Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Identification of fitness-defining drug resistance mutations in *Mycobacterium tuberculosis*

Master Thesis

Katerina Pratsinis
pratsink@ethz.ch

Department of Biosystems Science and Engineering
Computational Evolution
ETH Zürich

Supervisors:
Prof. Dr. Tanja Stadler
Dr. Etthel Windels

August 16, 2023

Acknowledgements

First and foremost, I would like to express my gratitude to my supervisor, Dr. Etthel Windels. Her insights and expertise offered tremendous support while navigating through my project and her constant guidance helped me progress this project meaningfully.
I am also particularly thankful to Dr. Bethany Allen and Cecilia Valenzuela Agüi, whose weekly peer feedback sessions greatly improved the quality of my work by pointing me toward useful software tools that eased many challenges I faced along the way.
My appreciation extends to Prof. Dr. Tanja Stadler and the entire cEVO team. Returning to an in-person setting for the first time since the beginning of my Master's program significantly enhanced my learning experience by working in an environment where everyone is interested in sharing their expertise. This contributed to an inspiring workplace and the success of this study.

Abstract

Tuberculosis (TB), primarily affecting the lungs, is an airborne infectious disease that can lead to death if left untreated. Standard TB is treated through combination therapy with four antibiotics. In recent years, it has been observed that genetic mutations in *Mycobacterium tuberculosis (Mtb)* render the bacteria resistant to first-line drugs, leading to drug-resistant TB (DR-TB). One of the drivers of these mutations is ineffective treatment due to the lack of information about a patient's drug susceptibility. Understanding which mutations cause DR-TB can lead to better TB control through early detection and effective treatment. This study investigates the fitness effects of individual DR *Mtb* genomes to improve our understanding of TB dynamics. An analysis of 1,134 genomes from Khayelitsha, South Africa (2008-2018), and 4,772 genomes from Georgia (2011-2021) provides insights into the disease's evolutionary trajectory. The methodology entailed phylogenetic tree inference, ancestral state reconstruction of specific mutations, and the use of a previously published computational framework, PhyloTensorFlow (phyloTF). PhyloTF is a maximum-likelihood approach to model the evolutionary dynamics of a structured population as a series of single-type birth-death processes. Detailed inspection of fitness estimates for the prevalent *rpoB* S450L revealed varying effects across data sets, suggesting a different base transmission dependent on the study population. The mutation *katG* S315T is common and estimated to be deleterious in all data sets. Other mutations with a relative fitness effect of > 1 were typically linked to one of these two mutations, compensating fitness costs from DR. One mutation in *Mtb* L4 in South Africa showed a high fitness estimate without being linked to any other DR mutation. The mutation is Q497R in the *embB* gene and has been associated with resistance toward the second-line drug ethambutol (EMB).
While phyloTF provides an efficient means of inferring the fitness effects of numerous mutations on large phylogenetic trees, it does not factor in uncertainties inherent to preliminary steps like tree inference or ancestral state reconstruction. Further, the model showed high sensitivity toward different birth, death and sampling rates. In future studies, incorporating a marginal fitness birth-death model within a Bayesian framework might offer improved accuracy.

Contents

Chapter 1

Introduction

1.1 Background on *Mycobacterium tuberculosis*

Mycobacterium tuberculosis (Mtb) is the causative agent of one of the oldest known and most lethal human infectious diseases [1]. It is classified into multiple distinct families of strains, also referred to as lineages [2]. L5 and L6, belonging to *M. africanum* (Maf), are only found in western Africa [3]. The recently identified L9 is very rare and closely related to L6 but has only been found in eastern Africa [4, 5]. Lineages 1, 2, 3, 4, 7, and the recently discovered lineage 8 belong to *M. tuberculosis* sensu stricto [3]. While L1 and L3 are mostly found around the Indian Ocean, L5 and L6 are confined to West Africa and L7 is only found in East Africa. In contrast, L2 and L4 are widespread globally and the most virulent compared to the other lineages [6]. Human-adapted *Mtb* has no known environmental or animal reservoir and thus infection is always through host-to-host transmission [3]. There are, however, related animal-adapted strains such as *M. bovis* in cattle and *M. caprae* in goats and sheep [3]. While there have been reported spillover events causing 1-3% of human TB cases, the transmission of these animal-adapted strains from human to human is rare [7].
The bacteria typically spread through airborne droplets released by individuals with active tuberculosis (TB). *Mtb* transmission depends on multiple factors such as the duration and proximity of exposure, pathogen virulence, and the susceptibility of an exposed person [8]. While TB typically affects the lungs (pulmonary TB), it can also impact other organs (extrapulmonary TB). Upon infection of a healthy individual, the bacteria begin to replicate in the alveoli of the lungs for the first few weeks until the person develops T-cell immunity, and thus, bacterial growth decreases [9]. *Mtb* has developed mechanisms to avoid immediate destruction and can survive within this hostile environment while remaining in a dormant state as a latent *Mtb* infection (LTBI) [8, 9]. Nearly a quarter of the global population is estimated to have LTBI, without disease onset [10]. Ten percent of latent infections are believed to lead to active TB disease, resulting in an estimated 10 million active cases every year, however many of these cases are not reported [11]. Active TB disease can present itself through various symptoms such as fever, fatigue, a persistent cough, or even coughing up blood [12].

Furthermore, with an annual death toll of over 1 million, TB is once again the leading cause of death among infectious diseases, after COVID-19 had become the most lethal infectious disease during 2020-2021 [11].
Active TB can be diagnosed using various techniques such as X-ray or PET-CT imaging, microscopy, or molecular tests [12]. In low-income and middle-income countries, where over 90% of TB patients live [13], sputum smear microscopy is the most widely used diagnostic method [12]. Due to its low sensitivity, the golden standard is to perform two tests on different days [14]. This method does not test for drug susceptibility. The introduction of the molecular assay Xpert MTB/RIF [15] to the market has the potential to improve TB diagnosis. This fully-automated Nucleic Acid Amplification Test has high sensitivity and can be used with sputum or other bodily specimens. In only two hours, it can confirm an active TB disease and detect resistance to rifampicin (RIF) [16].
With the rise of antibiotic resistance, it is increasingly more important to test for susceptibility to multiple drugs. In 2014, the World Health Assembly endorsed the World Health Organization (WHO)'s ambitious End TB Strategy, which aims to reduce TB deaths by 95% and cut new cases by 90% by 2035. Their vision includes the universal accessibility of multi-drug-susceptibility testing and highlights the need for intensified research towards TB interventions [11].

1.1.1 Tuberculosis Treatment

Treatment for TB typically involves a combination of antibiotics administered over several months. The standard regimen for drug-susceptible (DS) TB (DS-TB) patients consists of a two-month-long induction phase with RIF, isoniazid (INH), and pyrazinamide (PZA). Initially, EMB is also administered until the patient's resistance towards any of the former three antibiotics is excluded. Following this is a four-month-long consolidation phase with RIF and INH which concludes the so-called short-course regimen [17, 18].
In patients whose treatment proves effective, the induction phase showcases a biphasic kill curve in their sputum samples' bacteria, indicating at least two bacterial subpopulations with distinct treatment responses [19]. One subpopulation exhibits a quick reaction, while the other responds more gradually to the antibiotics. This notable observation underscores the theory that the efficacy of combination therapy, involving multiple antimycobacterial agents, can be attributed to these individual bacterial subpopulations. Despite most patients showing no signs of bacilli in their sputum after this initial treatment state, the consolidation phase remains a critical step in averting relapses. It has not yet been fully understood in which cases this is necessary, thus many patients undergo overtreatment to guarantee a cure [17].
In the case of DS infection, antituberculosis regimens have a nearly perfect treatment efficacy with only a 0% - 7% relapse rate in the subsequent two years [20]. Even intermittent regimens indicate similar results with a lower proportion of

adverse drug reactions (ADRs) [21, 22]. However, these long treatment periods impose multiple challenges. For one, the risk of drug toxicity can lead to ADRs and hence to an interruption or end of treatment. One cohort study observed at least one ADR in 15% of its cases, of which 7.7% were hospitalized, disabled, or died [23]. The most commonly reported side effects were liver dysfunction, gastrointestinal disorders, and allergic reactions. Secondly, its success is highly dependent on the patient's consistent adherence to the regimen. Among other reasons, this can be jeopardized by the high cost of treatment, stigma, or premature stopping of medication as soon as symptoms subside. Along with inadequate dosages, interruptions in antibacterial treatment regimens increase the selective pressure for drug-resistant (DR) mutants and are the main reason for their occurrence [19, 24]. In 1995, to counteract patient negligence, WHO launched DOTS – Directly Observed Treatment, Short-course, in which health workers at home or in a clinic, community volunteers, or family members watch a patient swallow every prescribed TB medication of their standard short-course regimen [18]. This strategy has had varying impacts [21, 25, 26].

1.1.2 Drug-resistant TB

DS-TB treatment already presents significant challenges, and these become even more grueling when dealing with DR infections. A standard six-month treatment can be adapted for INH monoresistance by substituting INH with a later-generation fluoroquinolone such as levofloxacin (LEV) or moxifloxacin (MXF) [17]. Resistance to INH typically arises from mutations in the *katG* or *inhA* genes [27, 28]. RIF-resistant TB is associated with mutations in the *rpoB* gene and involves alterations of RNA polymerase [29]. When TB displays resistance to both INH and RIF, it is classified as multidrug-resistant TB (MDR-TB) [30]. In the case where MDR-TB is resistant to a fluoroquinolone and at least one of three second-line drugs (amikacin (AMI), capreomycin (CAP), kanamycin (KAN)), it is classified as extensively drug-resistant (XDR) [31]. The recommended treatment for MDR-TB, previously lasting up to 20 months or longer, can now be reduced to six months thanks to the development of novel all-oral regimens [11]. Nonetheless, treatment of MDR-TB and XDR-TB remains a complex task due to the need for individualized treatment plans with drugs based on a patient's drug-susceptibility test, coupled with the increased likelihood of ADRs from the drugs.

The principal factors driving the spread of resistant TB are improper treatment resulting in amplification of resistance patterns, mutations during treatment, and transmission within communities [31, 32]. In populations with high TB incidence, transmission has been suggested to be the most likely cause of high rates of MDR-TB.

Initial predictions that DR mutations would carry a fitness cost, resulting in

reduced virulence and transmissibility, suggested that MDR-TB, while still a serious concern in affected populations, would remain a local issue [33]. However, the increased incidence rate over the past years contradicts this theory [34]. The factors underlying the spread of MDR-TB are likely far more complex and suggest that certain DR strains have the potential to outperform less fit DS strains [35]. The presence of secondary mutations is suggested, which compensate for this reduced fitness [36]. Compensatory mutations for RIF resistance conferring mutations have been identified in *rpoA*, *rpoB*, *rpoC*, where they encode the α, β and β' subunits of RNA polymerase [37]. Therefore, identifying MDR strains with increased transmission rates and employing targeted genotyping could bolster TB control through early detection and effective treatment.

1.1.3 Impact of COVID-19 on TB control

The COVID-19 pandemic has significantly disrupted global healthcare systems, introducing substantial challenges to the management and control of co-existing infectious diseases, especially TB [38]. As healthcare resources globally were diverted toward the COVID-19 crisis, routine TB care and prevention services were severely impacted. The extent of these disruptions ranges from interruption of TB diagnostic services and limited treatment access to slowed down progress in research, as well as diminished funding and political attention towards TB [39]. This intricate scenario created formidable obstacles for TB control, jeopardizing the progress toward WHO's End TB Strategy. Following a peak in reported new TB cases of 7.1 million in 2019, the pandemic's most immediate impact was a decrease to 5.8 million in 2020. Although a marginal recovery to 6.4 million reported cases occurred in 2021, the under-diagnosis during the pandemic's initial stages suggests a surge in undetected and therefore untreated TB cases [11]. Annual TB incidence for the years 2020-2021 is estimated to surpass 10 million and is accompanied by an increased estimate of DR-TB [11]. Besides the decrease in TB diagnoses, the number of patients treated for MDR-TB and global investment in essential TB services also suffered due to the pandemic [11].
In some circumstances, nations with pre-existing severe healthcare challenges were quick to respond to the pandemic, demonstrating the ability to efficiently reallocate their resources. A notable example is South Africa, which was able to use its robust infrastructure and experienced healthcare workforce for TB and HIV management to its advantage, redirecting efforts to tackle the COVID-19 crisis [40]. However, this success was also accompanied by a setback in the country's TB services and research and highlights the need for resilient healthcare systems.

1.1.4 Tuberculosis in South Africa

South Africa ranks among the nations with the highest TB incidence rates (615 cases per 100,000 population in 2019) [41]. Historical analysis suggests TB first emerged in South Africa in the 16th century upon the arrival of European immigrant workers [42]. Over the 20th century, incidence rose, exacerbated by the constraints of the apartheid era on the public health sector, thereby intensifying the country's susceptibility to the epidemic.
The introduction of DOTS bolstered the nation's TB control program. Nonetheless, the healthcare system struggled with an upsurge of HIV cases which subsequently fueled a steep rise in reported TB cases during the 1990s [43]. Since the start of the epidemic of HIV, there has been a shift in the location of the highest TB rates toward regions harboring high HIV incidence rates. One such example is Khayelitsha – a poor, peri-urban subdistrict in Cape Town that is severely impacted by poverty and high unemployment rates [44]. In 2006, this population of over half a million recorded an HIV prevalence of 33% among antenatal clinic patients, and in 2008, nearly 6,000 individuals were reported to have TB [44].
It has been suggested that higher TB notification rates associate with older age groups, however, there has been a discernable shift towards patients in their thirties, mirroring the age distribution of HIV infection. The consequences of poor control measures and low cure rates have resulted in drug resistance. A notable example is the 2006 outbreak in a KwaZulu-Natal hospital, where, alarmingly, 125 out of 183 confirmed TB patients were diagnosed with either MDR-TB or XDR-TB [45]. The outcome was generally fatal for those with XDR-TB, with each patient tested being HIV-positive. What was initially perceived as a local event has now escalated into a nationwide epidemic with surging DR-TB cases. Consequently, TB has emerged as the leading natural cause of death in South Africa [46].
The increasing rates of MDR-TB and XDR-TB have prompted increased research into the underlying genetic mutations. A study on TB strains in the Eastern Cape province identified that in RIF-resistant strains, in gene *rpoB*, S450L is the most common mutation, and the most prevalent INH resistance-conferring mutations are S315T and C15T in the *katG* and *inhA* genes respectively [47]. Another study applied Bayesian transmission tree inference methods and multivariate regression analysis to identify factors associated with RIF resistance *Mtb* strains [48]. The two most common lineages in the studied population, in Cape Town, South Africa, are L2 and L4. Compensatory mutations were found to coincide with an increased transmission rate of RIFresistant strains between hosts. Such insights, in combination with targeted genotype testing, hold the potential to immensely facilitate decisions on the treatment course in areas that are highly impacted by DR-TB.

1.1.5 Tuberculosis in Georgia

Georgia's reported cases of DR-TB in 2019 place it in ninth place among countries with the highest incidence rates for both MDR-TB (18.12% of TB cases) and XDR-TB (3.97% of TB cases) [49, 50]. Georgia's struggle with DR-TB is rooted in the country's history linked to the socioeconomic changes it underwent during the late 20th century.

The dissolution of the Soviet Union in 1991 led Georgia, among other Caucasian republics, into a period of economic, political, and social instability during its transition to a market economy [51]. These challenges had a profound impact on the country's health infrastructure, resulting in significant deterioration of public health services. In addition to the country's TB challenges due to socio-economic changes, it was found that most MDR-TB cases in Georgia are due to the ongoing transmission of MDR *Mtb* strains, particularly within the prison system [52]. This environment, marked by overcrowding, contributes to the high transmission rates and serves as a key driver of MDR-TB [52]. Furthermore, MDR *Mtb* frequently spills over from prisoners into the general public, resulting in up to 31% of MDR-TB cases in Georgia originating from detention centers and underlining the critical need for improved infection control measure [37].

In response to this healthcare crisis, WHO-supported and DOTS-based pilot projects were introduced to make TB control an integral part of Georgia's existing primary healthcare system. However, these pilot projects only had limited success resulting in high death and failure rates. These outcomes suggest a prevalence of MDR-TB in the region [51]. DOTS can prevent the development of MDR-TB as it improves treatment regimens and compliance with treatment. However, in regions where MDR-TB is already widespread due to transmission, additional interventions are necessary.

Recent studies have provided insights into the intricate dynamics of MDR-TB in Georgia, with bacterial genetics becoming a focal point. This involves the study of interactions between DR mutations, compensatory mutations, and other strain genetic backgrounds [35]. A comparison of strains from L2 with those of L4 revealed that L4 MDR strains transmit less than their DS counterparts. In contrast, MDR L2 strains appear to maintain their transmission fitness compared to DS strains. L2 in Georgia primarily consists of the two sublineages of 'Central Asia' and 'W149'. Their resilience towards drugs is thought to be a result of the RIF resistance mutation *RpoB* S450L, alongside compensatory mutations in the RNA polymerase that offset the fitness loss due to the DR mutation. A separate study underlines the alarming prevalence of PZA resistance among TB isolates in Georgia [53]. Out of 57 MDR-TB tested isolates, 57.9% exhibit phenotypic DR to PZA, raising serious concerns about the effectiveness of current and planned MDR-TB treatment regimens incorporating PZA.

Georgia faces a significant challenge in combatting MDR-TB, stemming from both genetic and socio-economic factors. Future efforts should prioritize the early identification of DR-TB cases and a detailed understanding of the genomic

factors driving this resistance. With this knowledge, treatment regimens can be continually updated to effectively manage and eventually eradicate MDR-TB in the region.

1.2 Modeling TB with Birth-Death Processes

The birth-death (BD) model is fundamental in phylodynamics and describes birth (proliferation) and death (extinction) rates of entities within a population [54]. Modeling a defined set of infected individuals $I(t)$ at time t as a compartment, each individual in that compartment has identical birth rates λ, death rates μ and sampling rates σ [54]. The sampling rate σ denotes the rate at which individuals from a compartment are sampled. In the context of *Mtb*, the birth rate signifies the transmission rate, and the death rate refers to a patient's recovery or death from TB. The trajectory of the population over time t is:

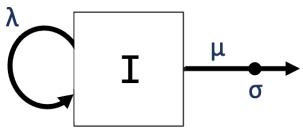

Figure 1.1: The constant rate BD model is represented by a single compartment of a population with a birth rate λ, death rate μ -and sampling rate σ.

$$I(t) = I(0) \cdot e^{(\lambda-\mu)t} \tag{1.1}$$

Equal rates for λ and μ lead to a constant population size, higher transmission rates yield exponential growth of the infected population, and higher death rates result in declining population size. In phylodynamics, this model describes the population dynamics of sampled entities on a phylogenetic tree, where each branch represents an individual [54]. The model typically starts with a single individual and concludes after a time T [54]. It portrays biological population dynamics and the evolutionary trajectory of genetic sequences. Here, the phylogenetic tree represents the shared ancestry and divergence of sequences [54].

1.2.1 Structured Birth-Death Model

In the presence of rate heterogeneity or structured population dynamics leading to differential transmission probabilities between individual pairs, a structured BD model is required. This model, also referred to as a multi-type BD (MTBD) model, partitions the population into discrete compartments, with each one characterized by its unique birth, death, and sampling rates [55, 56]. Additionally, migration rates γ describe how individuals transition between these compartments [54]. In TB, the MTBD model can reflect infections by DS and DR *Mtb*

strains and the two possible scenarios for the propagation of the DR strain (see Fig. 1.2): 1) DR-TB transmission from one patient to another, or 2) *de novo* DR, where DR-TB repeatedly arises through mutation in already infected individuals [54]. Structured BD models assume non-neutral evolution. When different com-

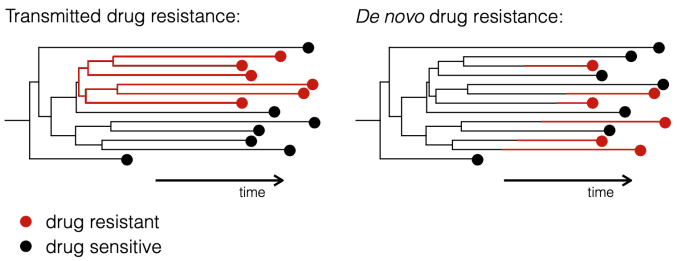

Figure 1.2: The tree on the left represents a phylogeny where drug resistance is transmitted in contrast to the scenario where drug resistance evolves on multiple occasions as illustrated on the right [54].

partments represent different mutant strains, this means that selection depends on the mutational process. MTBD models can offer a more comprehensive perspective, as they do not just capture genetic sequences but also the significance of other heritable traits, including spatial distribution, and phenotypic characteristics [54]. They accurately represent rapidly mutating entities like bacteria, since mutations significantly determine fitness variations among lineages. This variation subsequently impacts the branching structure of the phylogeny. The likelihood of a tree within this framework is based on a set of ordinary differential equations (ODEs) dependent on parameters λ, μ, σ, γ, and the state of a discrete character trait [56]. The computational realization of this model becomes increasingly complex as the number of features grows, complicating the estimation of site-specific effects for multi-site genotypes. Below, I describe two published models that approximate MTBD models in a computationally tractable manner.

1.2.2 Marginal Fitness Birth-Death Model

Addressing the computational constraints of MTBD models, Rasmussen and Stadler [57] presented the marginal fitness BD (MFBD) model. It assumes that the transmission rate of lineage n, denoted by λ_n, is proportional to the fitness f_g of its genotype g. The base transmission rate is given by λ_0.

$$\lambda_n = f_g \lambda_0 \tag{1.2}$$

Within this framework, $D_{n,k,i}$ represents the probability of lineage n at site k in state i producing a subtree consistent with observed tree tip states at site k. Site-specific fitness effects are estimated by incorporating the term $D_{n,k,i}$

in the likelihood function. Explained through a genotype g consisting of three nucleotides `ACT`, its methodology comprises three essential steps, conducted in all as part of the tree generating process [57]. For each node n observed in genotype g:

1. Marginal site probabilities are used to determine the likelihood of a lineage having a specific genotype:

$$\hat{\omega}_{n,g} = \hat{\omega}_{n,ACT} = \omega_{n,1,A} \times \omega_{n,2,C} \times \omega_{n,3,T} \tag{1.3}$$

2. Derivation of fitness f_n of a lineage n through the marginalization, weighted by its approximate genotype probability:

$$\hat{f}_{n,g} = \sum_{g \in G} f_g \hat{\omega}_{n,g} \tag{1.4}$$

Assuming a multiplicative fitness effect in the overall fitness of a lineage, $\hat{f}_{n,k,i}$ is computed as:

$$\hat{f}_{n,k,i} = \beta_{k,i} \prod_{l=1, l \neq k}^{L} \sum_{j=1}^{M} \beta_{l,j} \omega_{n,l,j} \tag{1.5}$$

where $\beta_{k,i}$ is the fitness effect of site k in state i. This approximation is especially useful for large site numbers L and high amounts of genotypes M.

3. Calculation of probability E_n, that a lineage leaves no sampled descendants.

These steps allow the computation of the marginal site densities $D_{n,k,i}$, which leads to the calculation of the full joint likelihood [57]. Updating $D_{n,k,i}$ requires numerical integration over a time step Δt for each site and state and can be a computational burden. This framework jointly infers the phylogenetic tree, the ancestral states, and fitness effects. The method is implemented within the Bayesian phylogenetic software BEAST2 in a package called Lumiere [58].
One of the model's constraints is its independence assumption across sites to estimate genotype probabilities, and may not hold in scenarios where epistatic interaction effects are present. This limitation could potentially affect the quality of the genotype approximations $\hat{\omega}_{n,g}$, as well as the overall accuracy of the model.

1.2.3 phyloTF

Another approach to modeling pathogen population dynamics via the structured BD process is demonstrated by phyloTensorFlow (phyloTF) which simplifies the MTBD model as a series of single-type BD processes [59]. It postulates that

fitness is directly proportional to a lineage's transmission rate (see Equation 1.2) and assumes constant rates for μ and σ across all lineages [59]. Further, fitness effects β_i of a lineage x_n are assumed to be multiplicative for each feature i in a set χ. Thus, the fitness mapping function is represented by $F(x)$:

$$F(x_n) = \prod_{i \in \chi} \beta_i x_{n,i} \tag{1.6}$$

PhyloTF iteratively refines its estimate to maximize the likelihood function. On a log scale, the fitness mapping function is presented as an additive linear model:

$$\log(F(x_n)) = \sum_{i \in \chi} \log(\beta_i) x_{n,i} \tag{1.7}$$

The model can be further refined by examining branch-specific fitness effects, represented as u_n, that capture variations in fitness originating from factors not explicitly incorporated in the model, such as genetic background effects at loci that have not been featured specifically [59]:

$$\log(F(x_n)) = \sum_{i \in \chi} \log(\beta_i) x_{n,i} + \log(u_n) \tag{1.8}$$

The likelihood of observing a phylogenetic tree T assuming a birth-death-sampling model given the fitness mapping function, base transmission rate λ_0, death rate μ (or recovery rate in the case of disease modeling), and sampling rate σ is given through the multiplication of the likelihood of a set of branching (transmission) events B, a set of sampling events S, and a set of lineages N [60]:

$$L(T|F(x), \lambda_0, \mu, \sigma) = \prod_{b \in B} L_{\text{branch}}(b) \prod_{s \in S} L_{\text{sample}}(s) \prod_{n \in N} L_{\text{line}}(n) \tag{1.9}$$

where μ and σ are assumed constant. The likelihood of an individual branching or transmission event b is:

$$L_{branch}(b) = F(x_{n(b)})\lambda_0 = \lambda_{n(b)} \tag{1.10}$$

where $n(b)$ refers to the parent lineage involved in a transmission event. The likelihood of a sampling event s is:

$$L_{sample}(s) = \begin{cases} \sigma\mu, & \text{if } t > 0 \\ \rho, & \text{if } t = 0 \end{cases} \tag{1.11}$$

where ρ is the present sampling probability. The likelihood of a lineage n evolving as observed is:

$$L_{line}(n) = D_n(\Delta_t) \tag{1.12}$$

where the term $D_n(\Delta_t)$ depends on the transmission and recovery rates [59]. phyloTF computes the maximum likelihood (ML) using gradient descent optimization for different values of β_i for each state i in χ [59]. Specifically, it uses keras' ADAM optimizer, which iteratively adapts learning rates (lr) based on gradients, allowing the algorithm to converge more efficiently to the optimal parameters [61].

Implementing the model in TensorFlow enables the efficient estimation of over 300 features and large phylogenies with more than 22,000 tip nodes [59]. The fast processing in this approach is an advantage compared to the MFBD model. However, unlike the MFBD model, phyloTF requires separate steps for tree inference, which is independent of mutation state, and ancestral state reconstruction. Thus, the model's uncertainty only considers fitness estimation.

In this study, phyloTF was used to efficiently compute fitness effects of mutations in large phylogenetic trees.

Chapter 2

Materials and Methods

2.1 Materials

This study utilized two substantial data sets provided by the Tuberculosis Research Unit at the Swiss Tropical and Public Health Institute (TPH). Both data sets, one from South Africa, the other from Georgia, with a collection of *Mtb* genomes collected from TB patients, amassed over a decade. The available samples had undergone Whole Genome Sequencing (WGS), facilitating their categorization by lineage. Additionally, WHO's catalog of known DR-conferring mutations was used and allowed the detection of drug resistance in each strain [62]. A collection of putative compensatory mutations for RIF-resistant strains was extracted from a previous study [52]. The number of invariant sites for each data set and its sub-sets was recorded separately.

2.1.1 Data Set from South Africa

The sequences in the South African data stem from primary care and hospitals in Khayelitsha, Cape Town, South Africa, and were collected during 2008 - 2018. Associated clinical data including the isolation date of each pathogen genome was made available from a prospectively maintained database from the Western Cape Provincial Health Data Centre. The available strains come from a retrospective cohort study of individuals that were routinely diagnosed with RIF-resistant or MDR-TB. The collection encompasses a total of 1,162 genomes of which 785 sequences belong to L2, 28 sequences belong to L3 and 349 sequences belong to L4. For the study, only L2 and L4 were analyzed as the amount of data for L3 is not sufficient for a meaningful analysis.

2.1.2 Data Set from Georgia

The study population of Georgia consists of pathogen genomes that caused DS- and MDR-TB from 2011 - 2021 in Georgia. DS *Mtb*, however, was only sequenced in 2013 - 2016. Mono-resistant Mtb strains were not included in this data set.

Patient-related clinical data collected through routine diagnostic work at the National Centre for Tuberculosis and Lung Diseases in Georgia provided additional patient information such as the isolation date of the pathogen genome. The available data set contains 4,772 TB sequences of which 2,607 belong to L2 and 2,165 belong to L4. In L2, there are 772 DS and 1835 DR strains, L4 has 1819 DS and 346 DR strains.

2.2 Methods

2.2.1 Phylogenetic Tree Inference

For both data sets, phylogenetic trees were derived for each lineage using an ML methodology to generate an optimal tree that maximizes the data's likelihood given the model. The task was performed using the software tool IQ-TREE (Intelligent Quick Tree) version 2.2.2.6, operating on the high-performance computing cluster Euler, managed by the High-Performance Computing group at ETH Zürich [63].
Here, DR mutations suspected to influence a genome's fitness may mutate at a different (faster) rate than the rest of the genome. Since these changes may not reflect the true evolutionary history of the bacteria, they were excluded from the tree inference, and their sites were considered invariant among all genomes. Since the alignments do not contain constant sites, the ascertainment bias correction was included in the inference model by annotating the number of invariant sites for each nucleotide. A *Mycobacterium canettii* sequence was used as an outgroup to facilitate root identification and direction of time in the phylogenetic tree.
As a substitution model, a General Time Reversible model with a Gamma distribution of rate variation among sites was chosen (GTR + G). It allows for different substitution rates for each pair of nucleotides, making it the most general model for nucleotide substitution. To estimate the reliability of the tree topology, ten bootstrap trees were generated that indicate the support of each branching point in the tree. For replication purposes, seed `'123'` was set. Among others, IQ-TREE returns a best-scoring ML tree and 10 bootstrap trees in Newick format whose branch lengths represent the number of nucleotide substitutions.

2.2.2 Tree Dating

Before transforming the branch lengths to represent time, the outgroup was removed from the phylogenetic tree and TempEst v1.5.3 was utilized to analyze the temporal signal at hand [64]. As the root-to-tip divergence did not correlate with the sampling time of the tips (see Fig. B.1 - B.4), a fixed clock rate of $5 \cdot 10^{-8}$ nucleotide changes/site/year was assumed for both *Mtb* L2 and L4 [65]. This was done in R version 4.1.0 with Least Squares Dating (LSD), using the

R package `Rlsd2` version 1.10 [66], which minimizes the total squared deviations from the molecular clock hypothesis [67]. The output tree is a time-scaled tree in nexus format which consequently was transformed into a Newick tree, as is required for the ancestral state reconstruction using PastML [68]. To avoid zero-branch-lengths, genomes that were sampled on the same date were set one day apart.

2.2.3 Mutations of Interest

In this part of the process, every genome in the alignment file, containing the variant genomic sites, was compared to a reference genome to create a binary table of all present mutations that are significantly associated with DR by the WHO [62]. This list was further constrained by only including DR and compensatory mutations that were present in at least 0.5% of the genomes, as the fitness effect of rare features under a birth-death model is underestimated [57]. For South Africa L4, where only 349 genomes were available, the threshold was increased to 1%. Further, DR mutations were checked for linkages. Any DR mutation that was $\geq 95\%$ linked to any other DR mutation was replaced with their interaction term. If the Pearson correlation between two DR mutations [69] was $\geq 95\%$, indicating a linkage in both directions, the two mutations were treated as one feature. In the case of compensatory mutations, only those were included in the analysis which were linked to a DR mutation with $\geq 95\%$. The number of genomes, the resulting mutation threshold, and the number of selected DR and compensatory mutations for each lineage of both data sets are seen in Tab. 2.1.

Data set	Genomes	Mutation threshold	DRM	CM
South Africa L2	785	8	43 (93)	8
South Africa L4	349	4	35 (93)	4
Georgia L2	2607	14	27 (86)	9
Georgia L4	2165	11	20 (105)	0

Table 2.1: Summary of selected DR mutations (DRM) and compensatory mutations (DM). The number of all present DR mutations before constraining the selection is in brackets next to the number of selected DR mutations.

Correlations and linkages of DR and compensatory mutations that mutated frequently are recorded for each data set in Tab. A.6 - A.12.

2.2.4 Ancestral State Reconstruction

Ancestral State Reconstruction (ASR) was performed using PastML (Phylogenetic Ancestral States Mapping and Localization) as a Python package in version

1.9.38 in ETH's high-performing computing cluster Euler [68]. PastML implements a fast likelihood method to map ancestral nodes in a time tree whose tip states are known.
Generally, a "character" can refer to a variety of traits or features. Here, it represented the mutation state according to the binary table. PastML takes a phylogenetic tree in Newick format and a table of observed tip states as inputs. For the character evolution model, Felsenstein's F81 model was used. This model was selected because of its property that the rate change from state i to j is proportional to j's equilibrium frequency π_j and therefore ensures differing rate changes between $0 \to 1$ and $1 \to 0$.
As the character states in this study were represented by their mutation state in the binary table, PastML was used to infer the ancestral states in a binary format. As a prediction method for the ancestral state, the ML method Marginal Posterior Probabilities Approximation (MPPA) was chosen. The table entries of PastML's output containing the marginal probabilities were then rounded to infer the binary ancestral states. Other ML approaches such as PastML's maximum a posteriori (MAP) or Joint method, which reconstructs the states of the scenario with the highest likelihood, were also tested. All three prediction methods returned the same binary ancestral states.

2.2.5 Maximum-Likelihood Fitness Inference

Model Parameters

For the fitness estimation, the MTBD model in phyloTF was used as described on the author's Github page [59, 70]. The model parameters for transmission rate and recovery rate were initially set to $\lambda_0 = 1/year$ and $\mu = 1/year$ and assumed to not vary over time [71]. The migration rate was set to $\gamma = 0/year$ for all data sets. The sampling rates upon removal and at present were set to be equal to one another $\sigma = \rho$. For the South African data set this was $\sigma_{SA} = \rho_{SA} = 0.36$, for the Georgian data set this was set to $\sigma_G = \rho_G = 0.9$ [35, 48]. PhyloTF analysis using a subset of ten selected features showed satisfactory convergence for the South African data (see Fig. B.6, B.7). The sampling rate provided for the Georgian data set hindered the model from converging to a meaningful result in a sensible time. Even after 50,000 iterations, the features *eis* C12T and *embA* C12T did not stabilize and were reaching fitness estimates >20 (see Fig. B.5). The sampling rate was thus reduced to a value that lead to the convergence of all features to a comparable fitness range as can be seen in the South African data set (see Tab. 2.2 and Fig. B.8, B.9).
For the fitness estimation of up to ten features, the only computational model specifications that needed to be set are the learning rate (lr) and amount of repetitions, epochs. The epochs were chosen based on how many iterations it

took for the model to converge to estimates maximizing the likelihood (see Eq. 1.9) (see Tab. 2.2, Fig. B.6 - B.9). For more than ten features, the L1 (Lasso) regularizer was used. The Lasso regularizer penalizes the model by the sum of the absolute values of the mutation fitness effects multiplied by a hyperparameter α and thus prevents estimated fitness effects from being too high or too low, was used [59]:

$$L1(i) = \sum |\alpha \cdot (i - offset)| \tag{2.1}$$

The offset was set to 1, as the regularizer is intended to penalize deviations from neutral fitness. As it is only recommended to include a regularizer when analyzing more than ten sites, the fitness estimates for up to ten sites were assumed to reflect the site effects most accurately. The value for hyperparameter α was chosen which yields the smallest Euclidean distance when modeled alone to when modeled as part of a larger set of features. Lastly, it was verified for convergence in the entire data set (see Fig. B.10 - B.13) before applying it toward further analysis. The subset of ten features was chosen to be the ten most common mutations in their respective data set with a linkage to one another of <80%.

Data set	**lr**	**epochs**	**offset**	α	$\sigma = \rho$[1 /year]
South Africa L2	0.001	10,000	1	5.9	0.36
South Africa L4	0.001	16,000	1	0	0.36
Georgia L2	0.001	10,000	1	3.1	0.4
Georgia L4	0.001	10,000	1	0	0.4

Table 2.2: Model parameters for each data set to infer fitness using phyloTF: The learning rate (lr) is the step size for each iteration of MLE, epochs are the number of iterations for the estimation, offset is set to 1 as it represents the neutral fitness effect, α is the hyperparameter of the L1 regression and the sampling fraction at present, ρ is set to the sampling fraction upon removal, σ.

Credible Intervals

To measure the uncertainty surrounding the ML estimates, their 95% credible intervals (CIs) were computed. The likelihood ratio (LR) test statistic was assumed to follow a χ^2_1 distribution [59], and was calculated for different values β_0 of the fitness effects.

$$\lambda_{LR} = -2 \cdot ln \left[\frac{sup_{\beta_0} L(T|F(x), \lambda_0, \mu, \sigma)}{sup_{\beta \in B} L(T|F(x), \lambda_0, \mu, \sigma)} \right] \tag{2.2}$$

Values of β_0 were included in the 95% CI if λ_{LR} did not surpass 3.841, which corresponds to the 95% threshold of the χ^2_1 distribution.

Sensitivity Analyses

To get an understanding of the model's robustness and the influence of assumptions about λ_0, μ, and $\sigma = \rho$, sensitivity analyses were conducted for different values of these parameters (Tab. 2.3). To keep the likelihood of a sampling event constant (see Eq. 1.11), changes in μ were combined with a change in σ and hence ρ. Further, the sensitivity analyses were expanded to test for estimated

Data Set	λ_0 **[1/year]**	μ **[1/year]**	$\sigma = \rho$ **[1/year]**
South Africa	2	1	0.36
South Africa	0.5	1	0.36
South Africa	1	2	0.18
South Africa	1	0.5	0.72
Georgia	2	1	0.4
Georgia	0.5	1	0.4
Georgia	1	2	0.2
Georgia	1	0.5	0.8

Table 2.3: For the sensitivity analyses, each data set was analyzed with altered set of rates for λ_0, μ and $\sigma = \rho$, as given by the rows. σ was set based on Eq. 1.11 to maintain the same likelihood of a sampling event.

transmission, recovery, and transmission rates $\hat{\lambda}_0$, $\hat{\mu}$, and $\hat{\sigma}$.

Reconstructed Fitness Tree

The fitness estimation does not only provide us with information about the fitness effects of individual mutations but combining the feature-specific fitness effects with PastML's output resulted in the transformation of the reconstructed binary table into a reconstructed fitness table. Multiplying all mutated features of a node then resulted in fitness per genotype and can be visualized in R with the annotated Nexus tree from the ASR analysis.

Chapter 3

Results

3.1 Phylogenetic and Ancestral State Reconstruction

This stuy analyzed 1,134 whole genome sequence alignments of MDR-*Mtb* in patients in South Africa and 4,772 MDR- and DS-*Mtb* whole genome sequence alignments from Georgia. Both data sets entailed genomes from L2 and L4, effectively creating 4 distinct data sets. For each of them, a time-calibrated ML phylogeny was reconstructed. All sampled bacteria genomes were first scanned for a set of mutations significantly associated with DR as specified by the WHO [62]. Then, they were checked for putative compensatory mutations found in *rpoA*, *rpoB*, and *rpoC*. Mutations in these genes only compensate for fitness costs associated with RIF resistance. The ancestral states were then reconstructed for all mutations through MPPA in PastML. Applying any of the other two ML methods, MAP and joint posterior probability estimation, and rounding results at 0.5 returned the exact same tree for each of the data sets. Fig. 3.1 shows the phylogeny of *rpoB* S450L on the top left, the most frequent mutation in South Africa L2, with its ancestral states. Its mutation is reconstructed to have occurred hundreds of years ago and shows most of the mutants belonging to the same two clades. The top right of Fig. 3.1 shows a less frequent mutation in South Africa L4, *embB* Q497R, which emerged more recently and is less prevalent than *rpoB* S450L in L2. The bottom left of Fig. 3.1 shows *fabG1* C15T in Georgia L2 which also shows fewer mutants than for *rpoB* S450L in South Africa L2 but its mutants go back several hundred years. Finally, the bottom right of Fig. 3.1 shows *katG* S315T in Georgia L4. Mutants of *katG* S315T show multiple small mutated clades in the phylogeny.

3.2 Fitness Effect of Mutations

Next, the fitness effect of the genetic features was estimated using phyloTF. For South Africa, 51 mutations were considered for L2 and 39 for L4. For Georgia, 27 mutations were studied in L2 and 20 in L4. In all data sets, several mutations show tight linkages to other DR mutations, as some mutations co-occur with

another one in $\geq 95\%$ of the cases (but not necessarily vice versa) (see Tab. A.6 - A.12). Fitness effects were estimated under a model where each feature has a multiplicative effect on the base transmission rate. A neutral feature has a fitness effect of 1 and deleterious or beneficial mutations have fitness effects less than or greater than 1, respectively. Since the death rate is fixed, these fitness effects directly quantify the features' effect on the effective reproductive rate R_e of lineages with the mutation:

$$R_e = \frac{\lambda_n}{\mu} = \frac{1}{\mu} \prod_{i \in \chi} \beta_i x_{n,i} \tag{3.1}$$

3.2.1 South Africa Lineage 2

In South Africa L2, twelve of the analyzed genetic features are deleterious and nine are beneficial relative to the baseline (Figure 3.2). Nine out of the twelve deleterious mutations occur in the *rpoB* gene (L430P, D435Y, D435V, H445Y, H445R, H445D, H445L, S450L, L452P), conferring resistance to RIF. Table A.1 summarizes mutation sites and their conferring drug resistance. The mutation *rpoB* S450L is especially interesting as it is the most common mutation in this data set. Also, the second and third most common mutations have a fitness cost, *fabG1* C15T and *katG* S315T, both conferring resistance to INH and the former also conferring resistance to ETH. The twelfth deleterious mutation is *rpsL* K88R and confers resistance to streptomycin (STM). All twelve of these mutations do not show a linkage to any other mutation. Of the beneficial mutations, one is a compensatory mutation in *rpoA* (T187A) and two are compensatory mutations in *rpoC* (D485Y and V483). These compensatory mutations are all linked to *rpoB* S450L. Mutations *rpoA* T187A and *rpoC* V483G are also linked to additional DR mutations that confer resistance to drugs other than RIF (see Tab. A.2 and A.7). Two beneficial mutations are found in *pncA*, Y103! and D8N, and confer resistance to PZA. Another two beneficial mutations are found in *embB*, M306V and M306I, conferring resistance to EMB, and the remaining two beneficial mutations are *inhA* I194T, conferring resistance to INH, and *gidB* L79S, conferring resistance to STM. All beneficial mutations are linked to at least one other DR mutation and represent an interaction term (see Tab. A.2, A.6, A.7). They are all linked to either *rpoB* S450L, *fabG1* C15T or *katG* S315T. Figure 3.3 shows the fitness tree and indicates that the beneficial fitness effects are not enough to compensate for the fitness costs.

3.2.2 South Africa Lineage 4

South Africa L4 contains 25 significant mutations of which 13 infer a fitness cost and 12 are beneficial mutations (see Fig. 3.4). Similarly to South Africa L2, ten of the deleterious mutations occur in the *rpoB* gene (L430P, D435Y, D435V,

S441L, H445D, H445N, H445Y, S450L, S450W, S452P), conferring resistance to RIF. Mutations *katG* S315T, *fabG1* C15T and *embB* G406V are also estimated to infer a fitness cost and the latter is the only one linked to another mutation, namely *katG* S315T. Mutations *rpoB* S450L and *katG* S315T represent the two most common mutations in the population in this data set. From the 12 beneficial mutations, one is compensatory (*rpoC* V1252L), and one is in *gyrA* (A90V) and confers resistance to the fluoroquinolones LEV and MXF. Another one is in *katG* (S315R) and four are in *embB* (C16G, G406A, M306V, Q497R). Two beneficial mutations conferring resistance to STM are in *rpsL* (K43R and K88Q) and the latter is correlated to *pncA* K96T, which is resistant to PZA. The remaining beneficial mutations are in *pncA* (G97C, Q10P, and H71Y). Apart from one, all beneficial mutations are linked to at least one other DR mutation (see Tab. A.3, A.8, A.9). Each of the ten interaction terms of the beneficial mutations either entails *rpoB* S450L or *katG* S315T. The fitness tree with all significant mutations can be seen in Fig. 3.5. Again, it shows a maximum strain fitness of $\sim$1. The only significantly beneficial mutation that is not linked to any other DR mutation is *embB* Q497R. It occurs in 5.2% of the South African L4 population.

3.2.3 Georgia Lineage 2

Similarly to South Africa L2, most of the Georgia L2 DR mutations have a neutral estimated fitness effect (see Fig. 3.6), including mutation *rpoB* S450L, which was deleterious in South Africa. Here, there are twelve deleterious mutations present of which two are located in *rpoB* (L430P and L452P), however, their most common location is in *embB* (M306I, G406S, G406D, G406A, D354A). The mutations *fabG1* C15T and *katG* S315T are also deleterious in this data set. While the Georgian data set contains DS *Mtb* genomes as well, *katG* S315T still mutated in $>$65% of the sequenced strains. Lastly, *gyrA* D94A, *eis* G10A, and the compensatory mutation *rpoC* D485N infer a fitness cost. From the eight beneficial mutations, three are compensatory mutations (*rpoB* L731P, *rpoC* G332C and *rpoC* V483G), two are in *gyrA* (D94G and D94Y), one is *rpsL* K88R, one is *pncA* T142M and the last one is *embA* C12T. All beneficial mutations are linked to at least one other DR mutation (see Tab. A.4, A.10, A.11). Here, they are all linked to *katG* S315T, the most frequent mutation in Georgia L2 with an estimated fitness effect of 0.352(0.343 - 0.360). Figure 3.7 shows that the beneficial fitness effect of strains that are linked to *katG* S315T is not enough to cause a positive fitness in any strain.
A previous study warned of an increasing amount of PZA-resistance mutations in Georgia [53]. PZA-resistance is largely due to mutations in gen *pncA* [53]. However, in this study for both Georgia lineages, only one mutation was found in this region, T142M in L2. It is mutated in $<$ 1.5% of the population and is linked to *rpoB* S450L.

3.2.4 Georgia Lineage 4

In contrast to the previous data sets, no putative compensatory mutations were detected in the data processing steps for Georgia L4 and the analysis was done for only DR mutations. Most of the features are deleterious (see Fig. 3.8), with five of them being located in *rpoB* (D435V, D435Y, H445D, H445Y, and S450L) and four in *embB* (M306I, M306V, G406D, and D1024N). Mutations *fabG1* C15T and *katG* S315T are also estimated to be deleterious in this data set. The twelfth mutation infers a fitness cost in *rpsL* K43R. Four out of 20 analyzed mutations are estimated to be significantly beneficial: *rpsL* K88R, *eis* C12T, *eis* C14T and *embA* C12T and they are all linked to the same mutation, *katG* S315T (see Tab. A.5). The mutation *katG* S315T again represents the most common mutation in this population with an estimated fitness cost of 0.383(0.36 0.406). As in the previous fitness trees, Fig. 3.9 does not show any strain with a positive fitness.

3.3 Model Evaluation

3.3.1 Bootstrap Trees

While the model does not account for uncertainty in the preliminary steps of tree inference and ancestral state reconstruction, bootstrap trees were produced to address uncertainty in the tree inference step. In most cases, the fitness MLE of the bootstrap trees lies near or in the 95% CI (see Fig. 3.2, 3.4, 3.6 and 3.8). Nonetheless, there were some cases where the values fall far outside the 95% CI and are widespread. This is usually the case for mutations with a wide 95% CI, but in the majority of the mutations the bootstrapped fitness MLE indicates a low uncertainty in the tree inference.

3.3.2 Sensitivity Analyses

To analyze the models' robustness to model parameters other than $\lambda_0 = 1$ and $\mu = 1$, sensitivity analyses were conducted and shown for South Africa L2 in Fig. 3.10 and 3.11. Changes in μ were combined with a change in σ and hence ρ (see Methods). The analyses show that both a decreased transmission rate $\lambda_0 = 0.5$ or an increased recovery rate $\mu = 2$ leads to a fitness estimate closer to 1. An increase to $\lambda_0 = 2$ or decrease to $\mu = 0.5$ mostly leads to divergence of the fitness estimate from 1 resulting in a more extreme fitness estimate than the original parameter values. Sensitivity analyses for the remaining data sets are in the appendix, Fig. B.14 - B.19. Further, phyloTF's capabilities were tested by estimating a model parameter in addition to the fitness as illustrated by the green stars in Fig. 3.10, 3.11 and B.14 - B.19. Values for $\hat{\lambda}$ and $\hat{\mu}$, in turn, were estimated to be lower than originally set and thus lead to a more neutral fitness

estimate in the case of $\hat{\lambda}$ and a more extreme fitness estimate in the case of $\hat{\mu}$. Attempts to estimate the sampling rate rendered $\hat{\rho} > 1$ and $\hat{\sigma} << 1$ and thus did not result in meaningful estimates.

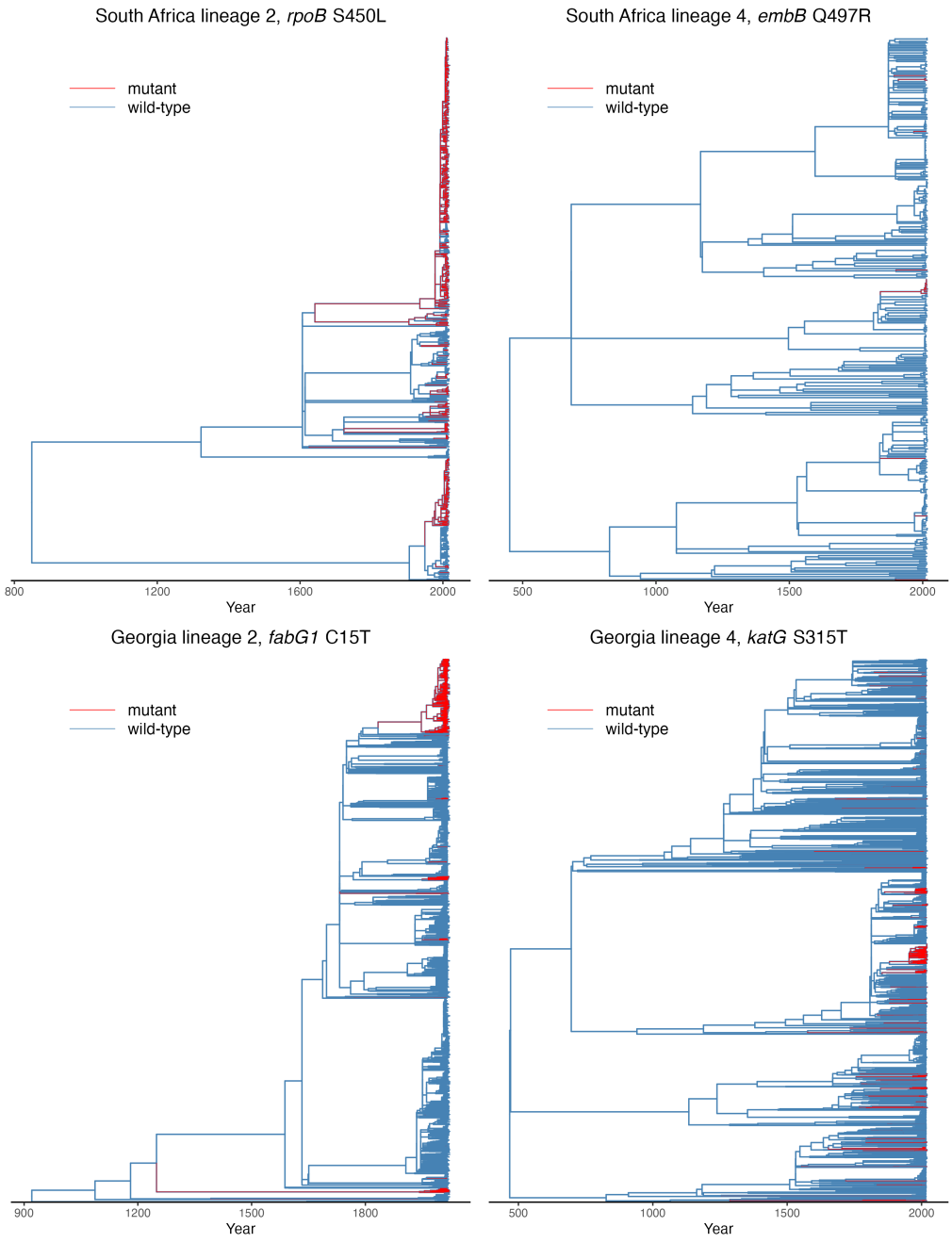

Figure 3.1: Phylogenetic trees with highlighted ancestral states of chosen mutations. Top left: *rpoB* S450L in South Africa L2 mutates often and originates in the distant past. Top right: *embB* Q497R in South Africa L4 mutates less often and emerged more recently. Bottom left: *fabG1* C15T has a low prevalence in Georgia L2 and has emerged multiple hundred years ago. Bottom right: *katG* S315T shows emergence on multiple occasions in multiple clades in Georgia L4.

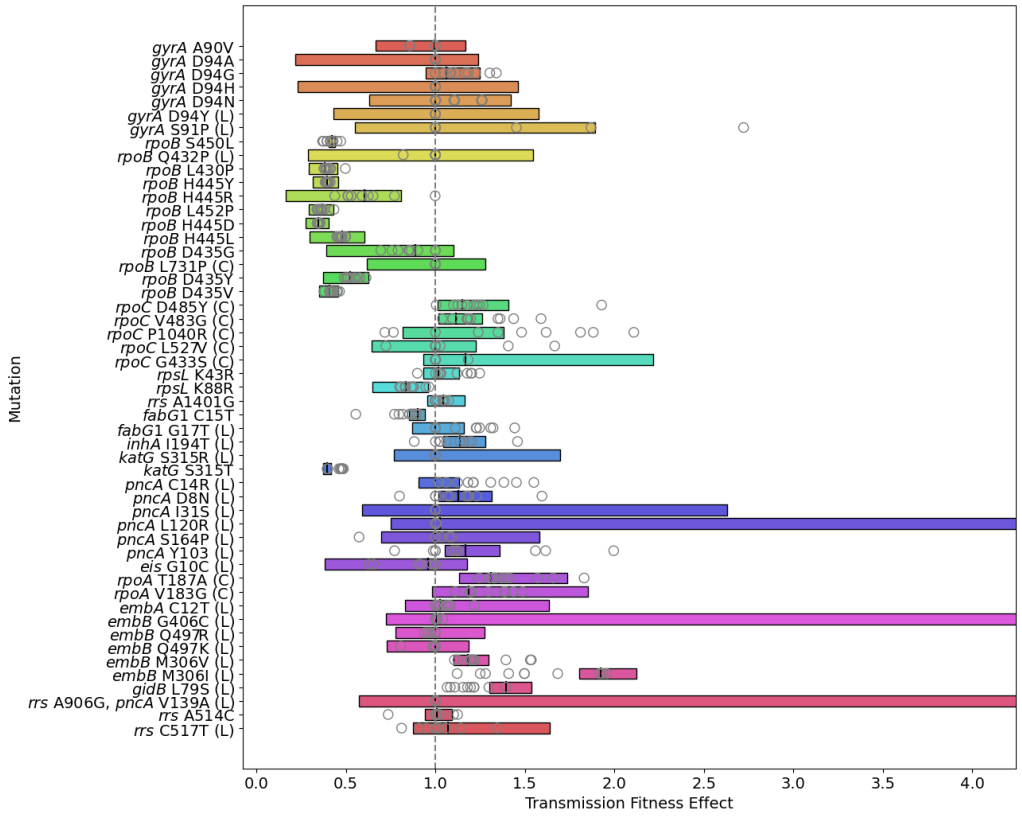

Figure 3.2: Estimated fitness effects for L2 in South Africa. The x-axis is limited to 4 for visualization purposes. Mutations are ordered by their position in the genome. ML Estimates (MLEs) are denoted by vertical lines, with 95% CIs represented by the surrounding boxes. For exact upper and lower limits, see Tab. A.2. For ten replicate bootstrap trees, the MLEs for each fitness effect are presented by grey circles. Where two mutations are listed, they correlate. Mutations ending with *(C)* are putative compensatory mutations, and mutations ending with *(L)* are linked to other mutations, but not necessarily vice versa (see Methods). Their linkages are documented in Tab. A.6 and A.7.

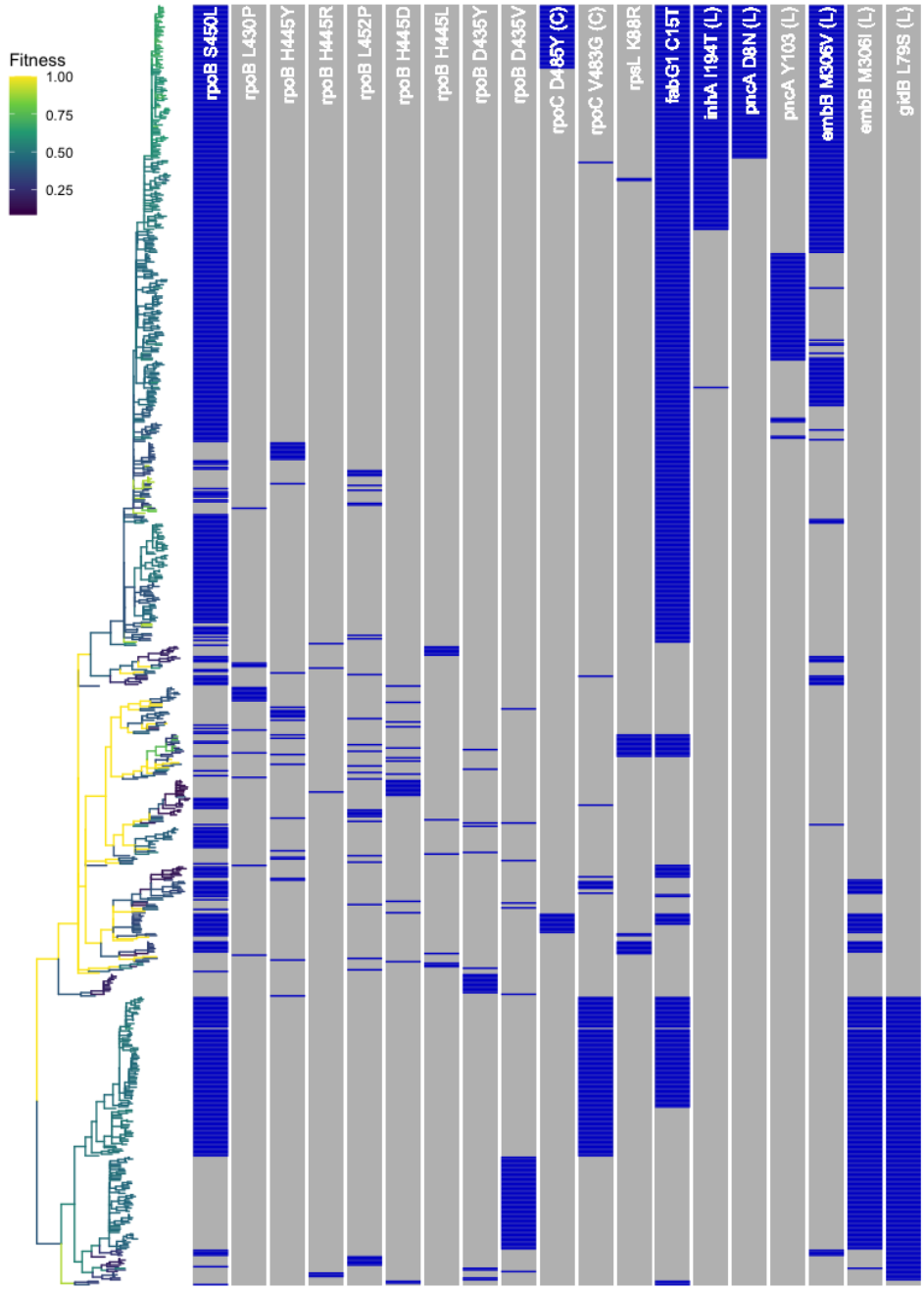

Figure 3.3: Fitness tree of South Africa L2 with lineages colored by their fitness in the tree. The positions of all significant mutations are marked in the columns to the right of the tree. Compensatory mutations are noted with a (C).

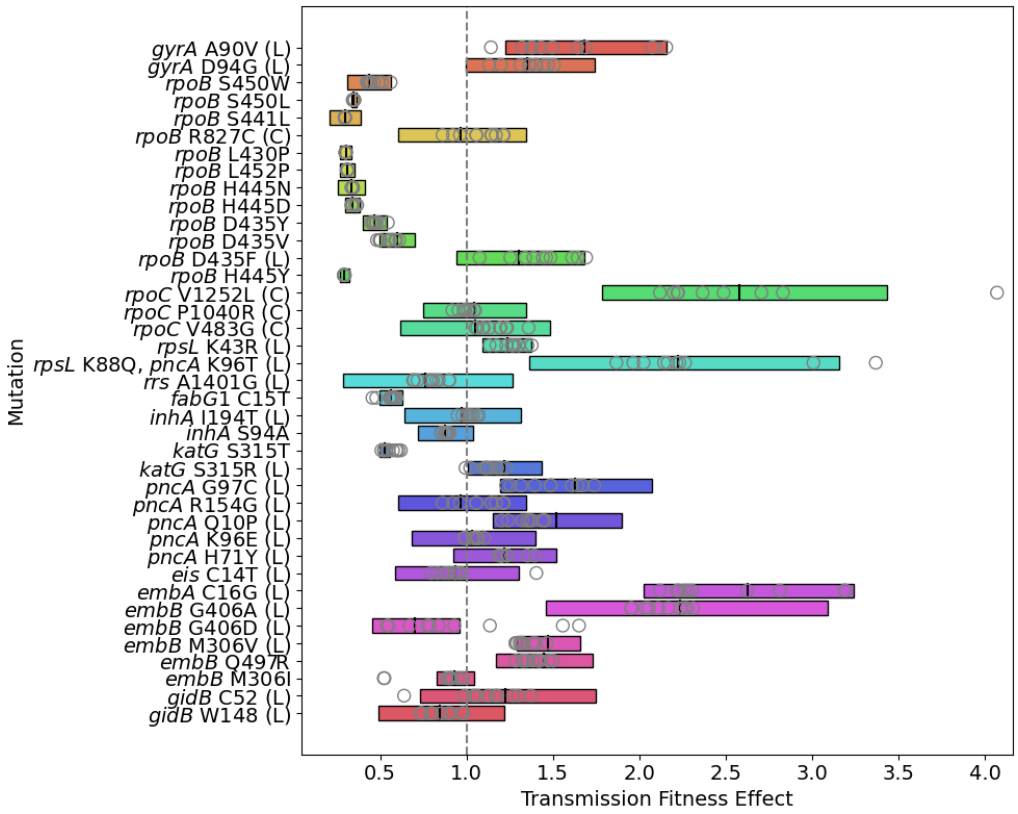

Figure 3.4: Estimated fitness effects for L4 in South Africa. The x-axis is limited to 4 for visualization purposes. Mutations are ordered by their position in the genome. MLEs are denoted by vertical lines, with 95% CIs represented by the surrounding boxes. For exact upper and lower limits, see Tab. A.3. For ten replicate bootstrap trees, the MLEs for each fitness effect are presented by grey circles. Where two mutations are listed, they correlate. Mutations ending with *(C)* are putative compensatory mutations, and mutations ending with *(L)* are linked to other mutations, but not necessarily vice versa (see Methods). Their linkages are documented in Tab. A.8 and A.9.

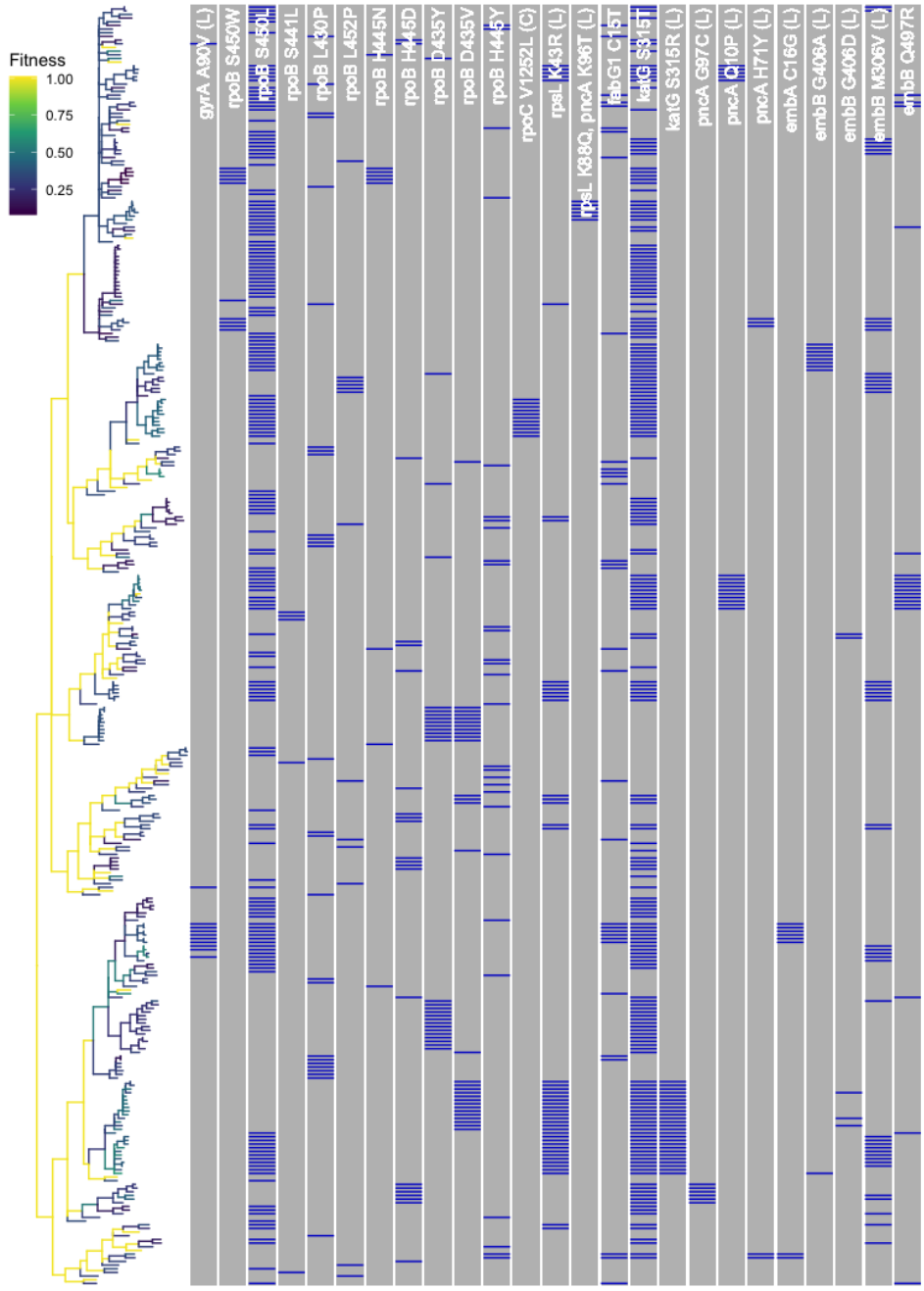

Figure 3.5: Fitness tree of South Africa L4 with lineages colored by their fitness in the tree. The positions of all significant mutations are marked in the columns to the right of the tree. Compensatory mutations are noted with a (C).

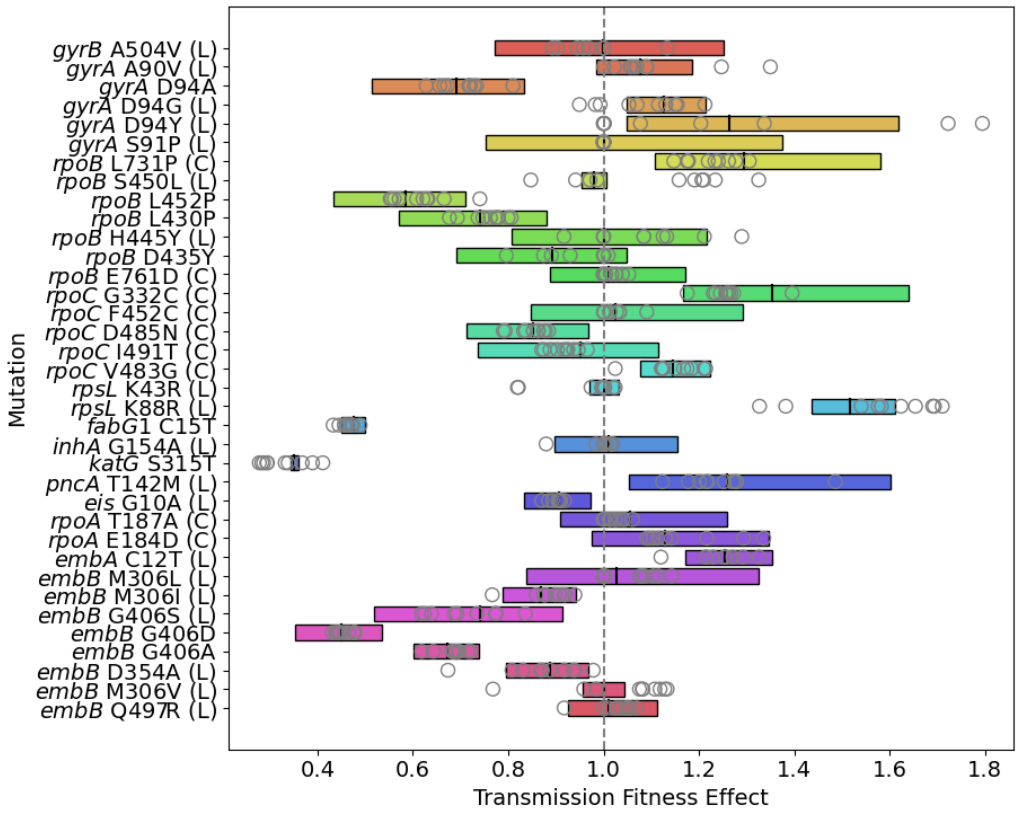

Figure 3.6: Estimated fitness effects for L2 in Georgia. Mutations are ordered by their position in the genome. MLEs are denoted by vertical lines, with 95% CIs represented by the surrounding boxes. For exact upper and lower limits, see Tab. A.4. For ten replicate bootstrap trees, the MLEs for each fitness effect are presented by grey circles. Where two mutations are listed, they correlate. Mutations ending with *(C)* are putative compensatory mutations, and mutations ending with *(L)* are linked to other mutations, but not necessarily vice versa (see Methods). Their linkages are documented in Tab. A.10 and A.11.

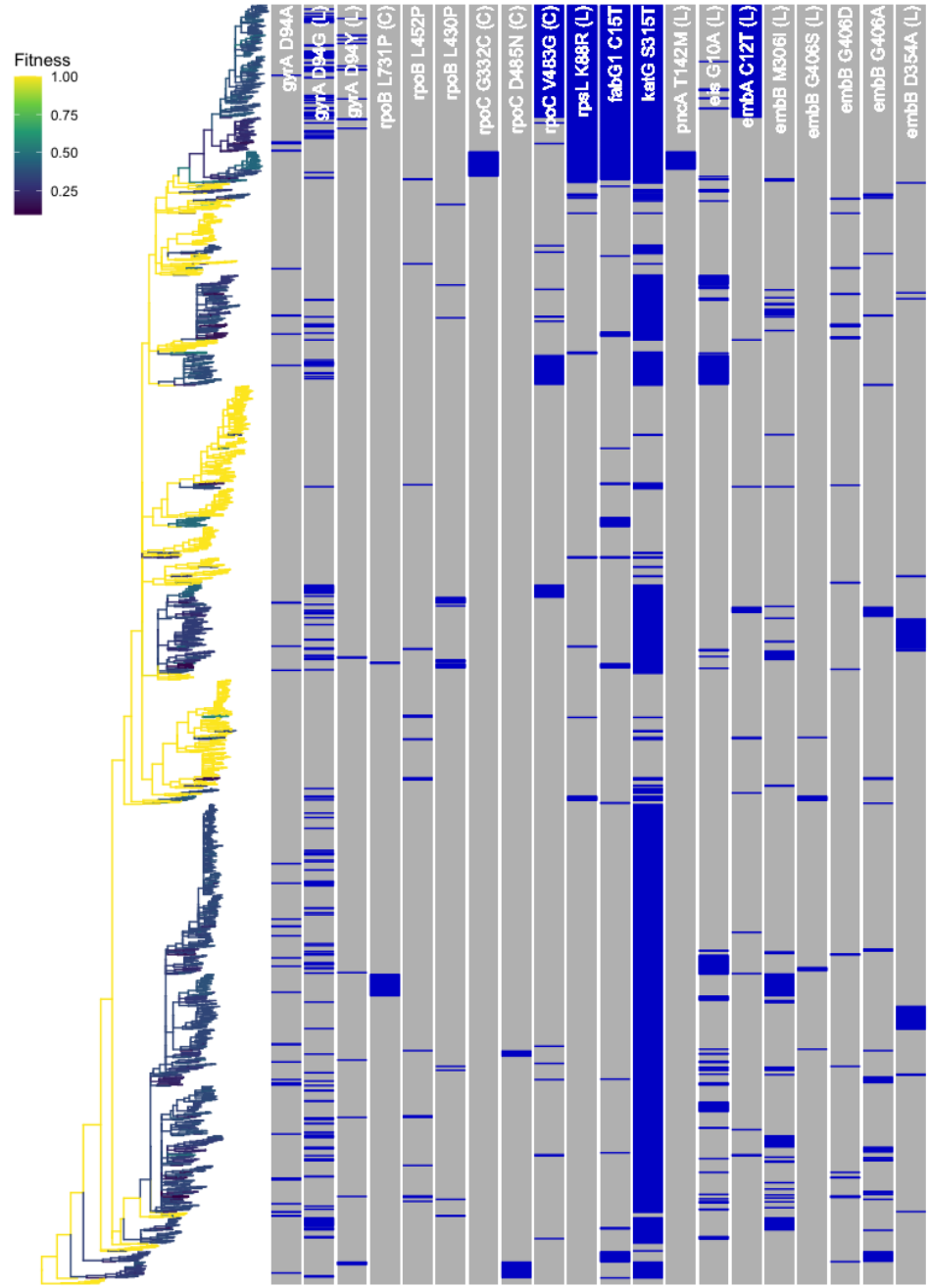

Figure 3.7: Fitness tree of Georgia 2 with lineages colored by their fitness in the tree. The positions of all significant mutations are marked in the columns to the right of the tree. Compensatory mutations are noted with a (C).

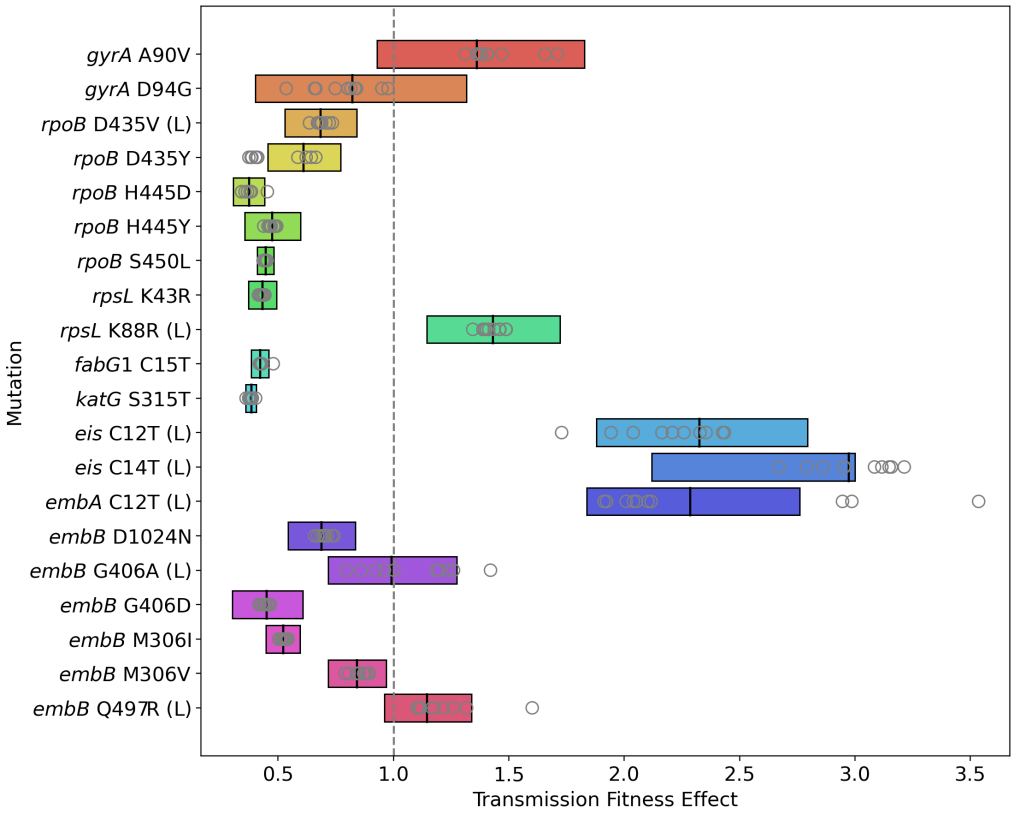

Figure 3.8: Estimated fitness effects for L4 in Georgia. ML Estimates (MLEs) are denoted by vertical lines, with 95% CIs represented by the surrounding boxes. For exact upper and lower limits, see Tab. A.4. For ten replicate bootstrap trees, the MLEs for each fitness effect are presented by grey circles. Where two mutations are listed, they correlate. Mutations ending with *(L)* are linked to other mutations, but not necessarily vice versa (see Methods). Their linkages are documented in Tab. A.4 and A.12.

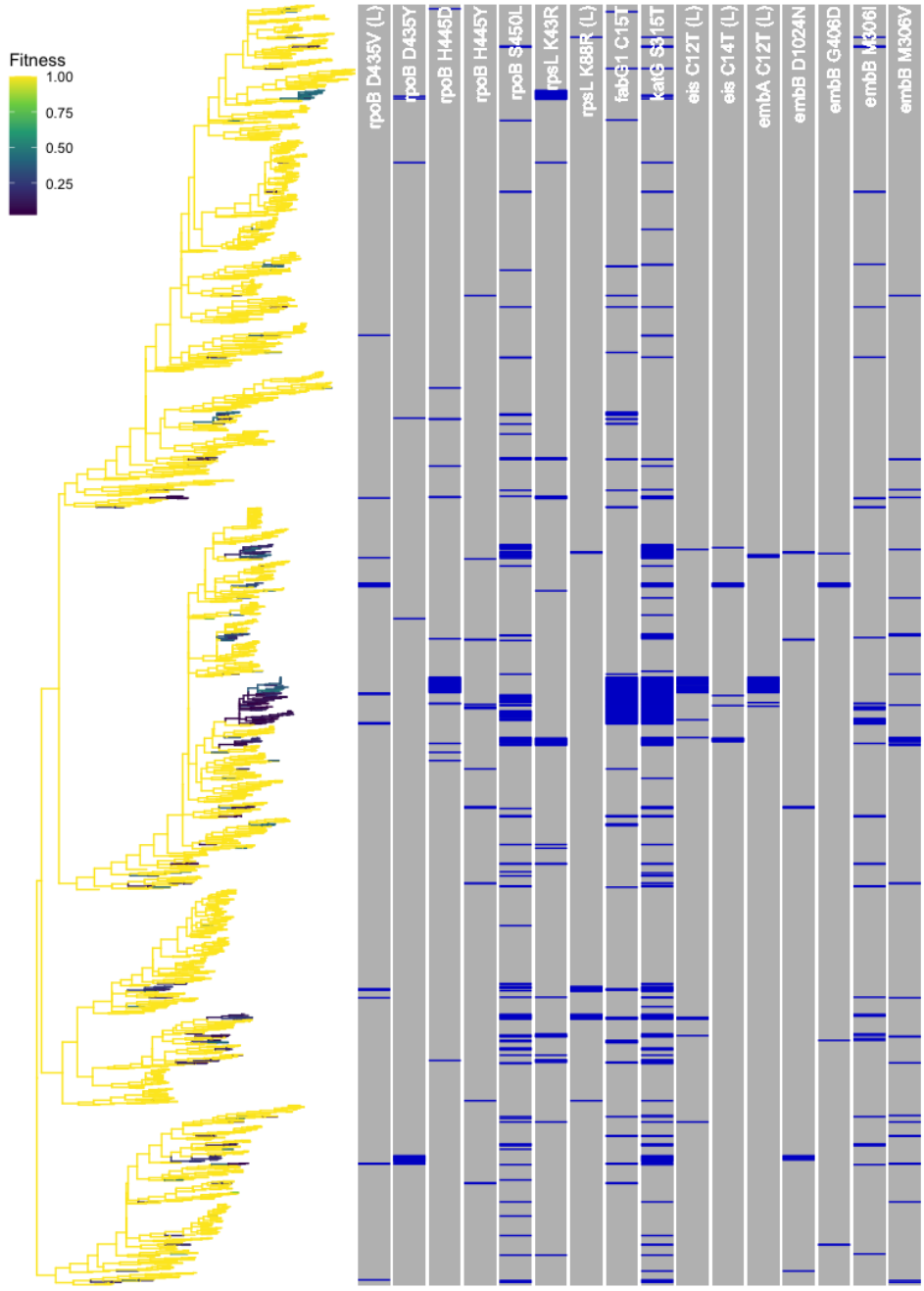

Figure 3.9: Fitness tree of Georgia L4 with lineages colored by their fitness in the tree. The positions of all significant mutations are marked in the columns to the right of the tree.

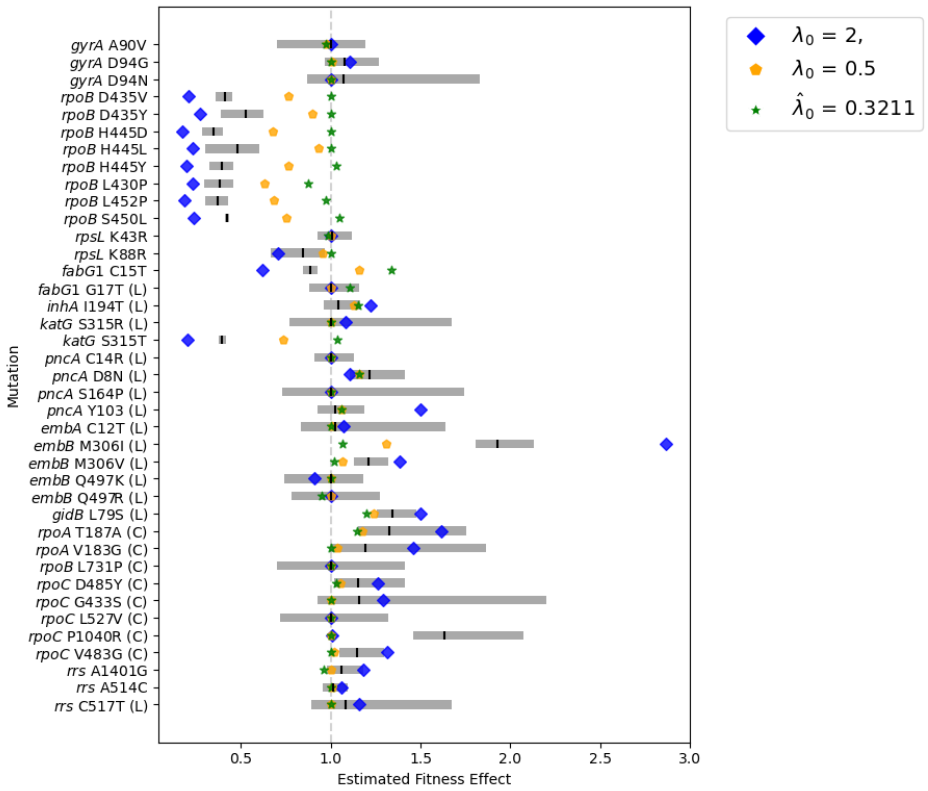

Figure 3.10: Sensitivity analyses and estimation of λ_0 in South Africa L2. The vertical black line represents the mutation's fitness MLE for $\lambda_0 = 1$, $\mu = 1$, and $\sigma = 0.36$ and its surrounding grey box is the 95% CI. Marked in blue and yellow are fitness estimates for different values of λ_0. Green represents the fitness estimates when estimating the transmission rate, $\hat{\lambda}_0$ as well.

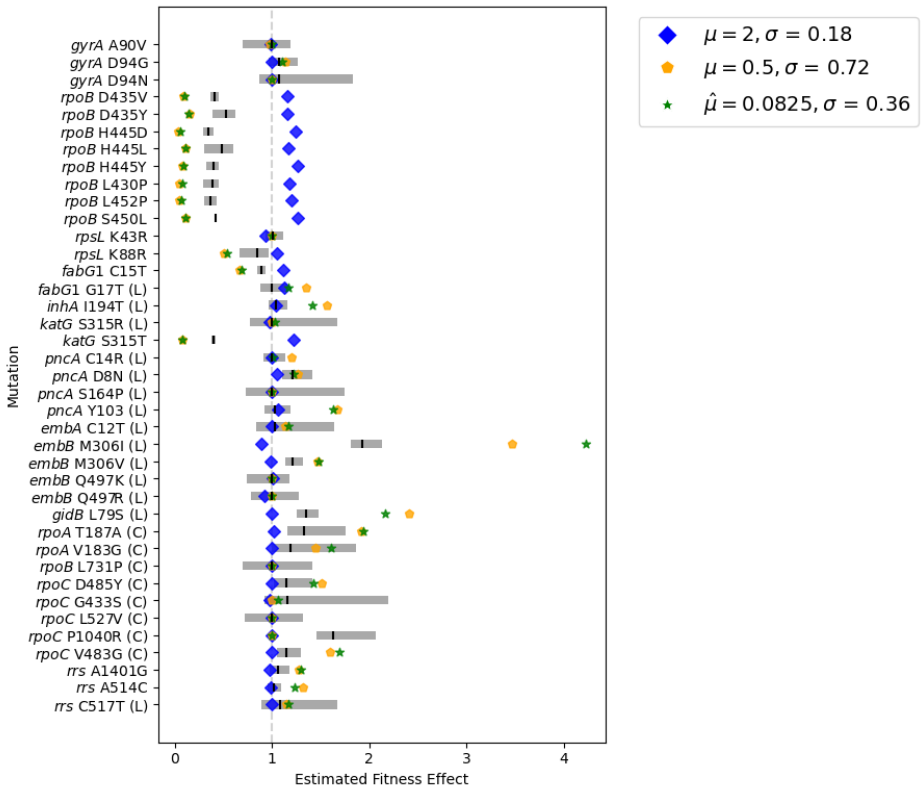

Figure 3.11: Sensitivity analyses and estimation of μ in South Africa L2. The vertical black line represents the mutation's fitness MLE for $\lambda_0 = 1$, $\mu = 1$, and $\sigma = 0.36$ and its surrounding grey box is the 95% CI. Marked in blue and yellow are fitness estimates for different values of μ. Green represents the fitness estimates when estimating the recovery rate, $\hat{\mu}_0$ as well.

Chapter 4

Discussion

This study aimed to explore the phyloTF framework for the estimation of fitness effects of individual DR mutations in *Mtb* genomes. ML methodologies were used to infer a phylogenetic tree, reconstruct ancestral states, and apply the phyloTF computational framework. The analysis covered 1,134 exclusively MDR and XDR genomes from Khayelitsha, South Africa, and 4,772 genomes from Georgia, of which roughly a third of L2 and the majority of L4 are DS. The findings underscore the consistent fitness cost of prevalent mutation *katG* S315T across all data sets and a varying fitness estimation for *rpoB* S450L. While phyloTF effectively estimates the fitness effects across large phylogenies, its limitations, especially in robustness testing, became apparent.

4.1 Fitness Estimates

The phylogenetic trees with ancestral states of chosen mutations in Fig. 3.1 show their evolutionary differences. While *rpoB* S450L, *fabG1* C15T and *katG* S315T have occurred hundreds of years ago, *embB* Q497R emerged more recently. Another interesting observation are the multiple instances of mutated *katG* S315T clades, suggesting its repeated *de novo* mutation in the phylogeny.
The fitness trees in Fig. 3.3, 3.5, 3.7, and 3.9 mostly indicate a strain fitness <1, likely due to the strong deleterious impacts of prevalent mutations such as *rpoB* S450L in South Africa and *katG* S315T across all data sets. Despite the presence of beneficial mutations, the cumulative fitness cost from these common mutations persists across both South African and Georgian populations. A simulation study of phyloTF has shown a high dependency on the overall sampling rate and the fraction of the population carrying a specific mutation [59]. The observed pattern may thus result from poor accuracy due to the low sampling rate in South Africa and Georgia.
The fitness effects of *rpoB* DR mutations differ across populations. *rpoB* encodes the RNA polymerase's β subunit and confers RIF resistance in mutants [72].
A previous study had found a high prevalence of PZA resistance in Georgia, predominantly stemming from mutations in *pncA* [53]. These findings could not

be observed for the Georgian data set of this study. Yet, the South African L2 revealed a notable amount of mutations in *pncA*, with D8N and Y103! being the most common and resulting in beneficial fitness effects. The gene *pncA* encodes the pyrazinamidase/nicotinamidase enzyme which is responsible for the conversion of PZA to its active form pyrazinoic acid (POA) [73, 74]. POA is toxic for *Mtb* as it disrupts its membrane potential and affects membrane transport [75]. Mutations in *pncA* are correlated with a loss of function of its encoding enzyme and thus with PZA resistance [73, 76].
It is important to highlight the co-occurrence trends observed: several compensatory and DR mutations with a beneficial fitness effect are linked to deleterious mutations. This was most pronounced for *katG* S315T, to which multiple DR mutations were linked across all data sets. In the South African data sets, similar linkages were seen for *rpoB* S450L and, in L2 specifically, for *fabG1* C15T. Across all data sets, *embB* Q497R stood out as the only mutation to demonstrate a beneficial fitness effect without any apparent linkages.

4.1.1 *katG* S315T

In this analysis, the mutation *katG* S315T appears frequently and is associated with a significantly deleterious fitness effect both in the South African and in the Georgian data sets. Other mutations that have a beneficial fitness estimate are often linked to *katG* S315T. The *katG* gene encodes a catalase-peroxidase enzyme and metabolizes INH [77]. Animal models have also shown the gene to be a virulence factor [78, 79]. Mutations in *katG* conferring to INH resistance are thus expected to also confer a fitness cost. Other studies, however, have found contradicting results to my study and refer to *katG* S315T as a 'no-cost' mutation [80, 81]. One possible explanation for the mutation's success in the other studies is the presence of compensatory mutations, which are not included in the list of compensatory mutations in this study. Mutations in *ahpC* lead to the gene's promoter expression and code for a protein which is functionally similar to *katG* and potentially acts as a compensatory mutation for *katG* mutations [82, 83]. Another explanation is based on the comparison of an S315T mutant to wild-type *katG* in *Escherichia coli* [84]. The study found that while the *katG* S315T mutant was less efficient at transforming INH to its active form of isonicotinic acid, the strains showed similar results in their enzymatic activity, indicating that *katG* S315T does not impose a fitness cost [84].

4.1.2 *fabG1* C15T

Similarly to *katG* S315T, *fabG1* C15T mutation is associated with INH resistance, and the WHO also associates it with resistance to the second-line drug ethionamide (ETH) [62]. In South African L2, it emerges as the second most common mutation and exhibits a deleterious fitness impact. Other, beneficial

mutations are often linked to it. Although it also demonstrates a fitness cost in South African L4 and in Georgia L2 and L4, its occurrence is lower in these data sets, suggesting specificity to South African L2. The gene *fabG1* codes for mycolic acid biosynthetic enzymes and thus plays an important role in the synthesis of the mycobacterial cell wall [85]. While varied levels of INH resistance have been observed in *fabG1* C15T mutants [86, 87, 88], WGS studies highlight an inconsistent link between *fabG1* C15T and ETH resistance, as the mutation does not always translate to phenotypic ETH resistance [86, 87, 88, 89]. The dynamics between the pathogen and host remain ambiguous and call for further investigation.

4.1.3 *rpoB* S450L

While mutations in *rpoB* predominantly display a deleterious effect in the South African data sets and Georgia L4, their impact is mostly insignificant in Georgia L2. One explanation for this could be a different base transmission in the different populations. The mutation*rpoB* S450L is present across all four data sets. In South Africa's L2 and L4, which consist exclusively of MDR- and XDR-TB data, it is one of the most frequent mutations and carries a fitness cost. Multiple mutations linked to *rpoB* S450L with minor compensatory effects are observed in both populations. However, Fig. 3.3 and 3.5 show that these links do not compensate sufficiently to result in mutants with positive fitness. Another study of the same population, however, found compensatory mutations and other mutations that are tightly linked to *rpoB* S450L to be correlated with a beneficial fitness [48]. In Georgia L2, which includes DS genomes and where compensatory mutations are found, the fitness impact of *rpoB* S450L is not deemed significant. In Georgia L4, which also includes DS genomes and where no compensatory mutations were found, the fitness effect is deleterious. This confirms a recent study's findings that strains carrying the *rpoB* S450L and a linked DR or compensatory mutation had a higher transmission fitness compared to strains carrying *rpoB* S450L without a compensatory mutation [35].

4.1.4 *embB* Q497R

Mutations in *embB* in the South African data sets mostly confer a neutral or even beneficial fitness effect and are linked to deleterious mutations such as *katG* S315T, and *rpoB* S450L. In contrast, *embB* mutations in Georgia are predominantly deleterious. In South Africa L4, the mutation *embB* Q497R infers a positive fitness effect and is not linked to any other DR mutation.
embB is part of the *embCAB* operon involved in the biosynthesis of the mycobacterial cell wall components arabinogalactan and lipoarabinomannan [90]. It is believed that EMB hinders arabinan synthesis and results in a lack of arabinan receptors for mycolic acid. The resulting accumulation of mycolic acid then re-

sults in cell death [91]. Multiple studies have correlated alterations in *embB*497, such as *embB* Q497R mutants, to moderate levels of resistance to EMB [92]. However, since these codons have also been found in EMB susceptible isolates, the correlation is uncertain and suggests EMB resistance may stem from multigenic mutations [92, 93]. Similarly to *fabG1* C15T, these observations underline the importance of further research to understand the role *embB* Q497R in DR-TB.

4.2 Evaluation of phyloTF

PhyloTF provides an efficient framework for the estimation of fitness effects within a phylogeny. Nevertheless, its effectiveness is challenged by sensitivity to model parameters and a lack of consideration for uncertainties related to tree inference and ancestral state reconstruction. The range of the fitness estimates is widespread for higher and lower values of λ_0 and μ. When comparing the *rpoB* S450L mutation between South Africa and Georgia, it becomes evident that including DS strains provides a more comprehensive estimation of fitness effects and underscores the importance of interpreting the results considering the respective baseline. Observed effects for *rpoB* S450L in Georgia L2 support previous studies [35]. The fitness estimates of *katG* S315T and the estimated fitness trees (see Fig. 3.3, 3.5, 3.7, 3.9), however, render the results from phyloTF questionable when comparing to previous studies [80, 81].
The long evolutionary history of *Mtb* has led to closely linked mutations and makes it challenging to pinpoint the fitness effects of individual mutations. The choice of regularization factor poses another challenge due to the absence of a benchmark for comparison. Ultimately, appropriate parameter choice and comprehensive inclusion of both DS and DR strains can optimize accuracy in this model.

4.3 Future Directions

This study opens up opportunities for further exploration of TB's evolution of DR mutations. Key areas of interest include investigating compensatory mutations for *katG* S315T and deepening our understanding of how *fabG1* C12T and *embB* Q497R contribute to DR-TB. A valuable analysis would be comparing these results with findings from the MFBD model using Lumiere in the Bayesian framework BEAST2 [57]. In the Bayesian framework, the difficulties with low sampling rates persist, and the independent genotype assumption remains a challenge as many mutations are tightly linked. Nonetheless, Lumiere's comprehensive approach, which considers tree estimation and ancestral state uncertainties, could improve accuracy. However, it is expected to have a higher computational burden as it estimates the posterior distribution of fitness effects using Bayesian

Markov Chain Monte Carlo. Depending on whether Lumiere supports or contradicts phyloTF's findings, it could point toward DR mutations that have the potential to spread or are already widespread in the population. Recognizing these mutations can enhance TB diagnostics through targeted testing for specific DR mutations, leading to efficient and effective patient treatment. Further, given the prevalent drug resistance to traditional TB antibiotics, understanding the implications of specific mutations can lead to the development of novel, effective drugs and support WHO's End TB Strategy.

Bibliography

[1] T. Sakai and Y. Morimoto, "The History of Infectious Diseases and Medicine," *Pathogens*, vol. 11, no. 10, p. 1147, 2022.

[2] F. Coll, R. McNerney, J. A. Guerra-Assunção, J. R. Glynn, J. Perdigão, M. Viveiros, I. Portugal, A. Pain, N. Martin, and T. G. Clark, "A robust SNP barcode for typing *Mycobacterium tuberculosis* complex strains," *Nature communications*, vol. 5, no. 1, p. 4812, 2014.

[3] S. Gagneux, "Ecology and evolution of *Mycobacterium tuberculosis*," *Nature Reviews Microbiology*, vol. 16, no. 4, pp. 202–213, 2018.

[4] M. Coscolla, S. Gagneux, F. Menardo, C. Loiseau, P. Ruiz-Rodriguez, S. Borrell, I. D. Otchere, A. Asante-Poku, P. Asare, L. Sánchez-Busó *et al.*, "Phylogenomics of *Mycobacterium africanum* reveals a new lineage and a complex evolutionary history," *Microbial genomics*, vol. 7, no. 2, p. 000477, 2021.

[5] C. A. Guerrero-Bustamante, R. M. Dedrick, R. A. Garlena, D. A. Russell, and G. F. Hatfull, "Toward a phage cocktail for tuberculosis: susceptibility and tuberculocidal action of mycobacteriophages against diverse *Mycobacterium tuberculosis* strains," *MBio*, vol. 12, no. 3, pp. 10–1128, 2021.

[6] M. Coscolla and S. Gagneux, "Consequences of genomic diversity in *Mycobacterium tuberculosis*," in *Seminars in immunology*, vol. 26, no. 6. Elsevier, 2014, pp. 431–444.

[7] B. Müller, S. Dürr, S. Alonso, J. Hattendorf, C. J. Laisse, S. D. Parsons, P. D. Van Helden, and J. Zinsstag, "Zoonotic *Mycobacterium bovis*–induced tuberculosis in humans," *Emerging infectious diseases*, vol. 19, no. 6, p. 899, 2013.

[8] I. Suárez, S. M. Fünger, S. Kröger, J. Rademacher, G. Fätkenheuer, and J. Rybniker, "The diagnosis and treatment of tuberculosis." *Deutsches Aerzteblatt International*, vol. 116, no. 43, 2019.

[9] H. Getahun, A. Matteelli, R. E. Chaisson, and M. Raviglione, "Latent *Mycobacterium tuberculosis* infection," *New England Journal of Medicine*, vol. 372, no. 22, pp. 2127–2135, 2015.

[10] R. M. Houben and P. J. Dodd, "The global burden of latent tuberculosis infection: a re-estimation using mathematical modelling," *PLoS medicine*, vol. 13, no. 10, p. e1002152, 2016.

[11] W. H. Organization, "Global Tuberculosis Report 2022," Tech. Rep., 2023.

[12] M. Pai, M. Behr, D. Dowdy, K. Dheda, M. Divangahi, C. Boehme *et al.*, "Tuberculosis," *Nature reviews disease primers*, vol. 2, p. 16076, 2016.

[13] K. R. Steingart, A. Ramsay, and M. Pai, "Optimizing sputum smear microscopy for the diagnosis of pulmonary tuberculosis," *Expert review of anti-infective therapy*, vol. 5, no. 3, pp. 327–331, 2007.

[14] R. A. Parker, "Implications of tuberculosis sputum culture test sensitivity on accuracy of other diagnostic modalities," *American journal of respiratory and critical care medicine*, vol. 199, no. 5, pp. 664–664, 2019.

[15] A. D. Harries, Y. Lin, A. M. Kumar, S. Satyanarayana, K. C. Takarinda, R. A. Dlodlo, R. Zachariah, and P. Olliaro, "What can National TB Control Programmes in low-and middle-income countries do to end tuberculosis by 2030?" *F1000Research*, vol. 7, 2018.

[16] M. Grobbelaar, G. E. Louw, S. L. Sampson, P. D. van Helden, P. R. Donald, and R. M. Warren, "Evolution of rifampicin treatment for tuberculosis," *Infection, Genetics and Evolution*, vol. 74, p. 103937, 2019.

[17] C. R. Horsburgh Jr, C. E. Barry III, and C. Lange, "Treatment of tuberculosis," *New England Journal of Medicine*, vol. 373, no. 22, pp. 2149–2160, 2015.

[18] H. S. Cox, M. Morrow, and P. W. Deutschmann, "Long term efficacy of DOTS regimens for tuberculosis: systematic review," *Bmj*, vol. 336, no. 7642, pp. 484–487, 2008.

[19] D. Mitchison and G. Davies, "The chemotherapy of tuberculosis: past, present and future," *The international journal of tuberculosis and lung disease*, vol. 16, no. 6, pp. 724–732, 2012.

[20] W. Fox, G. A. Ellard, and D. A. Mitchison, "Studies on the treatment of tuberculosis undertaken by the British Medical Research Council tuberculosis units, 1946–1986, with relevant subsequent publications," *The International Journal of Tuberculosis and Lung Disease*, vol. 3, no. 10, pp. S231–S279, 1999.

[21] M. F. Rabahi, J. L. R. d. Silva, A. C. G. Ferreira, D. G. S. Tannus-Silva, and M. B. Conde, "Tuberculosis treatment," *Jornal Brasileiro de Pneumologia*, vol. 43, pp. 472–486, 2017.

[22] G. Sotgiu, R. Centis, L. D'ambrosio, and G. B. Migliori, "Tuberculosis treatment and drug regimens," *Cold Spring Harbor perspectives in medicine*, vol. 5, no. 5, 2015.

[23] X. Lv, S. Tang, Y. Xia, X. Wang, Y. Yuan, D. Hu, F. Liu, S. Wu, Y. Zhang, Z. Yang *et al.*, "Adverse reactions due to directly observed treatment strategy therapy in Chinese tuberculosis patients: a prospective study," *PloS one*, vol. 8, no. 6, p. e65037, 2013.

[24] J. L. Khawbung, D. Nath, and S. Chakraborty, "Drug resistant Tuberculosis: A review," *Comparative immunology, microbiology and infectious diseases*, vol. 74, p. 101574, 2021.

[25] J. Karumbi and P. Garner, "Directly observed therapy for treating tuberculosis," *Cochrane database of systematic reviews*, no. 5, 2015.

[26] H. Zhang, J. Ehiri, H. Yang, S. Tang, and Y. Li, "Impact of community-based DOT on tuberculosis treatment outcomes: a systematic review and meta-analysis," *PloS one*, vol. 11, no. 2, p. e0147744, 2016.

[27] Y. Zhang, B. Heym, B. Allen, D. Young, and S. Cole, "The catalase—peroxidase gene and isoniazid resistance of *Mycobacterium tuberculosis*," *Nature*, vol. 358, no. 6387, pp. 591–593, 1992.

[28] A. S. Piatek, A. Telenti, M. R. Murray, H. El-Hajj, W. R. Jacobs Jr, F. R. Kramer, and D. Alland, "Genotypic analysis of *Mycobacterium tuberculosis* in two distinct populations using molecular beacons: implications for rapid susceptibility testing," *Antimicrobial agents and chemotherapy*, vol. 44, no. 1, pp. 103–110, 2000.

[29] A. Telenti, P. Imboden, F. Marchesi, L. Matter, K. Schopfer, T. Bodmer, D. Lowrie, M. Colston, and S. Cole, "Detection of rifampicin-resistance mutations in *Mycobacterium tuberculosis*," *The Lancet*, vol. 341, no. 8846, pp. 647–651, 1993.

[30] L. P. Ormerod, "Multidrug-resistant tuberculosis (MDR-TB): epidemiology, prevention and treatment," *British medical bulletin*, vol. 73, no. 1, pp. 17–24, 2005.

[31] K. J. Seung, S. Keshavjee, and M. L. Rich, "Multidrug-resistant tuberculosis and extensively drug-resistant tuberculosis," *Cold Spring Harbor perspectives in medicine*, vol. 5, no. 9, 2015.

[32] E. A. Kendall, M. O. Fofana, and D. W. Dowdy, "Burden of transmitted multidrug resistance in epidemics of tuberculosis: a transmission modelling analysis," *The Lancet Respiratory Medicine*, vol. 3, no. 12, pp. 963–972, 2015.

[33] C. Dye, B. G. Williams, M. A. Espinal, and M. C. Raviglione, "Erasing the world's slow stain: strategies to beat multidrug-resistant tuberculosis," *Science*, vol. 295, no. 5562, pp. 2042–2046, 2002.

[34] Q. Liu, T. Zuo, P. Xu, Q. Jiang, J. Wu, M. Gan, C. Yang, R. Prakash, G. Zhu, H. E. Takiff *et al.*, "Have compensatory mutations facilitated the current epidemic of multidrug-resistant tuberculosis?" *Emerging Microbes & Infections*, vol. 7, no. 1, pp. 1–8, 2018.

[35] C. Loiseau, E. M. Windels, S. M. Gygli, L. Jugheli, N. Maghradze, D. Brites, A. Ross, G. Goig, M. Reinhard, S. Borrell *et al.*, "The relative transmission fitness of multidrug-resistant *Mycobacterium tuberculosis* in a drug resistance hotspot," *Nature communications*, vol. 14, no. 1, p. 1988, 2023.

[36] A. Handel, R. R. Regoes, and R. Antia, "The role of compensatory mutations in the emergence of drug resistance," *PLoS computational biology*, vol. 2, no. 10, p. e137, 2006.

[37] A. P. Vargas, A. A. Rios, L. Grandjean, D. E. Kirwan, R. H. Gilman, P. Sheen, and M. J. Zimic, "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*, vol. 9, no. 2, pp. 121–137, 2020.

[38] C. F. McQuaid, A. Vassall, T. Cohen, K. Fiekert, R. White *et al.*, "The impact of COVID-19 on TB: a review of the data," *The International Journal of Tuberculosis and Lung Disease*, vol. 25, no. 6, pp. 436–446, 2021.

[39] L. Cilloni, H. Fu, J. F. Vesga, D. Dowdy, C. Pretorius, S. Ahmedov, S. A. Nair, A. Mosneaga, E. Masini, S. Sahu *et al.*, "The potential impact of the COVID-19 pandemic on the tuberculosis epidemic a modelling analysis," *EClinicalMedicine*, vol. 28, 2020.

[40] Q. Abdool Karim and C. Baxter, "COVID-19: impact on the HIV and tuberculosis response, service delivery, and research in South Africa," *Current HIV/AIDS Reports*, vol. 19, no. 1, pp. 46–53, 2022.

[41] W. H. Organization, "Global Tuberculosis Report 2020," Tech. Rep., 2020.

[42] R. M. Packard, "Tuberculosis and the development of industrial health policies on the Witwatersrand, 1902–1932," *Journal of Southern African Studies*, vol. 13, no. 2, pp. 187–209, 1987.

[43] S. S. A. Karim, G. J. Churchyard, Q. A. Karim, and S. D. Lawn, "HIV infection and tuberculosis in South Africa: an urgent need to escalate the public health response," *the Lancet*, vol. 374, no. 9693, pp. 921–933, 2009.

[44] H. S. Cox, C. McDermid, V. Azevedo, O. Muller, D. Coetzee, J. Simpson, M. Barnard, G. Coetzee, G. van Cutsem, and E. Goemaere, "Epidemic levels of drug resistant tuberculosis (MDR and XDR-TB) in a high HIV prevalence setting in Khayelitsha, South Africa," *PLoS One*, vol. 5, no. 11, p. e13901, 2010.

[45] N. R. Gandhi, A. Moll, A. W. Sturm, R. Pawinski, T. Govender, U. Lalloo, K. Zeller, J. Andrews, and G. Friedland, "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*, vol. 368, no. 9547, pp. 1575–1580, 2006.

[46] D. Bradshaw, V. Pillay-Van Wyk, R. Laubscher, B. Nojilana, P. Groenewald, N. Nannan, and C. Metcalf, "Cause of death statistics for South Africa: Challenges and possibilities for improvement," *South African MRC Burden of Disease Research Unit*, 2010.

[47] L. M. Faye, M. C. Hosu, N. Sineke, S. Vasaikar, A. Dippenaar, S. Oostvogels, R. M. Warren, and T. Apalata, "Detection of Mutations and Genotyping of Drug-Resistant *Mycobacterium tuberculosis* Strains Isolated from Patients in Rural Eastern Cape Province," *Infectious Disease Reports*, 2023.

[48] G. A. Goig, F. Menardo, Z. Salaam-Dreyer, A. Dippenaar, E. M. Streicher, J. Daniels, A. Reuter, S. Borrell, M. Reinhard, A. Doetsch *et al.*, "Effect of compensatory evolution in the emergence and transmission of rifampicin-resistant *Mycobacterium tuberculosis* in Cape Town, South Africa: a genomic epidemiology study," *The Lancet Microbe*, 2023.

[49] Q. Bu, R. Qiang, L. Fang, X. Peng, H. Zhang, and H. Cheng, "Global trends in the incidence rates of MDR and XDR tuberculosis: Findings from the global burden of disease study 2019," *Frontiers in Pharmacology*, vol. 14, p. 1156249, 2023.

[50] W. H. Organization, "Global Tuberculosis Report 2019," Tech. Rep., 2019.

[51] R. Zalesky, F. Abdullajev, G. Khechinashvili, M. Safarian, T. Madaras, M. Grzemska, E. Englund, S. Dittmann, and M. Raviglione, "Tuberculosis control in the Caucasus: successes and constraints in DOTS implementation," *The International Journal of Tuberculosis and Lung Disease*, vol. 3, no. 5, pp. 394–401, 1999.

[52] S. M. Gygli, C. Loiseau, L. Jugheli, N. Adamia, A. Trauner, M. Reinhard, A. Ross, S. Borrell, R. Aspindzelashvili, N. Maghradze *et al.*, "Prisons as ecological drivers of fitness-compensated multidrug-resistant *Mycobacterium tuberculosis*," *Nature medicine*, vol. 27, no. 7, pp. 1171–1177, 2021.

[53] S. Sengstake, I. L. Bergval, A. R. Schuitema, J. L. de Beer, J. Phelan, R. de Zwaan, T. G. Clark, D. van Soolingen, and R. M. Anthony, "Pyrazinamide resistance-conferring mutations in pncA and the transmission of multidrug resistant TB in Georgia," *BMC infectious diseases*, vol. 17, pp. 1–9, 2017.

[54] T. V. Tanja Stadler, Carsten Magnus, *Statistical and Computational Analysis of Genetic Sequence Data*. ETH Zurich, 2020.

[55] J. Scire, J. Barido-Sottani, D. Kühnert, T. G. Vaughan, and T. Stadler, "Improved multi-type birth-death phylodynamic inference in BEAST 2," *BioRxiv*, pp. 2020–01, 2020.

[56] T. Stadler and S. Bonhoeffer, "Uncovering epidemiological dynamics in heterogeneous host populations using phylogenetic methods," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 368, no. 1614, p. 20120198, 2013.

[57] D. A. Rasmussen and T. Stadler, "Coupling adaptive molecular evolution to phylodynamics using fitness-dependent birth-death models," *Elife*, vol. 8, p. e45562, 2019.

[58] D. Rasmussen, "Lumiere," https://github.com/davidrasm/Lumiere, 2021.

[59] L. Kepler, M. Hamins-Puertolas, and D. A. Rasmussen, "Decomposing the sources of SARS-CoV-2 fitness variation in the United States," *Virus Evolution*, vol. 7, no. 2, p. veab073, 2021.

[60] T. Stadler, "On incomplete sampling under birth–death models and connections to the sampling-based coalescent," *Journal of theoretical biology*, vol. 261, no. 1, pp. 58–66, 2009.

[61] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: A System for Large-Scale Machine Learning," in *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*. USENIX Association, 2016, p. 265–283.

[62] W. H. Organization *et al.*, "Catalogue of mutations in *Mycobacterium tuberculosis* complex and their association with drug resistance," 2021.

[63] L.-T. Nguyen, H. A. Schmidt, A. Von Haeseler, and B. Q. Minh, "IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies," *Molecular biology and evolution*, vol. 32, no. 1, pp. 268–274, 2015.

[64] A. Rambaut, T. T. Lam, L. Max Carvalho, and O. G. Pybus, "Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen)," *Virus evolution*, vol. 2, no. 1, p. vew007, 2016.

[65] F. Menardo, S. Duchêne, D. Brites, and S. Gagneux, "The molecular clock of *Mycobacterium tuberculosis*," *PLoS pathogens*, vol. 15, no. 9, p. e1008067, 2019.

[66] T.-H. To, "Rlsd2," https://github.com/davidrasm/phyloTF2/tree/main, 2020.

[67] T.-H. To, M. Jung, S. Lycett, and O. Gascuel, "Fast Dating Using Least-Squares Criteria and Algorithms," *Systematic Biology*, vol. 65, no. 1, pp. 82–97, 2015.

[68] S. A. Ishikawa, A. Zhukova, W. Iwasaki, and O. Gascuel, "A fast likelihood method to reconstruct and visualize ancestral scenarios," *Molecular biology and evolution*, vol. 36, no. 9, pp. 2069–2085, 2019.

[69] I. Cohen, Y. Huang, J. Chen, J. Benesty, J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," *Noise reduction in speech processing*, pp. 1–4, 2009.

[70] D. Rasmussen, "phyloTF," https://github.com/davidrasm/phyloTF2, 2021.

[71] C. Ozcaglar, A. Shabbeer, S. L. Vandenberg, B. Yener, and K. P. Bennett, "Epidemiological models of *Mycobacterium tuberculosis* complex infections," *Mathematical biosciences*, vol. 236, no. 2, pp. 77–96, 2012.

[72] V. Kapur, L.-L. Li, S. Iordanescu, M. R. Hamrick, A. Wanger, B. N. Kreiswirth, and J. M. Musser, "Characterization by automated DNA sequencing of mutations in the gene (rpoB) encoding the RNA polymerase beta subunit in rifampin-resistant Mycobacterium tuberculosis strains from New York City and Texas," *Journal of clinical microbiology*, vol. 32, no. 4, pp. 1095–1098, 1994.

[73] A. Scorpio, P. Lindholm-Levy, L. Heifets, R. Gilman, S. Siddiqi, M. Cynamon, and Y. Zhang, "Characterization of pncA mutations in pyrazinamide-resistant *Mycobacterium tuberculosis*," *Antimicrobial agents and chemotherapy*, vol. 41, no. 3, pp. 540–543, 1997.

[74] Y. Zhang, C. Vilchèze, and W. R. Jacobs JR, "Mechanisms of drug resistance in *Mycobacterium tuberculosis*," *Tuberculosis and the tubercle bacillus*, pp. 115–140, 2004.

[75] Y. Zhang, M. M. Wade, A. Scorpio, H. Zhang, and Z. Sun, "Mode of action of pyrazinamide: disruption of *Mycobacterium tuberculosis* membrane transport and energetics by pyrazinoic acid," *Journal of antimicrobial chemotherapy*, vol. 52, no. 5, pp. 790–795, 2003.

[76] S.-J. Cheng, L. Thibert, T. Sanchez, L. Heifets, and Y. Zhang, "pncA mutations as a major mechanism of pyrazinamide resistance in *Mycobacterium tuberculosis*: spread of a monoresistant strain in Quebec, Canada," *Antimicrobial agents and chemotherapy*, vol. 44, no. 3, pp. 528–532, 2000.

[77] K. Johnsson, W. A. Froland, and P. G. Schultz, "Overexpression, purification, and characterization of the catalase-peroxidase KatG from *Mycobacterium tuberculosis*," *Journal of Biological Chemistry*, vol. 272, no. 5, pp. 2834–2840, 1997.

[78] G. Middlebrook and M. L. Cohn, "Some observations on the pathogenicity of isoniazid-resistant variants of tubercle bacilli," *Science*, vol. 118, no. 3063, pp. 297–299, 1953.

[79] Z. Li, C. Kelley, F. Collins, D. Rouse, and S. Morris, "Expression of katG in *Mycobacterium tuberculosis* is associated with its growth and persistence in mice and guinea pigs," *Journal of Infectious Diseases*, vol. 177, no. 4, pp. 1030–1035, 1998.

[80] S. Gagneux, M. V. Burgos, K. DeRiemer, A. Enciso, S. Muñoz, P. C. Hopewell, P. M. Small, and A. S. Pym, "Impact of bacterial genetics on the transmission of isoniazid-resistant *Mycobacterium tuberculosis*," *PLoS pathogens*, vol. 2, no. 6, p. e61, 2006.

[81] S. Gagneux, "Fitness cost of drug resistance in *Mycobacterium tuberculosis*," *Clinical Microbiology and Infection*, vol. 15, pp. 66–68, 2009.

[82] D. R. Sherman, K. Mdluli, M. J. Hickey, T. M. Arain, S. L. Morris, C. E. Barry III, and C. K. Stover, "Compensatory ahp C gene expression in isoniazid-resistant *Mycobacterium tuberculosis*," *Science*, vol. 272, no. 5268, pp. 1641–1643, 1996.

[83] G. Napier, S. Campino, J. E. Phelan, and T. G. Clark, "Large-scale genomic analysis of *Mycobacterium tuberculosis* reveals extent of target and compensatory mutations linked to multi-drug resistant tuberculosis," *Scientific Reports*, vol. 13, no. 1, p. 623, 2023.

[84] N. L. Wengenack, J. R. Uhl, A. L. St. Amand, A. J. Tomlinson, L. M. Benson, S. Naylor, B. C. Kline, F. R. Cockerill III, and F. Rusnak, "Recombinant *Mycobacterium tuberculosis* KatG (S315T) is a competent catalase-peroxidase with reduced activity toward isoniazid," *Journal of Infectious Diseases*, vol. 176, no. 3, pp. 722–727, 1997.

[85] H. Marrakchi, S. Ducasse, G. Labesse, H. Montrozier, E. Margeat, L. Emorine, X. Charpentier, M. Daffé, and A. Quémard, "MabA (FabG1), a *Mycobacterium tuberculosis* protein involved in the long-chain fatty acid elongation system FAS-II," *Microbiology*, vol. 148, no. 4, pp. 951–960, 2002.

[86] Y.-X. Xiao, K.-H. Liu, W.-H. Lin, T.-H. Chan, and R. Jou, "Whole-genome sequencing-based analyses of drug-resistant *Mycobacterium tuberculosis* from Taiwan," *Scientific Reports*, vol. 13, no. 1, p. 2540, 2023.

[87] D. Liu, F. Huang, G. Zhang, W. He, X. Ou, P. He, B. Zhao, B. Zhu, F. Liu, Z. Li *et al.*, "Whole-genome sequencing for surveillance of tuberculosis drug resistance and determination of resistance level in China," *Clinical Microbiology and Infection*, vol. 28, no. 5, pp. 731–e9, 2022.

[88] J. Li, T. Yang, C. Hong, Z. Yang, L. Wu, Q. Gao, H. Yang, and W. Tan, "Whole-genome sequencing for resistance level prediction in multidrug-resistant tuberculosis," *Microbiology Spectrum*, vol. 10, no. 3, pp. e02 714–21, 2022.

[89] L. Chaidir, C. Ruesen, B. E. Dutilh, A. R. Ganiem, A. Andryani, L. Apriani, M. A. Huynen, R. Ruslami, P. C. Hill, R. van Crevel *et al.*, "Use of whole-genome sequencing to predict *Mycobacterium tuberculosis* drug resistance in Indonesia," *Journal of global antimicrobial resistance*, vol. 16, pp. 170–177, 2019.

[90] Z. Bakuła, A. Napiórkowska, J. Bielecki, E. Augustynowicz-Kopeć, Z. Zwolska, T. Jagielski *et al.*, "Mutations in the embB gene and their association with ethambutol resistance in multidrug-resistant *Mycobacterium tuberculosis* clinical isolates from Poland," *BioMed research international*, vol. 2013, 2013.

[91] A. Telenti, W. J. Philipp, S. Sreevatsan, C. Bernasconi, K. E. Stockbauer, B. Wieles, J. M. Musser, and W. R. Jacobs Jr, "The emb operon, a gene cluster of *Mycobacterium tuberculosis* involved in resistance to ethambutol," *Nature medicine*, vol. 3, no. 5, pp. 567–570, 1997.

[92] H. Safi, R. D. Fleischmann, S. N. Peterson, M. B. Jones, B. Jarrahi, and D. Alland, "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*, vol. 54, no. 1, pp. 103–108, 2010.

[93] E. Guerrero, D. Lemus, S. Yzquierdo, G. Vílchez, M. Muñoz, E. Montoro, and H. Takiff, "Association between embB mutations and ethambutol resistance in *Mycobacterium tuberculosis* isolates from Cuba and the Dominican Republic: reproducible patterns and problems," *Rev Argent Microbiol*, vol. 45, no. 1, pp. 21–26, 2013.

Appendix A

Data

A.1 Mutations

Locus	Conferring drug resistance	Abbreviation
gyrB	levofloxacin, moxifloxacin	LEV, MXF
gyrA	levofloxacin, moxifloxacin	LEV, MXF
rpoB	rifampicin	RIF
rpsL	streptomycin	STM
fabG1	ethionamide, isoniazid	ETH, INH
inhA G154A	rifampicin	RIF
inhA I194T	ethionamide, isoniazid	ETH, INH
inhA S94A	ethionamide, isoniazid	ETH, INH
katG	isoniazid	INH
pncA	pyrazinamide	PZA
eis C12T	kanamycin	KAN
eis C14T	capreomycin, kanamycin	CAP, KAN
eis G10A	kanamycin	KAN
eis G10C	kanamycin	KAN
embA	ethambutol	EMB
embB	ethambutol	EMB
gidB	streptomycin	STM
rrs A1401G	amikacin, capreomycin, kanamycin	AMI, CAP, KAN
rrs A514C	streptomycin	STM
rrs A906G	streptomycin	STM
rrs C517T	streptomycin	STM

Table A.1: This table denotes toward which antibiotic a mutation in a specific gene confers resistance. In case different mutations in the same gene confer resistance to different drugs in this study, the specific mutation is specified in the first column. It entails all genes analyzed in this study. The third column represents the abbreviation of the respective drugs. The genes are ordered by their

A.2 Fitness Estimates

Mutation	Linked with	Frequency	Fitness effect
gyrA A90V		0.037	0.996 (0.667 - 1.167)
gyrA D94A		0.008	1.000 (0.218 - 1.240)
gyrA D94G		0.080	1.060 (0.945 - 1.247)
gyrA D94H		0.006	1.000 (0.231 - 1.463)
gyrA D94N		0.018	1.000 (0.630 - 1.422)
gyrA D94Y	*rpoB* S450L, *fabG1* C15T	0.008	1.000 (0.432 - 1.575)
gyrA S91P	*rpoB* S450L, *fabG1* C15T *	0.009	1.000 (0.552 - 1.890)
rpoB S450L		0.690	0.423 (0.404 - 0.441)
rpoB Q432P	*fabG1* C15T	0.005	1.000 (0.289 - 1.544)
rpoB L430P		0.023	0.381 (0.291 - 0.453)
rpoB H445Y		0.041	0.394 (0.315 - 0.459)
rpoB H445R		0.008	0.604 (0.164 - 0.807)
rpoB L452P		0.043	0.368 (0.291 - 0.432)
rpoB H445D		0.029	0.347 (0.277 - 0.405)
rpoB H445L		0.015	0.481 (0.297 - 0.605)
rpoB D435G		0.006	0.889 (0.390 - 1.102)
rpoB L731P (C)	*rpoB* S450L, *rpsL* K43R *	0.014	1.000 (0.618 - 1.280)
rpoB D435Y		0.028	0.521 (0.374 - 0.627)
rpoB D435V		0.082	0.408 (0.351 - 0.457)
rpoC D485Y (C)	*rpoB* S450L	0.065	1.150 (1.011 - 1.409)
rpoC V483G (C)	*rpoB* S450L, *katG* S315T *	0.136	1.114 (1.015 - 1.262)
rpoC P1040R (C)	*rpoB* S450L, *fabG1* C15T	0.027	1.000 (0.815 - 1.383)
rpoC L527V (C)	*rpoB* S450L, *fabG1* C15T	0.020	1.000 (0.642 - 1.226)
rpoC G433S (C)	*rpoB* S450L, *embB* M396V	0.010	1.168 (0.933 - 2.217)
rpsL K43R		0.157	1.017 (0.934 - 1.132)
rpsL K88R		0.033	0.836 (0.650 - 0.958)
rrs A1401G		0.186	1.042 (0.956 - 1.163)
fabG1 C15T		0.627	0.900 (0.853 - 0.942)
fabG1 G17T	*rpoB* D435G, *rrs* A514C *	0.073	1.000 (0.870 - 1.161)
inhA I194T	*rpoB* S450L, *fabG1* C15T *	0.177	1.138 (1.042 - 1.279)
katG S315R	*rpsL* K88R, *katG* S315T	0.010	1.000 (0.770 - 1.698)
katG S315T		0.341	0.396 (0.373 - 0.418)
pncA C14R	*rpoB* S450L, *rrs* A514C *	0.122	1.001 (0.904 - 1.133)
pncA D8N	*rpoB* S450L, *fabG1* C15T *	0.120	1.129 (1.015 - 1.315)
pncA I31S	*rpoB* S450L, *rrs* A514C *	0.005	1.000 (0.589 - 2.631)
pncA L120R	*rpoB* H445L, *rpsL* K43R	0.008	1.025 (0.752 - 4.328)
pncA S164P	*rpoB* S450L, *fabG1* C15T *	0.012	1.000 (0.695 - 1.583)
pncA Y103!	*rpoB* S450L, *fabG1* C15T *	0.092	1.170 (1.053 - 1.359)
eis G10C	*rpoB* H445L, *rpsL* K43R *	0.008	0.959 (0.381 - 1.177)
rpoA T187A (C)	*rpoB* S450L, *fabG1* C15T	0.048	1.310 (1.133 - 1.737)
rpoA V183G (C)	*rpoB* S450L	0.026	1.187 (0.980 - 1.850)
embA C12T	*rpoB* S450L	0.023	1.027 (0.829 - 1.635)
embB G406C	*katG* S315T	0.008	1.001 (0.725 - 5.728)
embB Q497R	*rpoB* S450L	0.018	1.000 (0.775 - 1.276)
embB Q497K	*rpoB* S450L, *rrs* A514C *	0.028	1.000 (0.729 - 1.186)
embB M306V	*rpoB* S450L	0.264	1.183 (1.102 - 1.295)
embB M306I	*katG* S315T	0.233	1.925 (1.804 - 2.124)
gidB L79S	*katG* S315T	0.222	1.396 (1.301 - 1.536)
rrs A906G, *pncA* V139A	*rpoB* S450L, *fabG1* C15T *	0.005	1.000 (0.574 - 5.401)
rrs A514C		0.274	1.007 (0.941 - 1.091)
rrs C517T	*rpoB* S450L	0.031	1.069 (0.874 - 1.641)

Table A.2: Estimated fitness effects of mutations and their linkages in South Africa L2. Where two mutations are listed, they correlate. Mutations ending with *(C)* are putative compensatory mutations. The second column lists all linked DR mutations to the mutations in the first column. '*' in the second column indicates a linkage to more than two mutations. The entire list of linked mutations can be seen in Tab. A.6 and A.7. DR and compensatory mutations whose fitness effect is significantly larger than 1 are highlighted in yellow or gray respectively. DR mutations inferring a fitness cost are highlighted in red.

Mutation	Linked with	Frequency	Fitness effect
gyrA A90V	*katG* S315T	0.032	1.681 (1.225 - 2.155)
gyrA D94G	*katG* S315T	0.034	1.355 (0.994 - 1.742)
rpoB S450W		0.029	0.435 (0.310 - 0.564)
rpoB S450L		0.476	0.348 (0.331 - 0.365)
rpoB S441L		0.014	0.296 (0.205 - 0.388)
rpoB R827C (C)	*gyrA* A90V, *rpoB* S450L *	0.017	1.079 (0.677 - 1.508)
rpoB L430P		0.077	0.301 (0.269 - 0.333)
rpoB L452P		0.037	0.310 (0.270 - 0.349)
rpoB H445N		0.026	0.335 (0.257 - 0.414)
rpoB H445D		0.063	0.340 (0.300 - 0.379)
rpoB D435Y		0.080	0.469 (0.398 - 0.540)
rpoB D435V		0.086	0.599 (0.498 - 0.703)
rpoB D435F	*rpoB* D435Y, *rpoB* D435V	0.029	1.302 (0.940 - 1.681)
rpoB H445Y		0.083	0.294 (0.268 - 0.320)
rpoC V1252L (C)	*rpoB* S450L, *katG* S315T *	0.032	2.577 (1.783 - 3.434)
rpoC P1040R (C)	*rpoB* S450L	0.011	1.046 (0.750 - 1.345)
rpoC V483G (C)	*rpoB* S450L	0.011	1.048 (0.615 - 1.482)
rpsL K43R	*katG* S315T	0.135	1.235 (1.094 - 1.378)
rpsL K88Q, *pncA* K96T	*rpoB* S450L, *katG* S315T	0.017	2.225 (1.361 - 3.157)
rrs A1401G	*katG* S315T	0.012	0.758 (0.285 - 1.267)
fabG1 C15T		0.095	0.564 (0.499 - 0.629)
inhA I194T	*rpoB* S450L	0.014	0.975 (0.640 - 1.312)
inhA S94A		0.040	0.878 (0.721 - 1.038)
katG S315T		0.573	0.526 (0.498 - 0.554)
katG S315R	*rpsL* K43R, *katG* S315T	0.074	1.219 (1.010 - 1.435)
pncA G97C	*rpoB* H445D, *katG* S315T	0.017	1.630 (1.195 - 2.075)
pncA R154G	*gyrA* A90V, *rpoB* S450L *	0.017	1.079 (0.677 - 1.508)
pncA Q10P	*katG* S315T	0.043	1.517 (1.151 - 1.900)
pncA K96E	*rpoB* S450L, *rpsL* K43R *	0.017	1.033 (0.684 - 1.398)
pncA H71Y	*katG* S315T	0.014	1.510 (1.141 - 1.878)
eis C14T	*rpoB* S450L, *katG* S315T	0.020	0.925 (0.583 - 1.289)
embA C16G	*fabG1* C15T, *katG* S315T	0.023	2.114 (1.632 - 2.615)
embB G406A	*rpoB* S450L, *katG* S315T	0.026	2.238 (1.461 - 3.092)
embB G406D	*katG* S315T	0.014	0.698 (0.453 - 0.957)
embB M306V	*rpoB* S450L, *katG* S315T	0.129	1.472 (1.288 - 1.659)
embB Q497R		0.052	1.445 (1.168 - 1.727)
embB M306I		0.152	0.932 (0.825 - 1.042)
gidB C52!	*rpoB* S450L, *katG* S315T *	0.014	1.223 (0.734 - 1.746)
gidB W148	*rpoB* S450L, *katG* S315T *	0.011	0.693 (0.402 - 1.002)

Table A.3: Estimated fitness effects of mutations and their linkages in South Africa L4. Where two mutations are listed, they correlate. Mutations ending with *(C)* are putative compensatory mutations. The second column lists all linked DR mutations to the mutations in the first column. '*' in the second column indicates a linkage to more than two mutations. The entire list of linked mutations can be seen in Tab. A.8 and A.9. DR and compensatory mutations whose fitness effect is significantly larger than 1 are highlighted in yellow or gray respectively. DR mutations inferring a fitness cost are highlighted in red.

Mutations	Linked with	Frequency	Fitness effect
gyrB A504V	*rpoB* S450L, *inhA* G-154A *	0.010	1.000 (0.772 - 1.252)
gyrA A90V	*katG* S315T	0.074	1.076 (0.985 - 1.185)
gyrA D94A		0.017	0.690 (0.514 - 0.833)
gyrA D94G	*katG* S315T	0.091	1.126 (1.049 - 1.215)
gyrA D94Y	*katG* S315T	0.009	1.263 (1.048 - 1.619)
gyrA S91P	*rpoB* S450L, *katG* S315T	0.009	1.000 (0.752 - 1.374)
rpoB L731P (C)	*rpoB* S450L, *katG* S315T	0.017	1.294 (1.108 - 1.580)
rpoB S450L	*katG* S315T	0.593	0.979 (0.953 - 1.005)
rpoB L452P		0.011	0.584 (0.434 - 0.711)
rpoB L430P		0.013	0.741 (0.571 - 0.881)
rpoB H445Y	*rpsL* K43R	0.015	1.000 (0.806 - 1.217)
rpoB D435Y		0.013	0.892 (0.691 - 1.049)
rpoB E761D (C)	*rpoB* S450L, *katG* S315T *	0.020	1.012 (0.888 - 1.172)
rpoC G332C (C)	*rpoB* S450L, *rpsL* K88R *	0.019	1.353 (1.167 - 1.639)
rpoC F452C (C)	*rpoB* S450L, *katG* S315T *	0.011	1.025 (0.848 - 1.291)
rpoC D485N (C)	*rpoB* S450L, *rpsL* K43R *	0.016	0.851 (0.713 - 0.967)
rpoC I491T (C)	*rpoB* S450L, *rpsL* K43R *	0.011	0.951 (0.737 - 1.115)
rpoC V483G (C)	*rpoB* S450L, *katG* S315T	0.122	1.145 (1.077 - 1.223)
rpsL K43R	*katG* S315T	0.452	1.001 (0.971 - 1.032)
rpsL K88R	*katG* S315T	0.148	1.518 (1.437 - 1.612)
fabG1 C15T		0.159	0.476 (0.451 - 0.500)
inhA G154A	*katG* S315T	0.029	1.011 (0.897 - 1.155)
katG S315T		0.675	0.352 (0.343 - 0.360)
pncA T142M	*gyrA* A90V, *rpoB* S450L *	0.014	1.260 (1.052 - 1.603)
eis G10A	*katG* S315T	0.079	0.906 (0.832 - 0.972)
rpoA T187A (C)	*rpoB* S450L, *rpsL* K43R *	0.018	1.056 (0.908 - 1.259)
rpoA E184D (C)	*rpoB* S450L, *rpsL* K43R *	0.018	1.129 (0.974 - 1.347)
embA C12T	*rpoB* S450L, *katG* S315T	0.093	1.255 (1.171 - 1.354)
embB M306L	*katG* S315T	0.010	1.028 (0.837 - 1.326)
embB M306I	*katG* S315T	0.064	0.869 (0.788 - 0.941)
embB G406S	*katG* S315T	0.007	0.741 (0.519 - 0.913)
embB G406D		0.013	0.449 (0.352 - 0.535)
embB G406A		0.036	0.673 (0.602 - 0.739)
embB D354A	*rpoB* S450L, *katG* S315T	0.045	0.886 (0.796 - 0.967)
embB M306V	*katG* S315T	0.254	1.000 (0.957 - 1.044)
embB Q497R	*rpsL* K43R, *katG* S315T	0.054	1.010 (0.925 - 1.112)

Table A.4: Estimated fitness effects of mutations and their linkages in Georgia L2. Mutations ending with *(C)* are putative compensatory mutations. The second column lists all linked DR mutations to the mutations in the first column. '*' in the second column indicates a linkage to more than two mutations. The entire list of linked mutations can be seen in Tab. A.10 and A.11. DR and compensatory mutations whose fitness effect is significantly larger than 1 are highlighted in yellow or gray respectively. DR and compensatory mutations inferring a fitness cost are highlighted in red and blue respectively.

Mutations	Linked with	Frequency	Fitness effect
gyrA A90V		0.006	1.361 (0.929 - 1.827)
gyrA D94G		0.007	0.822 (0.401 - 1.317)
rpoB D435V	*katG* S315T	0.011	0.683 (0.530 - 0.842)
rpoB D435Y		0.008	0.610 (0.455 - 0.771)
rpoB H445D		0.019	0.373 (0.306 - 0.442)
rpoB H445Y		0.008	0.474 (0.355 - 0.597)
rpoB S450L		0.067	0.445 (0.410 - 0.481)
rpsL K43R		0.026	0.432 (0.372 - 0.494)
rpsL K88R	*katG* S315T	0.011	1.429 (1.145 - 1.722)
fabG1 C15T		0.062	0.421 (0.384 - 0.459)
katG S315T		0.118	0.383 (0.360 - 0.406)
eis C12T	*katG* S315T	0.016	2.325 (1.879 - 2.794)
eis C14T	*katG* S315T	0.008	2.972 (2.119 - 3.000)
embA C12T	*katG* S315T	0.016	2.286 (1.837 - 2.761)
embB D1024N		0.008	0.687 (0.543 - 0.836)
embB G406A	*katG* S315T	0.005	0.991 (0.718 - 1.274)
embB G406D		0.005	0.450 (0.301 - 0.607)
embB M306I		0.025	0.521 (0.448 - 0.596)
embB M306V		0.018	0.842 (0.718 - 0.968)
embB Q497R	*katG* S315T	0.023	1.145 (0.961 - 1.338)

Table A.5: Estimated fitness effects of mutations and their linkages in Georgia, L4. As the only linkages in here are to *katG* S315T, the second column represents the entire information on linked DR mutations in this data set. DR and compensatory mutations whose fitness effect is significantly larger than 1 are highlighted in yellow. DR mutations inferring a fitness cost are highlighted in red.

A.3 Linkages

South Africa Lineage 2

DR mutation(s)	Linked with
gyrA S91P	*rpoB* S450L, *fabG1* C15T, *inhA* I194T, *embB* M306V
gyrA D94Y	*rpoB* S450L, *fabG1* C15T
rpoB Q432P	*fabG1* C15T
fabG1 G-17T	*rpoB* D435V, *rrs* A514C, *katG* S315T, *embB* M306I, *gidB* L79S
inhA I194T	*rpoB* S450L, *fabG1* C15T, *embB* M306V
katG S315R	*rpsL* K88R, *katG* S315T
pncA S164P	*rpoB* S450L, *fabG1* C15T, *inhA* I194T, *embB* M306V
pncA Y103!	*rpoB* S450L, *rpsL* K43R, *fabG1* C15T
pncA C14R	*rpoB* S450L, *rrs* A514C, *katG* S315T, *katG* M306I, *gidB* L79S
pncA D8N	*rpoB* S450L, *fabG1* C15T, *inhA* I194T, *katG* M306V
pncA S164P	*rpoB* S450L, *fabG1* C15T, *inhA* I194T, *katG* M306V
pncA L120R	*rpoB* H445L, *rpsL* K43R, *katG* S315T, *katG* M306V
pncA I31S	*rpoB* S450L, *rrs* A514C, *katG* S315T, *katG* M306V; *gidB* L79S
eis G-10C	*rpoB* S450L, *rpsL* K43R, *fabG1* C15T, *pncA* Y103!
embA C12T	*rpoB* S450L
katG M306V	*rpoB* S450L
katG M306I	*katG* S315T
katG G406C	*katG* S315T
katG Q497R	*rpoB* S450L
katG Q497K	*rpoB* S450L, *rrs* A514C, *katG* S315T, *pncA* C14R, *katG* M306I, *gidB* L79S
gidB L79S	*katG* S315T
rrs A906G, *pncA* V139A	*gyrA* A90V, *rpoB* L452P, *katG* S315T, *gidB* L79S
rrs C517T	*rpoB* S450L

Table A.6: Entire list of DR mutations in South Africa L2 that are linked to another DR mutation in ≥95% (see Methods). *rrs* A906G and *pncA* V139A are correlated to each other, i.e. both linked to each other in ≥ 95%, and treated as an individual mutation.

Compensatory mutation	Linked DR mutations
rpoB L731P	*rpoB* S450L, *rpsL* K43R, *fabG1* C15T, *pncA* Y103!
rpoC G433S	*rpoB* S450L, *katG* M306V
rpoC V483G	*rpoB* S450L, *katG* S315T, *katG* M306I
rpoC D485Y	*rpoB* S450L
rpoC L527V	*rpoB* S450L, *fabG1* C15T
rpoC P1040R	*rpoB* S450L, *fabG1* C15T
rpoA, T187A	*rpoB* S450L, *fabG1* C15T
rpoA, V183G	*rpoB* S450L

Table A.7: Entire list of putative compensatory mutations for RIF analyzed in South Africa L2 that are linked to a DR mutation in ≥95% (see Methods).

South Africa Lineage 4

DR mutation(s)	Linked with
gyrA A90V	*katG* S315T
gyrA D94G	*katG* S315T
rpoB D435F	*rpoB* D435Y, *rpoB* D435V
rpsL K43R	*katG* S315T
rpsL K88Q, *pncA* K96T	*rpoB* S450L, *katG* S315T
rrs A1401G	*katG* S315T
inhA I194T	*rpoB* S450L
katG S315R	*rpsL* K43R, *katG* S315T
pncA R154G	*gyrA* A90V, *rpoB* S450L, *fabG1* C15T, *katG* S315T, *eis* C14T, *embA* C-16G, *katG* M306I
pncA G97C	*rpoB* H445D, *katG* S315T
pncA K96E	*rpoB* S450L, *rpsL* K43R, *katG* S315T, *katG* M306V
pncA H71Y	*katG* S315T
pncA Q10P	*katG* S315T
eis C14T	*rpoB* S450L, *katG* S315T
embA C-16G	*fabG1* C15T, *katG* S315T
katG M306V	*katG* S315T
katG G406D	*katG* S315T
katG G406A	*rpoB* S450L, *katG* S315T
gidB W148!	*rpoB* S450W, *katG* S315T, *katG* M306V
gidB C52!	*rpoB* S450L, *katG* S315T, *katG* M306V

Table A.8: Entire list of DR mutations in South Africa L4 that are linked to another DR mutation in ≥95% (see Methods). *rpsL* K88Q and *pncA* K96T are correlated to each other, i.e. both linked to each other in ≥ 95%, and treated as an individual mutation.

Compensatory mutation	Linked DR mutations
rpoB R827C (C)	*gyrA* A90V, *rpoB* S450L, *fabG1* C15T, *katG* S315T, *pncA* R154G, *eis* C14T, *embA* C-16G, *katG* M306I
rpoC V483G (C)	*rpoB* S450L
rpoC P1040R (C)	*rpoB* S450L
rpoC V1252L (C)	*rpoB* S450L, *katG* S315T, *katG* M306I

Table A.9: Entire list of putative compensatory mutations for RIF analyzed in South Africa L4 that are linked to a DR mutation in ≥95% (see Methods).

Georgia Lineage 2

DR mutation	Linked with
gyrB A504V	*rpoB* S450L, *inhA* G-154A, *katG* S315T, *katG* M306V
gyrA A90V	*katG* S315T
gyrA S91P	*rpoB* S450L, *katG* S315T
gyrA D94Y	*katG* S315T
gyrA D94G	*katG* S315T
rpoB H445Y	*rpsL* K43R
rpoB S450L	*katG* S315T
rpsL K43R	*katG* S315T
rpsL K88R	*katG* S315T
inhA G-154A	*katG* S315T
pncA T142M	*gyrA* A90V, *rpoB* S450L, *rpsL* K88R, *fabG1* C15T, *katG* S315T, *katG* M306V
eis G-10A	*katG* S315T
embA C12T	*rpoB* S450L, *katG* S315T
katG M306L	*rpoB* S450L, *rpsL* K43R, *katG* S315T
katG M306V	*katG* S315T
katG M306I	*katG* S315T
katG D354A	*rpoB* S450L, *katG* S315T
katG G406S	*katG* S315T
katG Q497R	*rpsL* K43R, *katG* S315T

Table A.10: Entire list of DR mutations in Georgia L2 that are linked to another DR mutation in ≥95% (see Methods).

Compensatory mutation	**Linked DR Mutations**
rpoA, T187A	*rpoB* S450L, *rpsL* K43R, *katG* S315T
rpoA, E184D	*rpoB* S450L, *rpsL* K43R, *katG* S315T, *katG* D354A
rpoB L731P	*rpoB* S450L, *katG* S315T
rpoB E761D	*rpoB* S450L, *katG* S315T, *katG* D354A
rpoC G332C	*rpoB* S450L, *rpsL* K88R, *fabG1* C15T; *katG* S315T, *katG* M306V
rpoC F452C	*rpoB* S450L, *rpsL* K43R, *katG* S315T, *katG* Q497R
rpoC V483G	*rpoB* S450L, *katG* S315T
rpoC D485N	*rpoB* S450L, *rpsL* K43R, *katG* S315T
rpoC I491T	*rpoB* S450L, *rpsL* K43R, *katG* S315T

Table A.11: Entire list of putative compensatory mutations for RIF analyzed in Georgia L2 that are linked to a DR mutation in ≥95% (see Methods).

Georgia Lineage 4

DR mutation	**Linked with**
rpoB D435V	*katG* S315T
rpsL K88R	*katG* S315T
eis C12T	*katG* S315T
eis C14T	*katG* S315T
embA C12T	*katG* S315T
embB G406A	*katG* S315T
embB Q497R	*katG* S315T

Table A.12: Entire list of DR mutations in Georgia L4 that are linked to another DR mutation in ≥95% (see Methods).

Appendix B

Plots

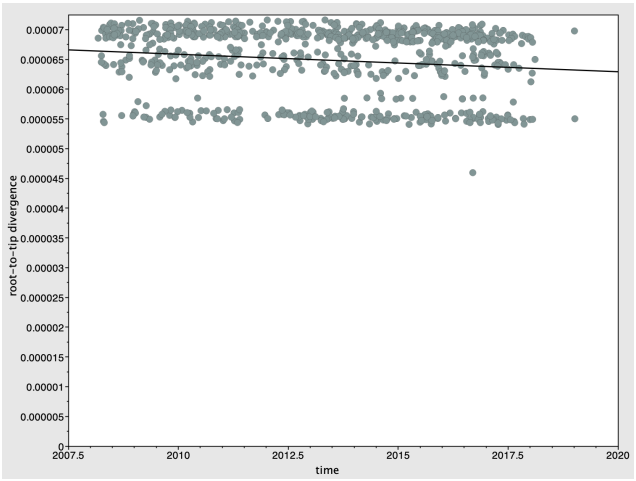

Figure B.1: Root-to-tip regression for South Africa L2 with $R^2 = 1.9884 \cdot 10^{-2}$, indicating that the temporal information does not fully explain the data variability.

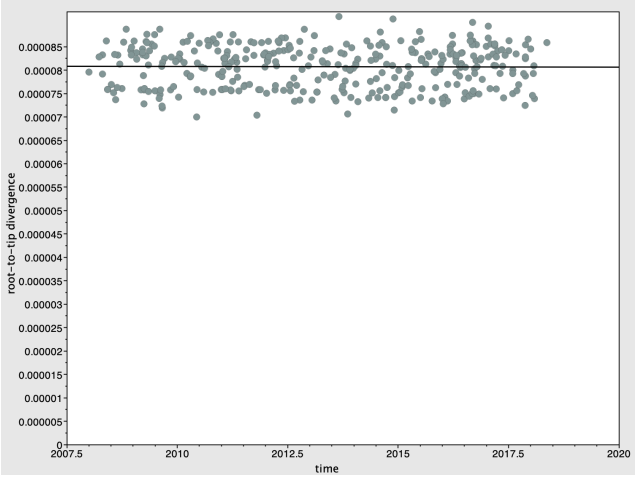

Figure B.2: Root-to-tip regression for South Africa L4 with $R^2 = 6.0736 \cdot 10^{-5}$, indicating that the temporal information does not fully explain the data variability.

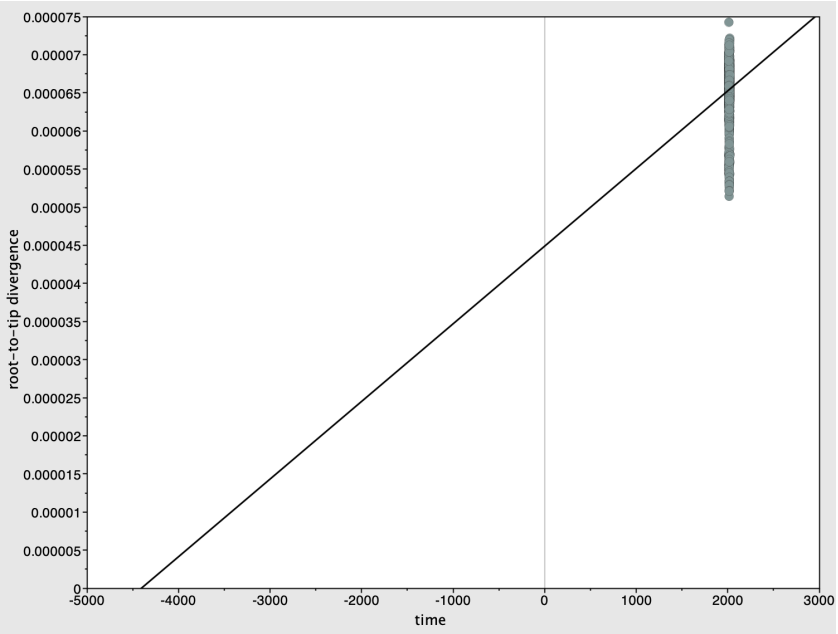

Figure B.3: Root-to-tip regression for Georgia L2 with $R^2 = 9.4909 \cdot 10^{-5}$, indicating that the temporal information does not fully explain the data variability.

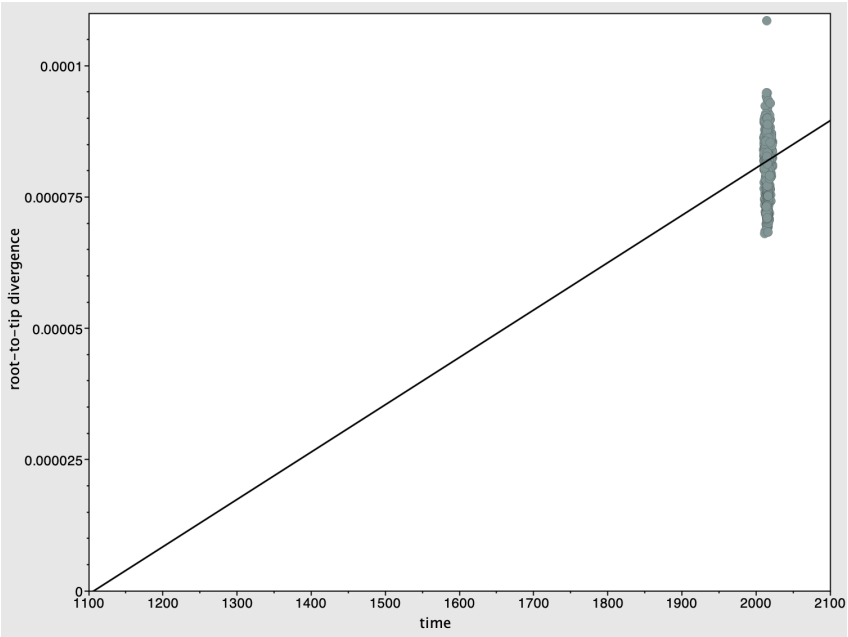

Figure B.4: Root-to-tip regression for Georgia L4 with $R^2 = 1.131 \cdot 10^{-3}$, indicating that the temporal information does not fully explain the data variability.

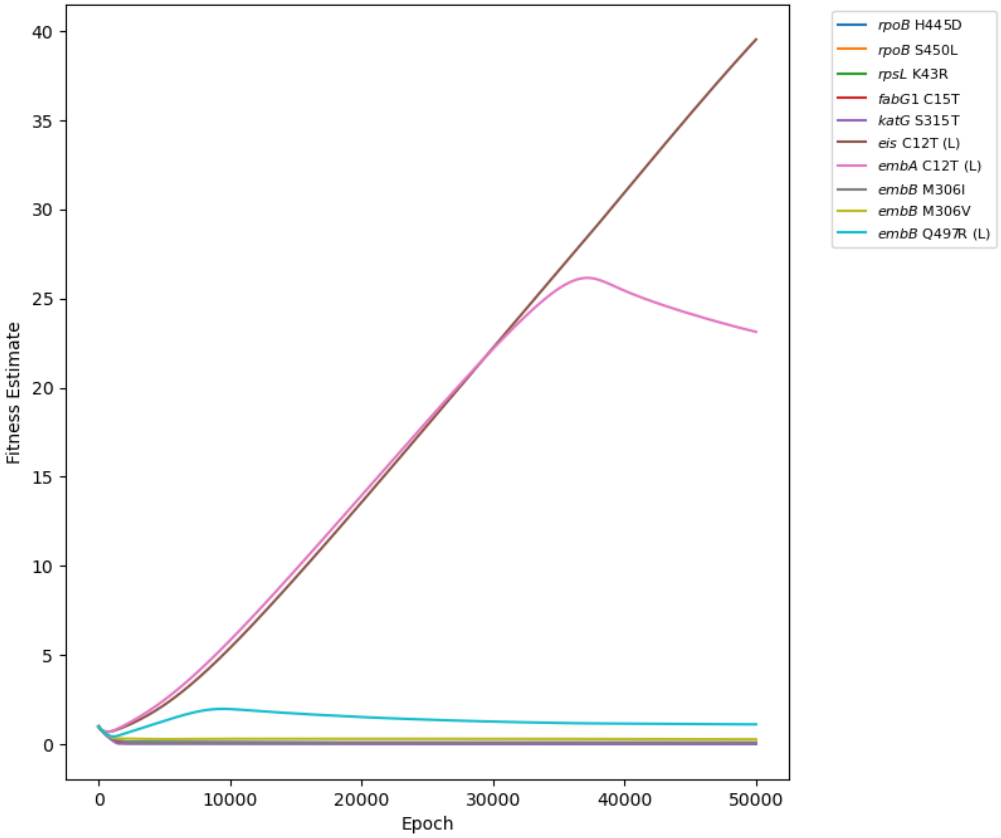

Figure B.5: Georgian L4 is unable to converge to meaningful value in a sensible amount of epochs with sampling rate $\sigma = 0.9$. The interaction terms of *eis* C12T with *katG* S315T and *embA* C12T with *katG* S315T (see Tab. A.12) are estimated to have fitness effects >20 and do not converge to a stable value in 50,000 epochs.

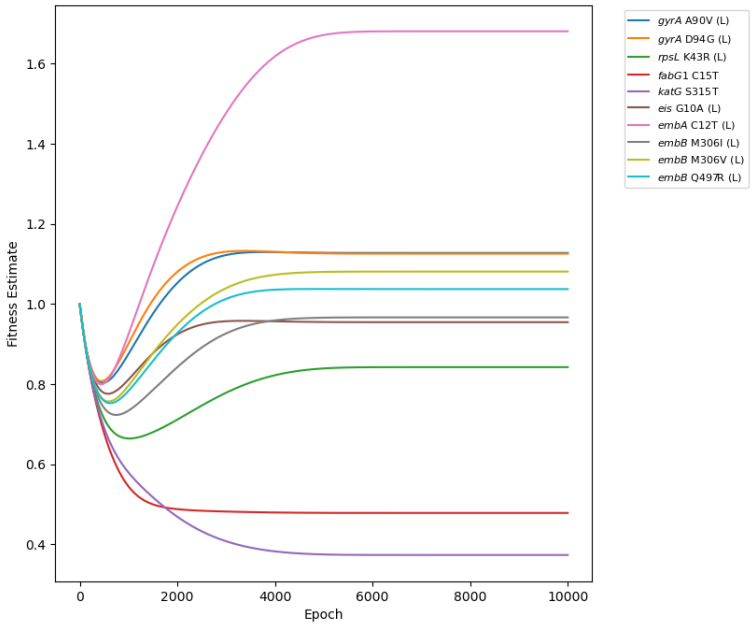

Figure B.6: A subset of ten common mutations in South African L2 converge to a stable fitness estimate in 10,000 epochs. The selected mutations are linked <80% to each other.

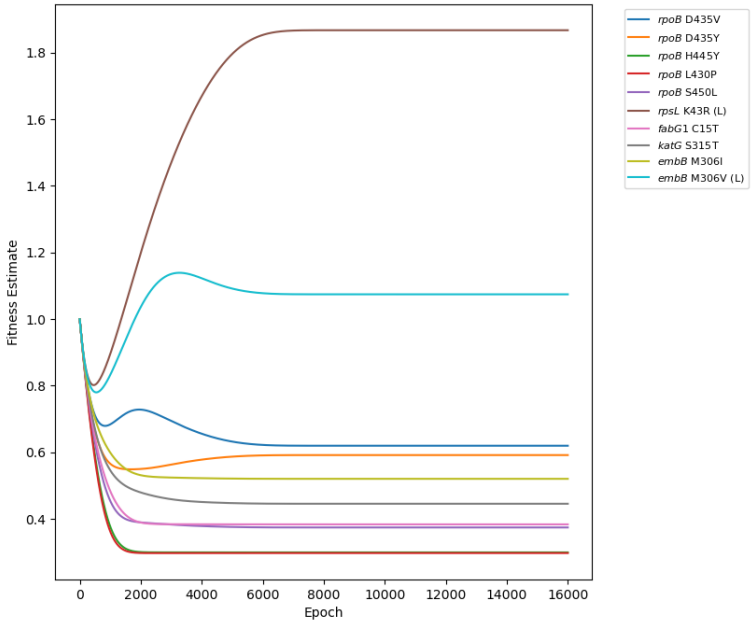

Figure B.7: A subset of ten common mutations in South African L4 converge to a stable fitness estimate in 16,000 epochs. The selected mutations are linked <80% to each other.

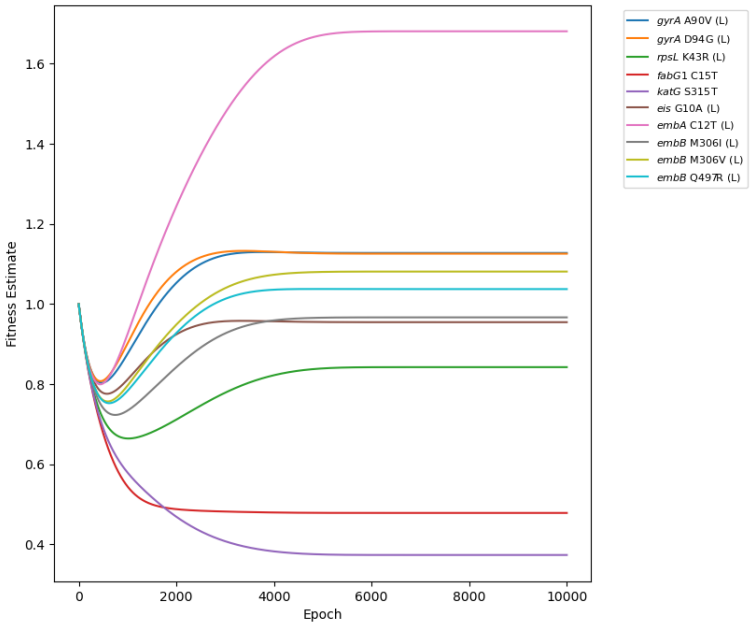

Figure B.8: A subset of ten common mutations in Georgia L2 converge to a stable fitness estimate in 10,000 epochs. The selected mutations are linked <80% to each other.

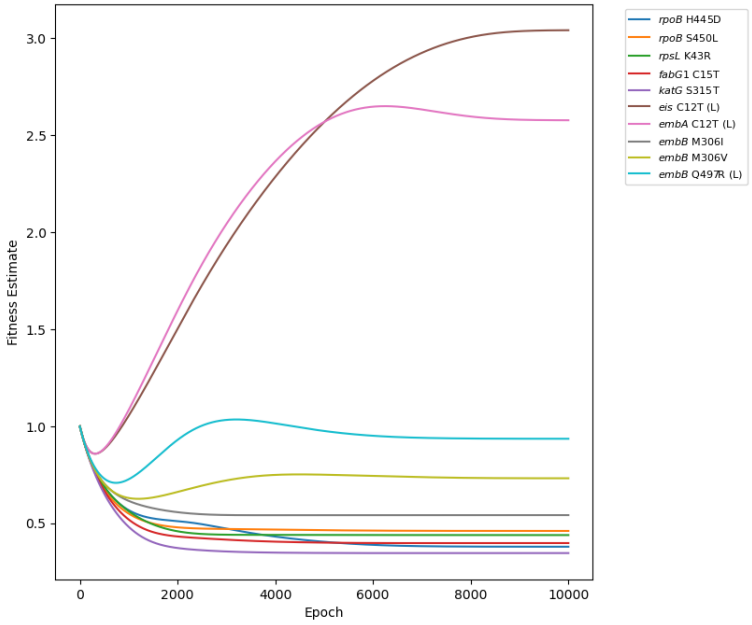

Figure B.9: A subset of ten common mutations in Georgia L4 converge to a stable fitness estimate in 10,000 epochs. The selected mutations are linked <80% to each other.

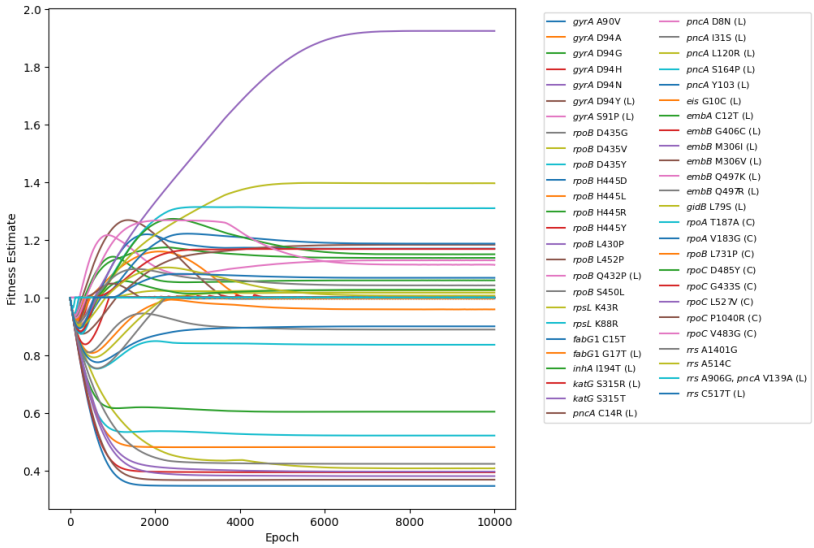

Figure B.10: DR and compensatory mutations in South African L2 converge in 10,000 epochs with L1 hyperparameter $\alpha = 5.9$. α was selected to result in the shortest Euclidean distance between the fitness estimates of a subset of ten mutations estimated alone without a regularizer and in the complete South Africa L2 data set with a regularizer (see Methods).

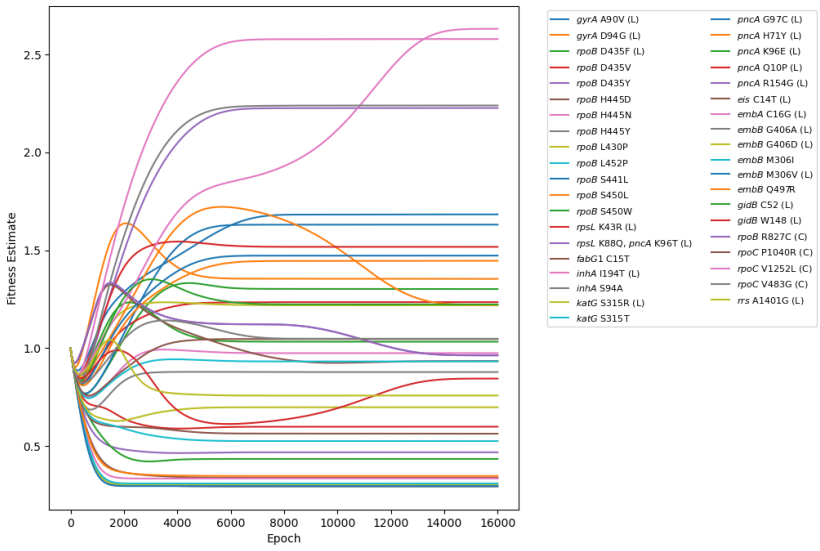

Figure B.11: DR and compensatory mutations in South African L4 converge in 16,000 epochs with L1 hyperparameter $\alpha = 0$. α was selected to result in the shortest Euclidean distance between the fitness estimates of a subset of ten mutations estimated alone without a regularizer and in the complete South Africa L4 data set with a regularizer (see Methods).

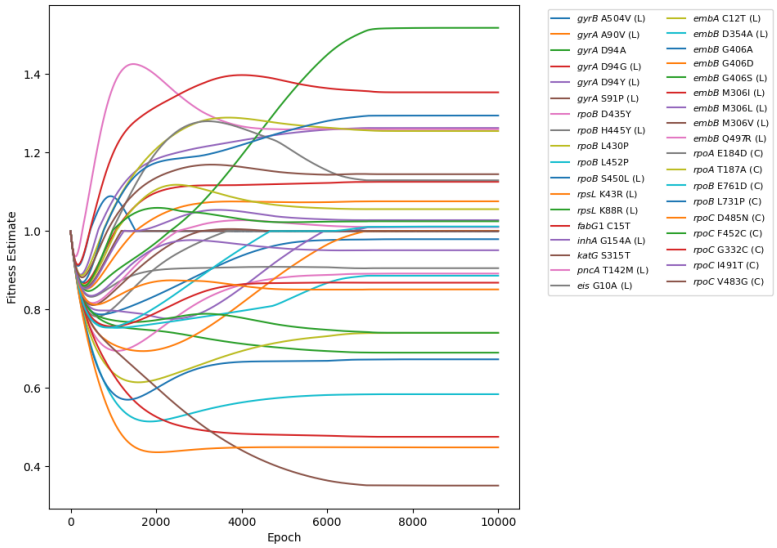

Figure B.12: DR and compensatory mutations in Georgian L2 converge in 10,000 epochs with L1 hyperparameter $\alpha = 3.1$. α was selected to result in the shortest Euclidean distance between the fitness estimates of a subset of ten mutations estimated alone without a regularizer and in the complete Georgia L2 data set with a regularizer (see Methods).

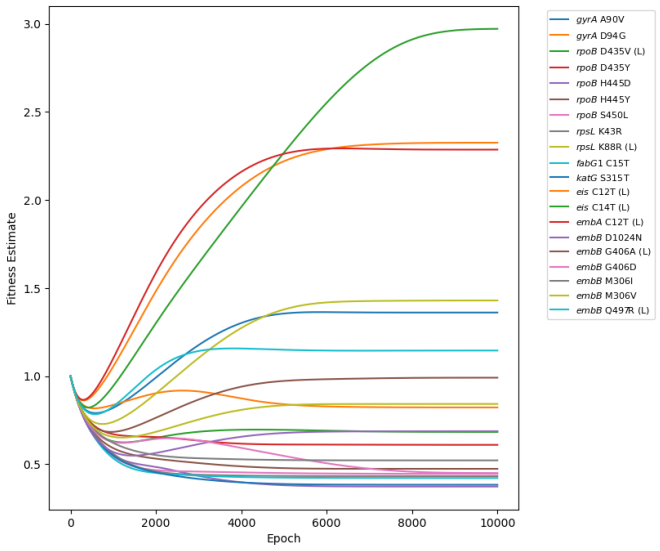

Figure B.13: DR and compensatory mutations in Georgian L4 converge in 10,000 epochs with L1 hyperparameter $\alpha = 0$. α was selected to result in the shortest Euclidean distance between the fitness estimates of a subset of ten mutations estimated alone without a regularizer and in the complete Georgia L4 data set with a regularizer (see Methods).

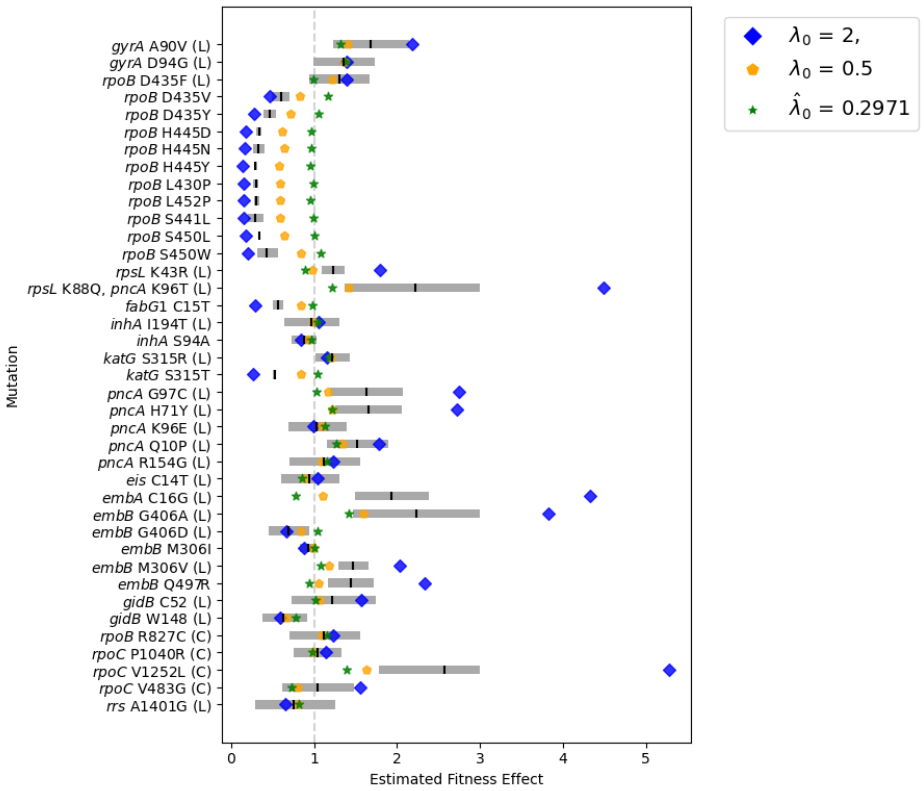

Figure B.14: Sensitivity analysis and estimation of λ_0 in South Africa L4. The vertical black line represents the mutation's fitness MLE for $\lambda_0 = 1$, $\mu = 1$, and $\sigma = 0.36$ and its surrounding grey box is the 95% CI. Fitness MLEs for increased/decreased values for λ are marked in blue and yellow. The green stars illustrate the fitness MLE for phyloTF's estimate of $\hat{\lambda}_0$.

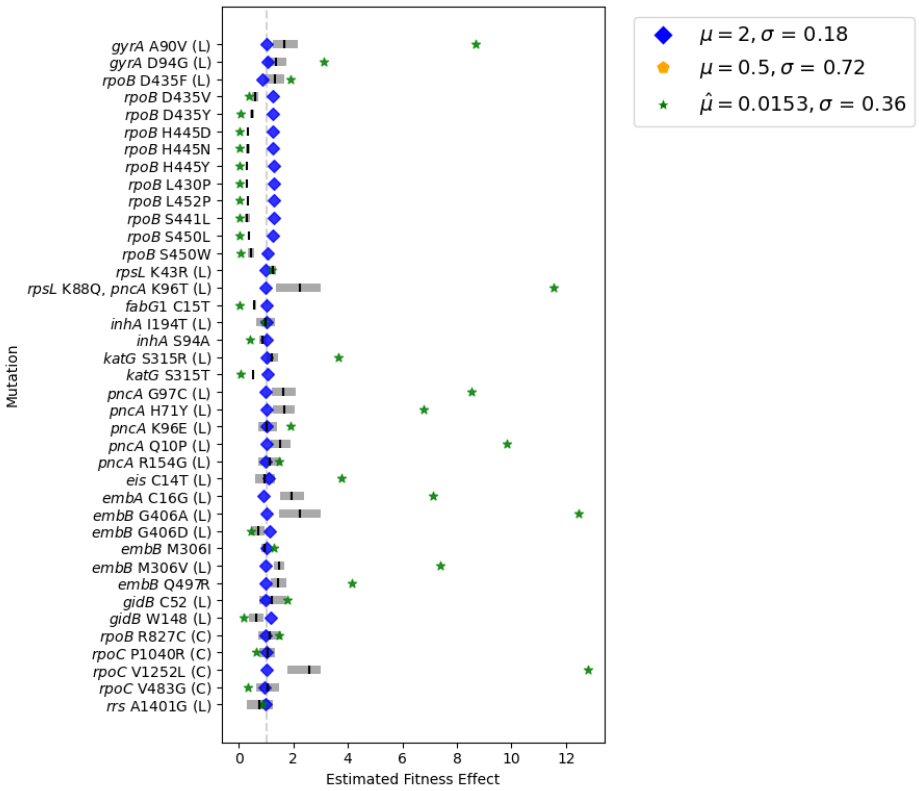

Figure B.15: Sensitivity analysis and estimation of μ in South Africa L4. The vertical black line represents the mutation's fitness MLE for $\lambda_0 = 1$, $\mu = 1$, and $\sigma = 0.36$ and its surrounding grey box is the 95% CI. Fitness MLEs for an increased value for μ are marked in blue. The green stars illustrate the fitness MLE for phyloTF's estimate of $\hat{\mu}$. $\mu = 0.5$ with $\sigma = 0.72$ did not produce any results as the model estimated at least one fitness effect to be $\beta_i = 0$ before the end of the iteration and ended the estimation process prematurely.

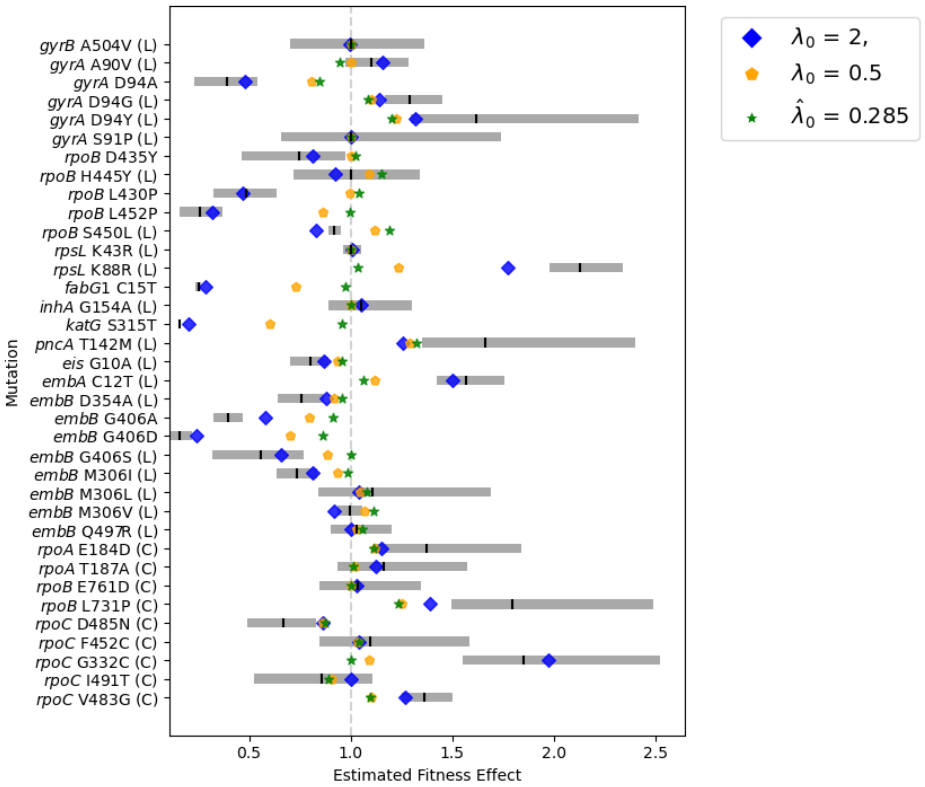

Figure B.16: Sensitivity analysis and estimation of λ_0 in Georgia L2. The vertical black line represents the mutation's fitness MLE for $\lambda_0 = 1$, $\mu = 1$, and $\sigma = 0.4$ and its surrounding grey box is the 95% CI. Fitness MLEs for increased/decreased values for λ are marked in blue and yellow. The green stars illustrate the fitness MLE for phyloTF's estimate of $\hat{\lambda}_0$.

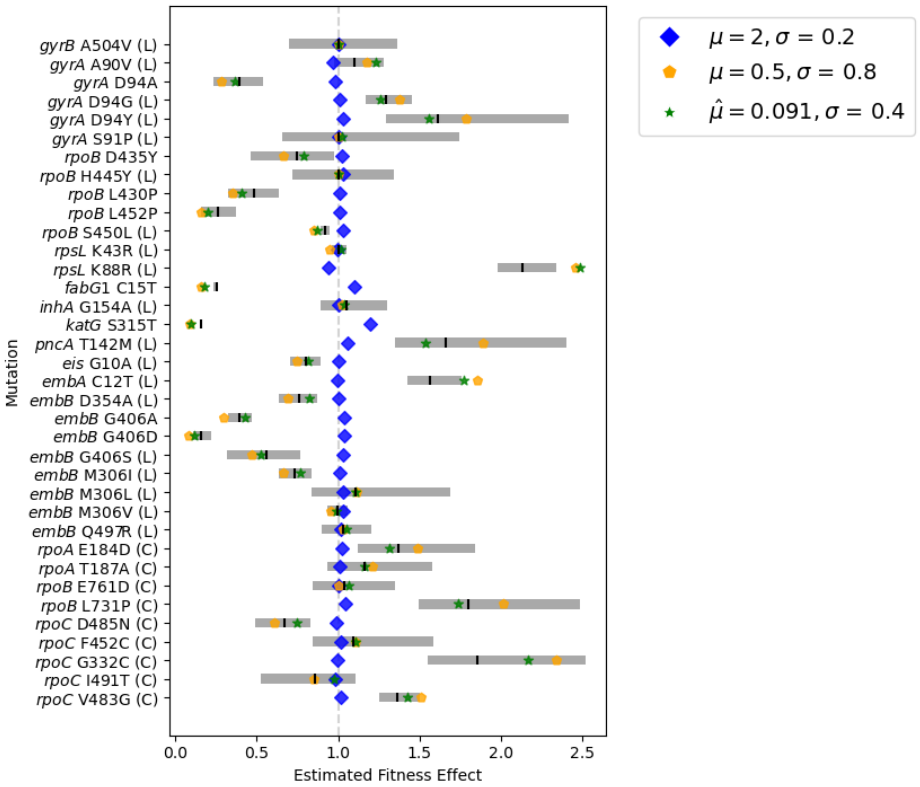

Figure B.17: Sensitivity analysis and estimation of μ in Georgia L2. The vertical black line represents the mutation's fitness MLE for $\lambda_0 = 1$, $\mu = 1$, and $\sigma = 0.4$ and its surrounding grey box is the 95% CI. Fitness MLEs for increased/decreased values for μ are marked in blue and yellow. The green stars illustrate the fitness MLE for phyloTF's estimate of $\hat{\mu}$.

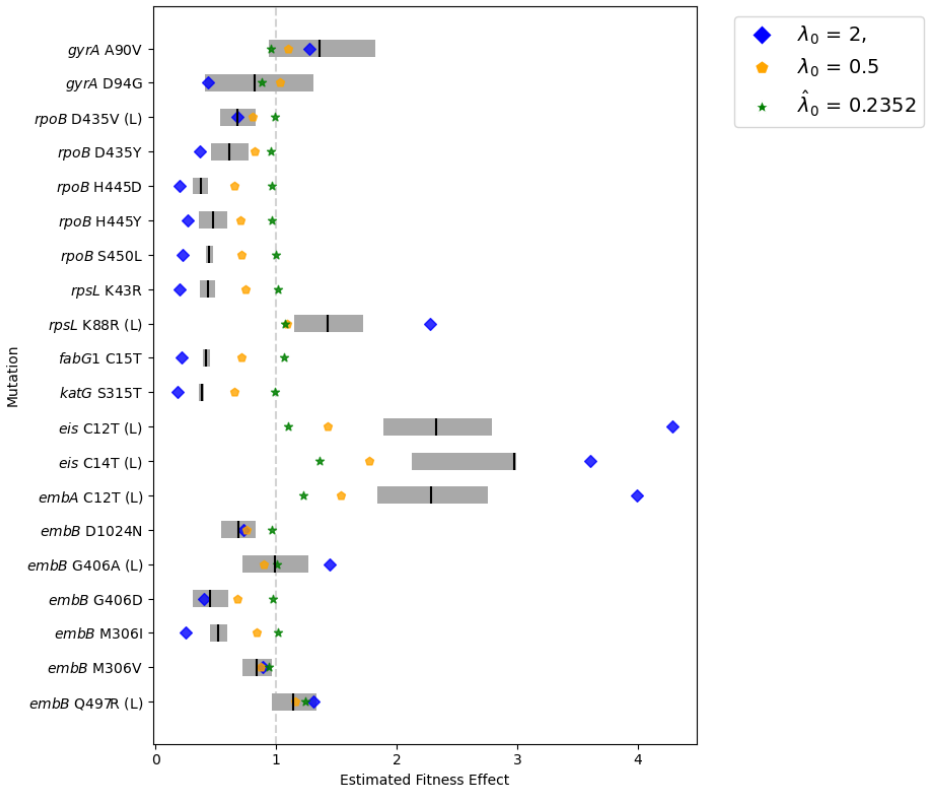

Figure B.18: Sensitivity analysis and estimation of λ_0 in Georgia L4. The vertical black line represents the mutation's fitness MLE for $\lambda_0 = 1$, $\mu = 1$, and $\sigma = 0.4$ and its surrounding grey box is the 95% CI. Fitness MLEs for increased/decreased values for λ are marked in blue and yellow. The green stars illustrate the fitness MLE for phyloTF's estimate of $\hat{\lambda}_0$.

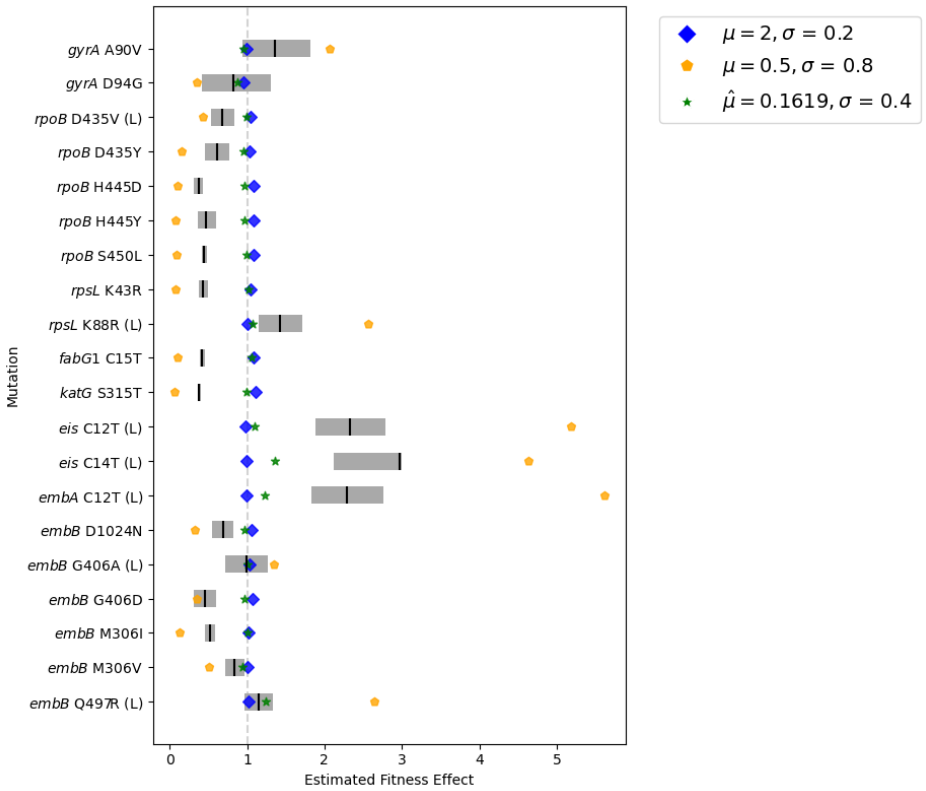

Figure B.19: Sensitivity analysis and estimation of μ in Georgia L4. The vertical black line represents the mutation's fitness MLE for $\lambda_0 = 1$, $\mu = 1$, and $\sigma = 0.36$ and its surrounding grey box is the 95% CI. Fitness MLEs for increased/decreased values for μ are marked in blue and yellow. The green stars illustrate the fitness MLE for phyloTF's estimate of $\hat{\mu}$.