707): p. 190-190.
52. Coscolla, M. and S. Gagneux. *Consequences of genomic diversity in Mycobacterium tuberculosis*. in *Seminars in immunology*. 2014. Elsevier.
53. Coscolla, M., et al., *M. tuberculosis T cell epitope analysis reveals paucity of antigenic variation and identifies rare variable TB antigens.* Cell host & microbe, 2015. **18**(5): p. 538-548.
54. Ogura Y, O.T., Iguchi A, Toh H, Asadulghani M, Oshima K, et al., *Comparative genomics reveal the mechanism of the parallel evolution of O157 and non-O157 enterohemorrhagic Escherichia coli.* Pro Natl Acad Sci USA, 2009. **106**: p. 6.
55. Pozo, P., et al., *Use of whole-genome sequencing to unravel the genetic diversity of a prevalent mycobacterium bovis spoligotype in a multi-host scenario in Spain.* Frontiers in Microbiology, 2022. **13**: p. 915843.

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

56. Tam, V., et al., *Benefits and limitations of genome-wide association studies.* Nature Reviews Genetics, 2019. **20**(8): p. 467-484.
57. Lew JM, K.A., Jones LM, Cole ST., *TubercuList-10 years after.* . Tuberculosis, 2011. **91**: p. 8.
58. Schleusener, V., et al., *Mycobacterium tuberculosis resistance prediction and lineage classification from genome sequencing: comparison of automated analysis tools.* Scientific Reports, 2017. **7**: p. 46327.
59. Organization, W.H., *The use of next-generation sequencing technologies for the detection of mutations associated with drug resistance in Mycobacterium tuberculosis complex: technical guide*. 2018, World Health Organization.
60. Vilchèze, C. and J. Jacobs, William R, *The mechanism of isoniazid killing: clarity through the scope of genetics.* Annu. Rev. Microbiol., 2007. **61**(1): p. 35-50.
61. Koleske, B.N., W.R. Jacobs, and W.R. Bishai, *The Mycobacterium tuberculosis genome at 25 years: lessons and lingering questions.* The Journal of clinical investigation, 2023. **133**(19).
62. Park, M., et al., *Evaluating the clinical impact of routine whole genome sequencing in tuberculosis treatment decisions and the issue of isoniazid mono-resistance.* BMC infectious diseases, 2022. **22**(1): p. 349.
63. Vogel, M., et al., *Implementation of whole genome sequencing for tuberculosis diagnostics in a low-middle income, high MDR-TB burden country.* Scientific reports, 2021. **11**(1): p. 15333.
64. Cox, H., et al., *Whole-genome sequencing has the potential to improve treatment for rifampicin-resistant tuberculosis in high-burden settings: a retrospective cohort study.* Journal of clinical microbiology, 2022. **60**(3): p. e02362-21.
65. Andries, K., et al., *A diarylquinoline drug active on the ATP synthase of Mycobacterium tuberculosis.* Science, 2005. **307**(5707): p. 223-227.
66. Koul, A., et al., *Diarylquinolines are bactericidal for dormant mycobacteria as a result of disturbed ATP homeostasis.* Journal of Biological Chemistry, 2008. **283**(37): p. 25273-25280.
67. CK, S., *A small-molecule nitroimidazopyran drug candidate for the treatment of tuberculosis.* Nature, 2000. **405**: p. 962-966.
68. Matsumoto, M., *OPC-67683, a nitro-dihydro-imidazo-oxazole derivatives with promising action against tuberculosis in vitro and in mice.* PLoS Medicine., 2006. **3**: p. 1-13.
69. Manjunatha, U.H., et al., *Identification of a nitroimidazo-oxazine-specific protein involved in PA-824 resistance in Mycobacterium tuberculosis.* Proceedings of the National Academy of Sciences, 2006. **103**(2): p. 431-436.
70. Singh, R., et al., *PA-824 kills nonreplicating Mycobacterium tuberculosis by intracellular NO release.* Science, 2008. **322**(5906): p. 1392-1395.
71. Ford, C.B., *The evolution of drug resistant Mycobacterium tuberculosis*. 2012, Harvard University.
72. Kato-Maeda, M., et al., *The nature and consequence of genetic variability within Mycobacterium tuberculosis.* The Journal of clinical investigation, 2001. **107**(5): p. 533-537.
73. Tsolaki, A.G., et al., *Genomic deletions classify the Beijing/W strains as a distinct genetic lineage of Mycobacterium tuberculosis.* Journal of clinical microbiology, 2005. **43**(7): p. 3185-3191.

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

74. Muzondiwa, D., *Exploring the evolution of drug resistance in Mycobacterium using whole genome sequencing data*. 2019: University of Pretoria (South Africa).
75. Olson, N.D., et al., *PrecisionFDA Truth Challenge V2: Calling variants from short and long reads in difficult-to-map regions.* Cell genomics, 2022. **2**(5).
76. Koboldt, D.C., *Best practices for variant calling in clinical sequencing.* Genome Medicine, 2020. **12**(1): p. 91.
77. Roberts, D.M., et al., *Two sensor kinases contribute to the hypoxic response of Mycobacterium tuberculosis.* Journal of Biological Chemistry, 2004. **279**(22): p. 23082-23087.
78. Pevzner, P.A., H. Tang, and M.S. Waterman, *An Eulerian path approach to DNA fragment assembly.* Proceedings of the national academy of sciences, 2001. **98**(17): p. 9748-9753.
79. Di Marco, F., et al., *Advantages of long-and short-reads sequencing for the hybrid investigation of the Mycobacterium tuberculosis genome.* Frontiers in Microbiology, 2023. **14**: p. 1104456.
80. Youmans, G.P. and A.G. Karlson, *Streptomycin sensitivity of tubercle bacilli: Studies on recently isolated tubercle bacilli and the development of resistance to streptomycin in vivo.* American review of tuberculosis, 1947. **55**(6): p. 529-535.
81. DUNNER, E., W.B. BROWN, and J. WALLACE, *The effect of streptomycin with para-amino salicylic acid on the emergence of resistant strains of tubercle bacilli.* Diseases of the Chest, 1949. **16**(6): p. 661-666.
82. Organization, W.H., *Meeting report of the WHO expert consultation on the definition of extensively drug-resistant tuberculosis, 27-29 October 2020*. 2021: World Health Organization.
83. Gandhi, N.R., et al., *Extensively drug-resistant tuberculosis as a cause of death in patients co-infected with tuberculosis and HIV in a rural area of South Africa.* The Lancet, 2006. **368**(9547): p. 1575-1580.
84. Coll, F., et al., *Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences.* Genome medicine, 2015. **7**: p. 1-10.
85. Yang TaeWon, Y.T., et al., *Side effects associated with the treatment of multidrug-resistant tuberculosis at a tuberculosis referral hospital in South Korea: a retrospective study.* 2017.
86. Shean, K., et al., *Drug-associated adverse events and their relationship with outcomes in patients receiving treatment for extensively drug-resistant tuberculosis in South Africa.* PloS one, 2013. **8**(5): p. e63057.
87. Walker, T., et al., *Whole genome sequencing for M/XDR tuberculosis surveillance and for resistance testing.* Clinical Microbiology and Infection, 2017. **23**(3): p. 161-166.
88. Trauner, A., et al., *Evolution of drug resistance in tuberculosis: recent progress and implications for diagnosis and therapy.* Drugs, 2014. **74**: p. 1063-1072.
89. Fauci, A.S. and N.T.W. Group, *Multidrug-resistant and extensively drug-resistant tuberculosis: the National Institute of Allergy and Infectious Diseases Research agenda and recommendations for priority research.* The Journal of infectious diseases, 2008. **197**(11): p. 1493-1498.
90. Connolly, L.E., P.H. Edelstein, and L. Ramakrishnan, *Why is long-term therapy required to cure tuberculosis?* PLoS medicine, 2007. **4**(3): p. e120.

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

91. Barry 3rd, C.E., et al., *The spectrum of latent tuberculosis: rethinking the biology and intervention strategies.* Nature Reviews Microbiology, 2009. **7**(12): p. 845-855.
92. Rittershaus, E.S., S.-H. Baek, and C.M. Sassetti, *The normalcy of dormancy: common themes in microbial quiescence.* Cell host & microbe, 2013. **13**(6): p. 643-651.
93. Koch, A., V. Mizrahi, and D.F. Warner, *The impact of drug resistance on Mycobacterium tuberculosis physiology: what can we learn from rifampicin?* Emerging microbes & infections, 2014. **3**(1): p. 1-11.
94. Dookie, N., et al., *Evolution of drug resistance in Mycobacterium tuberculosis: a review on the molecular determinants of resistance and implications for personalized care.* Journal of Antimicrobial Chemotherapy, 2018. **73**(5): p. 1138-1151.
95. Palomino, J.C. and A. Martin, *Drug resistance mechanisms in Mycobacterium tuberculosis.* Antibiotics, 2014. **3**(3): p. 317-340.
96. Almeida Da Silva, P.E. and J.C. Palomino, *Molecular basis and mechanisms of drug resistance in Mycobacterium tuberculosis: classical and new drugs.* Journal of antimicrobial chemotherapy, 2011. **66**(7): p. 1417-1430.
97. Rattan, A., A. Kalia, and N. Ahmad, *Multidrug-resistant Mycobacterium tuberculosis: molecular perspectives.* Emerging infectious diseases, 1998. **4**(2): p. 195.
98. Mitchison, D.A., *Basic mechanisms of chemotherapy.* Chest, 1979. **76**(6): p. 771-780.
99. Li, M.-c., et al., *rpoB mutations and effects on rifampin resistance in Mycobacterium tuberculosis.* Infection and Drug Resistance, 2021: p. 4119-4128.
100. Yuen, L.K., D. Leslie, and P.J. Coloe, *Bacteriological and molecular analysis of rifampin-resistant Mycobacterium tuberculosis strains isolated in Australia.* Journal of Clinical Microbiology, 1999. **37**(12): p. 3844-3850.
101. Traore, H., et al., *Detection of rifampicin resistance in Mycobacterium tuberculosis isolates from diverse countries by a commercial line probe assay as an initial indicator of multidrug resistance.* The international journal of tuberculosis and lung disease, 2000. **4**(5): p. 481-484.
102. Van Niekerk, K., et al., *Clade-specific distribution of antibiotic resistance mutations in the population of Mycobacterium tuberculosis. Prospects for drug resistance reversion.* Basic Biology and Applications of Actinobacteria, 2018: p. 79-98.
103. Banerjee, A., et al., *inhA, a gene encoding a target for isoniazid and ethionamide in Mycobacterium tuberculosis.* Science, 1994. **263**(5144): p. 227-230.
104. Larsen, M.H., et al., *Overexpression of inhA, but not kasA, confers resistance to isoniazid and ethionamide in Mycobacterium smegmatis, M. bovis BCG and M. tuberculosis.* Molecular microbiology, 2002. **46**(2): p. 453-466.
105. Gegia, M., et al., *Treatment of isoniazid-resistant tuberculosis with first-line drugs: a systematic review and meta-analysis.* The Lancet Infectious Diseases, 2017. **17**(2): p. 223-234.
106. Manson, A.L., et al., *Genomic analysis of globally diverse Mycobacterium tuberculosis strains provides insights into the emergence and spread of multidrug resistance.* Nature genetics, 2017. **49**(3): p. 395-402.

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

107. Takayama, K. and J.O. Kilburn, *Inhibition of synthesis of arabinogalactan by ethambutol in Mycobacterium smegmatis.* Antimicrobial agents and chemotherapy, 1989. **33**(9): p. 1493-1499.
108. Mikusova, K., et al., *Biogenesis of the mycobacterial cell wall and the site of action of ethambutol.* Antimicrobial agents and chemotherapy, 1995. **39**(11): p. 2484-2489.
109. Zhao, L.-l., et al., *Analysis of embCAB mutations associated with ethambutol resistance in multidrug-resistant Mycobacterium tuberculosis isolates from China.* Antimicrobial agents and chemotherapy, 2015. **59**(4): p. 2045-2050.
110. Hazbón, M.H., et al., *Role of embB codon 306 mutations in Mycobacterium tuberculosis revisited: a novel association with broad drug resistance and IS 6110 clustering rather than ethambutol resistance.* Antimicrobial agents and chemotherapy, 2005. **49**(9): p. 3794-3802.
111. Safi, H., et al., *Allelic exchange and mutant selection demonstrate that common clinical embCAB gene mutations only modestly increase resistance to ethambutol in Mycobacterium tuberculosis.* Antimicrobial agents and chemotherapy, 2010. **54**(1): p. 103-108.
112. Ahmad, S., A.-A. Jaber, and E. Mokaddas, *Frequency of embB codon 306 mutations in ethambutol-susceptible and-resistant clinical Mycobacterium tuberculosis isolates in Kuwait.* Tuberculosis, 2007. **87**(2): p. 123-129.
113. Giri, A., et al., *Polymorphisms in Rv3806c (ubiA) and the upstream region of embA in relation to ethambutol resistance in clinical isolates of Mycobacterium tuberculosis from North India.* Tuberculosis, 2018. **108**: p. 41-46.
114. Whitfield, M.G., et al., *A global perspective on pyrazinamide resistance: systematic review and meta-analysis.* PloS one, 2015. **10**(7): p. e0133869.
115. Shehzad, A., et al., *Challenges in the development of drugs for the treatment of tuberculosis.* The Brazilian journal of infectious diseases, 2013. **17**(1): p. 74-81.
116. Zhang, Y. and D. Mitchison, *The curious characteristics of pyrazinamide: a review.* The international journal of tuberculosis and lung disease, 2003. **7**(1): p. 6-21.
117. Zimhony, O., et al., *Pyrazinoic acid and its n-propyl ester inhibit fatty acid synthase type I in replicating tubercle bacilli.* Antimicrobial agents and chemotherapy, 2007. **51**(2): p. 752-754.
118. Peterson, N.D., et al., *Uncoupling environmental pH and intrabacterial acidification from pyrazinamide susceptibility in Mycobacterium tuberculosis.* Antimicrobial agents and chemotherapy, 2015. **59**(12): p. 7320-7326.
119. Dillon, N.A., et al., *Anti-tubercular activity of pyrazinamide is independent of trans-translation and RpsA.* Scientific reports, 2017. **7**(1): p. 6135.
120. Miotto, P., et al., *Drug resistance mechanisms and drug susceptibility testing for tuberculosis.* Respirology, 2018. **23**(12): p. 1098-1113.
121. Miotto, P., et al., *Mycobacterium tuberculosis pyrazinamide resistance determinants: a multicenter study.* MBio, 2014. **5**(5): p. 10.1128/mbio. 01819-14.
122. Whitfield, M.G., et al., *Mycobacterium tuberculosis pncA polymorphisms that do not confer pyrazinamide resistance at a breakpoint concentration of 100 micrograms per milliliter in MGIT.* Journal of clinical microbiology, 2015. **53**(11): p. 3633-3635.

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

123. Zhang, S., et al., *Mutations in panD encoding aspartate decarboxylase are associated with pyrazinamide resistance in Mycobacterium tuberculosis.* Emerging microbes & infections, 2013. **2**(1): p. 1-5.
124. Shi, W., et al., *Aspartate decarboxylase (PanD) as a new target of pyrazinamide in Mycobacterium tuberculosis.* Emerging microbes & infections, 2014. **3**(1): p. 1-8.
125. Gopal, P., et al., *Pyrazinoic acid inhibits mycobacterial coenzyme A biosynthesis by binding to aspartate decarboxylase PanD.* ACS infectious diseases, 2017. **3**(11): p. 807-819.
126. Pyle, M.M. *Relative numbers of resistant tubercle bacilli in sputa of patients before and during treatment with streptomycin*. in *Proceedings of the staff meetings. Mayo Clinic*. 1947.
127. Spies, F.S., et al., *Identification of mutations related to streptomycin resistance in clinical isolates of Mycobacterium tuberculosis and possible involvement of efflux mechanism.* Antimicrobial agents and chemotherapy, 2008. **52**(8): p. 2947-2949.
128. Okamoto, S., et al., *Loss of a conserved 7-methylguanosine modification in 16S rRNA confers low-level streptomycin resistance in bacteria.* Molecular microbiology, 2007. **63**(4): p. 1096-1106.
129. Maus, C.E., B.B. Plikaytis, and T.M. Shinnick, *Mutation of tlyA confers capreomycin resistance in Mycobacterium tuberculosis.* Antimicrobial agents and chemotherapy, 2005. **49**(2): p. 571-577.
130. Alangaden, G.J., et al., *Mechanism of resistance to amikacin and kanamycin in Mycobacterium tuberculosis.* Antimicrobial agents and chemotherapy, 1998. **42**(5): p. 1295-1297.
131. Suzuki, Y., et al., *Detection of kanamycin-resistant Mycobacterium tuberculosis by identifying mutations in the 16S rRNA gene.* Journal of clinical microbiology, 1998. **36**(5): p. 1220-1225.
132. Johansen, S.K., et al., *Capreomycin binds across the ribosomal subunit interface using tlyA-encoded 2'-O-methylations in 16S and 23S rRNAs.* Molecular cell, 2006. **23**(2): p. 173-182.
133. Georghiou, S.B., et al., *Evaluation of genetic mutations associated with Mycobacterium tuberculosis resistance to amikacin, kanamycin and capreomycin: a systematic review.* PloS one, 2012. **7**(3): p. e33275.
134. Zaunbrecher, M.A., et al., *Overexpression of the chromosomally encoded aminoglycoside acetyltransferase eis confers kanamycin resistance in Mycobacterium tuberculosis.* Proceedings of the National Academy of Sciences, 2009. **106**(47): p. 20004-20009.
135. Campbell, P.J., et al., *Molecular detection of mutations associated with first-and second-line drug resistance compared with conventional drug susceptibility testing of Mycobacterium tuberculosis.* Antimicrobial agents and chemotherapy, 2011. **55**(5): p. 2032-2041.
136. Dheda, K., et al., *The epidemiology, pathogenesis, transmission, diagnosis, and management of multidrug-resistant, extensively drug-resistant, and incurable tuberculosis.* The lancet Respiratory medicine, 2017. **5**(4): p. 291-360.
137. Drlica, K. and X. Zhao, *DNA gyrase, topoisomerase IV, and the 4-quinolones.* Microbiology and molecular biology reviews, 1997. **61**(3): p. 377-392.

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

138. Takiff, H.E., et al., *Cloning and nucleotide sequence of Mycobacterium tuberculosis gyrA and gyrB genes and detection of quinolone resistance mutations.* Antimicrobial agents and chemotherapy, 1994. **38**(4): p. 773-780.
139. Cheng, A.F., et al., *Multiplex PCR amplimer conformation analysis for rapid detection of gyrA mutations in fluoroquinolone-resistant Mycobacterium tuberculosis clinical isolates.* Antimicrobial agents and chemotherapy, 2004. **48**(2): p. 596-601.
140. Maruri, F., et al., *A systematic review of gyrase mutations associated with fluoroquinolone-resistant Mycobacterium tuberculosis and a proposed gyrase numbering system.* Journal of Antimicrobial Chemotherapy, 2012. **67**(4): p. 819-831.
141. Musser, J.M., *Antimicrobial agent resistance in mycobacteria: molecular genetic insights.* Clinical microbiology reviews, 1995. **8**(4): p. 496-514.
142. Aubry, A., et al., *Novel gyrase mutations in quinolone-resistant and-hypersusceptible clinical isolates of Mycobacterium tuberculosis: functional analysis of mutant enzymes.* Antimicrobial agents and chemotherapy, 2006. **50**(1): p. 104-112.
143. DeBarber, A.E., et al., *Ethionamide activation and sensitivity in multidrug-resistant Mycobacterium tuberculosis.* Proceedings of the National Academy of Sciences, 2000. **97**(17): p. 9677-9682.
144. Brossier, F., et al., *Molecular investigation of resistance to the antituberculous drug ethionamide in multidrug-resistant clinical isolates of Mycobacterium tuberculosis.* Antimicrobial agents and chemotherapy, 2011. **55**(1): p. 355-360.
145. Axelsen, P.H., *Essentials of antimicrobial pharmacology: a guide to fundamentals for practice*. 2002: Springer Science & Business Media.
146. Rengarajan, J., et al., *The folate pathway is a target for resistance to the drug para-aminosalicylic acid (PAS) in mycobacteria.* Molecular microbiology, 2004. **53**(1): p. 275-282.
147. Zhao, F., et al., *Binding pocket alterations in dihydrofolate synthase confer resistance to para-aminosalicylic acid in clinical isolates of Mycobacterium tuberculosis.* Antimicrobial agents and chemotherapy, 2014. **58**(3): p. 1479-1487.
148. Field, S.K., *Bedaquiline for the treatment of multidrug-resistant tuberculosis: great promise or disappointment?* Therapeutic advances in chronic disease, 2015. **6**(4): p. 170-184.
149. Ismail, N., et al., *In vitro study of stepwise acquisition of rv0678 and atpE mutations conferring bedaquiline resistance.* Antimicrobial agents and chemotherapy, 2019. **63**(8): p. 10.1128/aac. 00292-19.
150. Browne, S. and L. Hogerzeil, *"B 663" in the treatment of leprosy. Preliminary report of a pilot trial.* 1962.
151. Barry, V.C., et al., *A new series of phenazines (rimino-compounds) with high antituberculosis activity.* 1957.
152. Lu, Y., et al., *Activities of clofazimine against Mycobacterium tuberculosis in vitro and in vivo.* Zhonghua jie he he hu xi za zhi= Zhonghua jiehe he huxi zazhi= Chinese journal of tuberculosis and respiratory diseases, 2008. **31**(10): p. 752-755.
153. Lechartier, B. and S.T. Cole, *Mode of action of clofazimine and combination therapy with benzothiazinones against Mycobacterium tuberculosis.* Antimicrobial agents and chemotherapy, 2015. **59**(8): p. 4457-4463.

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

154. Xu, J., et al., *Primary clofazimine and bedaquiline resistance among isolates from patients with multidrug-resistant tuberculosis.* Antimicrobial agents and chemotherapy, 2017. **61**(6): p. 10.1128/aac. 00239-17.
155. Zhang, S., et al., *Identification of novel mutations associated with clofazimine resistance in Mycobacterium tuberculosis.* Journal of Antimicrobial Chemotherapy, 2015. **70**(9): p. 2507-2510.
156. Leach, K.L., et al., *Linezolid, the first oxazolidinone antibacterial agent.* Annals of the New York Academy of Sciences, 2011. **1222**(1): p. 49-54.
157. Zhang, Y., *The magic bullets and tuberculosis drug targets.* Annu. Rev. Pharmacol. Toxicol., 2005. **45**(1): p. 529-564.
158. Hillemann, D., S. Rüsch-Gerdes, and E. Richter, *In vitro-selected linezolid-resistant Mycobacterium tuberculosis mutants.* Antimicrobial agents and chemotherapy, 2008. **52**(2): p. 800-801.
159. Bloemberg, G.V., et al., *Acquired resistance to bedaquiline and delamanid in therapy for tuberculosis.* New England Journal of Medicine, 2015. **373**(20): p. 1986-1988.
160. Zimenkov, D.V., et al., *Examination of bedaquiline-and linezolid-resistant Mycobacterium tuberculosis isolates from the Moscow region.* Journal of Antimicrobial Chemotherapy, 2017. **72**(7): p. 1901-1906.
161. Sun, Z., et al., *Comparison of gyrA gene mutations between laboratory-selected ofloxacin-resistant Mycobacterium tuberculosis strains and clinical isolates.* International journal of antimicrobial agents, 2008. **31**(2): p. 115-121.
162. Carlos Palomino, J. and A. Martin, *Tuberculosis clinical trial update and the current anti-tuberculosis drug portfolio.* Current medicinal chemistry, 2013. **20**(30): p. 3785-3796.
163. Laskin, D.L., et al., *Macrophages, reactive nitrogen species, and lung injury.* Annals of the New York Academy of Sciences, 2010. **1203**(1): p. 60-65.
164. Grange, J.M., *Mycobacterium bovis infection in human beings.* Tuberculosis (Edinburgh), 2001. **81**: p. 7.
165. Fabre, M., Hauck, Y., Soler, C.,Koeck, J. L., van Ingen, J., and D. van Soolingen, Vergnaud, G. & Pourcel, C. , *Molecularcharacteristics of 'Mycobacterium canettii' the smooth Mycobacterium tuberculosis bacilli.* Infect. Genet. Evol., 2010. **10**: p. 9.
166. Koeck, J.-L.e.a., *Clinical characteristics of the smooth tubercle bacilli 'Mycobacterium canettii' infection suggest the existence of an environmental reservoir.* Clinical Microbiology and Infection, 2010. **17**: p. 7.
167. de Jong BC, A.M., Gagneux S., *Mycobacterium africanum-Review of an Important Cause of Human Tuberculosis in West Africa.* PLoS Neglected Tropical Diseases, 2010. **4**.
168. Gehre F, A.M., Fahun F, Odoun M, Uwizeye C, de Rijk P, et al., *The First Phylogeographic Population Structure and Analysis of Transmission Dynamics of M. africanum West African Combining Molecular Data from Benin, Nigeria and Sierra Leone.* PloS One, 2013. **8**(e77000).
169. Brynildsrud, O.B., et al., *Global expansion of Mycobacterium tuberculosis lineage 4 shaped by colonial migration and local adaptation.* Science advances, 2018. **4**(10): p. eaat5869.

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

170. Stucki, D., et al., *Mycobacterium tuberculosis lineage 4 comprises globally distributed and geographically restricted sublineages.* Nature genetics, 2016. **48**(12): p. 1535-1543.
171. Coscolla, M., et al., *Phylogenomics of Mycobacterium africanum reveals a new lineage and a complex evolutionary history.* Microbial genomics, 2021. **7**(2).
172. Napier, G., et al., *Robust barcoding and identification of Mycobacterium tuberculosis lineages for epidemiological and clinical studies.* Genome medicine, 2020. **12**: p. 1-10.
173. Menardo, F., et al., *Local adaptation in populations of Mycobacterium tuberculosis endemic to the Indian Ocean Rim.* F1000Research, 2021. **10**.
174. Couvin, D., Y. Reynaud, and N. Rastogi, *Two tales: Worldwide distribution of Central Asian (CAS) versus ancestral East-African Indian (EAI) lineages of Mycobacterium tuberculosis underlines a remarkable cleavage for phylogeographical, epidemiological and demographical characteristics.* PLoS One, 2019. **14**(7): p. e0219706.
175. O'Neill, M.B., et al., *Lineage specific histories of Mycobacterium tuberculosis dispersal in Africa and Eurasia.* Molecular ecology, 2019. **28**(13): p. 3241-3256.
176. Netikul, T., et al., *Whole-genome single nucleotide variant phylogenetic analysis of Mycobacterium tuberculosis Lineage 1 in endemic regions of Asia and Africa.* Scientific Reports, 2022. **12**(1): p. 1565.
177. Coll, F., et al., *A robust SNP barcode for typing Mycobacterium tuberculosis complex strains.* Nature communications, 2014. **5**(1): p. 4812.
178. Holt, K.E., et al., *Frequent transmission of the Mycobacterium tuberculosis Beijing lineage and positive selection for the EsxW Beijing variant in Vietnam.* Nature genetics, 2018. **50**(6): p. 849-856.
179. Maung, H.M.W., et al., *Geno-spatial distribution of Mycobacterium Tuberculosis and drug resistance profiles in Myanmar–Thai border area.* Tropical Medicine and Infectious Disease, 2020. **5**(4): p. 153.
180. Lempens, P., et al., *Isoniazid resistance levels of Mycobacterium tuberculosis can largely be predicted by high-confidence resistance-conferring mutations.* Scientific reports, 2018. **8**(1): p. 3246.
181. Palittapongarnpim, P., et al., *Evidence for host-bacterial co-evolution via genome sequence analysis of 480 Thai Mycobacterium tuberculosis lineage 1 isolates.* Scientific reports, 2018. **8**(1): p. 11597.
182. Thawornwattana, Y., et al., *Revised nomenclature and SNP barcode for Mycobacterium tuberculosis lineage 2.* Microbial genomics, 2021. **7**(11): p. 000697.
183. Liu, Q., et al., *Genetic features of Mycobacterium tuberculosis modern Beijing sublineage.* Emerging microbes & infections, 2016. **5**(1): p. 1-8.
184. Rad, M.E., et al., *Mutations in putative mutator genes of Mycobacterium tuberculosis strains of the W-Beijing family.* Emerging infectious diseases, 2003. **9**(7): p. 838.
185. Mokrousov, I., et al., *Origin and primary dispersal of the Mycobacterium tuberculosis Beijing genotype: clues from human phylogeography.* Genome Research, 2005. **15**(10): p. 1357-1364.

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

186. Mestre, O., et al., *Phylogeny of Mycobacterium tuberculosis Beijing strains constructed from polymorphisms in genes involved in DNA replication, recombination and repair.* PloS one, 2011. **6**(1): p. e16020.
187. Merker, M., et al., *Evolutionary history and global spread of the Mycobacterium tuberculosis Beijing lineage.* Nature genetics, 2015. **47**(3): p. 242-249.
188. Liu, Q., et al., *China's tuberculosis epidemic stems from historical expansion of four strains of Mycobacterium tuberculosis.* Nature ecology & evolution, 2018. **2**(12): p. 1982-1992.
189. Zhang, H., et al., *Genome sequencing of 161 Mycobacterium tuberculosis isolates from China identifies genes and intergenic regions associated with drug resistance.* Nature genetics, 2013. **45**(10): p. 1255-1260.
190. Shitikov, E., et al., *Evolutionary pathway analysis and unified classification of East Asian lineage of Mycobacterium tuberculosis.* Scientific reports, 2017. **7**(1): p. 9227.
191. Ajawatanawong, P., et al., *A novel Ancestral Beijing sublineage of Mycobacterium tuberculosis suggests the transition site to Modern Beijing sublineages.* Scientific reports, 2019. **9**(1): p. 13718.
192. Rutaihwa, L.K., et al., *Multiple introductions of Mycobacterium tuberculosis lineage 2–Beijing into Africa over centuries.* Frontiers in Ecology and Evolution, 2019. **7**: p. 112.
193. Conceição, E.C., et al., *Genomic Diversity of the rarely observed genotype of the Mycobacterium tuberculosis Central Asian (CAS) Lineage 3 from North Brazil.* Microorganisms, 2023. **11**(1): p. 132.
194. Shuaib, Y.A., et al., *Origin and global expansion of Mycobacterium tuberculosis complex lineage 3.* Genes, 2022. **13**(6): p. 990.
195. Demay, C., et al., *SITVITWEB–a publicly available international multimarker database for studying Mycobacterium tuberculosis genetic diversity and molecular epidemiology.* Infection, genetics and evolution, 2012. **12**(4): p. 755-766.
196. Rindi, L., N. Lari, and C. Garzelli, *Large Sequence Polymorphisms of the Euro-American lineage of Mycobacterium tuberculosis: a phylogenetic reconstruction and evidence for convergent evolution in the DR locus.* Infection, Genetics and Evolution, 2012. **12**(7): p. 1551-1557.
197. Streicher, E., et al., *Spoligotype signatures in the Mycobacterium tuberculosis complex.* Journal of clinical microbiology, 2007. **45**(1): p. 237-240.
198. Castets, M., et al., *Les bacilles tuberculeux de type africain: note préliminaire.* Rev. Tuberc. Pneumol, 1968. **32**: p. 179-184.
199. David, H.L., et al., *Numerical taxonomy analysis of Mycobacterium africanum.* International Journal of Systematic and Evolutionary Microbiology, 1978. **28**(4): p. 464-472.
200. Ates, L.S., et al., *Unexpected genomic and phenotypic diversity of Mycobacterium africanum lineage 5 affects drug resistance, protein secretion, and immunogenicity.* Genome biology and evolution, 2018. **10**(8): p. 1858-1874.
201. Asante-Poku, A., et al., *Mycobacterium africanum is associated with patient ethnicity in Ghana.* PLoS neglected tropical diseases, 2015. **9**(1): p. e3370.
202. Asante-Poku, A., et al., *Molecular epidemiology of Mycobacterium africanum in Ghana.* BMC infectious diseases, 2016. **16**: p. 1-8.

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

203. Mostowy, S., et al., *Genomic deletions suggest a phylogeny for the Mycobacterium tuberculosis complex.* The Journal of infectious diseases, 2002. **186**(1): p. 74-80.
204. Telford, M.J., *Phylogenomics.* Current Biology, 2007. **17**(22): p. R945-R946.
205. Yang, Z. and B. Rannala, *Molecular phylogenetics: principles and practice.* Nature reviews genetics, 2012. **13**(5): p. 303-314.
206. Saitou, N. and M. Nei, *The neighbor-joining method: a new method for reconstructing phylogenetic trees.* Molecular biology and evolution, 1987. **4**(4): p. 406-425.
207. Holder, M. and P.O. Lewis, *Phylogeny estimation: traditional and Bayesian approaches.* Nature reviews genetics, 2003. **4**(4): p. 275-284.
208. Mailund, T., et al., *Recrafting the neighbor-joining method.* BMC bioinformatics, 2006. **7**: p. 1-8.
209. McBroome, J., et al., *A daily-updated database and tools for comprehensive SARS-CoV-2 mutation-annotated trees.* Molecular biology and evolution, 2021. **38**(12): p. 5819-5824.
210. Quast, C., et al., *The SILVA ribosomal RNA gene database project: improved data processing and web-based tools.* Nucleic acids research, 2012. **41**(D1): p. D590-D596.
211. Thompson, J.D., D.G. Higgins, and T.J. Gibson, *CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.* Nucleic acids research, 1994. **22**(22): p. 4673-4680.
212. Edgar, R.C., *MUSCLE: multiple sequence alignment with high accuracy and high throughput.* Nucleic acids research, 2004. **32**(5): p. 1792-1797.
213. Wang, W., et al., *DecentTree: scalable Neighbour-Joining for the genomic era.* Bioinformatics, 2023. **39**(9): p. btad536.
214. Edwards, A.W. and L.L. Cavalli-Sforza, *The reconstruction of evolution.* 1963.
215. Thompson, E., *Likelihood and parsimony: Comparison of criteria and solutions.* Cladistics, 1986. **2**(1): p. 43-52.
216. Farris, J.S., *Methods for computing Wagner trees.* Systematic Biology, 1970. **19**(1): p. 83-92.
217. Fitch, W.M., *Toward defining the course of evolution: minimum change for a specific tree topology.* Systematic Biology, 1971. **20**(4): p. 406-416.
218. Kluge, A.G. and J.S. Farris, *Quantitative phyletics and the evolution of anurans.* Systematic Biology, 1969. **18**(1): p. 1-32.
219. Felsenstein, J., *A likelihood approach to character weighting and what it tells us about parsimony and compatibility.* Biological Journal of the Linnean Society, 1981. **16**(3): p. 183-196.
220. Chen, M.-H., L. Kuo, and P.O. Lewis, *Bayesian phylogenetics: methods, algorithms, and applications*. 2014: CRC Press.
221. Zhukova, A., et al., *Efficiently analysing large viral data sets in computational phylogenomics.* Phylogenetics in the Genomic Era, 2020: p. 5.3: 1--5.3: 43.
222. Whelan, S., P. Liò, and N. Goldman, *Molecular phylogenetics: state-of-the-art methods for looking into the past.* TRENDS in Genetics, 2001. **17**(5): p. 262-272.
223. Edwards, A.W.F., *Likelihood*, in *Time Series and Statistics*. 1972, Springer. p. 126-129.

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

224. Rogers, J.S., *On the consistency of maximum likelihood estimation of phylogenetic trees from nucleotide sequences.* Systematic biology, 1997. **46**(2): p. 354-357.
225. Chang, J.T., *Full reconstruction of Markov models on evolutionary trees: identifiability and consistency.* Mathematical biosciences, 1996. **137**(1): p. 51-73.
226. Rogers, J.S. and D.L. Swofford, *A fast method for approximating maximum likelihoods of phylogenetic trees from nucleotide sequences.* Systematic biology, 1998. **47**(1): p. 77-89.
227. Lewis, P.O., M.T. Holder, and D.L. Swofford, *Phycas: software for Bayesian phylogenetic analysis.* Systematic Biology, 2015. **64**(3): p. 525-531.
228. Bromham, L., *Substitution rate analysis and molecular evolution.* Phylogenetics in the Genomic Era, 2020: p. 4.4: 1--4.4: 21.
229. Zuckerkandl, E. and L. Pauling, *Evolutionary divergence and convergence in proteins*, in *Evolving genes and proteins*. 1965, Elsevier. p. 97-166.
230. Rambaut, A., *Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies.* Bioinformatics, 2000. **16**(4): p. 395-399.
231. Shankarappa, R., et al., *Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection.* Journal of virology, 1999. **73**(12): p. 10489-10502.
232. Drummond, A., O.G. Pybus, and A. Rambaut, *Inference of viral evolutionary rates from molecular sequences.* Adv Parasitol, 2003. **54**: p. 331-358.
233. Rambaut, A., et al., *Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen).* Virus evolution, 2016. **2**(1): p. vew007.
234. Langley, C.H. and W.M. Fitch, *An examination of the constancy of the rate of molecular evolution.* Journal of molecular evolution, 1974. **3**: p. 161-177.
235. Sanderson, M.J., *r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock.* Bioinformatics, 2003. **19**(2): p. 301-302.
236. To, T.-H., et al., *Fast dating using least-squares criteria and algorithms.* Systematic biology, 2016. **65**(1): p. 82-97.
237. Volz, E. and S.D. Frost, *Scalable relaxed clock phylogenetic dating.* Virus evolution, 2017. **3**(2): p. vex025.
238. Sagulenko, P., V. Puller, and R.A. Neher, *TreeTime: Maximum-likelihood phylodynamic analysis.* Virus evolution, 2018. **4**(1): p. vex042.
239. Jukes, T., *Evolution of protein molecules.* Mammalian Protein Metabolism, 1969. **3**.
240. Arenas, M., *Trends in substitution models of molecular evolution.* Frontiers in genetics, 2015. **6**: p. 319.
241. Kimura, M., *A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences.* Journal of molecular evolution, 1980. **16**: p. 111-120.
242. Hasegawa, M., H. Kishino, and T.-a. Yano, *Dating of the human-ape splitting by a molecular clock of mitochondrial DNA.* Journal of molecular evolution, 1985. **22**: p. 160-174.
243. Tavaré, S., *Some probabilistic and statistical problems in the analysis of DNA sequences,(1986).* Some mathematical questions in biology: DNA sequence analysis: American Mathematical Society: p. 57-86.

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

244. Tamura, K. and M. Nei, *Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees.* Molecular biology and evolution, 1993. **10**(3): p. 512-526.
245. Nascimento, F.F., M.d. Reis, and Z. Yang, *A biologist's guide to Bayesian phylogenetic analysis.* Nature ecology & evolution, 2017. **1**(10): p. 1446-1454.
246. Huelsenbeck, J.P., et al., *Bayesian inference of phylogeny and its impact on evolutionary biology.* science, 2001. **294**(5550): p. 2310-2314.
247. Felsenstein, J., *Inferring phylogenies*, in *Inferring phylogenies*. 2004. p. 664-664.
248. Drummond, A.J., et al., *Bayesian coalescent inference of past population dynamics from molecular sequences.* Molecular biology and evolution, 2005. **22**(5): p. 1185-1192.
249. Banner, K.M., K.M. Irvine, and T.J. Rodhouse, *The use of Bayesian priors in Ecology: The good, the bad and the not great.* Methods in Ecology and Evolution, 2020. **11**(8): p. 882-889.
250. Guindon, S. and O. Gascuel, *A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood.* Systematic biology, 2003. **52**(5): p. 696-704.
251. Drummond, A.J. and A. Rambaut, *BEAST: Bayesian evolutionary analysis by sampling trees.* BMC evolutionary biology, 2007. **7**: p. 1-8.
252. Ronquist, F., et al., *MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space.* Systematic biology, 2012. **61**(3): p. 539-542.
253. Höhna, S., et al., *RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language.* Systematic biology, 2016. **65**(4): p. 726-736.
254. Yang, Z., *PAML 4: phylogenetic analysis by maximum likelihood.* Molecular biology and evolution, 2007. **24**(8): p. 1586-1591.
255. Lartillot, N., T. Lepage, and S. Blanquart, *PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating.* Bioinformatics, 2009. **25**(17): p. 2286-2288.
256. Beerli, P., *Comparison of Bayesian and maximum-likelihood inference of population genetic parameters.* Bioinformatics, 2006. **22**(3): p. 341-345.
257. Hey, J. and R. Nielsen, *Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics.* Proceedings of the National Academy of Sciences, 2007. **104**(8): p. 2785-2790.
258. Rabosky, D.L., *Automatic detection of key innovations, rate shifts, and diversity-dependence on phylogenetic trees.* PloS one, 2014. **9**(2): p. e89543.
259. Pritchard, J.K., M. Stephens, and P. Donnelly, *Inference of population structure using multilocus genotype data.* Genetics, 2000. **155**(2): p. 945-959.
260. Nylander, J.A., et al., *AWTY (are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics.* Bioinformatics, 2008. **24**(4): p. 581-583.
261. Hastings, W.K., *Monte Carlo sampling methods using Markov chains and their applications.* 1970.
262. Roberts, G.O. and J.S. Rosenthal, *Optimal scaling for various Metropolis-Hastings algorithms.* Statistical science, 2001. **16**(4): p. 351-367.

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

263. Green, P.J., *MCMC in image analysis.* Markov chain Monte Carlo in practice, 1996: p. 381-400.
264. Lartillot, N. and H. Philippe, *Computing Bayes factors using thermodynamic integration.* Systematic biology, 2006. **55**(2): p. 195-207.
265. Team, R.C., *R Core Team (2023). _R: A language and environment for statistical computing, R Foundation for Statistical.* Computing, 2023.
266. Rambaut, A., et al., *Posterior summarization in Bayesian phylogenetics using Tracer 1.7.* Systematic biology, 2018. **67**(5): p. 901-904.
267. Lanfear, R., X. Hua, and D.L. Warren, *Estimating the effective sample size of tree topologies from Bayesian phylogenetic analyses.* Genome biology and evolution, 2016. **8**(8): p. 2319-2332.
268. Avise, J.C., et al., *Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics.* Annual review of ecology and systematics, 1987: p. 489-522.
269. Zhang, J. and M. Nei, *Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods.* Journal of molecular evolution, 1997. **44**: p. S139-S146.
270. Collins, T.M., P.H. Wimberger, and G.J. Naylor, *Compositional bias, character-state bias, and character-state reconstruction using parsimony.* Systematic Biology, 1994. **43**(4): p. 482-496.
271. Pupko, T., et al., *A fast algorithm for joint reconstruction of ancestral amino acid sequences.* Molecular biology and evolution, 2000. **17**(6): p. 890-896.
272. Gascuel, O. and M. Steel, *Predicting the ancestral character changes in a tree is typically easier than predicting the root state.* Systematic biology, 2014. **63**(3): p. 421-435.
273. Maddison, W. and D. Maddison, *Mesquite 2.* A modular system for evolutionary analysis, 2007. **3**.
274. Ishikawa, S.A., et al., *A fast likelihood method to reconstruct and visualize ancestral scenarios.* Molecular biology and evolution, 2019. **36**(9): p. 2069-2085.
275. Derendinger, B., et al., *Bedaquiline resistance in patients with drug-resistant tuberculosis in Cape Town, South Africa: a retrospective longitudinal cohort study.* The Lancet Microbe, 2023. **4**(12): p. e972-e982.
276. Nimmo, C., et al., *Detection of a historic reservoir of bedaquiline/clofazimine resistance-associated variants in Mycobacterium tuberculosis.* Genome Medicine, 2024. **16**(1): p. 34.
277. Andries, K., et al., *Acquired resistance of Mycobacterium tuberculosis to bedaquiline.* PLOS one, 2014. **9**(7): p. e102135.
278. WHO, G., *Documentation for World Health Assembly g7.* 2014.
279. Nations, U., *Political declaration of the UN general assembly high-level meeting on the fight against tuberculosis*. 2023, UNHQ New York.
280. Pai, H., et al., *Bedaquiline safety, efficacy, utilization and emergence of resistance following treatment of multidrug-resistant tuberculosis patients in South Africa: a retrospective cohort analysis.* BMC infectious diseases, 2022. **22**(1): p. 870.
281. Conradie, F., et al., *Clinical access to Bedaquiline Programme for the treatment of drug-resistant tuberculosis.* SAMJ: South African Medical Journal, 2014. **104**(3): p. 164-166.

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

282. Gler, M.T., et al., *Delamanid for multidrug-resistant pulmonary tuberculosis.* New England Journal of Medicine, 2012. **366**(23): p. 2151-2160.
283. Ghodousi, A., et al., *Acquisition of cross-resistance to bedaquiline and clofazimine following treatment for tuberculosis in Pakistan.* Antimicrobial agents and chemotherapy, 2019. **63**(9): p. 10.1128/aac. 00915-19.
284. Conradie, F., et al., *Treatment of highly drug-resistant pulmonary tuberculosis.* New England Journal of Medicine, 2020. **382**(10): p. 893-902.
285. Hameed, H.A., et al., *Molecular targets related drug resistance mechanisms in MDR-, XDR-, and TDR-Mycobacterium tuberculosis strains.* Frontiers in cellular and infection microbiology, 2018. **8**: p. 114.
286. Jang, J.G. and J.H. Chung, *Diagnosis and treatment of multidrug-resistant tuberculosis.* Yeungnam University journal of medicine, 2020. **37**(4): p. 277-285.
287. Xiao, Y.-X., et al., *Whole-genome sequencing-based analyses of drug-resistant Mycobacterium tuberculosis from Taiwan.* Scientific reports, 2023. **13**(1): p. 2540.
288. Salemi, M., et al., *High-resolution molecular epidemiology and evolutionary history of HIV-1 subtypes in Albania.* PloS one, 2008. **3**(1): p. e1390.
289. Babakir-Mina, M., et al., *The novel swine-origin H1N1 influenza A virus riddle: is it a domestic bird H1N1-derived virus?* The new microbiologica, 2010. **33**(1): p. 77.
290. Ayres, C.F.J., et al., *Zika virus detection, isolation and genome sequencing through Culicidae sampling during the epidemic in Vitória, Espírito Santo, Brazil.* Parasites & vectors, 2019. **12**: p. 1-9.
291. Monzón, S., et al., *Monkeypox virus genomic accordion strategies.* Nature Communications, 2024. **15**(1): p. 3059.
292. Loh, P.-R., M. Baym, and B. Berger, *Compressive genomics.* Nature biotechnology, 2012. **30**(7): p. 627-630.
293. Minh, B.Q., et al., *IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era.* Molecular biology and evolution, 2020. **37**(5): p. 1530-1534.
294. Nguyen, L.-T., et al., *IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies.* Molecular biology and evolution, 2015. **32**(1): p. 268-274.
295. Chernyaeva, E.N., et al., *Genome-wide Mycobacterium tuberculosis variation (GMTV) database: a new tool for integrating sequence variations and epidemiology.* BMC genomics, 2014. **15**: p. 1-8.
296. Laamarti, M., et al., *Afro-TB dataset as a large scale genomic data of Mycobacterium tuberuclosis in Africa.* Scientific Data, 2023. **10**(1): p. 212.
297. Olson, R.D., et al., *Introducing the bacterial and viral bioinformatics resource center (BV-BRC): a resource combining PATRIC, IRD and ViPR.* Nucleic acids research, 2023. **51**(D1): p. D678-D689.
298. Zhan, X., *checkVCF.* GitHub repository, 2016.
299. Martin, M., *Cutadapt removes adapter sequences from high-throughput sequencing reads.* EMBnet. journal, 2011. **17**(1): p. 10-12.
300. Andrews, S., *FastQC: a quality control tool for high throughput sequence data*. 2010, Cambridge, United Kingdom.
301. Ewels, P., et al., *MultiQC: summarize analysis results for multiple tools and samples in a single report.* Bioinformatics, 2016. **32**(19): p. 3047-3048.

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

302. Danecek, P., et al., *Twelve years of SAMtools and BCFtools.* Gigascience, 2021. **10**(2): p. giab008.
303. Li, H., et al., *The sequence alignment/map format and SAMtools.* bioinformatics, 2009. **25**(16): p. 2078-2079.
304. Garrison, E. and G. Marth, *Haplotype-based variant detection from short-read sequencing.* arXiv preprint arXiv:1207.3907, 2012.
305. Freschi, L., et al., *Population structure, biogeography and transmissibility of Mycobacterium tuberculosis.* Nature communications, 2021. **12**(1): p. 6099.
306. Verboven, L., et al., *TBProfiler for automated calling of the association with drug resistance of variants in Mycobacterium tuberculosis.* PLoS One, 2022. **17**(12): p. e0279644.
307. Danecek, P., et al., *The variant call format and VCFtools.* Bioinformatics, 2011. **27**(15): p. 2156-2158.
308. Marin, M., et al., *Benchmarking the empirical accuracy of short-read sequencing across the M. tuberculosis genome.* Bioinformatics, 2022. **38**(7): p. 1781-1787.
309. Ortiz, E.M., *vcf2phylip v2. 0: convert a VCF matrix into several matrix formats for phylogenetic analysis*. 2019, Zenodo Geneva.
310. Page, A.J., et al., *SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments.* Microbial genomics, 2016. **2**(4): p. e000056.
311. Menardo, F., et al., *Treemmer: a tool to reduce large phylogenetic datasets with minimal loss of diversity.* BMC bioinformatics, 2018. **19**: p. 1-8.
312. Kalyaanamoorthy, S., et al., *ModelFinder: fast model selection for accurate phylogenetic estimates.* Nature methods, 2017. **14**(6): p. 587-589.
313. Akaike, H., *Akaike's information criterion.* International encyclopedia of statistical science, 2011: p. 25-25.
314. Huddleston, J., et al., *Augur: a bioinformatics toolkit for phylogenetic analyses of human pathogens.* Journal of open source software, 2021. **6**(57).
315. Hoang, D.T., et al., *UFBoot2: improving the ultrafast bootstrap approximation.* Molecular biology and evolution, 2018. **35**(2): p. 518-522.
316. Organization, W.H., *Catalogue of mutations in Mycobacterium tuberculosis complex and their association with drug resistance*. 2023: World Health Organization.
317. Wang, L., et al., *Whole-genome sequencing of Mycobacterium tuberculosis for prediction of drug resistance.* Epidemiology & Infection, 2022. **150**: p. e22.
318. Chun, Z.M. and J.Q. Jun, *Erratum: Drug-Resistant Mycobacterium tuberculosis Isolates from New and Previously Treated TB Patients in China, 2017-2019.* Rev. Soc. Bras. Med. Trop, 2022: p. e0728B-2022.
319. Menardo, F., et al., *The molecular clock of Mycobacterium tuberculosis.* PLoS pathogens, 2019. **15**(9): p. e1008067.
320. Organization, W.H., *WHO Global Task Force on TB Impact Measurement: report of a subgroup meeting on methods used by WHO to estimate TB disease burden, 11-12 May 2022, Geneva, Switzerland.* 2022.
321. Filliol, I., et al., *Snapshot of moving and expanding clones of Mycobacterium tuberculosis and their global distribution assessed by spoligotyping in an international study.* Journal of Clinical Microbiology, 2003. **41**(5): p. 1963-1970.
322. Glynn, J.R., et al., *Worldwide occurrence of Beijing/W strains of Mycobacterium tuberculosis: a systematic review.* Emerging infectious diseases, 2002. **8**(8): p. 843.

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

323. Hlanze, H., A. Mutshembele, and O.N. Reva, *Universal Lineage-Independent Markers of Multidrug Resistance in Mycobacterium tuberculosis.* Microorganisms, 2024. **12**(7): p. 1340.
324. Napier, G., et al., *Large-scale genomic analysis of Mycobacterium tuberculosis reveals extent of target and compensatory mutations linked to multi-drug resistant tuberculosis.* Scientific Reports, 2023. **13**(1): p. 623.
325. Ngabonziza, J.C.S., et al., *A sister lineage of the Mycobacterium tuberculosis complex discovered in the African Great Lakes region.* Nature communications, 2020. **11**(1): p. 2917.
326. Vågene, Å.J., et al., *Geographically dispersed zoonotic tuberculosis in pre-contact South American human populations.* Nature communications, 2022. **13**(1): p. 1195.
327. Silva, M.L., et al., *Tuberculosis caused by Mycobacterium africanum: Knowns and unknowns.* PLoS pathogens, 2022. **18**(5): p. e1010490.
328. Nebenzahl-Guimaraes, H., et al., *Genomic characterisation of Mycobacterium tuberculosis lineage 7 and a proposed name: 'Aethiops vetus'.* Microbial Genomics, 2016. **2**(6): p. e000063.
329. Coscolla, M., et al., *Phylogenomics of Mycobacterium africanum reveals a new lineage and a complex evolutionary history.* Microbial genomics, 2021. **7**(2): p. 000477.
330. Zwyer, M., et al., *Back-to-Africa introductions of Mycobacterium tuberculosis as the main cause of tuberculosis in Dar es Salaam, Tanzania.* PLoS pathogens, 2023. **19**(4): p. e1010893.
331. Wirth, T., et al., *Origin, spread and demography of the Mycobacterium tuberculosis complex.* PLoS pathogens, 2008. **4**(9): p. e1000160.
332. Gutierrez, M.C., et al., *Ancient origin and gene mosaicism of the progenitor of Mycobacterium tuberculosis.* PLoS pathogens, 2005. **1**(1): p. e5.
333. Bagcchi, S., *WHO's global tuberculosis report 2022.* The Lancet Microbe, 2023. **4**(1): p. e20.
334. Fujiwara, M., et al., *Mechanisms of resistance to delamanid, a drug for Mycobacterium tuberculosis.* Tuberculosis, 2018. **108**: p. 186-194.
335. Leinonen, R., et al., *The European nucleotide archive.* Nucleic acids research, 2010. **39**(suppl_1): p. D28-D31.
336. Xie, J., et al., *Tree Visualization By One Table (tvBOT): a web application for visualizing, modifying and annotating phylogenetic trees.* Nucleic acids research, 2023. **51**(W1): p. W587-W592.
337. Brown, T.S., et al., *Pre-detection history of extensively drug-resistant tuberculosis in KwaZulu-Natal, South Africa.* Proceedings of the National Academy of Sciences, 2019. **116**(46): p. 23284-23291.
338. Murray, G.G., et al., *The effect of genetic structure on molecular dating and tests for temporal signal.* Methods in Ecology and Evolution, 2016. **7**(1): p. 80-89.
339. Bouckaert, R., et al., *BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis.* PLoS computational biology, 2019. **15**(4): p. e1006650.
340. Organization, W.H., *Companion handbook to the WHO guidelines for the programmatic management of drug-resistant tuberculosis*. 2014: World Health Organization.

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

341. Snobre, J., et al., *Bedaquiline-and clofazimine-selected Mycobacterium tuberculosis mutants: further insights on resistance driven largely by Rv0678.* Scientific Reports, 2023. **13**(1): p. 10444.
342. Liao, J., et al., *Serotype-specific evolutionary patterns of antimicrobial-resistant Salmonella enterica.* BMC evolutionary biology, 2019. **19**: p. 1-20.
343. Ho, S.Y. and B. Shapiro, *Skyline-plot methods for estimating demographic history from nucleotide sequences.* Molecular ecology resources, 2011. **11**(3): p. 423-434.
344. Agerton, T.B., et al., *Spread of strain W, a highly drug-resistant strain of Mycobacterium tuberculosis, across the United States.* Clinical infectious diseases, 1999. **29**(1): p. 85-92.
345. Caminero, J.A., et al., *Epidemiological evidence of the spread of a Mycobacterium tuberculosis strain of the Beijing genotype on Gran Canaria Island.* American journal of respiratory and critical care medicine, 2001. **164**(7): p. 1165-1170.
346. Cowley, D., et al., *Recent and rapid emergence of W-Beijing strains of Mycobacterium tuberculosis in Cape Town, South Africa.* Clinical infectious diseases, 2008. **47**(10): p. 1252-1259.
347. Zong, Z., et al., *Comparison of in vitro activity and MIC distributions between the novel oxazolidinone delpazolid and linezolid against multidrug-resistant and extensively drug-resistant Mycobacterium tuberculosis in China.* Antimicrobial agents and chemotherapy, 2018. **62**(8): p. 10.1128/aac. 00165-18.
348. Azimi, T., et al., *Linezolid resistance in multidrug-resistant mycobacterium tuberculosis: a systematic review and meta-analysis.* Frontiers in pharmacology, 2022. **13**: p. 955050.
349. Zhou, Z., et al., *Evolution and epidemic success of Mycobacterium tuberculosis in eastern China: evidence from a prospective study.* BMC genomics, 2023. **24**(1): p. 241.
350. Zhou, Y., et al., *The epidemic of multidrug resistant tuberculosis in China in historical and phylogenetic perspectives.* Journal of Infection, 2020. **80**(4): p. 444-453.
351. Seifert, M., et al., *Genetic mutations associated with isoniazid resistance in Mycobacterium tuberculosis: a systematic review.* PloS one, 2015. **10**(3): p. e0119628.
352. Cox, H.S., et al., *Emergence of extensive drug resistance during treatment for multidrug-resistant tuberculosis.* New England Journal of Medicine, 2008. **359**(22): p. 2398-2400.
353. Ahmad, N., et al., *Treatment correlates of successful outcomes in pulmonary multidrug-resistant tuberculosis: an individual patient data meta-analysis.* The Lancet, 2018. **392**(10150): p. 821-834.
354. Packard, R.M., *Tuberculosis and the development of industrial health policies on the Witwatersrand, 1902–1932.* Journal of Southern African Studies, 1987. **13**(2): p. 187-209.
355. Lawn, S.D., et al., *Impact of HIV infection on the epidemiology of tuberculosis in a peri-urban community in South Africa: the need for age-specific interventions.* Clinical Infectious Diseases, 2006. **42**(7): p. 1040-1047.
356. Hermans, S., C.R. Horsburgh Jr, and R. Wood, *A century of tuberculosis epidemiology in the northern and southern hemisphere: the differential impact of control interventions.* PloS one, 2015. **10**(8): p. e0135179.

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

357. Karim, S.S.A., et al., *HIV infection and tuberculosis in South Africa: an urgent need to escalate the public health response.* the Lancet, 2009. **374**(9693): p. 921-933.
358. Hallett, T.B., et al., *Declines in HIV prevalence can be associated with changing sexual behaviour in Uganda, urban Kenya, Zimbabwe, and urban Haiti.* Sexually transmitted infections, 2006. **82**(suppl 1): p. i1-i8.
359. Aldridge, B.B., et al., *The Tuberculosis Drug Accelerator at year 10: what have we learned?* Nature medicine, 2021. **27**(8): p. 1333-1337.
360. Kana, M.A., R. LaPorte, and A. Jaye, *Africa's contribution to the science of the COVID-19/SARS-CoV-2 pandemic.* BMJ Global Health, 2021. **6**(3): p. e004059.
361. Żukowska, L., et al., *An overview of tuberculosis outbreaks reported in the years 2011–2020.* BMC infectious diseases, 2023. **23**(1): p. 253.
362. Van Soolingen, D., et al., *Predominance of a single genotype of Mycobacterium tuberculosis in countries of east Asia.* Journal of clinical microbiology, 1995. **33**(12): p. 3234-3238.
363. Van Crevel, R., et al., *Mycobacterium tuberculosis Beijing genotype strains associated with febrile response to treatment.* Emerging infectious diseases, 2001. **7**(5): p. 880.
364. Mathema, B., et al., *Molecular epidemiology of tuberculosis: current insights.* Clinical microbiology reviews, 2006. **19**(4): p. 658-685.
365. Middelkoop, K., et al., *Molecular epidemiology of Mycobacterium tuberculosis in a South African community with high HIV prevalence.* The Journal of infectious diseases, 2009. **200**(8): p. 1207-1211.
366. Viegas, S.O., et al., *Mycobacterium tuberculosis Beijing genotype is associated with HIV infection in Mozambique.* PLoS One, 2013. **8**(8): p. e71999.
367. Luo, T., et al., *Southern East Asian origin and coexpansion of Mycobacterium tuberculosis Beijing family with Han Chinese.* Proceedings of the National Academy of Sciences, 2015. **112**(26): p. 8136-8141.
368. van Helden, P.D., et al., *Strain families of Mycobacterium tuberculosis.* TRENDS in Microbiology, 2002. **10**(4): p. 167-168.
369. Gona, P.N., et al., *Burden and changes in HIV/AIDS morbidity and mortality in Southern Africa Development Community Countries, 1990–2017.* BMC public health, 2020. **20**: p. 1-14.
370. Community, S.A.D., *Report of the Executive Secretary: Overview of the SADC Region's Performance on Regional Integration.* 2022-2023.
371. Faksri, K., et al., *In silico region of difference (RD) analysis of Mycobacterium tuberculosis complex from sequence reads using RD-Analyzer.* BMC genomics, 2016. **17**: p. 1-10.
372. Meehan, C.J., et al., *Whole genome sequencing of Mycobacterium tuberculosis: current standards and open issues.* NATURE reviews microbiology, 2019. **17**(9): p. 533-545.
373. Darriba, D., et al., *jModelTest 2: more models, new heuristics and parallel computing. Nat Meth 9: 772.* Go to original source, 2012.
374. Drummond, A.J., et al., *Relaxed phylogenetics and dating with confidence.* PLoS biology, 2006. **4**(5): p. e88.

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

375. Waddell, P.J. and M. Steel, *General time-reversible distances with unequal rates across sites: mixing Γ and inverse Gaussian distributions with invariant sites.* Molecular phylogenetics and evolution, 1997. **8**(3): p. 398-414.
376. Müller, N.F., D. Rasmussen, and T. Stadler, *MASCOT: parameter and state inference under the marginal structured coalescent approximation.* Bioinformatics, 2018. **34**(22): p. 3843-3848.
377. Newitt, M., *A History of Portuguese overseas expansion 1400-1668*. 2004: Routledge.
378. Frith, J., *History of tuberculosis. Part 1-phthisis, consumption and the white plague.* Journal of Military and Veterans Health, 2014. **22**(2): p. 29-35.
379. Daniel, T.M., *The history of tuberculosis.* Respiratory medicine, 2006. **100**(11): p. 1862-1870.
380. Barberis, I., et al., *The history of tuberculosis: from the first historical records to the isolation of Koch's bacillus.* Journal of preventive medicine and hygiene, 2017. **58**(1): p. E9.
381. Spencer, S.C., *British Liberty Stained:" Chinese Slavery," Imperial Rhetoric, and the 1906 British General Election.* Madison Historical Review, 2010. **7**(1): p. 3.
382. Crush, J., V. Williams, and S. Peberdy, *Migration in southern Africa.* Policy analysis and research programme of the Global Commission on International Migration, 2005.
383. Cobley, A., *"Lacking in Respect for Whitemen":"Tropical Africans" on the Witwatersrand Gold Mines, 1903–1904.* International Labor and Working-Class History, 2014. **86**: p. 36-54.
384. Baleta, A., *Southern African declaration targets TB in mining sector.* The Lancet, 2012. **380**(9849): p. 1217-1218.
385. Zhu, C., et al., *The global success of Mycobacterium tuberculosis modern Beijing family is driven by a few recently emerged strains.* Microbiology spectrum, 2023. **11**(4): p. e03339-22.
386. Organization, W.H., *Fourth WHO consultation on the translation of tuberculosis research into global policy guidelines: meeting report, 15 February 2024*. 2024: World Health Organization.
387. Organization, W.H., *WHO Bacterial Priority Pathogens List, 2024.* World Health Organization, 2024.
388. Coll, F., et al., *Genome-wide analysis of multi-and extensively drug-resistant Mycobacterium tuberculosis.* Nature genetics, 2018. **50**(2): p. 307-316.
389. Wang, F., et al., *Evolution and transmission patterns of extensively drug-resistant tuberculosis in China.* Antimicrobial agents and chemotherapy, 2015. **59**(2): p. 818-825.
390. Perdigao, J. and I. Portugal, *Genetics and roadblocks of drug resistant tuberculosis.* Infection, Genetics and Evolution, 2019. **72**: p. 113-130.
391. Casali, N., et al., *Evolution and transmission of drug-resistant tuberculosis in a Russian population.* Nature genetics, 2014. **46**(3): p. 279-286.
392. Gygli, S.M., et al., *Antimicrobial resistance in Mycobacterium tuberculosis: mechanistic and evolutionary perspectives.* FEMS microbiology reviews, 2017. **41**(3): p. 354-373.

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

393. Farhat, M.R., et al., *Genomic analysis identifies targets of convergent positive selection in drug-resistant Mycobacterium tuberculosis.* Nature genetics, 2013. **45**(10): p. 1183-1189.
394. Walker, T.M., et al., *Whole-genome sequencing for prediction of Mycobacterium tuberculosis drug susceptibility and resistance: a retrospective cohort study.* The Lancet infectious diseases, 2015. **15**(10): p. 1193-1202.
395. Nguyen, Q.H., et al., *Insights into the processes that drive the evolution of drug resistance in Mycobacterium tuberculosis.* Evolutionary applications, 2018. **11**(9): p. 1498-1511.
396. Salvatore, P.P., et al., *Fitness costs of drug resistance mutations in multidrug-resistant Mycobacterium tuberculosis: a household-based case-control study.* The Journal of infectious diseases, 2016. **213**(1): p. 149-155.
397. Huo, F., et al., *A 10-year comparative analysis shows that increasing prevalence of rifampin-resistant Mycobacterium tuberculosis in China is associated with the transmission of strains harboring compensatory mutations.* Antimicrobial Agents and Chemotherapy, 2018. **62**(4): p. 10.1128/aac. 02303-17.
398. Cohen, K.A., et al., *Evolution of extensively drug-resistant tuberculosis over four decades: whole genome sequencing and dating analysis of Mycobacterium tuberculosis isolates from KwaZulu-Natal.* PLoS medicine, 2015. **12**(9): p. e1001880.
399. Vargas, A.P., et al., *Determination of potentially novel compensatory mutations in rpoc associated with rifampin resistance and rpob mutations in Mycobacterium tuberculosis Clinical isolates from peru.* The International Journal of Mycobacteriology, 2020. **9**(2): p. 121-137.
400. Al-Saeedi, M. and S. Al-Hajoj, *Diversity and evolution of drug resistance mechanisms in Mycobacterium tuberculosis.* Infection and drug resistance, 2017: p. 333-342.
401. Cohen, K.A., et al., *Deciphering drug resistance in Mycobacterium tuberculosis using whole-genome sequencing: progress, promise, and challenges.* Genome medicine, 2019. **11**(1): p. 45.
402. Shea, J., et al., *Low-level rifampin resistance and rpoB mutations in Mycobacterium tuberculosis: an analysis of whole-genome sequencing and drug susceptibility test data in New York.* Journal of clinical microbiology, 2021. **59**(4): p. 10.1128/jcm. 01885-20.
403. Gagneux, S., *Strain variation in the Mycobacterium tuberculosis complex: its role in biology, epidemiology and control*. Vol. 1019. 2017: Springer.
404. Li, Q.-j., et al., *Positive epistasis of major low-cost drug resistance mutations rpoB531-TTG and katG315-ACC depends on the phylogenetic background of Mycobacterium tuberculosis strains.* International journal of antimicrobial agents, 2017. **49**(6): p. 757-762.
405. Fenner, L., et al., *Effect of mutation and genetic background on drug resistance in Mycobacterium tuberculosis.* Antimicrobial agents and chemotherapy, 2012. **56**(6): p. 3047-3053.
406. Consortium, C., *A data compendium associating the genomes of 12,289 Mycobacterium tuberculosis isolates with quantitative resistance phenotypes to 13 antibiotics.* PLoS biology, 2022. **20**(8): p. e3001721.

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

407. Fleiss, J.L., B. Levin, and M.C. Paik, *Statistical methods for rates and proportions*. 2013: john wiley & sons.
408. Schito, M. and D.L. Dolinger, *A collaborative approach for "ReSeq-ing" Mycobacterium tuberculosis drug resistance: convergence for drug and diagnostic developers.* EBioMedicine, 2015. **2**(10): p. 1262-1265.
409. Napier, G., et al., *Comparison of in silico predicted Mycobacterium tuberculosis spoligotypes and lineages from whole genome sequencing data.* Scientific Reports, 2023. **13**(1): p. 11368.
410. Kapopoulou, A., J.M. Lew, and S.T. Cole, *The MycoBrowser portal: a comprehensive and manually annotated resource for mycobacterial genomes.* Tuberculosis, 2011. **91**(1): p. 8-13.
411. Shannon, P., et al., *Cytoscape: a software environment for integrated models of biomolecular interaction networks.* Genome research, 2003. **13**(11): p. 2498-2504.
412. Dash, P., et al., *Three-dimensional models of Mycobacterium tuberculosis proteins Rv1555, Rv1554 and their docking analyses with sildenafil, tadalafil, vardenafil drugs, suggest interference with quinol binding likely to affect protein's function.* BMC structural biology, 2018. **18**: p. 1-13.
413. Abendroth, J., et al., *Mycobacterium tuberculosis Rv3651 is a triple sensor-domain protein.* Protein Science, 2018. **27**(2): p. 568-572.
414. Miotto, P., et al., *Transcriptional regulation and drug resistance in Mycobacterium tuberculosis.* Frontiers in Cellular and Infection Microbiology, 2022. **12**: p. 990312.
415. Kieser, K.J., et al., *Peptidoglycan synthesis in Mycobacterium tuberculosis is organized into networks with varying drug susceptibility.* Proceedings of the National Academy of Sciences, 2015. **112**(42): p. 13087-13092.
416. Morris, R.P., et al., *Ancestral antibiotic resistance in Mycobacterium tuberculosis.* Proceedings of the National Academy of Sciences, 2005. **102**(34): p. 12200-12205.
417. De Vos, M., et al., *Putative compensatory mutations in the rpoC gene of rifampin-resistant Mycobacterium tuberculosis are associated with ongoing transmission.* Antimicrobial agents and chemotherapy, 2013. **57**(2): p. 827-832.
418. Tasbiti, A.H., et al., *Comparing mRNA expression and protein abundance in MDR Mycobacterium tuberculosis: Novel protein candidates, Rv0443, Rv0379 and Rv0147 as TB potential diagnostic or therapeutic targets.* Biotechnology Reports, 2021. **30**: p. e00641.
419. Loarer, G., *The Approach to Characterizing Three S-Adenosyl-L-Methionine-Dependent Methyltransferases from Mycobacterium tuberculosis*. 2018: The University of Toledo.
420. Van der Spuy, G., et al., *Changing Mycobacterium tuberculosis population highlights clade-specific pathogenic characteristics.* Tuberculosis, 2009. **89**(2): p. 120-125.
421. Glynn, J.R., et al., *Changes in Mycobacterium tuberculosis genotype families over 20 years in a population-based study in Northern Malawi.* Plos one, 2010. **5**(8): p. e12259.
422. Klopper, M., et al., *Emergence and spread of extensively and totally drug-resistant tuberculosis, South Africa.* Emerging infectious diseases, 2013. **19**(3): p. 449.
423. Githui, W., et al., *Identification of MDR-TB Beijing/W and other Mycobacterium tuberculosis genotypes in Nairobi, Kenya.* The international journal of tuberculosis and lung disease, 2004. **8**(3): p. 352-360.

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

424. de Jong, B.C., et al., *Progression to active tuberculosis, but not transmission, varies by Mycobacterium tuberculosis lineage in The Gambia.* The Journal of infectious diseases, 2008. **198**(7): p. 1037-1043.
425. Cox, H., et al., *Drug-resistant tuberculosis in South Africa: history, progress and opportunities for achieving universal access to diagnosis and effective treatment.* South African Health Review, 2017. **2017**(1): p. 157-167.
426. Comas, I., et al., *Population genomics of Mycobacterium tuberculosis in Ethiopia contradicts the virgin soil hypothesis for human tuberculosis in Sub-Saharan Africa.* Current Biology, 2015. **25**(24): p. 3260-3266.
427. Cummins, S.L., *"VIRGIN SOIL"—AND AFTER: A WORKING CONCEPTION OF TUBERCULOSIS IN CHILDREN, ADOLESCENTS, AND ABORIGINES.* British Medical Journal, 1929. **2**(3575): p. 39.
428. Collins, T.F., *The history of southern Africa's first tuberculosis epidemic.* South African Medical Journal, 1982. **62**(21): p. 780-788.
429. Eldholm, V., et al., *Armed conflict and population displacement as drivers of the evolution and dispersal of Mycobacterium tuberculosis.* Proceedings of the National Academy of Sciences, 2016. **113**(48): p. 13881-13886.
430. Maloka, T., *Mines and labour migrants in Southern Africa.* Journal of historical sociology, 1997. **10**(2): p. 213-224.
431. Khoza, N., E. Tulisha, and S. Grey-Johnson, *Multi-concerted efforts to eliminate tuberculosis, silicosis, and other occupational lung diseases: legacies in the southern African region and policy reform efforts.* Occupational Health Southern Africa, 2024. **30**(1): p. 33-35.
432. Osewe, P.L. and B. Kistnasamy, *Tuberculosis Must Fall!: A Multisector Partnership to Address TB in Southern Africa's Mining Sector*. 2018: World Bank Publications.
433. Chirenda, J., *Migration and spread of drug resistant tuberculosis (DRTB) in Zimbabwe*. 2020, Stellenbosch: Stellenbosch University.
434. Dara, M., et al., *Minimum package for cross-border TB control and care in the WHO European region: a Wolfheze consensus statement.* European Respiratory Journal, 2012. **40**(5): p. 1081-1090.
435. Philalay, J.S., et al., *Genes required for intrinsic multidrug resistance in Mycobacterium avium.* Antimicrobial agents and chemotherapy, 2004. **48**(9): p. 3412-3418.

© University of Pretoria

© University of Pretoria

CHAPTER 8 APPENDIX A

8.1.1 Data and code availability

The supplementary data and code supporting this study can also be found on the UP research data repository under the following DOI: 10.25403/UPresearchdata.28417427 **. Below are the descriptions of the available resources.**

8.1.1.1 Chapter 2

- Supplementary File 2.1
  Description: Excel file with metadata of 9073 global Mtb isolates.
- Supplementary File 2.3
  Description: Text file with Unix commands and code used in this project.
- Supplementary file 2.2
  Description: BED file with LCRs that were excluded in phylogenetic inference
- DRM file
  Description: Text file with a catalogue of resistance mutations.
- Global.fasta
  Description: Snp alignment file
- Global.nwk
  Description: Tree file of global reference isolates in newick format

© University of Pretoria

8.1.1.2 Chapter 3

- Supplementary File 3.1:

  Description: Excel file with metadata on 6,058 African Mtb isolates.
- Supplementary File 3.2

  Description: Excel file with 5745 isolates after QC.
- Supplementary File 3.3

  Description: Excel file with 2168 isolates
- Supplementary File 3.4

  Description: Microsoft word file with BSP results for L1, L3 and sublineages of L4
- Africa.fasta

  Description: SNP alignment file
- Africa.tree

  Description: Tree file with the 5745 isolates
- (File).xml

  Description: BEAST2 file where (File) will be the name describing the exact data, eg Lineage2.xml

8.1.1.3 Chapter 4

- Supplementary File 4.1

  Description: Excel file with metadata on all isolates included in this chapter
- Supplementary file 4.2

  Description: JSON auspice visualization file for lineage 4 isolates

8.1.1.4 Chapter 5

- Supplementary File 5.1

Description: Excel file with metadata of all isolates analysed in chapter 5

© University of Pretoria

© University of Pretoria