Closing the Gaps - Improving Genetics-Based Predictions for Antimicrobial Resistance in *Mycobacterium tuberculosis* and *Escherichia coli*

Viktoria M. Brunner

University College, University of Oxford

A thesis submitted for the degree of *Doctor of Philosophy*

Michaelmas 2025

Contents

Abstract

"If we use antibiotics when not needed, we may not have them when they are most needed." — Tom Frieden, former director of the U.S. Centers for Disease Control and Prevention.

This statement remains acutely relevant as antimicrobial resistance (AMR) continues to rise globally. Inappropriate or delayed antibiotic use accelerates the spread of resistance, underscoring that effective diagnostics are as essential as the development of new antimicrobial agents. The central aim of this thesis is improving the prediction of antibiotic resistance, with particular emphasis on whole genome sequencing-based antimicrobial susceptibility testing (WGS-AST).

WGS-AST has transformed clinical microbiology by enabling rapid and comprehensive resistance prediction relative to traditional phenotypic testing, particularly for *M. tuberculosis*. However, discrepancies between phenotypic and genotypic results persist, limiting the clinical reliability of genomic prediction. This work investigates the sources of these discrepancies across three main areas: (i) the role of compensatory and fitness-related mutations in shaping resistance and their potential predictive value, (ii) the detection of resistant subpopulations and within-host diversity in both *M. tuberculosis* and *Enterobacteriaceae*, and (iii) the application of machine learning approaches, including graph-based models, to predict resistance from sequence, structural, and physicochemical data.

By combining evolutionary insights, quantitative analyses of within-sample diversity, and predictive modelling, this thesis outlines both the potential and the current limitations of WGS-AST. The findings demonstrate that compensatory mutations can serve as highly specific indicators of resistance, that resistant subpopulations have important clinical and epidemiological implications, and that machine learning models show promise but remain constrained by the underlying genetic architecture and available training data. Together, these results contribute to closing the gap between phenotypic and genotypic testing and advance the development of diagnostic frameworks that are biologically informed and clinically actionable.

1 Preface

1.1 Acknowledgements

Many people will tell you that a PhD will test your limits, often in ways that you won't expect. And even though you are warned, it can still take you by surprise when it hits you at some point during those four years. But I am now sitting at my desk, a week before submission, and can say that I have regretted neither my choice of doing a PhD, nor my move to the UK. And the experience of doing a DPhil in Oxford is something I would not want to have missed.

That I have made it through without any major issues is largely due to support from my supervisors, Nicole Stoesser and especially Philip Fowler. Still, the challenge of doing a PhD is not just an intellectual marathon. It is also testing your mental endurance, and many of the hurdles I had to take were a combination of stress caused by my project and life in general, which does not always want to take the backseat while you are busy working away on your thesis.

With this in mind, I am grateful for all the support I received from my friends in Oxford during this sometimes bumpy ride. Especially my colleagues: Eli, for being one my closest friends from the start, Kev, for being a kind voice of support and a great gym partner, Hermione, for taking the lead when things needed to get done, Dot, for inspiring us all to do better, Mel, for being my favourite seat neighbour, and Dylan & Dylan for the productive and entertaining company in our little Fowler-lab subgroup. And thank you to Bede, for cheering me on more the closer I got to the finish line, and for being one of the only people besides my supervisors and examiners to actually read my entire thesis. :)

I am also grateful for all the people that have made my life in Oxford so fulfilling. Thank you to Sofia, for sharing my striving for being organised and on time, to Elina, for being a role model in persisting when facing adversity, to Jeanne, for the many hours on the river and the emotional support, to Joe, for joining me on countless bike rides, to Vanessa, for making me miss the motherland a little less by enjoying peak German humour with me, thank you to my flat mates Mel, Tom and Charlotte, for taking my mind off work with our Midsomer Murders marathons, and thank you to Annina and Hans, for welcoming me to London during my internship and for the many evenings spent discussing things big and small. I

appreciate each one of you.

I am incredibly happy to still have my friends from undergrad close to me, even if we have not been close in geographical space during the last four years. My UK adventures would not have been possible without you helping me through those intense years at ETH. Lena, Daria, Jenny, Helena and Silvana, I am so lucky to have each one of you by my side.

I want to thank my sister for being a steady point of reference throughout my life and through different life stages. And most importantly, I want to thank my parents for giving me the self-confidence and sense of security to dare take a chance on things in life. I don't take it for granted and will always be indebted to you for supporting each of my decisions along the way.

1.2 Funding

My DPhil was supported by funding from the Biotechnology and Biological Sciences Research Council (UKRI-BBSRC, grant number BB/T008784/1). University College Oxford has supported my DPhil by awarding me the Oxford-Radcliffe scholarship, which has covered my living expenses for the last four years.

1.3 List of Figures

1.4 List of Tables

1.5 Papers arising from this thesis

Peer-reviewed and published

Brunner, V.M. Fowler, P.W. (2024). Compensatory mutations are associated with increased *in vitro* growth in resistant clinical samples of *M. tuberculosis*. *Microbial genomics* **10**:2057-5858.

Brunner, V.M. Fowler, P.W. (2025). Subpopulations in clinical samples of *M. tuberculosis* can give rise to rifampicin resistance and shed light on how resistance is acquired. *JAC-Antimicrobial resistance* **7**:dlaf175..

1.6 Contributions to thesis

I, Viktoria Marie Brunner, conducted all of the analyses presented in this thesis with appropriate support from my supervisors and colleagues. I hereby declare and gratefully acknowledge the assistance I have received from others in relation to this work below.

The work by the CRyPTIC consortium has been integral for the analyses throughout my thesis. They collected, phenotyped and whole genome sequenced all *Mycobacterium tuberculosis* samples, processed all the phenotypic and sequencing data and calculated epidemiological cut-off values to segregate resistant and susceptible isolates. Further, there are several people that have supported me in realising the work that forms the basis of individual chapters:

Chapter 3

I wrote the manuscript on which the chapter is based and produced all analyses and figures myself. I am grateful for discussions with Tim Peto, David Eyre, Daniel Wilson and Kerri Malone.

Chapter 4

The analyses for the manuscript on which the first part of the chapter is based was conducted by myself. The manuscript was written and the figures created by myself and edited by Philip Fowler. For the work on *E. coli* and *K. pneumoniae* samples, I would like to thank Kumeren Govender for providing access to his direct-from-blood culture metagenomic sequencing dataset. Samples were taken from patients with blood stream infections at the Oxford University Hospitals NHS Foundation Trust between 2020-2021. I would like to thank Tim Peto, Nicole Stoesser, Sam Lipworth and David Eyre for helpful suggestions.

Chapter 5

I am grateful to Philip Fowler and Dylan Dissanayake for the joint work on the Python package 'sbmlsim', used for generating the synthetic datasets. I also valued our close collaboration during the development of the GCN model for resistance prediction. Equally, I am grateful to Philip Fowler, Dylan Adlard and Charlotte Lynch for their work on the Python package 'sbmlcore', which was used for feature design.

1.7 Abbreviations

AMR	Antimicrobial resistance
AST	Antimicrobial susceptibility testing
AUC	Area under the curve
CARD	Comprehensive Antibiotic Resistance Database
CLSI	Clinical and Laboratory Standards Institute
CM	Compensatory mutation
CRyPTIC	Consortium for Resistance Prediction in Tuberculosis: an International Consortium
CNN	Convolutional Neural Network
ECOFF	Epidemiological cut-off value
ENA	European Nucleotide Archive
EUCAST	European Committee on Antimicrobial Susceptibility Testing
FN	False negative
FNR	False negative rate
FP	False positive
FRS	Fraction of read support
GCN	Graph Convolutional Network
ISO	International Organization for Standardization

LD	Linkage disequilibrium
LR	Logistic regression
MALDI-TOF	Matrix-Assisted Laser Desorption/Ionization Time-of-Flight
ME	Major error
MGIT	Mycobacteria Growth Indicator Tube
MDR-TB	Multidrug-resistant tuberculosis
MIC	Minimum inhibitory concentration
ML	Machine learning
MRSA	Methicillin resistant *Staphylococcus aureus*
NGS	Next generation sequencing
NPV	Negative predictive value
ONT	Oxford Nanopore Technologies
PCR	Polymerase chain reaction
PPV	Positive predictive value
RAV	Resistance-associated variant
RAST	Rapid antimicrobial susceptibility testing
RNAP	RNA polymerase
ROC	Receiver operating characteristic
RRDR	Rifampicin resistance determining region

RR-TB	Rifampicin resistant tuberculosis
SE	Sensitivity
SP	Specificity
SNP	Single nucleotide polymorphism
TB	Tuberculosis
TN	True negative
TP	True positive
VME	Very major error
VRE	Vancomycin resistant Enterococci
VRSA	Vancomycin resistant *Staphylococcus aureus*
WHO	World Health Organization
WGS	Whole genome sequencing
XDR-TB	Extensively drug-resistant tuberculosis
XGBoost	Extreme Gradient Boosting Decision Trees

2 Introduction

This introductory chapter establishes the conceptual and clinical context for this thesis by outlining the global challenge of antimicrobial resistance (AMR) and the essential role of diagnostics in its containment. It introduces the biological and evolutionary basis of resistance, the transition from traditional phenotypic testing to genomic and computational approaches, and the relevance of diagnostic innovation in infectious disease management. Finally, it highlights pathogen-specific characteristics and the resulting diagnostic challenges in *Mycobacterium tuberculosis (M. tuberculosis)* and *Enterobacteriaceae*.

2.1 Antimicrobial resistance

2.1.1 From epidemics to early diagnostics

Infectious diseases have been spreading through human populations for at least as long as written records exist. The oldest written reports of devastating plagues date from the ancient Roman and Greek empires.[1] But given what we know about the co-evolution of humans and animals, and the pathogens that spread through and between them through zoonosis, it is likely that large-scale epidemics have existed ever since humans started settling down in larger communities during the neolithic revolution.[2;3] As settlement size and population density increased, the close contact between domesticated animals and other humans, as well as poor sanitation and difficulty of accessing clean water made human settlements ideal for the spread of infectious agents.[4]

As a result, pandemics like the 'Black Death' (the bubonic plague) ravaged through medieval Europe and killed up to a third of its population.[5] As with previous plagues, the responsible pathogen was only identified centuries later: Both the Justinian plague in the ancient Roman empire, as well as the bubonic plague for instance were caused by the bacterium *Yersinia pestis*. This pathogen would continue to spread through India and China, before being isolated for the first time in Hongkong by Alexandre Yersin in 1894.[5]

It was around the same time that the bacterium responsible for the illness tuberculosis was discovered, also often called 'consumption', due to the substantial weight loss, or 'White Plague', due to the symptoms of

severe anaemia, both associated with late stages of the disease.[6] Robert Koch had previously isolated the agents responsible for the Cholera pandemics and Anthrax infections, before becoming interested in and successfully isolating the causative agent of tuberculosis, *M. tuberculosis*.[7;8]

Arguably, these early days of pathogen identification marked the start of modern-day diagnostics. Koch's discovery of major disease-causing microorganisms went hand in hand with the development of essential microbiological techniques by his co-workers, such as the use of agar and Petri dishes for growing bacterial cultures.[9] In his postulates, Koch also attempted to establish a causal relationship between the presence of specific microbes and a specific disease.[10] This causal relationship is now recognised to be much more nuanced than described in his postulates, as is reflected in current diagnostic methods. However, the logical framework used in his work is still applicable to problems in modern-day diagnostics, such as in antimicrobial resistance, where the presence of certain genetic variants predicts resistance.[11]

2.1.2 The golden age of antibiotics

The identification of specific agents responsible for disease in the late 19th century, along with the progressive introduction of sanitation and better nutrition, constituted important steps towards fighting infectious disease. Nevertheless, the 20th century started with a surge in the spread of many diseases, especially airborne viral and bacterial diseases, such as the Spanish flu and tuberculosis. Increased globalisation had enabled new routes of transmission, and the European population was reeling from the repercussions of the First World War.[12;13]

But in 1928, an accidental discovery heralded the coming of the golden age of antibiotics: Alexander Fleming observed the bactericidal effects of the beta-lactam penicillin against *Staphylococcus aureus (S. aureus)* growing in a Petri dish.[14] Even prior to this, synthetic antimicrobial agents had been employed successfully for treatment of infectious disease,[15] but the natural product penicillin sparked the idea of investigating antibiotics from natural sources, and isolating and adapting them for clinical use.[1]

In the 1940s, around the same time as penicillin was introduced for routine clinical use, Waksman *et al.* discovered that the genus *Streptomyces* from the phylum Actinomycetota is a producer of several

antibiotics,[16] and since then *Streptomyces* genomes have been heavily mined for potential natural product gene clusters. One of the first drugs discovered to be produced by Streptomyces was streptomycin, an aminoglycoside antibiotic, which was also the first drug with notable activity against *M. tuberculosis* infections.[16;17]

The discovery of the first antibiotic-producing microbes by Waksman *et al.* prompted pharmaceutical companies to invest heavily in screening of promising microbes and fungi for compounds with activity against pathogenic bacteria.[18] This led to the discovery of many major antibiotic classes like tetracyclines and macrolides within the span of only 20 years (Figure 1). The discovery of new antibiotics slowed down considerably after 1965. Most new drugs introduced were derivatives of existing classes and the rediscovery of already known compound became ubiquitous.[17] Even new approaches to drug discovery such as heterologous expression of silent biosynthetic gene clusters and genome mining did not lead to any major breakthroughs. This in turn led to a slow divestment from antibiotics discovery in the pharmaceutics space, with natural product discovery increasingly focussing on screening for anti-cancer agents and drugs for treating heart and neurological diseases.[18]

The antimicrobial drugs available to date can be divided into three classes based on their mode of action. Disruption of cell wall synthesis (e.g. beta-lactams and vancomycin), disruption of protein synthesis (e.g. tetracyclines and macrolides) and disruption of nucleic acid synthesis (e.g. fluoroquinolones and rifampicin).[19] The lack of diversity in drug targets is another problem in antibiotics discovery, since cross-resistance becomes a major concern.[20]

2.1.3 From penicillin resistance to multidrug-resistance

Given the wealth of antibiotic classes discovered between 1945-1965, there should have been ample choice for clinical treatment of bacterial infectious diseases. But the use of antibiotics has been continually hindered by the target pathogens developing the ability to inactivate, export or otherwise avoid the effect of the drug, i.e. become resistant.[21] There is even evidence for resistance pre-dating the use of antibiotics by humans.[22]

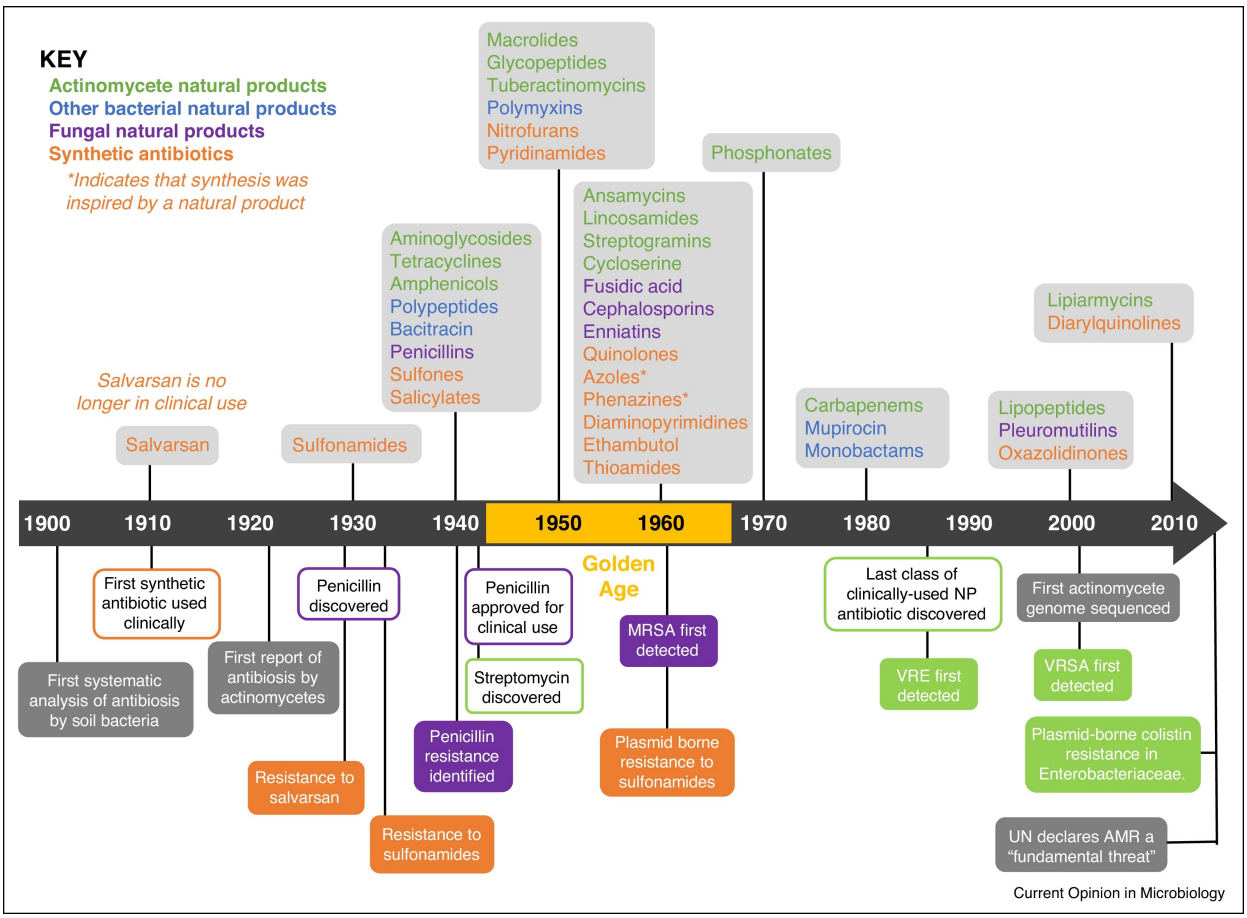


Figure 1: The timeline of antibiotic classes discovery and resistance. The colours in the key indicate the source of the antibiotics. MRSA = methicillin resistant *S. aureus*, NP = natural product, VRE = vancomycin resistant Enterococci, VRSA = vancomycin resistant *S. aureus*. This figure was reproduced from Hutchings *et al.*[17]

The first formal description of penicillin resistance dates back to 1940, before the official approval of the drug for clinical use (Figure 1).[23] Clinical penicillin resistance in *S. aureus* was first confirmed in 1942, sparking efforts to produce modified versions of the beta-lactam antibiotic, such as methicillin.[17] However, it soon became apparent that resistance would develop and spread within a few years once an antibiotic was extensively used.

The possibility of widespread penicillin resistance was alluded to as early as 1945, when Fleming warned of the misuse of penicillin in his Nobel prize speech: 'There may be a danger, though, in underdosage. It is not difficult to make microbes resistant to penicillin in the laboratory by exposing them to concentrations not sufficient to kill them, and the same thing has occasionally happened in the body.'[24] Penicillin resistance was shortly followed by methicillin resistance in *S. aureus* (MRSA),[25] and other clinically relevant pathogens quickly followed suit, like penicillin resistance in *Escherichia coli (E. coli)*[26] and

streptomycin resistance in *M. tuberculosis*.[21]

The discovery of resistance associated with mobile genetic elements further intensified this crisis, since mobile elements allow resistance to spread horizontally through bacterial populations and even species.[27] The short generation times, high mutation rates and frequent genetic recombination present in many pathogenic bacteria make it possible for antimicrobial resistance to arise and spread throughout bacterial populations rapidly. This is reinforced by human interference: In addition to the misuse and overuse of antibiotics in the clinical setting alluded to by Fleming, problems arise from extensive use of antibiotics as 'growth supplements' in livestock, regulatory barriers and inappropriate prescribing.[28] The resulting constant exposure to ineffective or sub-therapeutic levels of antibiotics leads to directional selection, encouraging the faster emergence and spread of resistant strains.[29;30] What followed, was the discovery of combined resistance to multiple antibiotics within bacterial populations, termed multidrug-resistance (MDR). This is now especially relevant in *M. tuberculosis* infection (MDR-TB), as well as in many nosocomial infections with e.g. *E. coli* and *Klebsiella pneumoniae (K. pneumoniae)*.[21]

As mentioned in the previous section, new antibiotic drugs are being developed but the pace is slow and the rate at which resistance to new drugs develops appears higher.[31] With the increasing prevalence of antibiotic resistance and the decreasing discovery rate of potent new drugs, we are heading towards a human-fabricated global health crisis.[29;32]

2.1.4 The threat of a post-antibiotic era

The rise of multidrug-resistant bacteria is one of the grand challenges we are facing as a global society. Although the scenarios from pre-20th century epidemics might seem like ancient history, the uncontrolled spread of MDR pathogens could render almost all treatment options ineffective and lead to a positive feedback loop in the spread of resistance.

Already in 2019, out of the 8.9 million deaths due to bacterial infections, an estimated 4.71 million (52.8%) were associated with AMR.[33;34] And as per the O'Neill report from 2016,[35] the number of AMR associated deaths per year could reach 10 million by 2050, although more recent studies set it slightly

lower at 8.22 million AMR associated deaths per year.[34] It is estimated that the financial repercussions associated with AMR could then cost the global economy upwards of £66 trillion. There will also be a non-negligible impact of AMR on routine surgeries, such as invasive surgeries and organ transplants, due to the high risk of drug-resistant infection that would no longer be preventable with prophylactic antibiotics. The same is true for cancer treatments with chemotherapy, which leads to immunosuppression, and therefore also typically requires prophylactic antibiotics.[35]

To prevent these worst-case scenarios from becoming reality is the goal of AMR research, whether through development of new drugs with new mechanisms to fight bacterial infections, or by ensuring the antibiotics we have available are being put to the best possible use through antibiotic stewardship, supported by rapid diagnostic tools.[35]

2.1.5 The case for improving diagnostics

Drug development in the antibiotic space has stalled, both due to difficulty in identifying new candidates as well as decreasing investment incentives.[18;28;35] Even if this were to change, the history of antibiotic development has taught us that the discovery of new drugs alone is not sufficient, since resistance evolves and spreads quickly, due to high mutation rates and selection pressure.[29;30]

Infectious disease diagnostics can help clinicians select the optimal treatment for each patient, by identifying the responsible pathogen and the resistance profile of the infection. This can help mitigating the influence of inappropriate prescribing as part of antibiotic stewardship, and slow down resistance spread by preventing the exertion of further selective pressure on resistant populations.[28] In addition, routine diagnostics allow monitoring the spread of resistance worldwide. Developing new, rapid diagnostic tools is hence one of the main recommendations of the O'Neill report.[35]

But just like antibiotics discovery on its own is necessary but not sufficient, simply optimising the use of available antibiotics can only take us so far if there are insufficient treatment options available for highly drug-resistant cases. It is clear that drug development and diagnostics need to work hand in hand, and progress in one field warrants progress in the other in order to slow down the progressive spread of AMR.

2.2 Infectious disease diagnostics

2.2.1 Pathogen identification and phenotypic antimicrobial susceptibility testing

The field of infectious disease diagnostics has come a long way since Koch first isolated and cultivated the *M. tuberculosis* bacterium. The innovations that accompanied this process, such as the use of solid medium made from agar and advanced microscopy methods, remain cornerstones of modern bacteriology for many pathogens.[9] Together with the invention of Gram staining in 1882,[36] these techniques were used as early diagnostic tools for identifying and differentiating between bacterial species.

While early diagnostics focused on identifying the pathogen responsible for the infection, the advent of antimicrobial resistance led to the necessity to determine the antibiotic susceptibility profile of the pathogen as well. This would enable a tailored therapeutic choice and prevent further resistance spread.[28] To achieve this, different methods of phenotypic antimicrobial susceptibility testing (AST) were developed, where the growth arrest or death of the microbe is observed when grown in the presence of different antimicrobial agents.[37]

The first method for phenotypic AST, which is still common in modern microbiology laboratories, was developed in 1940 and is known as disc diffusion.[38] This involves placing paper discs containing a defined antibiotic concentration on an agar plate inoculated with the test organism at a defined density. The resistance level is assumed to be proportional to the zone of inhibition on the agar around the discs. However, disc diffusion only gives a qualitative result, such as resistant, intermediate and susceptible. It does not report on the minimum inhibitory concentration (MIC) of the antibiotic. Agar dilution[39] and later broth dilution methods filled this gap by inoculating the test organism on either solid (agar dilution) or liquid medium (broth dilution) with serially diluted concentrations of antibiotic, usually in doubling dilutions. MIC can also be determined through gradient diffusion (E test), which involves placing a strip impregnated with an antibiotic at varying concentrations onto an agar plate inoculated with a known concentration of the test organism.[40] For all these methods, the lowest antibiotic concentration with no observable bacterial growth is the MIC, hence represented as an interval rather than a point estimate.[37] The resistance call is then made by assessing the MIC in relation to the epidemiological cut-off (ECOFF)

value, which is defined for different antibiotics and species. The ECOFF for a given species and drug describes the MIC above which bacterial isolates are assumed to have phenotypically detectable acquired resistance mechanisms, and below which the isolates are part of the wild-type population.[41] It hence only considers *in vitro* phenotypic data, in sharp contrast to clinical breakpoints. Clinical breakpoints are based on a mixture of *in vitro* and *in vivo* data.[41] They attempt to relate MICs to clinical outcome and are used for decision making in clinical practice. ECOFF and clinical breakpoints are hence predictive of two different scenarios (phenotypic resistance *in vitro* vs treatment success). In practice, ECOFF and clinical breakpoints are often identical, mostly due to difficulty in collecting data for determining clinical breakpoints.

Although both disc diffusion and dilution methods have since been optimised and refined and are still the gold-standard AST method for most pathogens, there have also been various attempts to introduce automation and parallelisation. Machines such as the VITEK 2 by bioMérieux and the BD Phoenix automated microbiology system reduced turnaround time for phenotypic AST significantly. The latter system uses parallelised microdilution to both identify species and measure MICs for a wide range of bacteria and antibiotics.[37] But despite efforts to shorten the overall turnaround time and optimise sample processing, phenotypic AST remains a labour-intensive and time-consuming process for many pathogens, mostly due to the need to grow the microbes to sufficient densities in order to test the effect of the antibiotic.[42] Additionally, certain microbes are less amenable to automated phenotypic AST systems due to their slow growth and fastidious nature, such as *M. tuberculosis*.[43]

For pathogen identification, rapid new methods include the Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry (MALDI-TOF MS). This employs classic mass spectrometry to detect specific peptide mass fingerprints, which allow identification down to the species level in many cases.[44] While this is very useful for pathogen identification, it is not yet widely used for determining antimicrobial susceptibility profiles, due to technological limitations.[45]

2.2.2 The advent of molecular diagnostics for antimicrobial susceptibility testing

Parallel to the diversification of phenotypic AST methods, genotypic approaches have been explored. Initially, these focused on the identification of species based on the detected presence of certain pathogen-specific DNA sequences, e.g. using DNA probes.[46] However, the abundance of the unamplified target DNA is often low and even signal-amplification cannot rescue the readout.[47]

The invention of the polymerase chain reaction (PCR) in 1985 opened up a whole range of possibilities for species identification and genotypic AST.[48] The targeted *in vitro* amplification of specific DNA sequences in multiple rounds of PCR leads to increased sensitivity, and the choice of specific primers allows for high specificity for any target pathogen sequence.[47] PCR can be combined with novel techniques for efficiently isolating DNA from diverse microbial targets,[49] and various methods for the detection of the amplified product, such as gel electrophoresis[50] or real-time detection using fluorescent probes.[47] The introduction of microarrays as a downstream readout of the PCR has further improved the technique by allowing simultaneous assessment of large numbers of microbial genetic targets.[51]

This suite of new molecular diagnostics tools, all relying on the initial amplification step by PCR, allows the underlying genetic cause of resistance to be detected, i.e. specific resistance genes or mutations, even if the initial signal is weak. Based on this, many molecular tests were developed for rapid detection of various pathogens, which was especially useful for viral pathogens[52], as well as molecular tests for genotypic AST, which was a viable alternative for slow-growing species like *M. tuberculosis*, where phenotypic AST is very time-consuming.[53]

2.2.3 The sequencing revolution

This breakthrough in molecular diagnostics involved the incorporation of DNA sequencing technologies. DNA sequencing can be used to perform a whole range of useful diagnostic tasks: It can elucidate the taxonomic background of a pathogen, e.g. by examining the ribosomal RNA and house-keeping genes.[54] In the same vein, it can be used to link cases to outbreaks by identifying and comparing single nucleotide polymorphisms (SNPs).[55;56] And, most important in the context of this thesis, it can be used to detect resistance associated mutations and genes as part of genotypic AST.[57]

However, prior to the invention of Sanger sequencing, DNA sequencing was a very labour-intensive process. Efforts for elucidating the sequence of the *lac* operator, for example, progressed very slowly at 1 base per month.[58] In 1976, Frederick Sanger came up with the chain terminator procedure, which relies on the addition of four different, fluorescently marked nucleotides to a traditional PCR, which terminate the transcription upon incorporation into the nascent DNA chain.[59] The readout then proceeds through electrophoresis on polyacrylamide slab gels, enabling single-base resolution.[60] This development sped up sequencing efforts significantly, and automated versions of Sanger sequencing enabled sequencing of up to 1000 bases per day by 1987.[61] Further improvements in speed followed and enabled large sequencing endeavours like the elucidation of the human genome.[62] Up to 2007, cost of sequencing decreased roughly according to Moore's law, i.e. the cost halved every two years as performance doubled (Figure 2). But it was clear that even with these improvements, Sanger sequencing would not be efficient enough to conduct regular routine sequencing of larger genomes.[63]

This limitation was overcome with the introduction of next-generation sequencing (NGS) methods. These methods do not rely on electrophoretic separation as a readout step, hence avoiding the hitherto most time-consuming step of the sequencing process. Instead, NGS methods employ sequencing-by-synthesis, enabling real-time readout, and multiplexing, which enables massive parallelisation.[63] Importantly, this increase in sequencing speed also reduced the cost of sequencing per base significantly, which made the sequencing cost decrease dramatically from 2007 (Figure 2).[64] With a maximum length of 300-500 base pairs, the resulting sequencing reads are shorter than the previous Sanger sequencing reads. This is due to technical requirements of the NGS methods, where the DNA is broken up into shorter sequences, hence also the name 'short-read sequencing'. Due to the higher error rate, short read sequencing also requires high sequencing depth, so errors can be diluted and avoided.[65] Out of the different NGS technologies competing for the market, the Illumina platform has managed to establish a near monopoly on the short-read sequencing market in recent years.[66]

The DNA sequencing toolbox was further extended when single molecule real-time sequencing became possible. This is the first approach that does not require DNA amplification as part of the sequencing process.[67] This significant advancement is hence often referred to as the third generation sequencing or

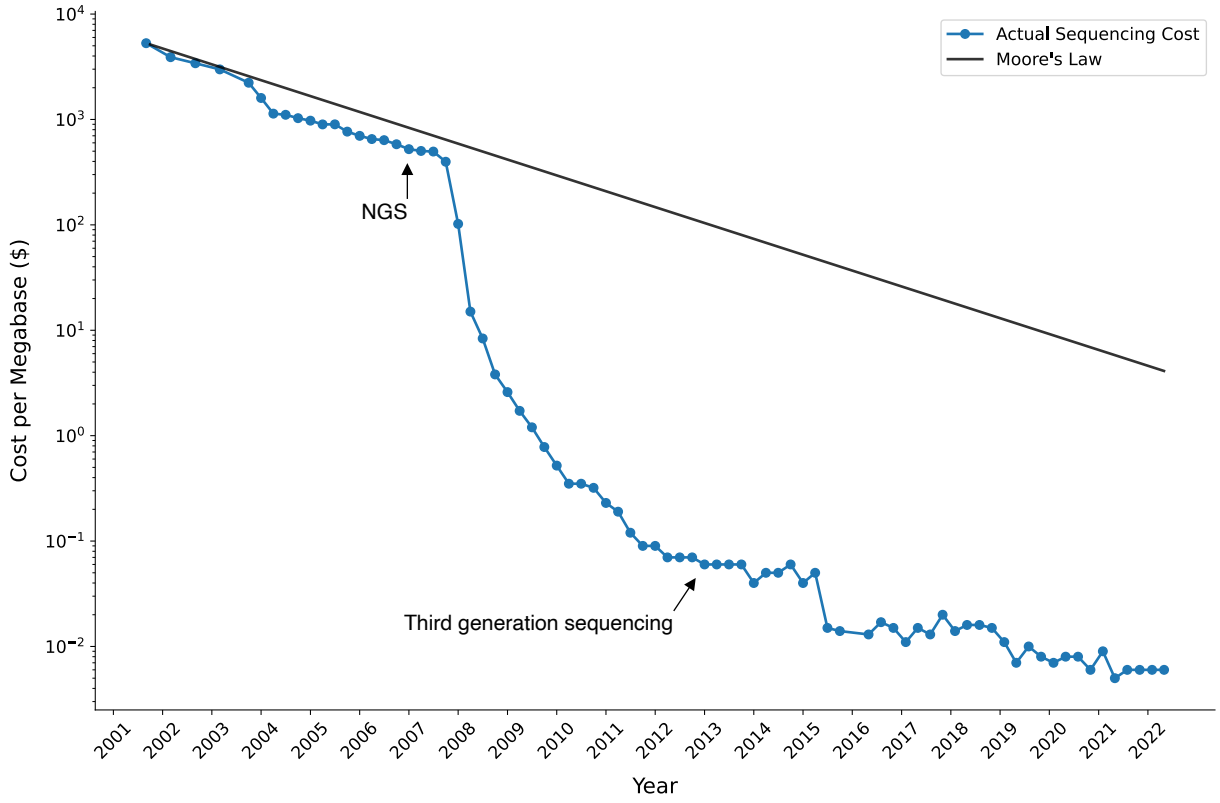


Figure 2: The evolution of DNA sequencing cost since 2001. Sequencing cost roughly decreases according to Moore's law until the advent of NGS. Moore's law was first described with reference to the doubling of computing power roughly every two years. It was then extended to technologies in general, which are regarded as successful if they follow this law. Data shown in the figure was obtained from the National Human Genome Research Institute.[64]

'long-read sequencing' (Figure 2). Although initially having higher error rates than e.g. Illumina sequencing, long-read sequencing was able to establish itself by enabling sequencing with read lengths well over 10 kb.[63] The most promising approaches in the long-read sequencing field are the real-time polymerase-mediated synthesis implemented by PacBio[68] and the nanopore-based sequencing introduced by Oxford Nanopore Technologies (ONT).[69] The latter proved especially useful due to the portability of the sequencer itself, enabling sequencing to be conducted in the field.[70] In addition, long-read sequencing opened up the possibility of reliable *de novo* assembly, even of highly repetitive sequences.[71] *De novo* assembly is especially useful for organisms with large accessory genomes, that are therefore not amenable to mapping to a reference genome.[72] The capabilities of long-read sequencing in this realm were initially shown in a study from 2015 which attempted to assemble a complete *E. coli* genome *de novo* using nanopore reads only.[73] The successful assembly also showed that despite the higher error rate, nanopore sequencing can achieve high quality assemblies even without hybrid sequencing, i.e. without

including short-read data for correction. Since then, error rates of both long- and short-read sequencing have been reduced significantly. For bacterial genomes, sequencing quality scores can reach Phred quality scores of Q15 and above for nanopore sequencing, corresponding to a base call accuracy of 96.8%. Illumina reaches even higher accuracy and hence is more reliable for single nucleotide polymorphism (SNP) detection and epidemiological surveillance.[74]

2.2.4 The role of whole genome sequencing in infectious disease diagnostics

With rapid and cost-efficient short and long-read technologies at hand, whole genome sequencing (WGS) for both genome assembly and variant detection has become possible. This opened up new avenues for the surveillance of resistance spread and enabled WGS-AST as a form of untargeted genotypic AST.[57] The advantage of performing WGS instead of using targeted sequencing, multiplexed PCR or microarrays to detect resistance is that in theory, the entire genome of the organism is available. Hence there is no built-in limit to the number of relevant sites or genes that can be assessed for resistance.[75] Resistance prediction can then be performed by cross-referencing the detected variants using databases of antibiotic resistance determinants, i.e. known AMR genes or mutations, and then applying this current, curated knowledge in a 'rules-based' classification.[75] Databases for AMR genes are available for multiple species and include CARD,[76;77] ResFinder,[78] as well as specific catalogues for single species, such as the catalogue of mutations in the *M. tuberculosis* complex and their association with drug resistance, published by the WHO.[79] Having the entire genetic information available and stored securely also allows for later in-depth investigations of resistance and virulence associated regions of the genome, thereby improving diagnostic capabilities and our understanding of the underlying molecular causes of resistance.

Rapid genotypic AST is especially promising for slow-growing pathogens such as *M. tuberculosis*, where phenotypic AST takes a long time.[80;81] Delaying treatment or starting empiric treatment are the consequence, possibly leading to treatment failure and increased resistance spread. But genotypic AST, including WGS-AST, is still mainly used for drug discovery, resistance monitoring or research, and is not usually available in routine clinical microbiology laboratories.[37] In order to establish WGS-AST as a routine diagnostic tool, it is necessary to demonstrate that performance is comparable to the gold stan-

dard phenotypic methods. The European Committee on Antimicrobial Susceptibility Testing (EUCAST) suggests establishing this genotypic–phenotypic concordance based on ECOFF values.[82] They also emphasise the need for standardised bioinformatic pipelines to process sequencing reads, in order to reduce bias introduced in this step. Overall, it is clear that further work is needed to make WGS-AST a reliable diagnostic tool for many pathogens.

2.3 Antimicrobial resistance in specific pathogens

2.3.1 *M. tuberculosis*: Infection characteristics and treatment

Studies suggest that *M. tuberculosis*, the aetiological agent of tuberculosis (TB), has been co-evolving with the modern human since before the out-of-Africa migration.[83] There are records of high mortality rates from TB throughout human history, and, although it is treatable, the disease to this day remains responsible for the deaths of about 1.25 million people in 2023, with a global TB incidence rate of 134/100,000.[84] This is in part due to funding for prevention, diagnosis and treatment of TB continuously falling short of the needed amount to implement national strategic plans. In 2021, this shortfall reportedly amounted to 1.6 billion US dollars.[85]

M. tuberculosis spreads via aerosols from individuals with active pulmonary disease and hence first infects the alveoli in a new host. From here, it can spread to other organs, which can lead to extrapulmonary TB, but disease symptoms most often manifest in the lungs.[86] Instead of causing active disease directly, infection with *M. tuberculosis* will, in the majority of cases, lead to a latent infection.[87] Latent TB can only be detected by eliciting a specific immune response, such as through the tuberculin skin test,[88] in an otherwise healthy host, and not by a chest X-ray and sputum smear, which are used to diagnose active TB.[89] Latent disease can, in an estimated 10% of cases, turn into active disease through reactivation.[87] Only then can *M. tuberculosis* bacteria spread to close contacts or be isolated from the host,[88] which creates problems for diagnostics. It is estimated that around a quarter of the world population have latent TB, constituting a large reservoir for potential reactivation TB.[90]

Streptomycin was the first active agent for the treatment of TB,[16;17] which was used as monotherapy for the first few years after its discovery. After drug resistance emerged, it was combined with other antibiotics for treatment,[91] in order of their introduction to the clinic: isoniazid, ethambutol and rifampicin.[92] But in 2019, the WHO treatment guidelines were updated to recommend limited use of streptomycin only,[93] due to widespread resistance dating back to its use as a monotherapy and the serious side effects for the patient. Today, the first-line antibiotics used for drug-susceptible TB treatment are hence isoniazid, rifampicin, pyrazinamide and ethambutol. The most effective and best tolerated among those are

isoniazid and rifampicin. Shorter treatment regimens might include the use of moxifloxacin.[94] For drug-resistant TB, treatment depends on the exact combination of resistance types, but will generally include second-line antibiotics, such as bedaquiline and fluoroquinolones, and such regimens often display higher toxicity, are less effective and more expensive.[95]

2.3.2 *M. tuberculosis*: AMR and diagnostics

Unfortunately, despite little evidence of horizontal gene transfer,[96] *M. tuberculosis* is prone to developing antimicrobial resistance very rapidly through chromosomal mutations.[97] Resistance also becomes fixed quickly due to additionally arising fitness-related mutations, among other reasons.[98] As a result, the proportion of people infected with *M. tuberculosis* strains resistant to the major first-line antibiotic drugs is rising each year.[85;99;100] Multi-drug-resistant (resistant to both isoniazid and rifampicin) and rifampicin resistant *M. tuberculosis* (MDR/RR-TB) infections alone were responsible for an estimated 150,000 deaths per year in 2023. Combined with the fact that the annual number of people who developed MDR/RR-TB was 400,000 in the same year,[84] it emphasises the increased risk of death in those cases with drug resistance. In line with this observation, chances for treatment success decrease once a RR/MDR-TB treatment regimen is necessary, from 88% down to 68%.[84] These factors were recognized through the recent addition of RR-TB to the WHO pathogen priority list.[101]

In addition to RR- and MDR-TB, there are two additional official definitions of drug-resistant TB: pre-extensively drug-resistant TB (pre-XDR-TB), which is TB that is resistant to rifampicin and any fluoroquinolone; and XDR-TB, where additionally either bedaquiline or linezolid resistance is detected.[84] Both of these require more complicated treatment regimes, with more side effects, higher cost and lower rates of treatment success.[95] Understanding how resistance mutations emerge, how they become fixed and how they spread is hence of high importance. It is also why the first step towards successfully treating *M. tuberculosis* infections is fast and reliable diagnostics, including AST.

Phenotypic AST in *M. tuberculosis* is routinely performed either on solid medium such as Löwenstein-Jensen (LJ) medium, or in liquid broth macrodilutions, for example with Mycobacteria Growth Indicator Tubes (MGIT^{TM}), developed by Becton Dickinson. This method relies on the fluorescence triggered by

the conversion of oxygen to carbon dioxide in the presence of *M. tuberculosis* bacteria. The MGIT tube hence allows deducing the presence of the pathogen and, if supplemented with antibiotics, the resistance phenotype from the presence of fluorescence in the tube.[102] MIC can be inferred by using different levels of antibiotic dilution in the liquid medium, but it is a very cost-intensive and time-consuming process.

This is a very reliable approach for most drugs, and additionally has very good sensitivity for detecting resistant subpopulations, with e.g. MGIT being able to detect heteroresistance for rifampicin down to 1% resistant bacteria.[103] But, since *M. tuberculosis* is a slow-growing species, phenotypic AST is a time-consuming approach, despite efforts to reduce culture time.[104] It takes around 10 days from initial culture to obtaining AST results in liquid medium[105] and even longer for phenotypic AST methods that make use of solid media, with 20–42 days until results.[106] In addition, many clinics lack appropriate in-house diagnostic assays and therefore rely on sending the isolates to a reference laboratories, increasing the turn-around time further.

Genotypic AST instead detects the presence of genetic variants to infer whether a sample is resistant and can be much quicker than phenotypic AST.[107] The prerequisites are that the underlying mechanism is genetic and the exact genetic variants conferring resistance are known. *M. tuberculosis* lends itself to approaches involving sequencing or hybridisation since its lack of horizontal gene transfer means that resistance is mainly conferred by chromosomal mutations in a few genes[96] and the detected variants correlate strongly with phenotypic resistance to common agents.[79] One can hence look for the presence of specific, resistance-associated genetic variation using rapid molecular tests such as GeneXpert MTB/RIF[108]. This is a nucleic acid amplification test enabling real-time PCR-based molecular testing directly from sputum within 2 hours.[109] This test is restricted to rifampicin resistance detection, based on mutations found in $>$99.5% of all rifampicin resistant strains. But since it is commonly assumed that rifampicin resistance is strongly correlated with isoniazid resistance, the detection of rifampicin resistance will generally lead to initiation of MDR-TB treatment regimens based on second-line drugs.[94;95]

An increasingly common approach is to use WGS-AST to scan the entire *M. tuberculosis* genome for resistance-associated variants (RAVs) for different antibiotics.[110] This approach allows for testing all

known resistance mechanisms and hence is more comprehensive than GeneXpert. It involves WGS of the patient-derived sample, most often based on an early positive liquid culture,[111] followed by application of a high-confidence catalogue of RAVs, enabling the sample to be classified as susceptible or resistant to a panel of antibiotics.[112] *M. tuberculosis* has a single, small chromosome,[113] and is inherently slow-growing which makes phenotypic AST difficult. It is hence an ideal candidate for WGS-AST, which is already used by the WHO for drug resistance surveillance of *M. tuberculosis*.[114] In 2017, Public Health England was the first national administration to introduce routine WGS for mycobacterial cultures, which included drug susceptibility testing.[115] The success of this approach led to the WHO recommending the application of targeted sequencing technologies for genotypic AST in *M. tuberculosis* respiratory samples to diagnose resistance to rifampicin, isoniazid, fluoroquinolones, pyrazinamide and ethambutol, instead of using culture-based phenotypic AST.[116]

The strong correlation of variants in relatively few resistance-associated genes in *M. tuberculosis* with phenotypic resistance to common agents also allows for the straightforward construction of resistance mutation catalogues for *M. tuberculosis*, based on statistics and expert rules. The most comprehensive of catalogues is the second edition of the WHO catalogue of mutations in the *M. tuberculosis* complex and their association with drug resistance.[79] It contains the most complete list of genetic markers of phenotypic resistance in *M. tuberculosis* to date, and achieves high sensitivity and specificity of resistance prediction for most drugs. However, catalogues cannot predict the effect of unknown mutations, which is especially detrimental in recently introduced drugs with novel mechanisms of action like bedaquiline.[117] Some of these unknown mutations can be categorised using expert rules, such as the assumption that any non-synonymous mutation in the *rpoB* gene and any premature stop codon, insertion, or deletion in *ethA*, *gid*, *katG*, or *pncA* are associated with resistance.[118] One caveat is that these general rules may cause false positives.

A side effect of the overwhelming use of liquid culture in AST for *M. tuberculosis*, for both phenotypic AST and as an initial culture step in WGS-AST, is that it is often not possible to use pure single colony cultures. For WGS-AST, this is because cells in liquid culture cannot be separated efficiently and therefore when a sample is taken from an early positive liquid culture, one ends up with a population of

different cells in so-called 'crumbs'.[111] The results from both MGIT and sequencing will hence always reflect the resistance profile of multiple, possibly genetically distinct clones of *M. tuberculosis*. This is especially useful in light of the fact that *M. tuberculosis* infections are often genetically diverse. Mixed infections as well as within-host evolution can explain the existence of these discrete subpopulations.[119]

2.3.3 Differences between *M. tuberculosis* and *Enterobacteriaceae* infections and consequences for diagnostics

The *Enterobacteriaceae* family comprises several clinically prominent opportunistic pathogens and commensals. Two of the most relevant opportunistic pathogens in this family are *E. coli* and *K. pneumoniae*, both being a common source of community and hospital-acquired infections and showing wide-spread antimicrobial resistance. This severely restricts treatment options for Gram-negative infections.[120;121] Both *E. coli* and *K. pneumoniae* have much shorter generation times than *M. tuberculosis*, around 20-30 minutes compared to around 16 hours.[122;123] In contrast to *M. tuberculosis*, they also have a diverse accessory genome augmented by conjugative plasmids. These enable frequent horizontal gene transfer, which aids the acquisition of resistance.[124] Due to plasmid-mediated resistance and short generation times, resistance can spread within *E. coli* and *K. pneumoniae* populations much more quickly than via *de novo* chromosomal mutations and vertical spread,[125] as is the case in *M. tuberculosis*.

In particular, this led to the currently observed high levels of extended-spectrum beta-lactamase (ESBL) and carbapenemase producing *E. coli* and *K. pneumoniae*,[124;126] which can cleave most beta-lactam antibiotics. Carbapenems and third-generation cephalosporins are central to the treatment of Gram-negative infections, especially the carbapenems, which are a class of last-resort antibiotics with relatively few adverse effects.[127] *E. coli* and *K. pneumoniae* that are resistant to those two antibiotic classes are hence at the top of the WHO bacterial priority pathogens list in 2024, together with rifampicin-resistant *M. tuberculosis*, indicating that they pose a critical threat to public health.[101] While beta-lactam resistance is the most concerning for treatment, quinolone resistance also readily occurs in *Enterobacteriaceae*. Initially, this was believed to be purely chromosomally evolved, since the protein target DNA gyrase is an essential protein encoded in the main chromosome of the bacteria, but plasmid-mediated quinolone resistance has

been reported in *K. pneumoniae* and *E. coli* as well.[126]

The reference method for phenotypic AST in *E. coli* and *K. pneumoniae* is broth microdilution for obtaining MICs.[128] This is often performed by automated microbiology systems such as the BD Phoenix,[129] which can return an antibiogram within 16 hours due to the short generation time of these bacteria.[130] This system requires pure culture isolates as a starting point, necessitating pre-culturing of the bacteria to obtain these. Single-colony picks from solid media cultures are hence the basis for phenotypic AST in *Enterobacteriaceae*, and they also form the basis for WGS-AST. This is in contrast to *M. tuberculosis*, where liquid cultures are typically used for both phenotypic and genotypic AST.

That being said, there are other approaches to phenotypic testing for bloodstream infections, such as the EUCAST recommended rapid disc diffusion method.[131] This direct rapid antimicrobial susceptibility testing (RAST) approach has been validated for use in *E. coli*, *K. pneumoniae* and other pathogens in a multicentre study and out-performs the routine approaches in terms of time to result with only 4-6 hours of incubation needed.[132] In contrast to the single-colony-pick based approaches, direct RAST is performed directly from blood culture and should hence allow within-host heterogeneity to be captured. On the other hand, direct RAST is only available for a limited number of antibiotics.

While WGS-AST is used in surveillance and research of Gram-negative infections,[133;134] it is not yet an accepted diagnostic tool. However, it has been shown that WGS-AST can reach a sensitivity and specificity comparable to routinely used phenotypic methods.[135] To process the sequencing reads, the genetic makeup and AST approaches specific to the bacterial species warrant the use of divergent methods. Since *M. tuberculosis* is largely chromosomal, mapping approaches are common. These map the reads to a reference strain for variant calling (most often H37Rv).[112] For *E. coli* and *K. pneumoniae*, their large accessory genome warrants approaches using *de novo* assembly.[72] These do not detect subpopulations, since assembly algorithms seek a single consensus sequence. Due to single-colony picks being the basis of AST in *E. coli* and *K. pneumoniae*, subpopulations are also not expected in the first place. In *M. tuberculosis* on the other hand, subpopulations can be present due to picking 'crumbs' and not pure colonies for sequencing, and the mapping approach allows detecting these through the fraction of read support in

divergent loci. Both phenotypic and WGS-AST in *M. tuberculosis* are hence somewhat metagenomic, while the diversity of *E. coli* and *K. pneumoniae* infections is not captured as part of routine AST.

However, efforts are underway to establish direct-from-blood culture bottle metagenomic sequencing in *E. coli* and *K. pneumoniae*,[136] which should allow capturing within-host diversity in the sample. This would be the genotypic equivalent of the direct RAST phenotypic method for AST.

2.4 Thesis outline: Closing the gaps - Improving genetics-based resistance prediction

According to recommendations from EUCAST, the capabilities of WGS-AST should be evaluated based on the categorial agreement of the resistance prediction based on WGS-AST and an established phenotypic AST method.[82] Categorial agreement between the two should be as high as possible, ideally over the required minimum threshold of 95% with respect to both sensitivity and specificity to pass the ISO standard for antimicrobial susceptibility test devices.[137] This implies few false positives and false negatives. In the context of medical testing, the CLSI standard is to describe these as major errors (ME, false positives) and very major errors (VME, false negatives), respectively.[57] This distinction is made because VMEs imply a resistant phenotype accidentally called susceptible by the tested method, which would result in treatment failure and possible onward transmission. MEs on the other hand can lead to treating a susceptible infection with a different antibiotic, which may incur more side-effects for the patient and should hence be avoided, but will likely not result in treatment failure. The false negative rate (FNR) will hence be an important performance measure for WGS-AST. Likewise, sensitivity will be used as a readout, since it is 1 - FNR.

There are multiple reasons why the results of phenotypic AST and genotypic AST might show categorial disagreement. Phenotypic AST directly assesses whether the pathogen is resistant to the drug it is exposed to, by examining if there is growth or not. However, identifying bacterial growth and hence resistance can be subjective, especially when the MIC of a sample is very close to the ECOFF. For many drug-pathogen combinations, the MIC distributions are not bimodal, but rather approximate a normal distribution. The ECOFF may then not be able to clearly separate the wild type from the resistant phenotype, hence possibly leading to major and very major errors. We will also inevitably encounter human error in the phenotyping procedure, such as mislabelling and swapping of samples. And for some drug-pathogen combinations, the phenotyping can be irreproducible due to measurement errors.

The above limitations are difficult to resolve, since they are either based on the inherent properties of the MIC distribution of the drugs, or are due to unresolvable factors, such as the inevitability of human error. But there are several aspects of WGS-AST which can be improved upon to achieve higher categorial agreement with phenotypic AST.

Firstly, while error rates associated with short- and long-read sequencing technologies might have significantly decreased in recent years, there are still a considerable number of sequencing errors in any given sequencing dataset. This is problematic if these sequencing errors happen to be in resistance-associated regions, where they might influence the resistance prediction result. The sequencing error could then be at the site of a resistance-associated variant (RAV), leading to a very major error. In the first results chapter (Chapter 3), we will produce a comprehensive list of compensatory mutations (CMs) in the RNA polymerase, and investigate their influence on *in vitro* fitness. CMs do not cause resistance but are associated with it, by definition, so could be used to expand the current list of RAVs. This idea will be explored in the first results chapter (Chapter 3), as part of the investigation of CMs in *M. tuberculosis*. The majority of Chapter 3 has been published in a peer-reviewed journal.[138]

Secondly, resistance might be present in a subpopulation of the sample taken for testing. If based on diverse inocula as opposed to single-colony picks, phenotypic methods can detect resistant subpopulation due to the use of selective conditions, often resolving down to a single viable organism.[47] Sequencing approaches on the other hand often ignore subpopulations, despite having the ability to resolve them, due to the need to exclude sequencing errors. We will hence investigate the effect of considering resistant subpopulations in WGS-AST, which becomes apparent when evaluating their influence on the categorial agreement. In the second results chapter (Chapter 4), we will look at this in the context of rifampicin resistance in *M. tuberculosis*. This part of Chapter 4 has been published in the journal of antimicrobial chemotherapy - AMR.[139] We will also evaluate whether subpopulations might be relevant in *Enterobacteriaceae* infections, where they are currently not routinely resolved due to culture conditions, using the example of *E. coli* and *K. pneumoniae* bloodstream infections.

Lastly, databases of antibiotic resistance determinants are not exhaustive, since they can by definition only contain mutations with sufficient evidence supporting their effect on drug susceptibility. As a result, they cannot contain all possible genetic causes of resistance, and WGS-AST cannot provide a definitive result for samples containing novel mutations. There is hence a need to develop methods for predicting resistance for unknown, novel alleles. In the third results chapter (Chapter 5), we will test the ability of machine learning models to predict resistance based on the knowledge of resistance determinants and their

influence on structural and physicochemical properties of the affected amino acids. Using the example of fluoroquinolone resistance in both *M. tuberculosis* and *E. coli*, we will demonstrate the possibilities and limitations of these model-based approaches for resistance prediction.

Throughout this thesis, we will be using different versions of the CRyPTIC dataset of clinical samples of *M. tuberculosis*.[140;141] For Chapters 3 and 5, we used dataset version 1.1.1 from January 2021,[142] which is the initial release. This version includes 77,860 WGS samples, 41,130 of which with matching phenotypic AST data. For Chapter 4, we used version 3.0.0 of the dataset.[143] This is because we needed to resolve subpopulations through the fraction of read support, which is only supported from this version forward. It contains 41,575 samples with both WGS and phenotypic AST data.

3 Rifampicin resistance and compensation in *M. tuberculosis*

This chapter investigates compensation in rifampicin resistance. Specifically, this involves identification of compensatory mutations (CMs) and where they occur on the rifampicin target protein. We will also assess the usefulness of CMs for improving rifampicin resistance prediction and the influence of resistance mutation and CM presence on *in vitro* growth of *M. tuberculosis* as a proxy for bacterial fitness.

3.1 Introduction

Rifampicin is one of the four first-line antibiotics for treating tuberculosis (TB). It binds to the RNA polymerase (RNAP), which is involved in the transcription of DNA into RNA. Due to the fundamental nature of this process for the cell's survival, the RNAP is an essential protein. The RNAP holoenzyme in *M. tuberculosis* is comprised of five main subunits ($\alpha_2\beta\beta'\gamma$) and the σ factor (Figure 3A).[144] The drug rifampicin binds close to the active site in the β subunit (*rpoB* gene). In susceptible bacteria, amino acids within the so-called 'rifampicin resistance determining region' (RRDR) form hydrogen bonds with rifampicin (Figure 3B). The bound rifampicin sterically obstructs the elongation of newly synthesized RNA, thereby stalling protein production in the bacteria.[145]

Compensation of rifampicin resistance fitness cost

Since *M. tuberculosis* exhibits very little evidence of horizontal gene transfer,[96] resistance to this drug mostly arises through chromosomal mutations within or close to the RRDR that prevent rifampicin from binding.[146–148] These resistance-conferring mutations introduce different amino acid side chains close to the active site of the RNAP, hence it is unsurprising that they also introduce a fitness cost.[149;150] This fitness cost in rifampicin-resistant bacteria manifests as decreased performance in competition assays with susceptible bacteria.[150] While the molecular basis of the fitness deficit is not entirely understood, it is suspected to be related to decreased stability of the RNAP open-promoter complex[151] as well as steric hindrance of RNA exit by the mutated amino acids.[152]

The existence of a fitness cost was long seen as an indicator that by restricting antibiotic use in general or by periodically cycling different antibiotics, resistance would not readily spread throughout the human

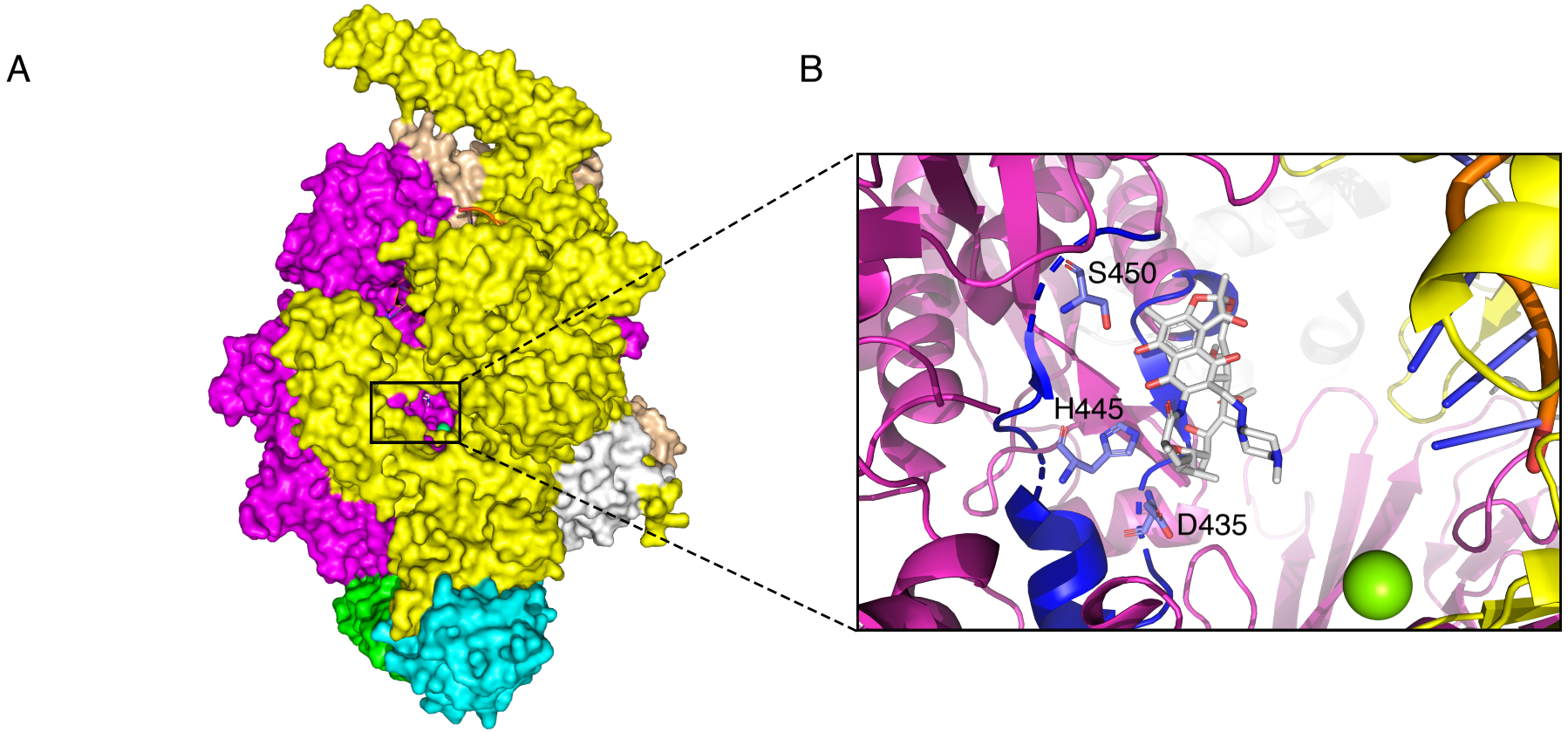


Figure 3: The drug rifampicin interrupts RNA synthesis by binding to the β subunit of the RNA polymerase (RNAP).[145] **(A)** Overview of the entire RNAP. The β subunit of the RNAP is shown in magenta, the β' subunit in yellow, the two α subunits in light blue and green, the ω subunit in white and the σ factor in light orange. The active site is framed in black. It can be seen through the secondary channel, with the active site magnesium depicted in green and the drug rifampicin in white. **(B)** Close-up of the active site. The DNA strand used as a template for transcription is shown on the right in orange and dark blue. The β subunit of the RNAP is shown in magenta, with the 'rifampicin resistance determining region' (RRDR) highlighted in dark blue. The protruding amino acids D435, H445 and S450 are reported to form hydrogen bonds or van der Waals interactions with the drug rifampicin.[145] Due to the proximity of the RRDR to the RNAP active center, the binding of rifampicin causes a disruption of RNA synthesis through steric clash.

population. This unfortunately was not the case and it was initially suspected that the fitness cost, especially of the most common resistance mutation *rpoB* S450L, is simply not high enough to stop resistance spread.[149] While this is a viable explanation, especially in real-world scenarios where intermittent exposure to antibiotics as part of treatment can favour the resistant strains, the main reason for the spread of rifampicin resistance despite the fitness cost is likely more complex.

The low fitness phenotype has been observed to be partially or completely rescued by the presence of so-called CMs.[98;153;154] These CMs emerge in various subunits of the RNAP[153] and have been shown to partially restore RNAP activity in *Mycobacterium smegmatis*.[154] The existence of CMs also gives a plausible explanation for the persistence of resistance mutations long after antibiotic treatment is stopped, when the fitness cost should lead to genetic reversion to the wild-type phenotype.

Identification of compensatory mutations in large datasets

Due to their potential to lead to fixation of resistance mutations[152] and the attendant epidemiological consequences, CMs have been extensively investigated. Many candidates have been identified in previous studies,[155–161] but the number of samples available to these studies is often small. In addition, the effects on growth and/or polymerase activity have only been experimentally confirmed for a small subset of putative CMs.[154;160] To further our understanding of resistance spread and persistence, it would be useful to construct a more comprehensive list of CMs and, if possible, to dissect their direct influence on the fitness of *M. tuberculosis*.

The identification of CMs is not trivial and depends on how they are defined. In some publications, due to the low number of available samples, CMs are simply assumed to be all mutations in the RNAP that co-occur with resistance mutations, even if these mutations are also seen in samples that are susceptible.[157] Higher numbers of samples allowed CMs to be defined as mutations that *exclusively* occur with resistance mutations.[155;159;161] With the dataset of 77,860 *M. tuberculosis* sample genomes we have at hand, where sequencing errors and sample mislabeling are likely to lead to a considerable amount of false positives, it is possible and necessary to employ statistical association testing to identify CMs.

Role of compensation in disease transmission

Whether CMs are directly associated with increased transmission rates is an ongoing discussion in the field. While there is strong evidence of compensation being associated with higher transmission rates in *M. tuberculosis*,[162;163] it has also been suggested that the final transmissibility is determined by an epistatic effect consisting of strain background, CMs and the low-fitness cost resistance mutation *rpoB* S450L.[164] One investigation however claimed to have found very limited influence on transmission by compensation, or, in fact, by any other commonly assumed characteristics of high-fitness phenotypes such as Lineage 2 association or the presence of low fitness cost resistance mutations.[165] This study however did not assess transmission duration or success directly but instead looked only at clustering of samples as a proxy for transmissibility. Additionally, they used a very conservative cut-off for defining these clusters (5 SNPs or fewer), which will fail to associate related samples in case a hypothetical common ancestor was not sequenced. Overall, the overwhelming evidence still points towards CMs being associated with increased transmission.

Recent evidence suggests that *in vivo* fitness might be enhanced directly by compensatory evolution,[166] which makes our *in vitro* growth data in combination with the identification of a comprehensive list of CMs very timely. If we could show increased growth in samples with compensation, this would agree with the recent finding of increased *in vivo* fitness in compensated samples. This would imply that compensatory evolution leads to higher *in vitro* growth, which translates into higher *in vivo* fitness, and ultimately increased transmission.

3.2 Aims of this chapter

In this Chapter, we shall identify a comprehensive set of 51 putative CMs based on a Genetics dataset comprising 77,860 *M. tuberculosis* sample genomes collated by the international CRyPTIC project.[141] Our dataset is about 10-100 times larger than those used in previous studies,[155–161] increasing statistical power and robustness. This high-confidence list of CMs will allow us to propose compensatory mechanisms by mapping our putative CMs onto the RNA polymerase structure. More importantly, the list contributes significantly to the field by potentially forming the basis for future research into compensation and fitness.

In the context of this thesis, our list of highly resistance-associated compensatory mutations might help us improve resistance prediction based on whole genome sequencing (WGS) by allowing us to flag resistance in samples with sequencing errors in the relevant resistance-conferring regions. Therefore, we will also assess the usefulness of using this list for resistance prediction.

Another advantage of our investigation is that we have *in vitro* growth data available for a subset of 13,990 samples derived from photographs of 96-well plates after two weeks incubation – we shall call this the Growth dataset.[167] This will allow us to investigate the association between the observed growth phenotypes and the respective genotype, i.e. the presence of resistance-conferring and compensatory mutations. Doing so can give us valuable insight into the fitness changes incurred by the fitness cost and the compensation thereof, and allow us to draw conclusions towards the influence on *M. tuberculosis* infection transmission.

The aims of this Chapter are hence three-fold:

1. Identify new compensatory mutations in the RNA Polymerase of *M. tuberculosis* and propose possible compensatory mechanisms

2. Use the list of compensatory mutations to improve resistance prediction by association

3. Investigate the influence of compensatory mutations on fitness by using *in vitro* growth data

3.3 Methods

Most figures and tables in this Chapter can be reproduced using a GitHub repository and attendant Python3 Jupyter notebook available online.[168] The relevant analysis can be rerun in the browser using the corresponding google colab button in the README, or it can be run locally but this requires all dependencies to be installed.

Dataset sources

The datasets used as a basis for the analysis in this Chapter were already available in the format of a series of data tables, and I was hence not involved in the data collection or curation process.

The Genetics dataset comprised 77,860 whole-genome sequenced patient-derived samples of *M. tuberculosis* collected and collated by the CRyPTIC project.[141] The specific version used was v1.1.1, released in January 2021.[142] Mutations with respect to version 3 of the H37Rv reference genome were aggregated and are available in a series of data tables, along with other data, such as the predicted resistance to various antibiotics, ENA accession numbers, lineage association and the respective amount of growth in 96-well plates.[169] In brief, following read filtering with kraken2 (v2.1.3),[170] speciation via competitively mapping reads to a manifest of mycobacterial reference genomes using minimap2 (v2.24-r1122),[171] and lineage determination with mykrobe (v0.12.1),[172] genetic variants were called using clockwork v0.12.5 which incorporates minos[173]. Version 3 of the H37Rv *M. tuberculosis* reference genome is used throughout. The resulting variant call format (vcf) files were then parsed using gnomonicus v3.0.6[174] to produce a series of data tables containing all the genetic information from these samples.

A subset of 13,990 samples (the Growth dataset) were inoculated onto 96-well broth microdilution plates, incubated for two weeks and then a photograph was taken using a Thermo Fisher $Sensititre^{TM}$ $Vizion^{TM}$ Digital MIC Viewing System by the CRyPTIC project.[140;175] Each photograph was subsequently analysed and the detected bacterial growth, as measured by AMyGDA, was recorded as described elsewhere (Figure 4).[167] In short, the bacterial growth is defined as the percentage of dark pixels in a circle centred on the well but with half the radius (to avoid edge artifacts and shadows) and therefore can reach 100%.

Here, we only considered mean growth of the two positive control wells, which grow bacteria in absence of antibiotic compounds (Figure 4A, bottom right). The Pearson correlation coefficient of growth in the control wells was 0.72, indicating good correlation but the divergence between wells is noticeable (Figure 4B). The skew towards higher growth detected in well 2 might be related to this well being more prone to shadows, since it is located at the bottom right of the plate. The shadows can in turn lead to more growth being detected as a result of darker pixels.

Processing the images produces a relative growth value for each well and a mean positive control value for each plate. The dataset distribution of mean positive control growth is shown in Figure 4C. We shall refer to this dataset as the Growth dataset.

Pairwise statistical association testing

Using the Genetics dataset we compiled a list of putative CMs by performing statistical association testing for all mutations in the RNA polymerase (RNAP) genes that were observed to co-occur with any mutations listed as conferring resistance to rifampicin in the second edition of the WHO catalogue of mutations in *M. tuberculosis*.[79] We excluded any samples where sequencing was inconclusive. The statistical test used was Fisher's exact test, which determines categorical independence of two variables. It is based on the hypergeometric distribution and uses the binomial coefficients to calculate exact test statistics, hence it has a high computational cost for datasets of our size.

To minimise the effect of this, we used a Python3 package with an optimised implementation of Fisher's test (fishers-exact-test). This is faster than the standard implementation and enabled us to run Fisher's exact test for every pair of resistance and co-occurring mutation within 2.5 hours on a modern workstation.

Evaluating and filtering results from Fisher's exact test

For a largely clonal species like *M. tuberculosis*, linkage disequilibrium (LD) will artificially inflate the p-values in any SNP-phenotype association test, as can be seen in most genome wide association studies of microbial species.[176] LD can hence mask true causal mutation-phenotype relationships, like the one between rifampicin resistance and CM presence. Furthermore, traditional stratification corrections alone

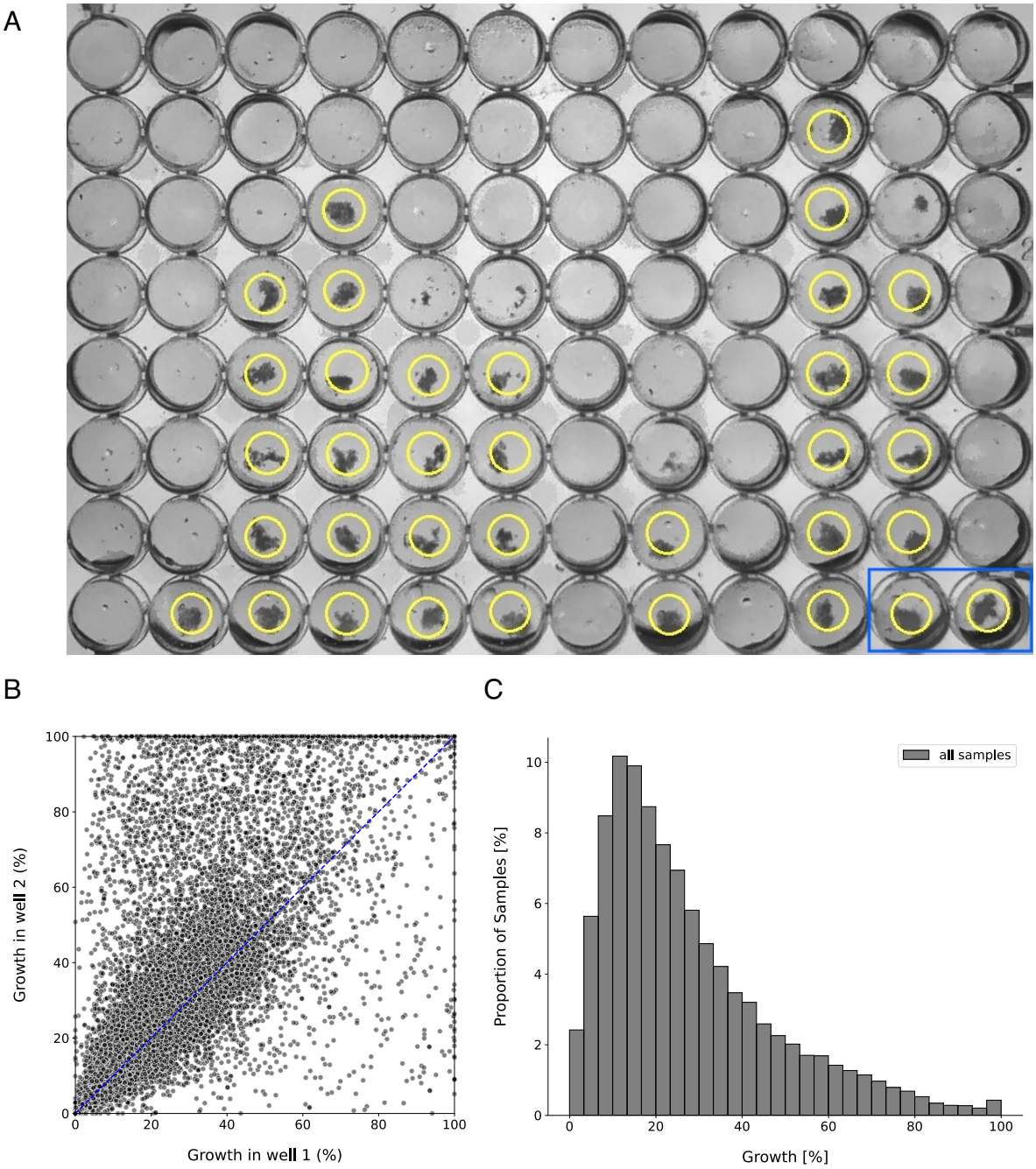


Figure 4: Growth data acquisition via detection of covered well area. **(A)** Example photograph of the 96-well broth microdilution plates processed by AMyGDA.[167] Growth is detected in all wells with a yellow circle, and detection is restricted to the area within the circle. The plate has a dilution series of different antibiotics in each column, and two positive control wells on the bottom right (blue box). Picture courtesy of Prof. Philip W. Fowler. **(B)** Correlation of growth in the positive control wells over all samples with growth data. The line of perfect correlation is shown in blue. The Pearson correlation coefficient of growth in paired control wells across all samples was 0.72. **(C)** The distribution of growth in percent of covered well-area as measured in the CRyPTIC project[167] was plotted as a histogram against the proportion of samples that display this amount of growth.

will not be sufficient to identify causal variants in species with strong LD and the use of homoplasy as a criterion has been proposed instead.[177]

We hence decided to use a more conservative p-value correction than simply applying a Bonferroni correction. To assess the performance of the p-value correction, we compiled a list of reference CMs from the literature. All included reference CMs have either been confirmed experimentally or have been identified by at least three published studies, hence we shall assume that they are high-confidence CMs (Table 1). We shall use this compiled reference list to assess our arbitrary choice of p-value cut-off by considering the true positive rate (TPR, Figure 5). Moving forward, we decided to only include hits with a p-value above the 98th percentile, since this still recovers 16/17 (94.1%) of high-confidence CMs (Table 1) and gives us a reasonable number of hits that we can process using our downstream filters (Figure 5). The entire list without p-value cut-off is available online.[168]

To make sure we filter out any phylogenetic mutations, we additionally removed any synonymous mutations and tested the remaining for homoplasy. Synonymous mutations tend to be phylogenetic markers since they do not change the protein structure and hence should mostly not be subject to any selectional pressure. Homoplasy is an indicator of selective pressure that causes similar mutations to arise in genetically and evolutionary distinct populations. Accordingly, we use it as a necessary and sufficient criterion for deciding on the validity of our hits.

Our resulting preliminary list of CMs, including their respective associated resistance mutation is shown in Table 2.

putative CM	Exp. evidence	155	157	156	161	158	154	159	160	sum(ref)	fisher
rpoC V483G	✓	✓	✓	✓	✓	✓	✓	✓		7	✓
rpoC V483A	✓	✓	✓	✓	✓			✓		5	✓
rpoC I491T		✓	✓	✓		✓	✓			5	✓
rpoC L527V		✓	✓		✓	✓				4	✓
rpoA T187A		✓	✓			✓			✓	4	✓
rpoC G332R				✓	✓	✓		✓		4	✓
rpoC L507V			✓		✓		✓	✓		4	✓
rpoC N698S		✓	✓	✓	✓					4	✓
rpoC W484G		✓	✓						✓	3	✓
rpoC I491V	✓	✓	✓		✓					3	✓
rpoC V517L		✓	✓		✓					3	✓
rpoC L516P	✓	✓			✓			✓		3	✓
rpoC H525Q			✓	✓	✓					3	
rpoC N698K	✓	✓			✓		✓			3	✓
rpoC V1252L	✓				✓		✓	✓		3	✓
rpoC F452L	✓						✓			1	✓
rpoC P1040R	✓				✓					1	✓

Table 1: Reference compensatory mutations (CMs) used for evaluating our approach to identifying new CMs. Putative CMs are shown on the left. We included reference CMs that were either proven experimentally or identified in at least three of the reference papers. Fisher indicates if the CM came up as significantly resistance associated in our statistical association test and was located below the heuristic p-value threshold and showed homoplasy.

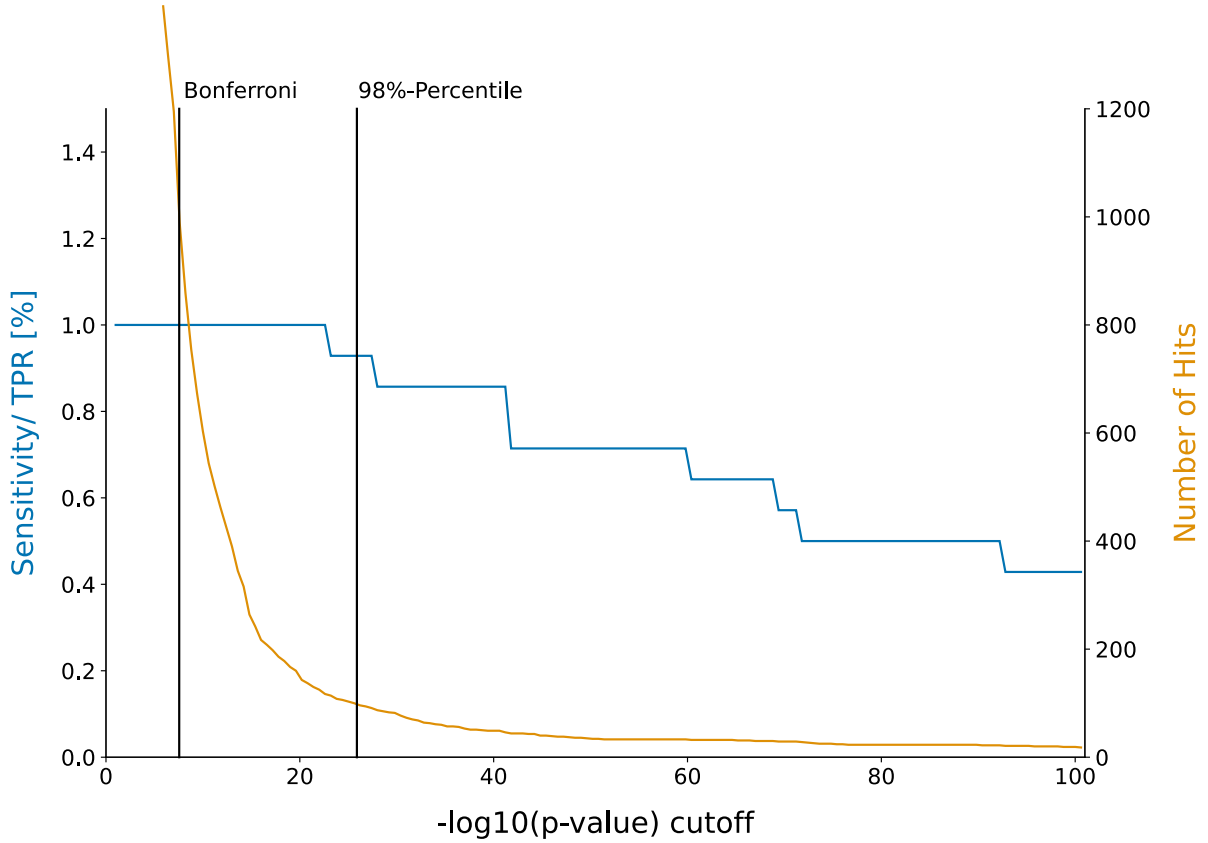


Figure 5: Sensitivity and number of significant hits (putative compensatory mutations) depending on p-value. The graph shows the number of significant hits and reference hits detected depending on the $\log_{10}$ p-value cut-off shown on the x-axis. The left y-axis refers to the percentage of found reference hits from a compiled list, also termed sensitivity or true positive rate (TPR). The right y-axis shows the number of mutations that were classified as significantly resistance associated under the respective cut-off. The vertical lines indicate the p-value cut-off with Bonferroni correction and our heuristic p-value cut-off at the 98% quantile, respectively.

resistance	putative CM	only CM	both	-$\log_{10}$(p-value)	literature evidence	homoplasy
rpoB S450L	*rpoC* E1092D	2012	1989	inf		
rpoB S450L	*rpoC* V483G	37	1206	inf	✓	✓
rpoB S450L	*rpoC* I491V	19	665	inf	✓	✓
rpoB S450L	*rpoC* V483A	33	586	inf	✓	✓
rpoB S450L	*rpoC* I491T	10	457	293.06	✓	✓
rpoB S450L	*rpoC* P1040R	32	396	225.09	✓	✓
rpoB S450L	*rpoC* F452S	2	345	230.58	✓	✓
rpoB S450L	*rpoB* E761D	0	304	207.05		
rpoB S450L	*rpoB* L731P	1	226	151.44	✓	✓
rpoB S450L	*rpoC* N698S	2	205	135.23	✓	✓
rpoB S450L	*rpoC* D485Y	8	194	118.89	✓	✓
rpoB S450L	*rpoC* V517L	1	184	122.87	✓	✓
rpoB S450L	*rpoC* G332S	16	179	100.19		✓
rpoB S450L	*rpoC* V1252L	3	175	113.24	✓	✓
rpoB S450L	*rpoA* T187A	3	171	110.54	✓	✓
rpoB S450L	*rpoC* D485N	2	166	108.82	✓	✓
rpoB S450L	*rpoC* L516P	3	144	92.37	✓	✓
rpoB S450L	*rpoC* G433S	3	141	90.35	✓	✓
rpoB S450L	*rpoB* R827C	8	123	72.06		✓
rpoB S450L	*rpoC* P1040S	3	118	74.93		✓
rpoB S450L	*rpoC* L527V	3	113	71.58	✓	✓
rpoB S450L	*rpoC* G332R	5	113	68.95	✓	✓
rpoB S450L	*rpoC* P1040A	1	110	72.70	✓	✓
rpoB L452P	*rpoB* I1106T	3	103	212.49		
rpoB D435G	*rpoB* I1106T	3	103	263.70		
rpoB S450L	*rpoC* K445R	0	98	66.49		✓
rpoB S450L	*rpoC* F452L	3	96	60.24	✓	✓
rpoB S450L	*rpoC* L547V	2	80	50.94		
rpoB S450L	*rpoC* W484G	11	80	41.69	✓	✓
rpoB S450L	*rpoB* A692T	16	79	37.45		
rpoB S450L	*rpoB* I480V	3	78	48.27	✓	✓
rpoB S450L	*rpoA* V183G	10	77	40.62	✓	✓
rpoB S450L	*rpoB* K891E	0	74	50.18		✓
rpoB S450L	*rpoC* N416S	3	72	44.30	✓	✓
rpoB S450L	*rpoB* Q409R	14	69	32.77		✓

rpoB S450L	*rpoC* V1039A	1	68	44.37		✓
rpoB S450L	*rpoB* P45S	5	65	37.49	✓	✓
rpoB S450L	*rpoB* Q975H	17	65	28.57		
rpoB S450L	*rpoC* A521D	2	65	40.93	✓	✓
rpoB S450L	*rpoC* L507V	1	64	41.68	✓	✓
rpoB S450L	*rpoC* N826T	0	64	43.39		
rpoB S450L	*rpoC* V431M	4	62	36.58	✓	✓
rpoB S450L	*rpoB* V496A	0	60	40.68		
rpoB S450L	*rpoB* I488V	2	59	36.94		✓
rpoB S450L	*rpoC* V1252M	2	59	36.94	✓	✓
rpoB S450L	*rpoB* A286V	4	56	32.68	✓	✓
rpoB S450L	*rpoC* T812I	3	51	30.48	✓	✓
rpoB S450L	*rpoC* K1152Q	0	51	34.57		
rpoB S450L	*rpoB* R827L	0	50	33.89	✓	✓
rpoB S450L	*rpoC* L449V	0	50	33.89	✓	✓
rpoB S450L	*rpoC* N698K	4	48	27.50	✓	✓
rpoB S450L	*rpoC* F452C	1	48	30.94	✓	✓
rpoB S450L	*rpoA* A180V	0	48	32.54		✓
rpoB I491F	*rpoC* E1033A	25	46	95.60		✓
rpoB S450L	*rpoA* D190G	0	46	31.18	✓	✓
rpoB S450L	*rpoA* G31S	0	44	29.82	✓	✓
rpoB S450L	*rpoC* P434R	3	44	25.91	✓	✓
rpoB S450L	*rpoA* E184D	0	40	27.11		✓
rpoB H445R	*rpoC* S561P	2	39	98.27	✓	✓
rpoB V170F	*rpoB* V168A	2	25	64.98		
rpoB L452P	*rpoB* H1028R	1	23	46.47		
rpoB S450W	*rpoA* P25R	1	20	45.52		
rpoB H445D	*rpoC* G388A	1	17	32.33	✓	✓
rpoB D435G	*rpoB* I491L	18	17	32.89		✓
rpoB D435Y	*rpoB* R167C	1	15	29.72		
rpoB S450W	*sigA* A223T	0	15	35.07		
rpoB H445Y	*rpoB* E207K	0	15	29.05		
rpoB V170F	*rpoC* G571R	11	14	30.91	✓	
rpoB S441A	*rpoC* L405M	28	12	36.05		
rpoB S441A	*rpoB* L464M	3	11	39.12		
rpoB Q432P	*rpoC* T853A	3	10	28.72		
rpoB Q432K	*rpoZ* T107I	1	10	30.13		
rpoB S441A	*sigA* E385Q	14	9	27.77		

rpoB S441A	*sigA* G380A	3	9	31.34		
rpoB S441A	*sigA* I382V	3	9	31.34		
rpoB S441A	*sigA* L386M	4	9	30.83		
rpoB S441A	*rpoB* E460D	5	8	26.66		
rpoB S441A	*rpoC* I128V	3	8	27.56		
rpoB S441A	*rpoB* R791T	0	7	25.92		

Table 2: Hit list resulting from Fisher's exact test for association of resistance with co-occuring mutations, after removing synonymous mutations. The first column indicates the resistance mutation that the putative compensatory mutation (CM) in the second column is associated to. 'Only CM' indicates how often the CM occurs on its own, without the corresponding resistance mutation, and 'both' indicates how often we see the two mutations occur together. The last two columns indicate whether the CM has been mentioned in the literature and if it shows homoplasy, respectively.

Constructing binary homoplasy index

The CRyPTIC consortium constructed a phylogenetic tree from 15,211 isolates,[141;178] which is the number of samples collected that have both genetic data and minimum inhibitory concentrations available, although some of these were not photographed and processed using AMyGDA. They constructed the tree based on the pairwise genetic distance matrix, where a neighbourhood-joining tree was visualised and annotated in R using the *ggtree* library and the lineages were assigned using mykrobe.[172;179] We mapped our putative hits on the tree using the publicly available software iTOL.[180] Any samples originating from organisms other than human-adapted *M. tuberculosis* (e.g. *Mycobacterium caprae* and *Mycobacterium bovis*) and lineages other than 1-4 were removed. We also removed any samples where sequencing was inconclusive, since these were initially excluded from our statistical association testing. We inferred homoplasy if a mutation appeared in at least three genetically distant samples without common ancestors in the tree. This number was chosen to account for the possibility of false positives due to sequencing errors or hemiplasies. It should be noted that if a mutation was not classified as homoplasic, this does not exclude the possibility of this mutation being a CM, since the phylogenetic tree does not include all samples (15,211 out of 77,860). A negative homoplasy call could hence also be caused by under-representation of the samples with the respective mutation in the tree.

Plotting of growth data

The average growth, as measured by AMyGDA[167] of the two positive control wells for each sample was obtained from the CRyPTIC dataset. To quantify the difference in growth of the different sample types (pan-susceptible, resistant or resistant with CMs), the pairwise Mann-Whitney p-value for each two respective growth distributions was calculated. In addition, the medians and confidence intervals of the medians were approximated using a bootstrapping approach and included in notched box-plot figures. This is a robust method, which considers the non-normal distribution of the growth data. For evaluating significance, we always referred to Bonferroni-corrected Mann-Whitney p-values.

Mapping of putative compensatory mutations on RNA polymerase structure

High-confidence homoplasic hits were mapped in PyMOL onto the crystal structure of the *M. tuberculosis* RNAP in complex with rifampicin[145] to identify clusters and thereby elucidate potential compensatory mechanisms.

3.4 Results

3.4.1 Rifampicin resistant *M. tuberculosis* samples show lower *in vitro* growth densities than pan-susceptible samples

First, we analysed the effect of resistance-conferring mutations on the *in vitro* growth of *M. tuberculosis* bacteria. Based on our Growth dataset of 13,990 whole genome sequenced patient-derived *M. tuberculosis* samples with associated 96-well growth data, we defined two sets of samples: One resistant to rifampicin and another that is susceptible. Resistance was defined as the presence of any mutation associated with resistance to rifampicin according to a published catalogue,[169] whilst a sample was assumed to be susceptible if the catalogue classified it as pan-susceptible (susceptible to all four first line tuberculosis drugs – rifampicin, isoniazid, ethambutol and pyrazinamide). We excluded samples containing any other non-synonymous mutations in the RNA polymerase (RNAP), since they could have a secondary effect on growth. That being said, we cannot exclude an effect due to mutations in other genes. We shall further assume that synonymous mutations in the RNAP have no effect on the growth phenotype, and hence need not be excluded. The resulting sample sizes can be seen in Table 3.

There is a significant difference between the growth distributions of our sets (Figure 6A), as confirmed by a significant Mann-Whitney p-value of 2.95e-12 (Table 3). The median growth density in resistant samples was significantly lower than in pan-susceptible samples (4.5% absolute growth difference), hence we conclude that we see higher growth in non-resistant bacteria.

This is consistent with work by Gagneux *et al.* who reported that rifampicin resistance introduces a fitness cost in *M. tuberculosis*, which they showed by *in vitro* competitive fitness experiments.[150] Their study indicates that the magnitude of the fitness cost in *M. tuberculosis* depends on the mutation, with *rpoB* S450L, the most prevalent rifampicin resistance mutation observed clinically, being associated with the lowest fitness cost.[150] We had sufficient samples to investigate growth for the three most common resistance mutations in the dataset: S450L, H445Y and D435V. Also consistent with Gagneux *et al.*, the median growth of *rpoB* S450L mutants was the highest among the three resistance mutations (0.3% and 1.2% absolute growth difference, respectively), while still growing to significantly lower densities

mutation	median growth [%]	CI low	CI high	p-value$_s$	n
pan-susceptible	22.1	21.6	22.7		5283
any resistance	17.6	16.6	18.5	2.95e-12	795
specific resistance mutation:					
rpoB S450L	16.7	15.1	18.9	4.37e-03	196
rpoB H445Y	15.5	12.0	19.2	1.26e-02	42
rpoB D435V	16.4	15.5	17.2	3.92e-14	325

Table 3: Median growth of samples with and without indicated resistance mutations, where growth is represented by the percentage of a well containing bacterial growth as measured by the CRyPTIC project. The confidence interval (CI) for the median is calculated using bootstrapping where 'CI low' indicates the lower threshold and 'CI high' the upper threshold. The Mann-Whitney p-value is calculated in reference to pan-susceptible sample growth and *n* indicates the sample size.

than pan-susceptible samples (5.4% absolute growth difference, p = 4.37e-03, Figure 6B, Table 3). However, the difference in median growth between the three resistance mutations is non-significant (Table 4), presumably due to the number of eligible samples being too low.

When analysing the growth in samples with specific resistance mutations, it was striking that samples with the mutation *rpoB* S450L, despite being claimed to have a low fitness cost, had lower growth than the average resistant sample (Table 3). Here it is important to note that we are still looking at samples with exclusively resistance mutations in the RNAP complex, and no other non-synonymous mutations present. This makes it possible to largely rule out compensation as a factor influencing growth and explains the low sample size (Table 3).

We hence decided to look at the reason for the comparatively high growth in resistance only samples, and found that a few, 'disputed' resistance mutations (*rpoB* L430P, *rpoB* L452P, *rpoB* D435Y)[181] are likely responsible for the higher median growth. Samples with either of these mutations exhibit unusually high growth, comparable to susceptible growth levels (Figure 7). These specific mutations have been shown to test susceptible in phenotypic testing in up to 43% of cases,[181] so we decided to check the phenotypes of affected samples. Indeed we found that only 47% of samples with *rpoB* L430P were phenotypically resistant, with the other 53% of samples being unknown or susceptible calls. The same was true for 35% of samples with *rpoB* L452P and 30% of samples with *rpoB* D435Y. This demonstrates that it is non-trivial to assemble a high-confidence list of resistance mutations in *M. tuberculosis*. However, this does not have a large influence on our overall data analysis, since samples with these mutations make up

mutation	median growth [%]	p_{S450L}	p_{H445Y}	p_{D435V}	n
rpoB S450L	16.7		0.4193	0.2664	196
rpoB H445Y	15.5	0.4193		0.6445	42
rpoB D435V	16.4	0.2664	0.6445		325

Table 4: Median growth of resistant samples with specific resistance mutations. Mann-Whitney p-values are given with respect to the subscripted sample type and n indicates the sample size. Growth distributions of these samples are shown in Figure 6B.

less than 2% of the overall dataset. The effect is much more pronounced in the analysis in Figure 6, since we restricted the dataset to samples that only contain resistance mutations and no other mutations, which excluded a large number of resistant samples.

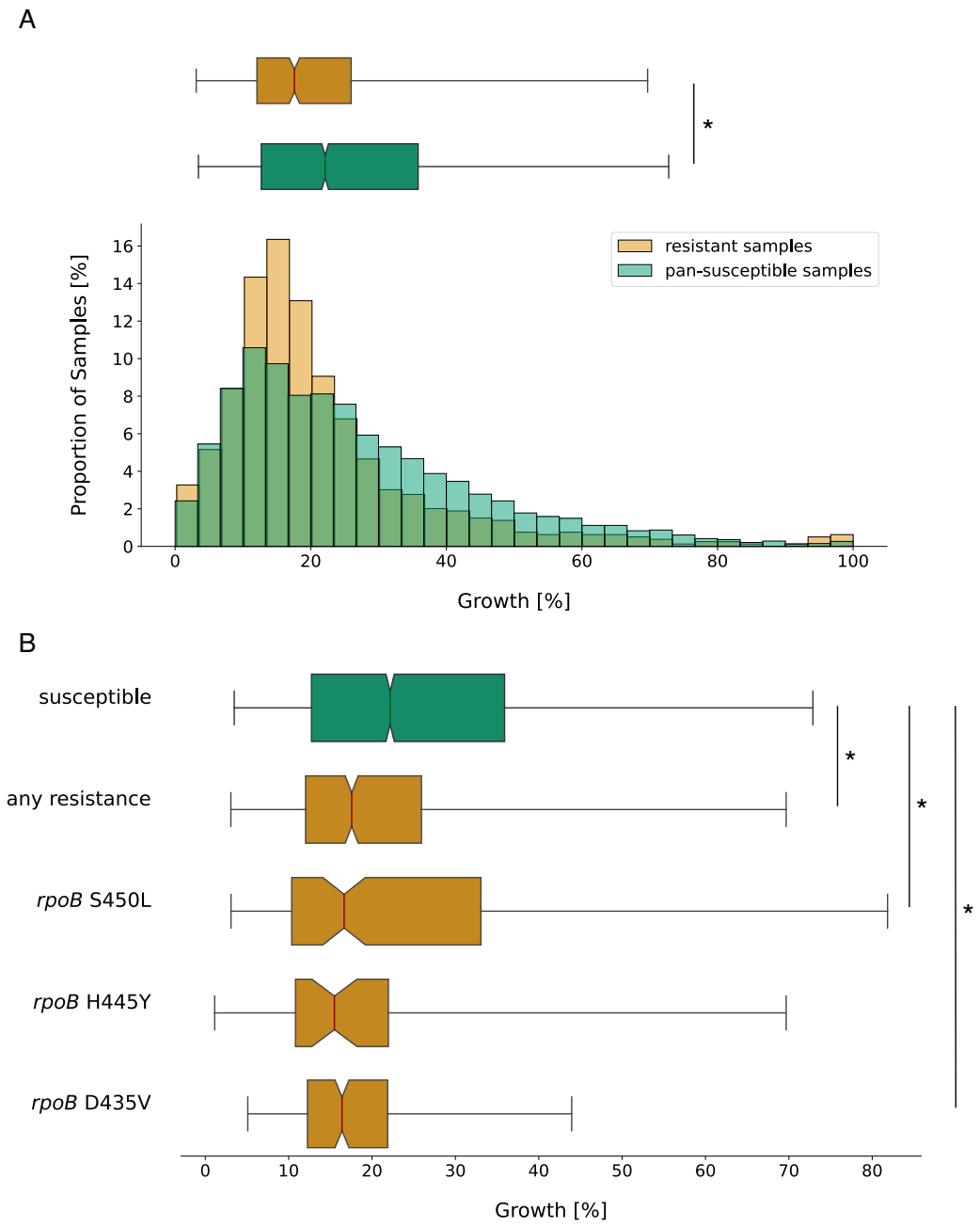


Figure 6: Presence of rifampicin resistance-conferring mutations in the RNA polymerase (RNAP) of *M. tuberculosis* is associated with lower median growth compared to pan-susceptible samples. (A) Distributions of growth in percent of covered well-area as measured in the CRyPTIC project[167] were plotted as a histogram against the proportion of samples that display this amount of growth (bottom) and as a notched box plot reflecting the distribution quantiles (top). Samples with rifampicin resistance mutations but no other potentially interfering mutations are plotted in red, samples that were classified as pan-susceptible are plotted in green. For the box plot, half of the data lies within the area of the box and 95% in the area covered by the whiskers. Outliers (5% of the data) were removed to achieve a cleaner representation. Indented areas close to the medians indicate their respective confidence intervals, while the star (*) indicates a significant Bonferroni-corrected Mann-Whitney p-value ($p < 0.05/10\%$). The respective medians, confidence intervals and the Mann-Whitney p-value are listed in Table 3. **(B)** Plot structure equivalent to the box plot in A, but the red bars represent subsets of rifampicin resistant samples that exhibit only the resistance mutation indicated to their left and no other potentially interfering mutations. The medians, their confidence intervals (CI) and Mann-Whitney p-values of the distributions are listed in Table 3. For a histogram representation of the data please refer to Supplementary Figure S1.

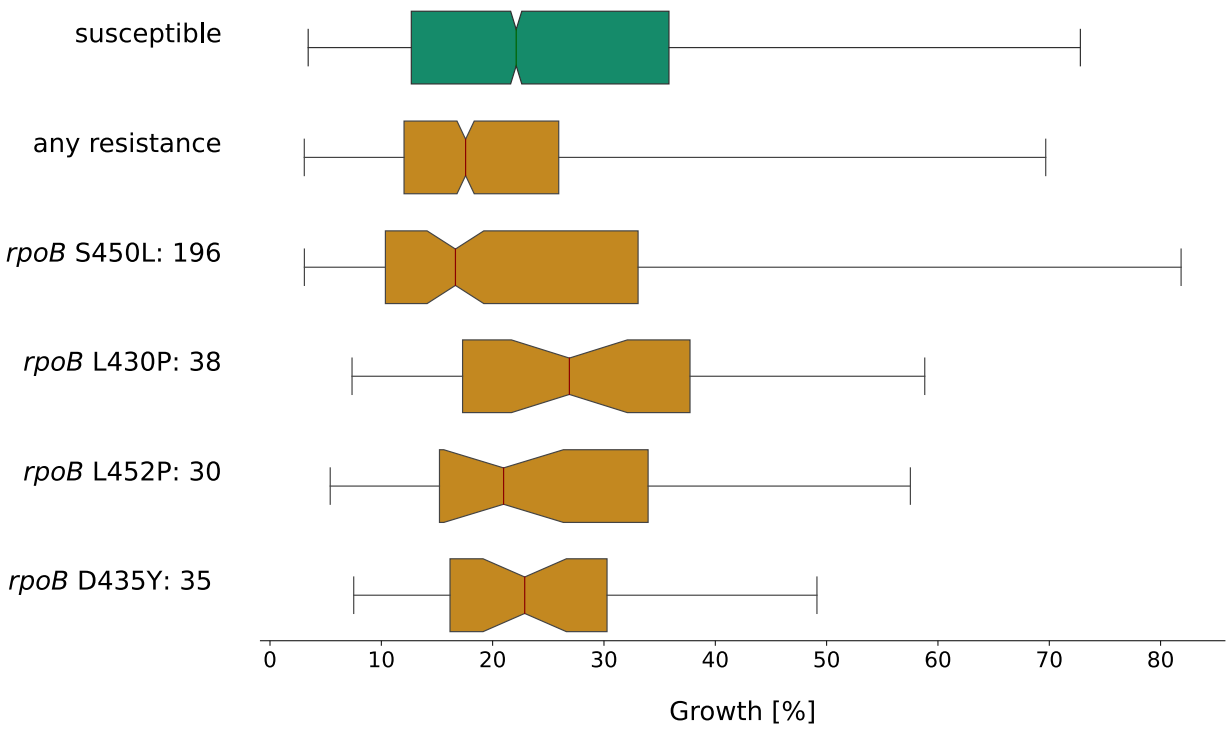


Figure 7: Growth of resistant samples with 'disputed' resistance calls is higher than in samples with *rpoB* S450L. Plot setup identical to Figure 6B, but here comparing the growth of samples with the 'disputed' resistance mutations *rpoB* L430P, *rpoB* L452P and *rpoB* D435Y to samples with the most common resistance mutation *rpoB* S450L. The numbers after the mutation name indicate the number of samples with this specific mutation in the dataset.

3.4.2 Constructing a list of high confidence compensatory mutations through resistance association and homoplasy

Given samples with resistance-conferring mutations grow less well compared to wild type, one expects to detect additional mutations in the population that have subsequently evolved in resistant samples to restore, at least partially, this apparent drop in fitness. To identify these mutations we performed a Fisher's exact test for each pair of resistance and co-occurring RNAP mutation in the Genetics dataset, to determine if the latter is significantly associated with resistance (Methods).

Despite being generally accepted as a conservative approach to correct for multiple testing, naively applying a Bonferroni-corrected p-value of 0.01% (Figure 5) to the results of the pairwise Fisher's exact test leads to hundreds of hits (Table in Appendix) and a TPR of 100% when compared to a compiled list of reference CMs (Table 1). This is most likely due to the inflating effect of linkage disequilibrium (LD) on p-values, as explained in the Methods section.

We hence decided to first apply a more conservative and arbitrary p-value cut-off at the 98% percentile to counter the p-value inflation, and later additionally excluded any hits that are not homoplasic. A compiled list of reference CMs was used to evaluate the performance (Table 1).

We were able to recover 94.1% (16/17, Table 1) of these putative high-confidence CMs using our arbitrary p-value cut-off set at the 98% interval, equivalent to a p-value cut-off at $p = 10^{-25.9}$. This results in a preliminary list of mutations that are significantly associated with rifampicin resistance. Synonymous mutations are unlikely to have a strong effect on the growth phenotype, as they do not change the structure of the final protein and thus they were removed from the list. The resulting list contained 78 putative CMs (Table 2). Of note, about 70% of hits were significantly associated with the resistance mutation S450L, which indicates a general strong association of this particular resistance mutation with the presence of CMs. It should be noted that the choice of the p-value cut-off was arbitrary, although informed by the true positive rate (Figure 5). Hence it is likely that there are other putative CMs below the threshold we set. But the sharp increase in the number of hits after the 98th percentile (Figure 5) would render it difficult to test all possible mutations for homoplasy using our approach of mapping them to the phylogenetic tree.

To test for homoplasy, we mapped these 78 putative CMs onto a phylogenetic tree of our *M. tuberculosis* samples (Methods). The majority of the most frequent putative CMs in our dataset indeed show homoplasy (Figure 8). But the putative CM *rpoC* E1092D for example exclusively occurred in a single clade of Lineage 2 and is thus highly likely to be a phylogenetic marker rather than a CM. We observed clustering in clades of Lineage 2 for other CMs (Figure 8: P1040R, I491V: blue trapezoid and V483A: red trapezoid), but in contrast to E1092D, these CMs also occurred in other, distant parts of the tree without a common ancestor. This makes convergent evolution of compensation more likely than homology.

After filtering out hits that do not exhibit homoplasy, we arrived at a final list of 51 putative CMs. Out of these 51 hits, 12 have to our knowledge not been previously described (Figure 9, Table 2). All these hits would be valid starting points for further investigations of compensatory effects and mechanisms. In this final list, the association of CMs with the specific resistance mutation *rpoC* S450L is even more apparent, with this mutation accounting for 92% of CM associations.

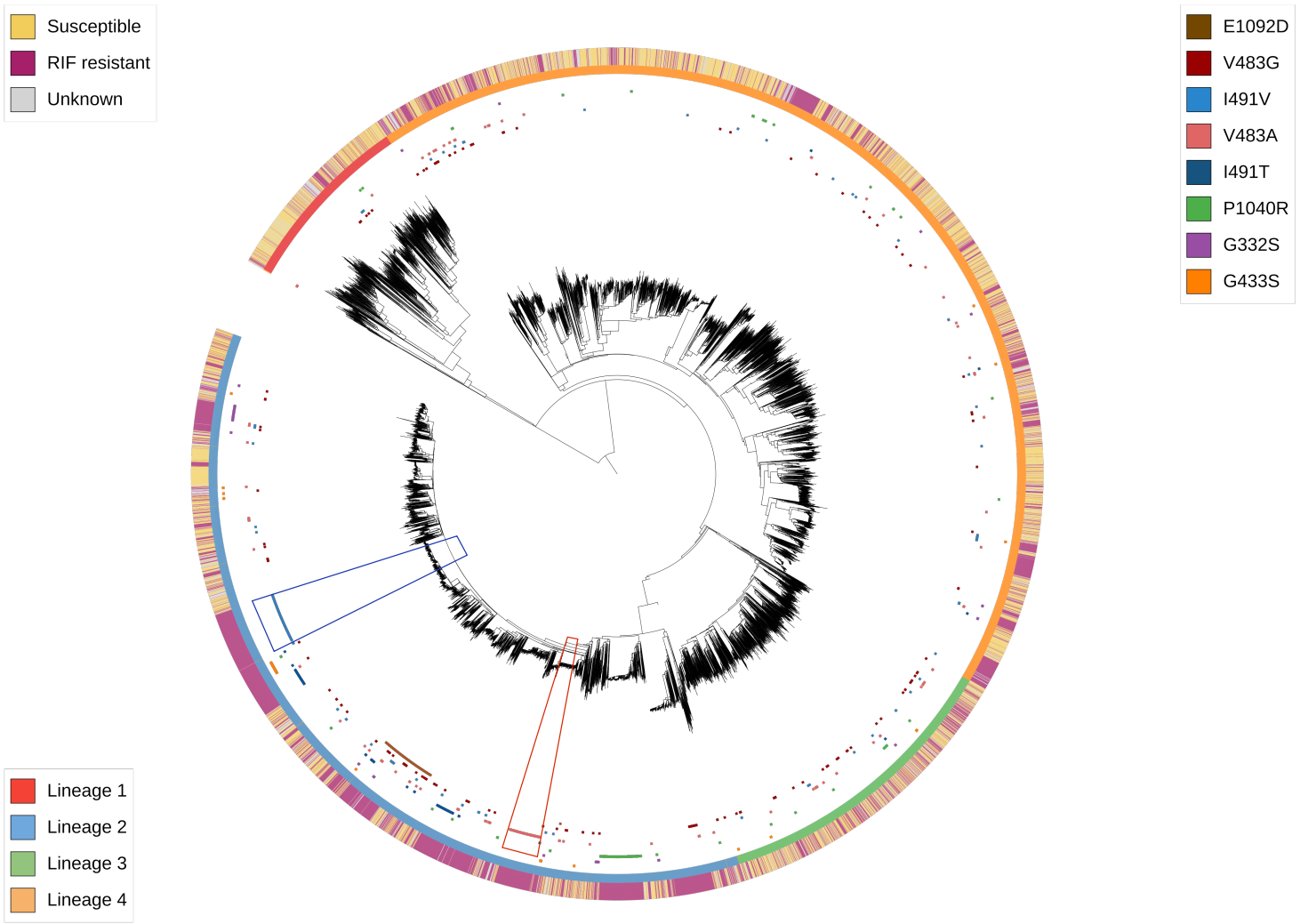


Figure 8: Putative compensatory mutations (CMs) are distributed widely across the phylogenetic tree. Phylogenetic tree assembled from single nuceotide polymorphism (SNP)-distances of about 15,000 samples. The resistance level of the samples is indicated on the outermost ring, while the second ring indicates the lineage. The most common CMs (all on *rpoC*) were mapped on the innermost ring, with the trapezoids indicating clades within Lineage 2 that show a cluster of a specific CM (I491V: blue trapezoid and V483A: red trapezoid).

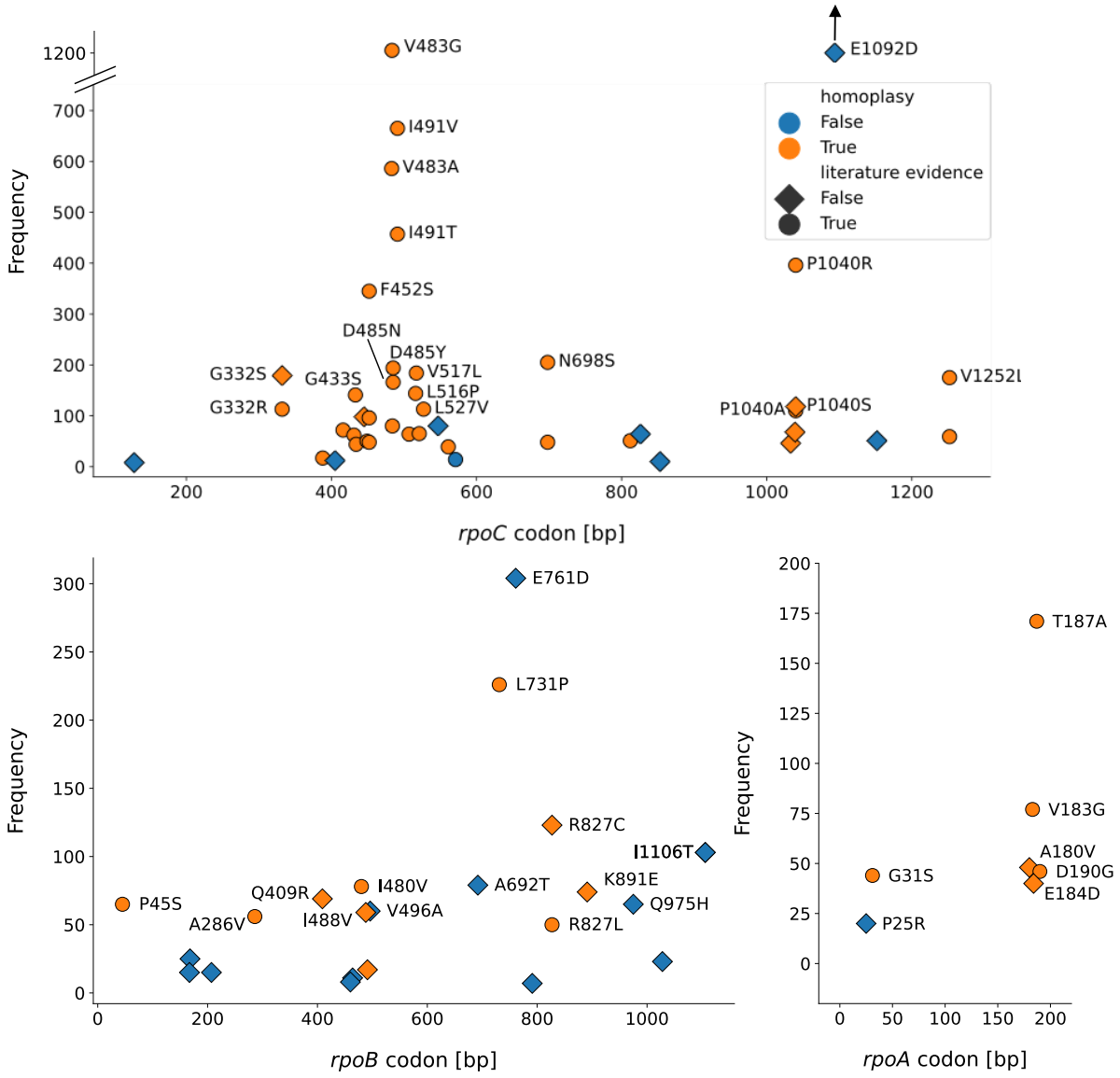


Figure 9: Putative compensatory mutations (CMs) are found in all RNA polymerase genes. The putative CMs were mapped according to their position in the respective gene and their frequency of co-occurrence with resistance mutations. There were five hits on the σ factor that are not shown, as well as one hit on the *rpoZ* gene (Table 2). All of these do not show homoplasy. E1092D is outside of the plotting range due to its high frequency (1989 observations).

3.4.3 Most compensatory mutations are found at interfaces between the RNA polymerase subunits

To evaluate their position on the protein structure, we mapped all 51 non-synonymous putative CMs that show homoplasy (Table 2) onto the RNAP in complex with rifampicin.[145] The hits clustered in four different regions of the RNAP (Figure 10A): i) at the interfaces between subunits (Figure 10B), ii) on the β and β' subunits close to the "rifampicin resistance determining region" (RRDR) and the active centre (Figure 10C), iii) around the secondary channel in the β' subunit (Figure 10D) and iv) at the DNA entry channel (Figure 10E). We found putative CMs in the β subunit of the RNAP, which had long been overlooked as a possible location for CMs, as most efforts in identifying CMs initially focused on the β' subunit. There were putative CMs in most subunits of the RNAP, however only the ones in the *rpoA*, *rpoB* and *rpoC* genes show homoplasy (Figure 9, Table 2) and most of them were found on the β' (*rpoC* gene) subunit, which supports the widespread hypothesis that it plays the major part in compensation.

The majority of CMs, including those most common in our dataset (Table 2: V483G/A and I491V/T), were located close to the interface of the β, β' and α RNAP subunits (Figure 10B). The amino acid changes in some of these locations suggest possible interactions between charged side chains, prompting a conformational change in the structure (Figure 11). Several hits also clustered close to the RRDR, where rifampicin binds to the RNAP of susceptible bacteria (Figure 10C). The RNAP secondary channel (Figure 10D) and the DNA entry channel (Figure 10E) are locations for CMs that, to our knowledge, have not been discussed in the literature. The latter is especially interesting, since we observed CMs very close to where the DNA strand enters the protein (Figure 10F).

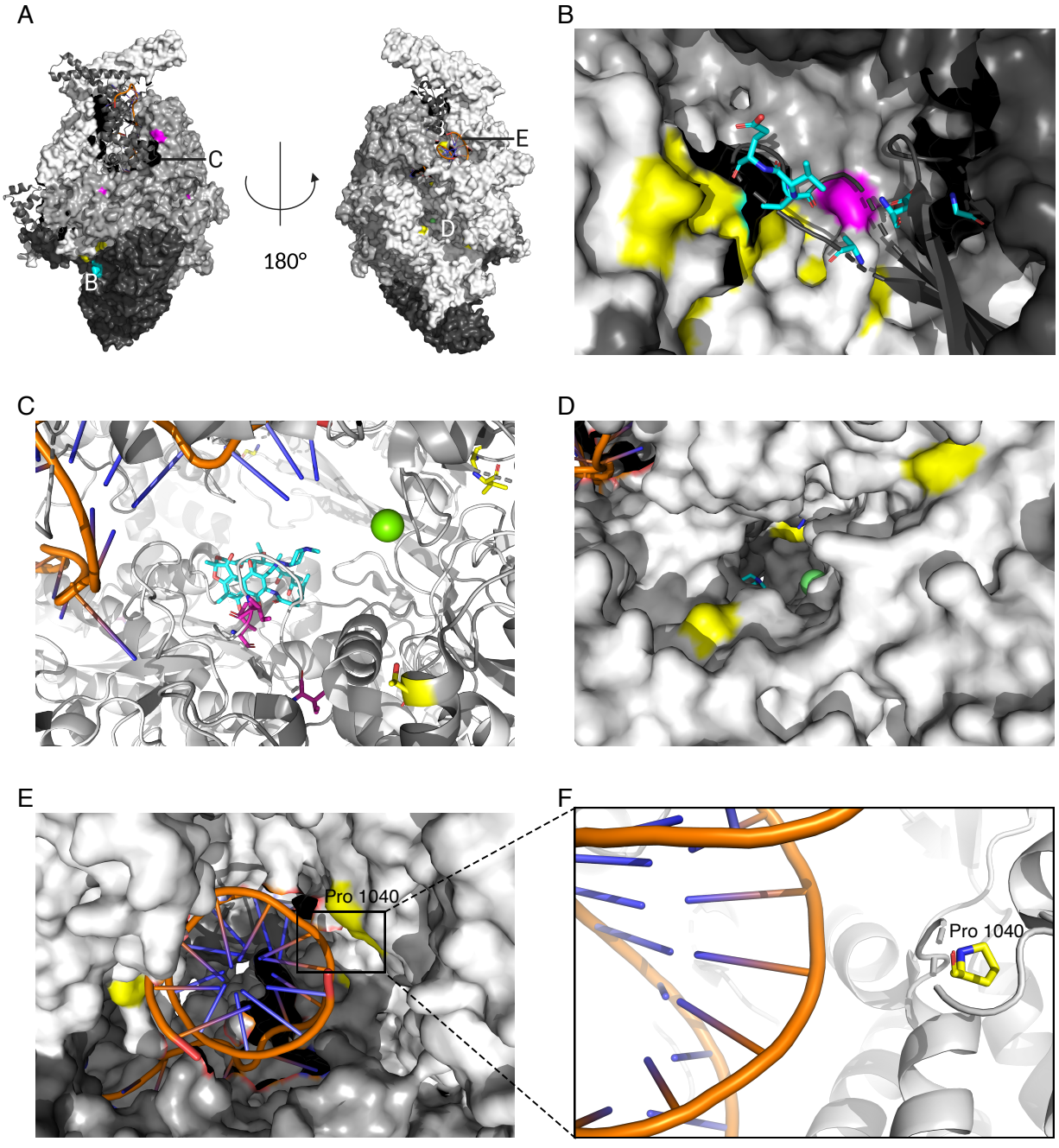

A
C
B
180°
E
D
B
C
D
E
Pro 1040
F
Pro 1040

Figure 10: Compensatory mutations (CMs) map to various subunits of the RNA polymerase. **(A)** Overview of clustering regions for CMs. Letters indicate where the CM clustering regions are located. **(B)** The interaction region of subunits α (black), β (dark grey) and β' (light grey). CMs can be found in all these subunits and are highlighted in color (β subunit: magenta, β' subunit: yellow, α subunit: light blue, stick representation). **(C)** CMs close to the 'rifampicin resistance determining region' (RRDR) on the β and β' subunits. Rifampicin (light blue) is shown bound to the RRDR. The DNA strand is visible on the top left in dark blue and orange stick representation. The active centre magnesium ion is shown in green. CMs are highlighted in colour (β subunit: magenta and β' subunit: yellow) and by stick representation. **(D)** CMs close to the RNAP secondary channel in the β' subunit (light grey) are shown in yellow. The location of the active site inside the protein can be deduced through the active site magnesium ion, indicated in green. **(E)** CMs close to the DNA entry channel are shown in yellow. The DNA helix is shown in dark blue and orange stick representation. **(F)** Close-up of the location of a putative CM (yellow stick representation, CM mutates Proline to Arginine) close to the DNA backbone. This CM might change interactions of the RNAP with the DNA strand.

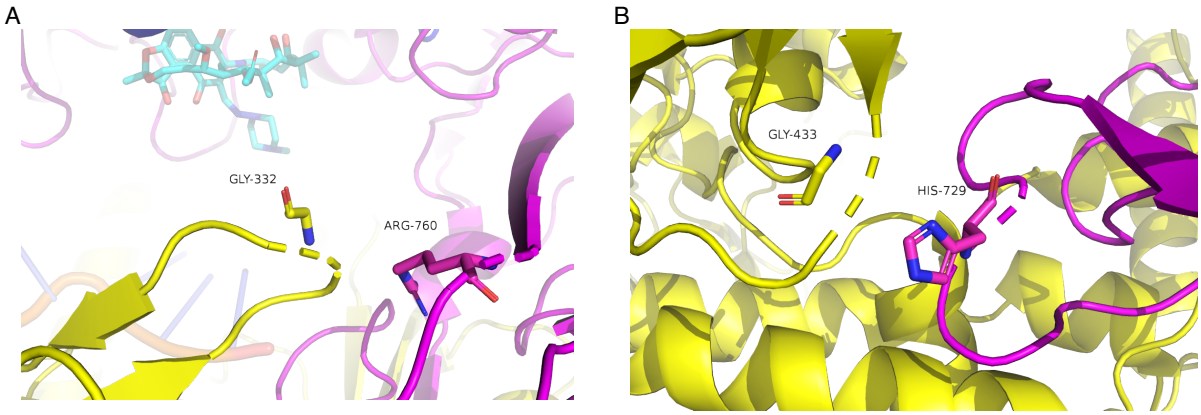


Figure 11: Two high-confidence compensatory mutations (CMs) on the RNA polymerase (RNAP) might change the subunit interaction through electrostatic interactions. **(A)** The CM G332S is located on the β' subunit, in a contact region to the β subunit (magenta). The change from Glycine (stick representation) to Serine (polar side chain) might enable an interaction with the close-by Arginine (positively charged side chain) on the β subunit. The bound drug rifampicin (light blue) can be seen in the background. **(B)** The CM G433S is located on the β' subunit, in a contact region to the β subunit. The change from Glycine (stick representation) to Serine might enable an interaction with the close-by Histidine (positively charged side chain) on the β subunit.

3.4.4 Compensatory mutations can identify rifampicin resistance by association with high specificity

A cause for false negative calls in phenotypic resistance prediction with WGS-AST can be low coverage or sequencing errors in genomic regions associated with resistance, hence it would be useful if there was redundancy when predicting resistance using genetic features. This can be especially relevant in rifampicin resistance, where resistance is very often caused by a single mutation, *rpoB* S450L.

The previously identified 51 CMs (Table2) are by definition only present in resistant samples. Catalogues of resistance mutations usually only contain mutations that are assumed to be causative of resistance. But there is often no explicit experimental evidence and the phenotypic data is inconclusive. Hence a high association with resistance, as is given for our list of CMs, could be a valid alternative to predicting resistance. By allowing samples to be predicted as rifampicin-resistant if they contain either a resistance-associated variant (RAV) and/or a CM one might even expect to reduce the false negative rate. To test this, we appended these 51 CMs to the WHO catalogue of RAVs for rifampicin.

First, we checked how many samples show compensation. We found that 4,289 out of our 35,538 samples contained at least one CM. Including CMs in addition to RAVs for resistance prediction reduced the number of false negative calls from 591 to 568 (Table 5), corresponding to an absolute decrease of 0.22% in false negative rate (FNR, Figure 12A). The resulting improvement in sensitivity is not significant, probably due to the small number of affected samples compared to the size of the dataset.

To evaluate how well CMs can identify rifampicin resistance on their own, we checked the specificity of using exclusively CMs to predict the resistance phenotype. As expected, only moderate sensitivity is achieved (40.5%, Figure 12B), since only a fraction of resistant samples exhibit compensation. But interestingly, the specificity of predicting resistance is significantly higher (99.7%) than when simply applying the WHOv2 catalogue of mutations (98.1%, Figure 12B). This suggests that while we could use CMs to predict rifampicin resistance in the absence of a RAV, we should not rely on them exclusively.

RAVs	CMs	TP	FP	TN	FN	SE	SP
✓		9819	488	24640	591	94.3%	98.1%
✓	✓	9842	490	24638	568	94.5%	98.0%
	✓	4221	73	25055	6189	40.5%	99.7%

Table 5: Contingency table values and performance metrics for different scenarios of the catalogue-based predictions for rifampicin resistance. Scenarios are shown for either using RAVs and/or CMs for prediction. "SE" shows sensitivity and "SP" the specificity of the resistance prediction.

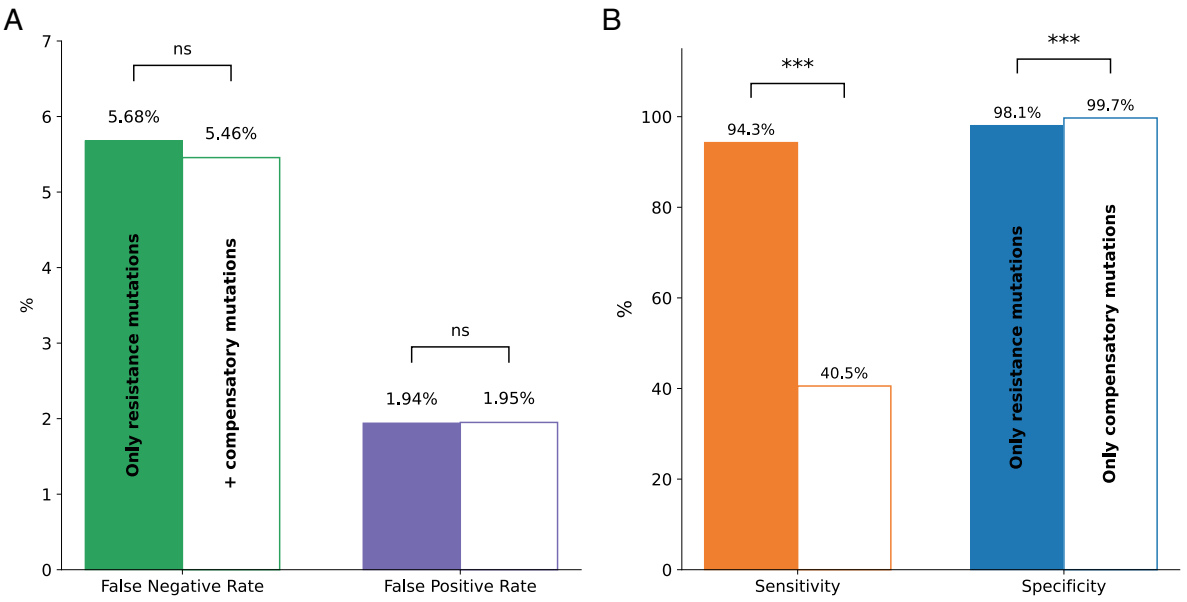


Figure 12: Using compensatory mutations (CMs) as resistance indicators lowers the false negative rate. (A) False negative rate (FNR) and false positive rate (FPR) of rifampicin resistance prediction based on either only resistance mutations, or resistance mutations and CMs. The effect of including CMs is clearly visible when looking at the FNR, but insignificant. The effect on the FPR is negligible. **(B)** Sensitivity and specificity of rifampicin resistance prediction based on either presence of resistance mutations, **or** presence of CMs in the sample. Sensitivity of the CM only prediction is low as expected, since only about 40% of our resistant samples show compensation, but the increase in specificity is significant too.

3.4.5 Compensatory mutations are associated with higher growth levels *in vitro*

If we have correctly identified CMs, growth densities in the resistant samples with CMs should be higher than in resistant samples without CMs. To test this we compared the growth distributions of pan-susceptible samples, rifampicin resistant samples that contain at least one homoplasic CM from our complete list (Table 2), and resistant samples without any CMs in our Growth dataset. Resistant samples with CMs grew to significantly higher densities (7.9% absolute growth difference, p = 3.92e-57) than resistant samples without CMs (Figure 13A, Table 6). Surprisingly, the median growth of resistant samples with CMs also significantly out-performed the growth of pan-susceptible samples (4.5% absolute growth difference, p = 6.25e-26, Figure 13A, Table 6). The association of CMs with such high growth levels could mean that CMs increase fitness above wild-type levels, or hint at confounding factors.

One confounding factor could be rooted in the differences in virulence between *M. tuberculosis* lineages. Lineage 2 is thought to be associated with higher transmission rates than other lineages, [182] which could be reflected in higher *in vitro* growth densities. Indeed, considering only pan-susceptible samples, Lineage 2 showed significantly higher growth than Lineage 4 (5% absolute growth difference, p = 7.99e-33) and higher growth than Lineage 1 (3% absolute growth difference, p = 7.17e-02), whilst still being out-performed by Lineage 3 (6% absolute growth difference, p = 8.36e-16, Figure 13B, Table 7). In addition, 60% of resistant samples in Lineage 2 contained at least one CM from our final list, compared to 34% in other lineages. Higher growth densities of Lineage 2 combined with CM accumulation could therefore lead to inflated growth phenotype association with CMs.

We hence need to examine lineage contribution to the enhanced growth phenotype in compensated samples. In Lineage 2 we still found a significantly higher growth density for samples with CMs than for pan-susceptible samples (5.3% absolute growth difference, p = 3.23e-02, Figure 14). For all other lineages, the absolute growth difference between compensated resistant samples and pan-susceptible samples was either insignificant, or pan-susceptible samples grew to higher densities (Figure 14, Table 8).

In Lineage 3 we observed a slightly higher growth density of resistant samples with CMs when compared to resistant samples without CMs (1.9% absolute growth difference, p = 2.08e-02), but this was insignif-

sample type	median growth [%]	CI low	CI high	p-value$_r$	p-value$_s$	n
pan-susceptible	22.1	21.6	22.7			5283
resistant and no CMs	18.7	18.0	19.3		3.92e-15	2869
resistant and CMs	26.6	25.5	27.4	3.92e-57	6.25e-26	2667

Table 6: Median growth of resistant samples with compensatory mutations compared to pan-susceptible samples and samples with only resistance mutations. The confidence interval (CI) for the median is calculated using bootstrapping where 'CI low' indicates the lower threshold and 'CI high' the upper threshold. P-values are given with respect to resistant (p-value$_r$) and pan-susceptible sample growth (p-value$_s$) and n indicates the sample size.

Lineage	median growth [%]	CI low	CI high	p-value$_1$	p-value$_2$	p-value$_3$	n
Lineage 1	23.1	20.6	25.5				534
Lineage 2	26.1	25.3	27.0	7.17e-02			1331
Lineage 3	32.1	29.6	34.8	8.68e-14	8.36e-16		706
Lineage 4	18.1	17.6	18.6	1.22e-05	7.99e-33	3.31e-64	2656

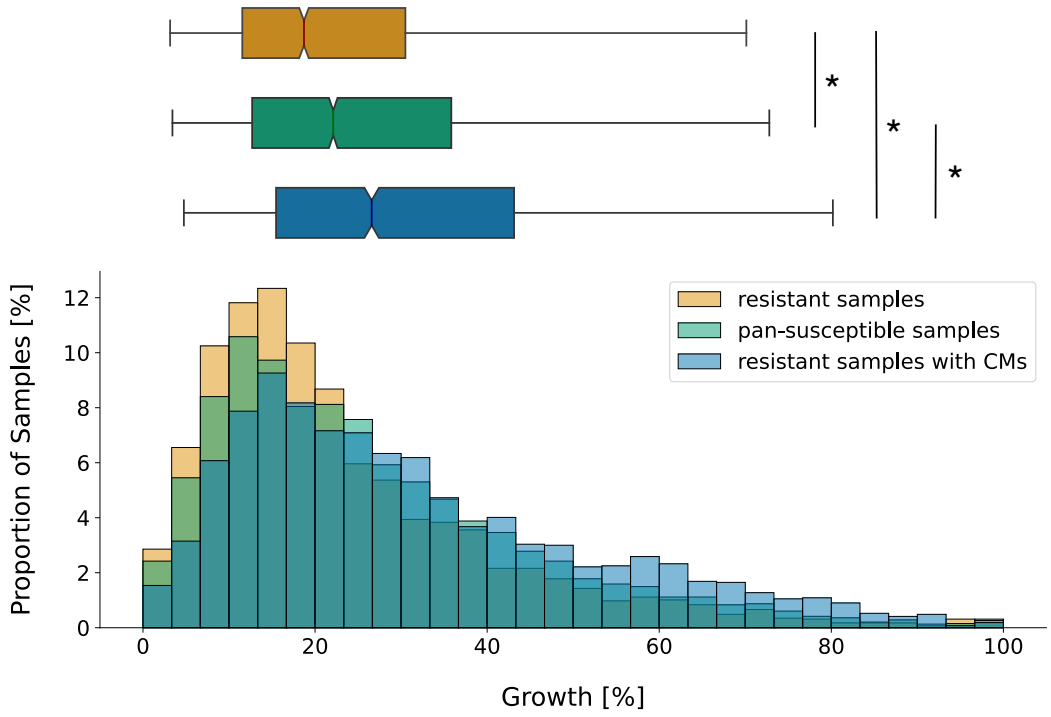
Table 7: Median growth of pan-susceptible samples from different *M. tuberculosis* lineages. The confidence interval (CI) for the median is calculated using bootstrapping where 'CI low' indicates the lower threshold and 'CI high' the upper threshold. P-values are given with respect to each lineage, indicated by the subscript x (p-value$_x$) and n indicates the sample size.

icant after applying Bonferroni correction, probably due to low sample size. This indicates that while CMs partially restore fitness, the high growth phenotypes are not caused by CMs alone. There might for instance be other growth-associated mutations in Lineage 2, acting as confounding factors. We do not observe an effect on growth in Lineages 1 and 4. In Lineage 1, the sample size might be too low to detect an effect, since we could not even observe the initial fitness cost in resistant samples without CMs (Figure 14, Table 8). Lineage 4 on the other hand showed very low overall growth (Figure 13B) which makes it difficult to detect the positive influence of CMs due to the relatively small initial effect size of the fitness cost as measured by *in vitro* growth (Figure 14).

sample type	median growth [%]	CI low	CI high	p-value$_r$	p-value$_s$	n
Lineage 1:						
pan-susceptible	23.1	20.6	25.3			534
resistant and no CMs	22.9	19.6	28.7		0.784	126
resistant and CMs	23.3	19.6	25.2	0.516	0.596	58
Lineage 2:						
pan-susceptible	26.1	25.4	27.2			1331
resistant and no CMs	21.6	20.4	22.5		2.46e-05	1103
resistant and CMs	31.4	30.4	32.5	5.65e-43	3.20e-24	1788
Lineage 3:						
pan-susceptible	32.1	29.6	34.9			706
resistant and no CMs	24.4	22.4	27.2		8.79e-09	370
resistant and CMs	26.3	23.7	33.2	2.07e-02	4.38e-02	190
Lineage 4:						
pan-susceptible	18.1	17.6	18.6			2656
resistant and no CMs	14.4	13.7	15.2		6.17e-17	1252
resistant and CMs	14.6	13.7	15.4	0.995	3.44e-11	626

Table 8: Lineage-wise median growth of samples with different compensatory mutations compared to pan-susceptibles and samples with only resistance. The confidence interval (CI) for the median is calculated using bootstrapping where 'CI low' indicates the lower threshold and 'CI high' the upper threshold. P-values are given with respect to resistant (p-value$_r$) and pan-susceptible sample growth (p-value$_s$) and n indicates the sample size.

A


*
*
*
resistant samples
pan-susceptible samples
resistant samples with CMs
Proportion of Samples [%]
0
2
4
6
8
10
12
0
20
40
60
80
100
Growth [%]

B

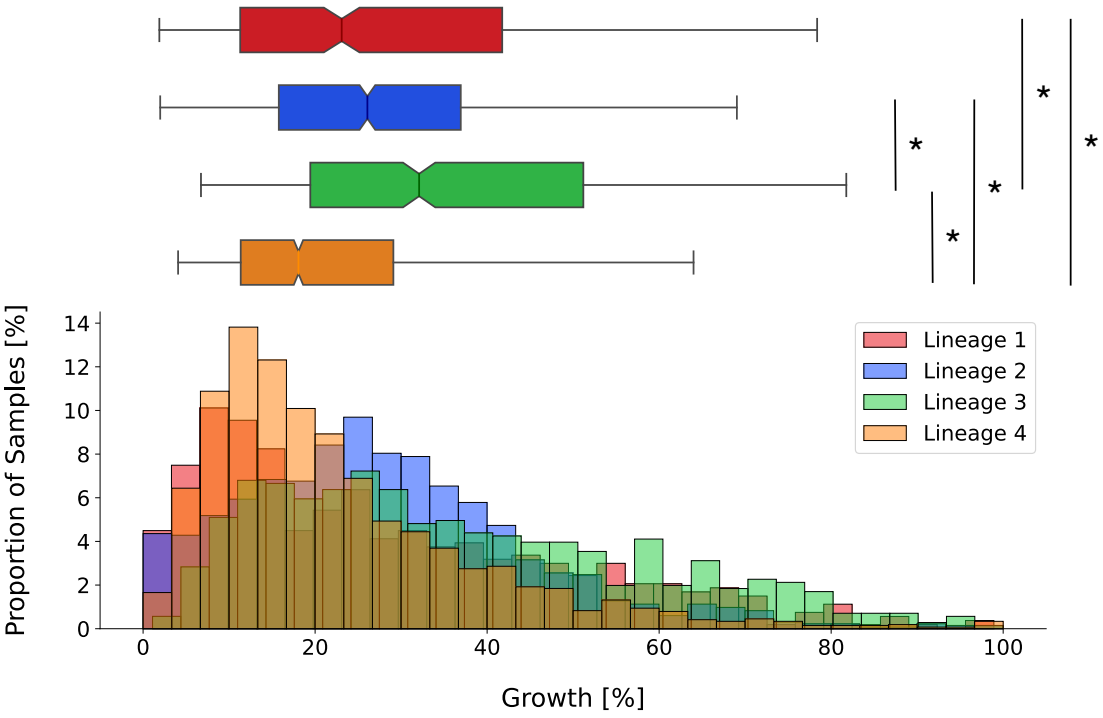

*
*
*
*
*
Lineage 1
Lineage 2
Lineage 3
Lineage 4
Proportion of Samples [%]
0
2
4
6
8
10
12
14
0
20
40
60
80
100
Growth [%]

Figure 13: Presence of compensatory mutations (CMs) in samples with rifampicin resistance-conferring mutations in the RNA polymerase of *M. tuberculosis* is associated with higher growth densities. (A) Distributions of growth in percent of covered well-area as measured in the CRyPTIC project[167] were plotted as a histogram against the proportion of samples that display this amount of growth (bottom) and as a notched box plot reflecting the distribution quantiles (top). Samples with rifampicin resistance mutations but no putative CMs are plotted in red, samples that were classified as pan-susceptible are plotted in green. Samples that have rifampicin resistance mutations and at least one CM are shown in blue. For the box plot, half of the data lies within the area of the box and 95% in the area covered by the whiskers. Outliers (5% of the data) were removed to achieve a cleaner representation. Indented area close to the medians indicate their respective confidence interval, while the star (*) indicates a significant Bonferroni-corrected Mann-Whitney p-value ($p < 0.05/3\%$). The respective medians, confidence intervals and the Mann-Whitney p-values are listed in Table 6.**(B)** Plot structure equivalent to the box plot in A, but the bars represent pan-susceptible samples from different lineages, plotted in the respective colours indicated in the legend. The star (*) indicates a significant Bonferroni-corrected Mann-Whitney p-value ($p < 0.05/6\%$). The respective medians, confidence intervals and the Mann-Whitney p-value are listed in Table 7.

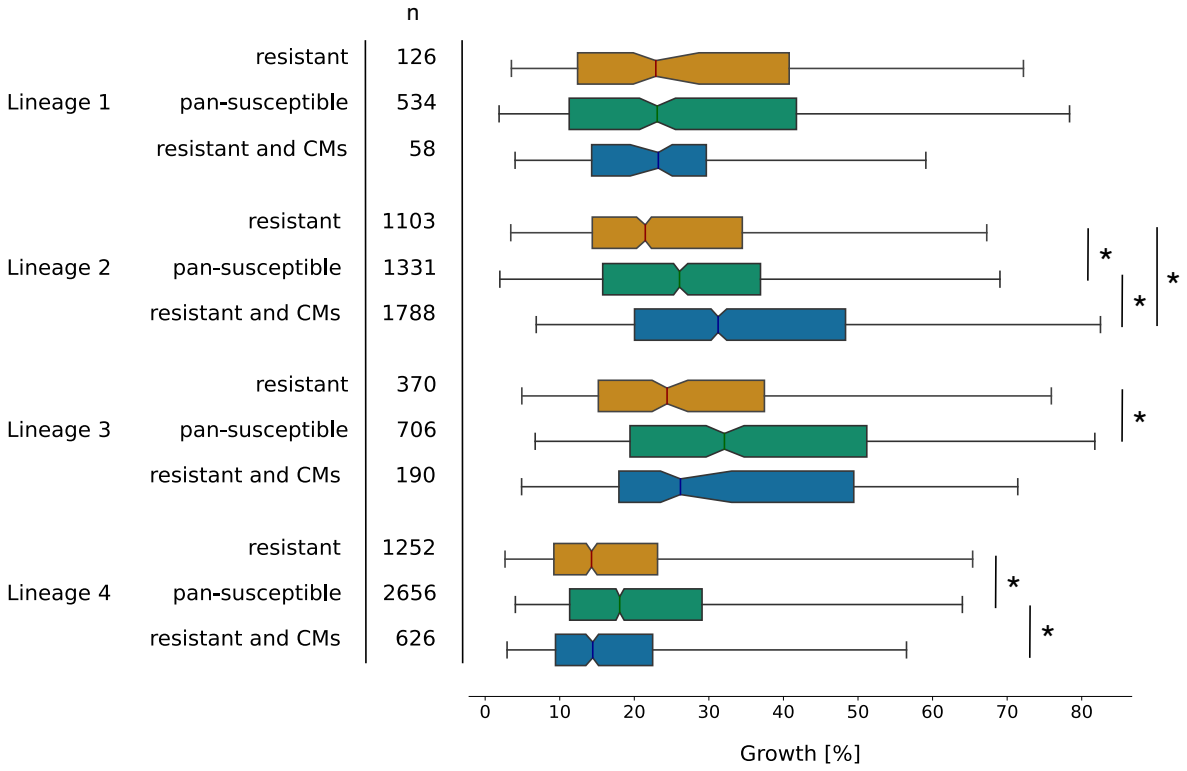


Figure 14: Presence of compensatory mutations (CMs) in samples with rifampicin resistance-conferring mutations in the RNA polymerase of *M. tuberculosis* is associated with higher growth densities in some lineages. Plot structure equivalent to the box plot in Figure 13, but the box plots represent subsets of samples that belong to the lineage displayed on the left. The sample size is shown in the column marked with 'n'. For a histogram representation of the same data refer to Supplementary Figure S2. The star (*) indicates a significant Bonferroni-corrected Mann-Whitney p-value ($p < 0.05/3\%$). The respective medians, confidence intervals and the Mann-Whitney p-values are listed in Table 8.

3.4.6 The effect of compensatory mutations on *in vitro* growth is confounded with lineage and clade affiliation

Lineage 2 is the only lineage that showed significantly higher growth in resistant samples with CMs than in pan-susceptible samples and also accumulates CMs more than any other lineage in our Growth dataset. While this might indicate that CMs play an important role in this Lineage, we suspected that there may be other mutations co-occurring with CMs that enhanced growth densities to the reported high levels.

We hence wanted to investigate if the high growth densities observed in samples with CMs within Lineage 2 reflect the growth advantage of a few high-fitness clades. We aimed to dissect the clade influence for two representative CM clusters with sufficient available samples (Figure 8), specifically setting a minimum cut-off at n = 50 for each sample group to obtain meaningful statistics. The first cluster showed accumulation of the CM *rpoC* I491V (blue trapezoid in Figure 8). The average SNP-distance of 23 within the cluster indicates that this cluster is not the result of recent transmission according to UKHSA guidelines,[183] although the maximum SNP-distance of 44 highlights that samples are nonetheless very closely related. Most likely the cluster is indeed an established clade within Lineage 2 and not an over-sampled local outbreak. We saw that the clade with the *rpoC* I491V cluster showed significantly higher growth than pan-susceptible samples (11.1% absolute growth difference, p = 6.52e-24), but the effect outside of this clade was just as strong (11.5% absolute growth difference, p = 1.12e-07, Figure 15A, Table 9).

The second cluster showed accumulation of the CM *rpoC* V483A (red trapezoid in Figure 8). The average SNP-distance was higher than in the other cluster, with an average of 31 pairwise SNP differences between samples in the cluster and a maximum SNP-distance of 149. We saw that median growth in the clade with the *rpoC* V483A cluster was significantly higher than in samples with this CM outside of the clade (5.4% absolute growth difference, p = 6.79e-03). But median growth of the latter was still significantly higher than the median growth of resistant samples without any CMs (8.1% absolute growth difference, p = 4.24e-03, Figure 15B, Table 9), which suggests that a small positive effect of the CM on growth is conserved after removing the confounding influence of the high-fitness clade. However, the increase of growth over pan-susceptible growth levels is not significant outside of the high-fitness clade

for the CM V483A (3.6% absolute growth difference, p = 0.263, Figure 15B, Table 9), which implies that the clade association is a confounding factor in this case.

The association of CM clusters with specific high-fitness clades within Lineage 2 hence could be a confounding factor in our growth data analysis and might explain a part of the observed increase of growth densities to levels higher than the wild-type growth densities. Further work is needed to completely disentangle the influence of these high-fitness clades on the growth phenotype. However, the small growth advantage of resistant sample with CMs over resistant sample without CMs appears to be conserved even when these clades are removed from the analysis. This small growth advantage compared to resistant samples without CMs is observed similarly in Lineage 3 (Figure 14). This would agree with the hypothesis of CMs leading to a partial restoration of fitness, while other secondary mutations increase growth to levels surpassing even susceptible growth densities.

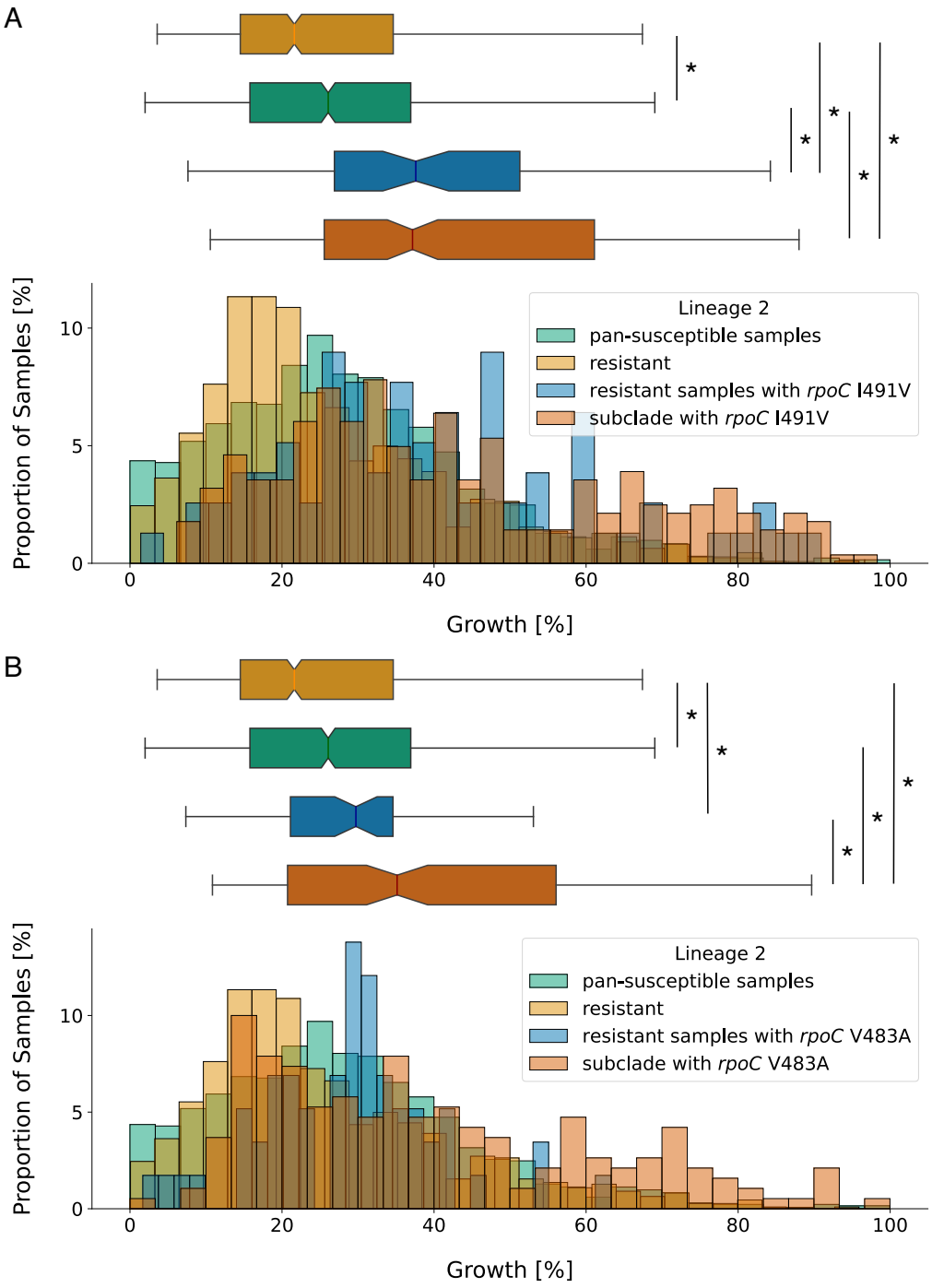


Figure 15: *M. tuberculosis* clades with clusters of compensatory mutations (CMs) explain some of the high growth densities associated with CMs in Lineage 2. (A) Growth distributions (percentage of covered well area) in Lineage 2 were plotted as a histogram against the proportion of samples that display this amount of growth (bottom) and as a notched box plot reflecting the distribution quantiles (top). Lineage 2 samples were classified as pan-susceptible (green), rifampicin resistant (red), resistant and showing the CM I491V outside of the CM cluster clade (blue) and as part of the Lineage 2 clade where all samples show the CM I491V (dark red). For the box plot, half of the data lies within the area of the box and 95% in the area covered by the whiskers. Outliers (5% of the data) were removed to achieve a cleaner representation. Indented area close to the medians indicate their respective confidence interval, while the star (*) indicates a significant Bonferroni-corrected Mann-Whitney p-value ($p < 0.05/6\%$). The respective medians, confidence intervals and the Mann-Whitney p-values are listed in Table 9. **(B)** Plot structure equivalent to the box plot in A, but the CM in question is V483A. Distribution medians and their confidence intervals are shown in Table 9.

sample type	growth [%]	CI low	CI high	p-value$_r$	p-value$_s$	p-value$_{rCM}$	n
pan-susceptible	26.08	25.21	26.99				1331
resistant and no CMs	21.6	20.5	22.4		2.46e-05		1103
CM *rpoC* I491V:							
resistant	37.6	31.5	42.3	9.57e-11	1.12e-07		78
resistant and subclade	37.2	33.9	41.4	2.14e-33	6.52e-24	0.499	282
CM *rpoC* V483A:							
resistant	29.7	26.9	31.7	4.24e-03	0.263		58
resistant and subclade	35.1	30.7	38.9	9.45e-17	9.89e-11	6.79e-03	190

Table 9: Median growth of samples from Lineage 2 with different compensatory mutations (CMs), compared to growth of pan-susceptible samples, growth of resistant samples without CMs and growth of a resistant clade that shows a cluster of the respective CM. Median growth is represented by the median percentage coverage of a well containing bacterial growth as measured by the CRyPTIC project. The confidence interval (CI) for the median is calculated using bootstrapping where 'CI low' indicates the lower threshold and 'CI high' the upper threshold. P-values are given with respect to resistant sample growth (p-value$_r$), pan-susceptible sample growth (p-value$_s$) and sample growth of resistant samples with CMs outside of the cluster subclade (p-value$_{rCM}$). n indicates the sample size.

3.5 Discussion

In this Chapter, we have investigated the emergence of CMs associated with rifampicin resistance in *M. tuberculosis*. We demonstrate how these mutations influence bacterial growth *in vitro* and discuss their potential relevance for improving resistance prediction in WGS-AST.

3.5.1 Main conclusions

A key factor in the spread of rifampicin resistance are CMs, named due to their potential to compensate for the fitness cost that resistance mutations introduce in *M. tuberculosis*. In this Chapter, we have identified a comprehensive set of high-confidence CMs in *M. tuberculosis* using our Genetics dataset of 77,860 samples and also captured and dissected their influence on *in vitro* growth of the bacteria using our Growth dataset comprising 13,990 samples. Overall, we identified 78 unique hits, of which 51 exhibit homoplasy (Supplementary Table 2). This set includes 94.1% of the compiled reference CMs (Table 1) and 12 hits that are, to our knowledge, novel. We used this list of high-confidence putative CMs to lower the false negative rate for rifampicin resistance prediction, which resulted in an absolute decrease of 0.22% in the false negative rate (FNR, 5.68% to 5.46%). Additionally, through mapping these high-confidence putative CMs on the RNAP structure, we identified four CM clustering regions (Figure 10A). Two of them have, to our knowledge, not been described before: the DNA entry tunnel and the RNAP secondary channel.

Gagneux *et al.* have first shown the fitness cost of rifampicin resistance using competitive fitness assays.[150] We found that the decrease in fitness upon acquisition of resistance is reflected in the *in vitro* growth distributions of our samples (Figure 6A). Concerning the impact of CMs on growth, our findings indicate that while CMs recover *in vitro* growth to a certain extent, they are probably associated with and/or may act synergistically with other growth-enhancing factors, which can be clade-specific, that boost growth yet further. The high growth densities in Lineage 2, for example, appear to be at least partially caused by CM clusters associated with clades with increased growth phenotypes, such as for the CM *rpoC* V483A (Figure 15B). Samples with this CM showed a high-growth cluster in Lineage 2, which reached growth levels above those of susceptible samples. While resistant samples with this CM outside

of the high growth clade still grew to higher densities than uncompensated resistant samples, they did not grow to significantly higher densities than susceptible samples. This indicates that the high-growth clade is confounding the growth data. For other CMs, these high-growth clusters do not appear to be the only explanation for the high growth levels, e.g. for the CM *rpoC* I491V (Figure 15A). Hence there may be additional factors associated with CM presence that enhance growth in Lineage 2. This growth data analysis also made it clear that such investigations will have to be conducted lineage- or even sublineage-wise, to account for population structure.

CMs are likely to exert their effect on fitness in different ways, depending on their location in the protein structure. The different clustering regions we observed might hence represent distinct mechanisms of action. The secondary channel for instance serves as a direct connection from the outside of the protein to the active centre and is presumed to facilitate the diffusion of substrate nucleotides into the protein for incorporation into the nascent mRNA. It has been proposed that molecules entering through this channel could regulate RNAP activity as well.[184] CMs at this location could therefore modify diffusion in and out of the RNAP. This might play a role in fitness regeneration as well as in establishing resistance to rifampicin, which might well enter the RNAP through the secondary channel.

The other novel location with a CM cluster was the DNA entry channel, and mutated residues here could alter interactions with the DNA helix. As an illustrative example, the mutation P1040R on the β' subunit changes a neutral, bulky Proline side-chain to a positively charged, elongated Arginine and is located very close to the DNA backbone in the crystal structure (Figure 10F). This could cause the Arginine to flip out of the averted position and interact with the negatively charged DNA backbone, thereby changing the way the DNA helix is positioned upon entering the protein.

However, most of the conformational changes caused by CMs affect the interface between the subunits and the space around the active site. CMs in the interfacial region might alter binding of the subunits, perhaps leading to higher protein stability, without necessarily affecting the active site.[161] Mutations close to the RRDR have been suspected to alter the conformation of the active site, possibly increasing enzyme activity.[160] Most CMs are hence likely to enhance transcriptional activity of the RNAP either

through changing the conformation of the active site or through increasing the overall stability of the protein complex. The increased activity could, in turn, lead to increased transmissibility.

In this context, our *in vitro* results for increased growth in presence of CMs complement recent evidence showing that *in vivo* fitness might be enhanced by compensatory evolution.[166] Specifically, our results support a positive influence of CMs on growth and, consequently, on fitness in samples with CMs in Lineages 2 and 3. We failed to show a similar effect in Lineages 1 and 4, which is a similar outcome to a study that looked at transmission fitness. They reported a higher relative transmission fitness for L2 MDR strains showing the resistance mutation S450L with compensation than without, but could not show this for any strains outside of Lineage 2.[164] While this is an ambiguous observation, it could be due to the strain background having a strong epistatic effect on the transmissibility of *M. tuberculosis* infections. The results from this study also indicate that the detected increased *in vitro* growth could in fact directly translate into higher transmission rates at population level. This is especially relevant for Lineage 2, since it currently out-competes other lineages at population level, presumably due to higher transmission success.[185]

Allowing the presence of CMs in the RNA polymerase to identify rifampicin resistance by association seems a trivial way to boost sensitivity in WGS-AST, provided the list of CMs is of high confidence. The improvement in sensitivity on this dataset was not significant, probably due to the moderate read depth ensuring there were relatively few resistant samples with a CM where the resistance mutations could not be resolved but the CMs could. Still, the inclusion of CMs lowered the FNR. In scenarios with much lower read depth (for example in multiplexed samples), allowing CMs to predict rifampicin resistance could provide a necessary boost to the sensitivity and FNR. The validity of using CMs as additional indicators of rifampicin resistance is supported by the fact that the specificity of resistance calls made based solely on the presence of CMs is very high at 99.7%.

Lastly, their strong association with high-fitness clades makes CMs interesting candidates for further studies. There are two possible scenarios, firstly it is possible that CMs may have been preceded by evolutionary older secondary growth-associated mutations that push growth above wild-type levels. This

is supported by the fact that the high-growth CM clusters were located in sublineages 2.2 and 2.2.7, respectively, which are mostly part of the modern Beijing sublineages. These lineages are known to exhibit increased virulence.[186] Secondly, it is possible that by partially restoring fitness after resistance acquisition, CMs could have facilitated the emergence of further mutations and hence the establishment of these high-fitness clades. This would make CMs a key determinant in the increased transmissibility observed in Lineage 2. To uncover which of these scenarios is more likely, one would have to identify the underlying genetic cause of the higher fitness in the affected sublineages of Lineage 2 and determine if this pre-dates the emergence of resistance and CMs.

In both scenarios, the notion that CMs play an important role in the spread of the increasingly virulent Lineage 2 is supported by the fact that this lineage accumulated CMs more than any other lineage, presumably also due to the reported higher mutation rates.[187] It has been suggested that such compensated multi-drug-resistant (MDR) mutants might be selected for in environments where many patients are treated with antituberculosis drugs.[186] This emphasises the need for appropriate antibiotic stewardship to prevent further spread of MDR *M. tuberculosis*.

3.5.2 Limitations

The inclusion of CMs for rifampicin resistance prediction has a negligible effect on performance when compared to using resistance mutations only. We used a very conservative list of CMs, which may have additionally lowered sensitivity. The arbitrarily low p-value cut-off at the 98% percentile could easily be reduced to a more permissive threshold, potentially detecting additional CMs. While this potential improvement of the FNR comes at a cost for the false positive rate, the trade-off could easily be minimised by assessing different p-value cut-offs. Importantly, we did manage to lower the FNR with our current CM list, which is here equivalent to the very major error rate. The very major error refers to resistant samples wrongfully classified as susceptible. This is the most detrimental type of error to make in resistance prediction, as it will lead to treatment failure and resistance spread. Keeping the FNR low is hence of highest priority.

The nature of our growth data is a general limitation of our approach to capturing the fitness effects of

CMs. The photographs taken by the Vizion instrument had moderate resolution and occasionally were affected by shadows and other artifacts, such as condensation.[167] In addition, due to uncertainty in algorithmically locating the wells, growth is only measured in the centre of each well. Taken together, the measured percentage growth is therefore somewhat noisy and should only be compared between large number of samples, as here. Also, growth was measured after two weeks of incubation and therefore represents the stationary phase, denying us any information on fitness advantages visible during the exponential growth phase, such as an increased growth rate. This could be changed by including growth monitoring during the exponential growth phase of the bacteria. Since this is difficult for a slow-growing species like *M. tuberculosis*, polymerase activity assays could potentially close this gap. By testing laboratory-derived *M. tuberculosis* mutants with an engineered combination of resistance mutations and CMs, the direct influence of CMs on RNAP activity could be tested. Even so, as the growth data was acquired *in vitro*, any fitness advantages that arise from the interaction of *M. tuberculosis* bacteria with their host environment will not be captured. We hence cannot necessarily directly translate the results of this study to the epidemiological reality of *M. tuberculosis* spread in human populations. As this was outside the scope of this study, we also did not account for possible associations between mutations outside of the RNA polymerase (RNAP) and rifampicin resistance. One could however adapt our statistical association testing method for other genes.

On a similar note, it would be possible to model the growth data using a regression model, which would allow including known confounding factors as variables to explain the observed phenotypes. This could help disentangle the lineage and clade associations of our high-growth phenotypes. However, explicitly modelling the growth data could be difficult given the high noise levels, exemplified by the divergence between positive controls from the same sample (Figure 4).

Using a multi-variate model could also solve the problem of other resistance alleles confounding our growth analysis, such as isoniazid resistance which also comes with a fitness cost and is highly associated with rifampicin resistance.[??] Disentangling this association is really difficult, since the sample size for rifampicin mono-resistance is very small, reducing the dataset from 795 samples to only 94 samples. The Mann-Whitney p-value of the growth deficit in rifampicin mono-resistant samples compared to pan-

susceptible samples with this smaller sample size is however still just about significant at p=0.049.

Another limitation is the strong linkage disequilibrium (LD) that is ubiquitous for mutations in largely clonal species like *M. tuberculosis*.[176] As discussed in the methods, LD will lead to artificially inflated p-values in any SNP-phenotype association test and might therefore mask causal variants. This makes it difficult to choose a sensible p-value threshold, as even conservative approaches like the Bonferroni-correction will not hold. There are various methods to correct for population stratification, such as applying an inflation factor calculated as the chi-square test median ratio[188] or using multivariate rather than Fisher's exact or Chi-square tests.[189;190] But since a recent paper shows that population stratification corrections alone do not solve for the confounding influence of LD,[177] we opted for combining a very conservative p-value cut-off with the homoplasy criterion to construct our list of putative CMs. The conservative cut-off implies that we will have a significant number of false negatives, but should have a very low rate of false positives. For the purpose of our investigation, this was the desired outcome.

Lastly, due to limited coverage of our data, we did not expand our analyses to *M. tuberculosis* Lineages other than 1-4, and did not evaluate any of the non-tuberculous mycobacteria either. This implies that we cannot necessarily extrapolate our findings to all members of the *Mycobacterium tuberculosis* complex.

3.5.3 Outlook and future work

Overall, the whole-genome sequencing data enabled the construction of a tractable, high-confidence list of 51 putative CMs, which can form the basis of future investigations. In addition, we derived further insights by combining the sequencing data with our *in vitro* growth data, allowing us to estimate the changes in growth phenotype following emergence of resistance and CMs. CMs arise in a similar fashion in other *M. tuberculosis* genes, such as *ahpC* and *gyrA*, following resistance to isoniazid and fluoroquinolones, respectively.[191] The fitness cost following mutations in the *katG* gene (isoniazid resistance) for example has recently been formally captured and the authors suggest that compensatory evolution in genes related to oxidative stress tolerance (*sodA, ahpC*) could relieve this cost.[192]

Even in other bacterial species, such as *Salmonella typhimurium* and *Escherichia coli*, there is evidence

for a fitness cost due to antibiotic resistance and compensatory mechanisms arising to reduce its impact.[193] Our approach to identifying CMs could hence be applied in a similar fashion to other microorganisms and drugs where resistance mutations are annotated and known to introduce a fitness cost.

This is especially useful in light of the finding that CMs can serve as highly specific resistance markers in the absence of a direct resistance mutation. They hence improve genetics-based resistance prediction, especially in settings with low read depth. In future work we will investigate the read depths at which it becomes beneficial to consider CMs for calling rifampicin resistance routinely by randomly down-sampling sequencing reads.

3.6 Supplementary

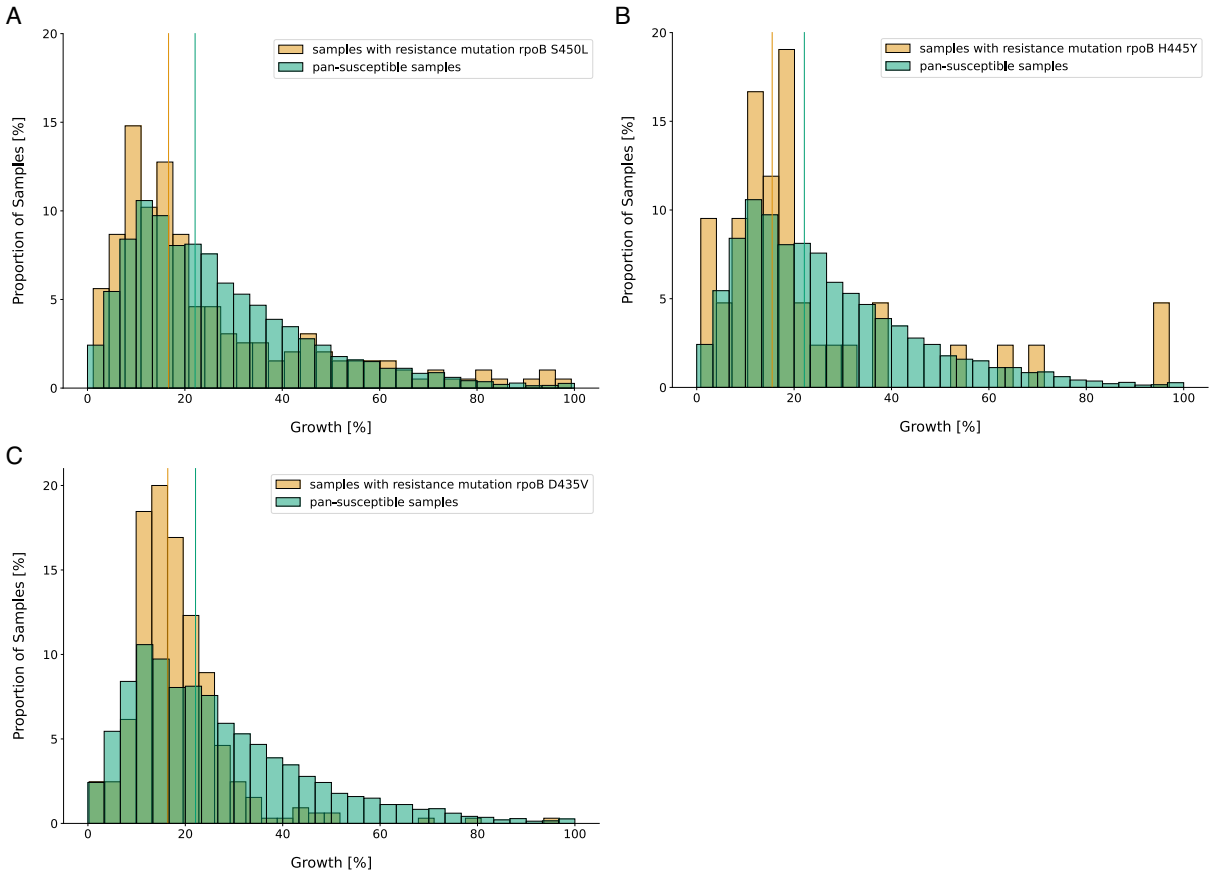


Figure S1: Growth distributions for pan-susceptible samples vs samples with specific rifampicin resistance mutations in *M. tuberculosis* (A-C) Distributions of growth in percent of covered well-area as measured in the CRyPTIC project[167] were plotted as a histogram against the proportion of samples that display this amount of growth. Samples with the resistance mutation indicated in the legend and no other potentially interfering mutations are plotted in red, samples that were classified as pan-susceptible are plotted in green. Vertical lines indicate the respective medians. The medians and Mann-Whitney p-values of the distributions are listed in Table 3.

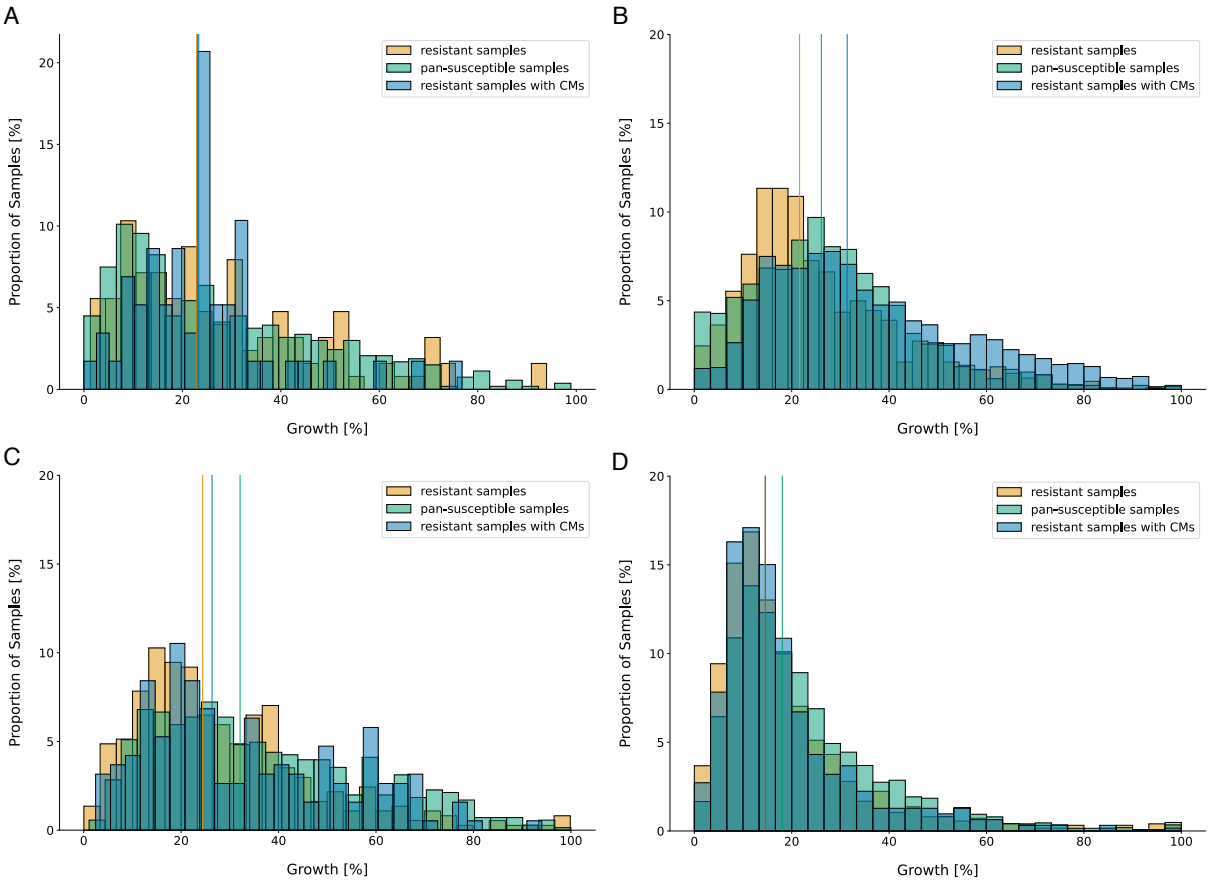


Figure S2: Growth distributions of *M. tuberculosis* samples within different lineages (A) Distribution of growth in *M. tuberculosis* Lineage 1 in percent of covered well-area as measured in the CRyPTIC project [167] were plotted as a histogram against the proportion of samples that display this amount of growth. Samples with rifampicin resistance mutations but no putative compensatory mutations (CMs) are plotted in red, samples that were classified as pan-susceptible are plotted in green. Samples that have rifampicin resistance mutations and at least one CM are shown in blue. Vertical lines indicate the respective medians. The medians and Mann-Whitney p-values of the distributions are shown in Table 8. **(B)** Plot layout as in (A), but samples derive from *M. tuberculosis* Lineage 2. **(C)** Plot layout as in (A), but samples derive from *M. tuberculosis* Lineage 3. **(D)** Plot layout as in (A), but samples derive from *M. tuberculosis* Lineage 4.

4 Subpopulations in bacterial infections

This chapter investigates the relevance of genetic heterogeneity in bacterial subpopulations in clinical microbiological samples of *M. tuberculosis*, *E. coli* and *K. pneumoniae* infections. Specifically, we will assess the ability of established methods in whole genome sequencing (WGS)-based antimicrobial susceptibility testing (AST) to detect clinically relevant subpopulations and the attendant consequences for resistance prediction.

4.1 Introduction

Infections with clinically relevant pathogens are seldom entirely homogeneous.[194–196] Many different scenarios can lead to the emergence of genetically distinct subpopulations in infections, from within-host evolution to secondary infections. Hence, while resistance often reaches fixation quickly during antibiotic treatment, it is still possible that at the time when the sample is taken, the patient has a heterogeneous infection, i.e. an infection with a resistant subpopulation. This phenomenon has been termed 'heteroresistance' and has been described in many pathogens.[197] We will therefore examine the relevance of genetic subpopulations for resistance in both *M. tuberculosis* and *E. coli*/*K. pneumoniae* infections. For *M. tuberculosis*, we will directly assess the influence of resistant subpopulations on how well WGS-based resistance testing performs in predicting rifampicin resistance. For *E. coli* and *K. pneumoniae*, we will investigate whether there is evidence of substantial genetic diversity within infection-associated patient samples. Detecting genetic heterogeneity in infections with these organisms would have substantial implications for testing practice.

Relevance of resistant subpopulations in *M. tuberculosis* resistance prediction

The second edition of the WHO catalogue of mutations contains the most comprehensive list of resistance associated variants (RAVs) to date for predicting rifampicin resistance in *M. tuberculosis* and achieves 93.3% sensitivity and 96.9% specificity on its training dataset compared to standard phenotypic AST results.[79] Despite being one of the best-performing drugs in WGS-based resistance prediction, the sensitivity for detecting rifampicin resistance remains below the 95% threshold proposed for antimicrobial susceptibility test devices by the International Standards Organization (ISO).[137] As introduced in Chapter

2, it is unrealistic to expect perfect agreement between the results of phenotypic and genotypic AST. But the discrepancies create problems for diagnostics, specifically for WGS-based resistance prediction.[198] If WGS-AST is to complement or even replace phenotypic AST in some settings, it needs to achieve comparable performance. Hence there is a strong need to close the gap between phenotypic and genotypic AST results.

In the following, we investigate the potential of resistant subpopulations to explain a part of this gap in rifampicin resistance in *M. tuberculosis*. In *M. tuberculosis*, numerous studies have identified significant genotypic within-host diversity.[199;200] In addition, considering subpopulations when testing for resistance has already been shown to significantly improve the sensitivity of genotypic resistance prediction for fluoroquinolones in *M. tuberculosis*.[196] The hypothesis is that resistant subpopulations are not identified by many WGS bioinformatics pipelines, yet phenotypic AST methods (e.g. Mycobacteria Growth Indicator Tube, MGIT) will flag samples as resistant if as little as 1% of the bacteria are resistant.[103] Hence the genotypic approach is under-calling resistance (Figure 16). However, for most bacterial pathogens, resistant subpopulations are not well-characterised and heteroresistance is assumed to be rare. In addition, the standard laboratory protocol is to pick single colonies for downstream AST. Hence any possibly present genetic diversity is filtered out prior to sequencing. This is not true for *M. tuberculosis* where e.g. an aliquot taken for DNA extraction from a MGIT tube or a sweep taken from LJ medium usually contains multiple 'crumbs'[201] and so could readily harbour resistant subpopulations, reflecting more accurately the within-host diversity.

WGS-based resistance prediction relies on characterising genetic variation in resistance-associated loci. Hence the general workflow for WGS-based resistance prediction in *M. tuberculosis* involves sequencing the sample and aligning the reads to a reference genome. The reads then form a pile-up at each locus, which is the basis for variant calling. We define the fraction of read support (FRS) as the proportion of reads at a genetic locus that supports a specific genetic variant. Bioinformatics pipelines for *M. tuberculosis* resistance prediction conventionally specify a minimum FRS for a genetic variant to be called (Figure 16). The first edition of the WHO *M. tuberculosis* mutation catalogue[202] and the CRyPTIC consortium[141] both used a conservative FRS threshold of 0.90 when calling genetic variants, due to relying on the same

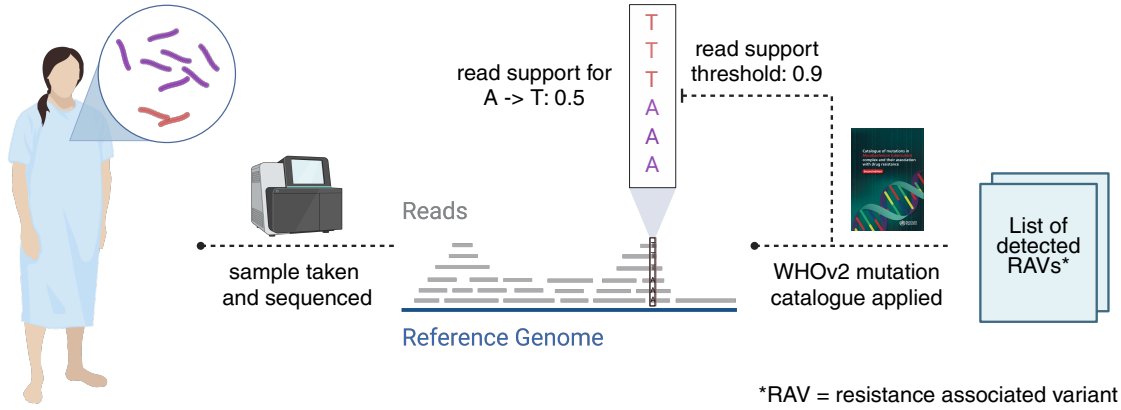


Figure 16: Established workflow for WGS-based resistance prediction and the impact of the read support threshold. The sample, here taken from a patient with a resistant *M. tuberculosis* subpopulation (red), is sequenced and reads are aligned to an *M. tuberculosis* reference genome. Variant calling is performed against the reference, and the WHOv2 mutation catalogue is applied to classify mutations as resistant or susceptible. The final list of detected RAVs will only contain variants that scored above a previously applied read support threshold, which here is conservative at 0.9.

pipeline for variant calling (Clockwork).[173] This high threshold prevents sequencing errors leading to spurious variant calls, but as sequencing technologies have improved and error rates decreased, this now seems potentially too conservative since it also ensures genetic subpopulations are not detected. Such subpopulations can show resistance, e.g. if an infection has only recently evolved resistance, perhaps in response to treatment, or if there has been a secondary resistant infection.

Emergence scenarios for resistant subpopulations

In addition to their potential to evade detection, samples with resistant subpopulations capture an interesting state in bacterial infections, which may arise through multiple different scenarios (Figure 17). While it is obviously a complex phenomenon, there are three main scenarios that could lead to the presence of resistant subpopulations in a sample. The first possibility is within-host evolution of resistance, where the resistant subpopulation gradually takes over a susceptible population during drug treatment (Figure 17A). Secondly, the evolved drug resistance can revert back to a susceptible phenotype or be outcompeted after drug treatment is stopped (Figure 17B); or thirdly the resistant subpopulation is acquired through a secondary infection (Figure 17C). Distinguishing between these possibilities requires that we know whether the sample was taken before, during or after drug treatment. Unfortunately, precise data on this is scarce in WGS datasets and hence we can seldom distinguish between scenarios 1 and 2 (Figure 17A, B). The

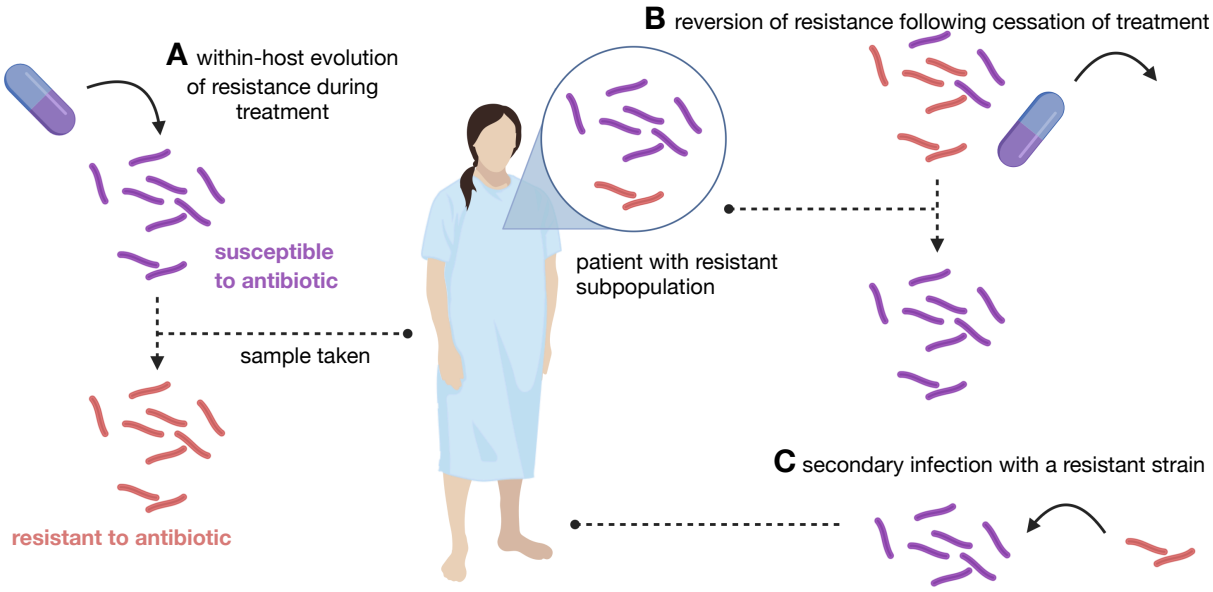


Figure 17: Three ways a patient sample can show an infection with a resistant subpopulation. The susceptible subpopulation is shown in purple, the resistant subpopulation in red. Samples with resistant subpopulations can originate from multiple different scenarios, among them (**A**) within-host-evolution, (**B**) reversion to susceptible phenotype following cessation of treatment, and (**C**) secondary infection.

second unknown variable is the source of resistance: within-host evolution or a secondary infection. We aim to discern the source of resistance by quantifying the amount of genetic diversity in *M. tuberculosis* samples with resistant subpopulations, which might allow us to define a lower bound for the amount of resistant samples arising through secondary infection.

The single colony pick paradigm for WGS-based resistance testing

As mentioned previously, for most pathogens, resistant subpopulations are not expected to be captured during WGS since the standard laboratory practice in routine diagnostics is to pick single colonies and hence heteroresistance is overlooked. While there are alternative WGS-based approaches including metagenomics and plate sweeps,[136] these are less commonly used. Mostly, samples are taken from a patient, undergo incubation in media designed to promote growth, and if culture-positive, a single colony is picked for further purification ahead of phenotypic AST and/or sequencing (Figure 18A). Hence potential subpopulations are not characterised in the sequencing output (Figure 18B).

One should therefore test whether single colony picks as part of routine diagnostics pathways accurately represent the genetic diversity present in an infection. By purifying a single colony from the original

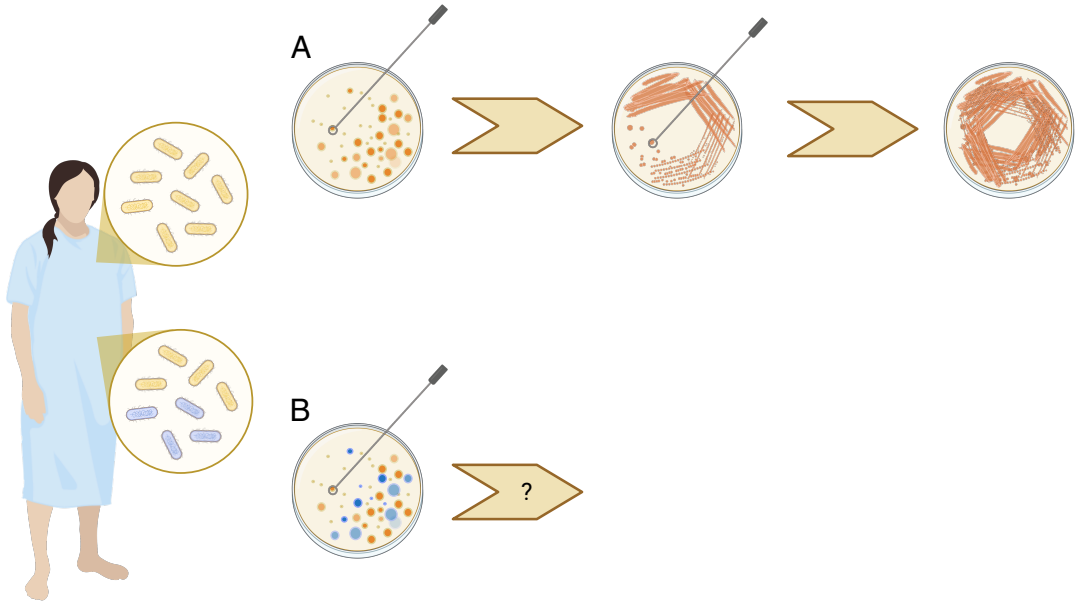


Figure 18: Illustration of the single colony pick paradigm, commonly used for bacterial isolate processing in diagnostic microbiology. To prepare samples for downstream processing, including sequencing, single colonies from a primary plate derived from a sample inoculum (e.g. a blood culture) are processed to obtain a single clone of the bacteria. If a single species was identified diagnostically, one colony is picked from the initial plate, purified in a second plating step, and then streaked out generously on the final plate to obtain material for further diagnostic evaluation. The initial plate will contain multiple different colonies, which may include genetically diverse populations. In A, we assume they are all from the same strain of bacteria, and hence genetically almost identical. In B, we assume that there are multiple strains of the same bacterial species present, possibly with different resistance markers. Picking different single colonies could hence result in multiple different results for resistance testing by downstream phenotypic as well as genotypic AST.

streak-out plate, any within-host diversity of the original infection will be missed. This could include resistant subpopulations. *M. tuberculosis* infections are much more likely to show resistant subpopulations, due to infection characteristics discussed in the introduction: Reactivation of latent infections, low mutation rates and lack of horizontal gene transfer all contribute to the fact that resistant subpopulations are more frequent and persistent in *M. tuberculosis* than in *E. coli* or *K. pneumoniae* infections. The extent to which resistant subpopulations in the latter two are relevant is not yet fully defined, but recent data suggest an effect on outcomes for *E. coli* blood stream infections.[203] There is also evidence for significant within-host genotypic and phenotypic diversity in both *E. coli* and *K. pneumoniae* infections.[194;195]

4.2 Aims of this chapter

In this Chapter, we will show that the performance of WGS-based AST for rifampicin can be significantly improved by permitting rifampicin resistant subpopulations to contribute to the final resistance classification in *M. tuberculosis*. The use of the large CRyPTIC dataset of 35,538 *M. tuberculosis* samples with both WGS data and a phenotypic AST measurement for rifampicin allows us to statistically evaluate the influence of different FRS thresholds on resistance prediction performance. We end up significantly improving the sensitivity of rifampicin resistance prediction by using a less conservative FRS threshold.

By leveraging the available data on the prevalence of heterogeneous mutations (mutations with FRS$<$0.90) in samples with rifampicin resistant subpopulations, we will also be able to quantify within-sample diversity. This will allow us to discern between different resistance emergence scenarios, as shown in Figure 17, for a third of the sample subset. We also make some interesting observations regarding geographical mixing of lineages.

In addition, we will assess the suitability of single colony picks for capturing resistance in *E. coli* and *K. pneumoniae* infections. Specifically, we will assess whether substantial within-sample diversity can be seen in a pool of 52 sterile bloodstream infection samples. If we can e.g. confirm the presence of multiple strains of a pathogenic species, or at least show significant diversity within the samples, it is possible that there are divergent, clinically relevant resistance markers present as well.

The aims of this Chapter are hence:

1. Improve rifampicin resistance prediction in *M. tuberculosis* by considering resistant subpopulations

2. Use within-sample diversity to investigate the source of resistance in samples with subpopulations

3. Confirm the presence of genetic heterogeneity within *E. coli* and *K. pneumoniae* bloodstream infection samples

4.3 Methods

Most figures and tables in this chapter can be reproduced using a GitHub repository and attendant Python3 Jupyter notebook available online.[168] The relevant analysis can be rerun using the corresponding google colab button in the README.

Dataset sources: *M. tuberculosis*

The samples used for investigating rifampicin resistance in *M. tuberculosis* are from an aggregated source. They hence might differ in how samples were initially processed. CRyPTIC samples were recommended to be sub-cultured either using Löwenstein–Jensen tubes, 7H10 agar plates or MGIT tubes, with MGIT tubes being the most common approach.[140] The Comprehensive Resistance Prediction for Tuberculosis: an International Consortium (CRyPTIC) project collected >20,000 *M. tuberculosis* samples, each of which underwent WGS and AST using one of two bespoke 96-well broth microdilution plates.[140;141] They also aggregated *M. tuberculosis* samples with WGS and/or AST data that had been previously published. An early version of this dataset with heterogeneous AST methods[142] was used to build the first edition of the WHO catalogue of tuberculosis resistance-associated variants.[118] For a detailed description of how the samples collected by the CRyPTIC project were cultured, sequenced and phenotyped, please refer to the original publications.[140;141] Throughout this project, we used v3.0.0 of the CRyPTIC project dataset, which we will refer to as 'the CRyPTIC dataset'.[143]

WGS data processing: *M. tuberculosis*

A subset of 41,575 publicly available samples from the CRyPTIC dataset, which are also stored in the European Nucleotide Archive, with short-read paired-end FASTQ files were uploaded to the EIT Global Pathogen Analysis Service (GPAS) (https://www.gpas.global) and processed using version d5f9cd0 of the Mycobacterial pipeline.[204] The pipeline for processing the CRyPTIC dataset consists of the same steps described in the methods section of Chapter 3: read filtering, speciation, lineage determination and lastly variant calling using clockwork v0.12.5 which incorporates minos.[173] The variant caller, Clockwork, calculates the FRS for each putative call and, by default, only calls genetic variants where FRS$\geq$0.90 with a filter applied to the remainder.[173] To protect against spurious calls due to sequencing errors, variants

also needed to be supported by at least three reads.

For this analysis, we instructed the downstream tool, gnomonicus, to ignore the FRS filter and record all potential genetic variants so that we could use the sequence data to detect *M. tuberculosis* subpopulations.[174] Most samples (n=37,594/41,575 (90.4%)) had one or more of 101,020 mutations in total detected in the RNA polymerase (genes *rpoA*, *rpoB*, *rpoC*, *rpoZ* and *sigA*). This included 3,260 so-called null mutations where there were insufficient (< 3) reads to reliably call a variant.

Examining *rpoB* in more detail we found 31,347/41,575 (75.4%) samples containing 58,789/101,020 (58.2%) mutations with a FRS$\geq$0.90 and 1,372/41,575 (3.3%) samples containing 1,940/101,020 (1.9%) mutations supported by a FRS$<$0.90. In both cases the most common *rpoB* mutations were A1075A, the most common phylogenetic mutation, and S450L, the most common resistance-associated mutation, as expected. Mutations were flagged as being associated with resistance to rifampicin if they were included in the second edition of the WHO catalogue of mutations in *M. tuberculosis*.[79] We used the published list of high-confidence compensatory mutations (CMs)[138], that we derived in Chapter 3, to annotate which mutations are compensatory.

Phenotypic AST data processing: *M. tuberculosis*

A total of 52,148 samples from the CRyPTIC dataset had one or more binary rifampicin drug susceptibility test results using a range of methods; the most common were broth microdilution plates (n=24,172/52,148 (46.3%)) and mycobacterial growth indicator tubes (n=23,682/52,148 (45.4%)). Resistance was called based on the MIC of the respective sample and the predefined epidemiological cut-off value (ECOFF) for rifampicin. If a sample had been tested using more than one method and all methods produced the same S/R result then, if present, the CRyPTIC result was retained since it has richer data and has undergone additional quality control. If the phenotypic methods disagreed on the outcome then the resistant result was retained. This resulted in 48,031/52,148 (92.5%) samples with a single rifampicin AST result.

Merging with the genetic data resulted in a combined dataset of 35,538 samples which have both WGS

data and a single, binary, phenotypic AST result for rifampicin: this dataset forms the basis of our subsequent analysis.

Statistical Analyses: *M. tuberculosis*

The sensitivity and specificity of the genotypic resistance call were calculated with respect to the phenotypic drug susceptibility result, defined as the gold standard. Significance was tested using a proportions z-test.

Dataset sources: *E. coli* and *K. pneumoniae*

To overcome the limitations of routine diagnostic workflows in producing sequencing datasets relevant to evaluating heterogeneity in infecting bacterial populations, we used a direct-from-blood culture metagenomic sequencing dataset based on blood stream samples. The samples were collected and sequenced by Govender *et al.*[136] They collected 273 blood culture samples from patients with blood stream infections at the Oxford University Hospitals NHS Foundation Trust (2020-2021), incubating each culture bottle in the BD BACTEC™ FX systems until it flagged positive or for five days (Figure 19). DNA was extracted from 1.5 mL aliquots of the blood culture samples, using the BiOstic Bacteraemia kit (MoBio, Qiagen, USA), with minor adjustments as stated in their preprint.[136] An internal control was added (*Thermus thermophilus* DNA, 2% of the DNA concentration) and DNA libraries were prepared with the Rapid Sequencing Kit with barcoding (SQK-RBK004). Long-read sequencing (Oxford Nanopore Technologies) was performed on the GridION platform with R9.4.1 flow cells for 24 hours, with live basecalling using Guppy (v3.2.6).

The resulting sequencing reads were demultiplexed using Guppy (v6.4.8) and low quality reads and reads below a read length of 1000 bp were removed using prinseq-lite (v0.20.4). We took these data for further evaluation of heterogeneity in *E. coli* and *K. pneumoniae* bloodstream infections. We excluded any samples where the species identification performed as part of routine diagnostic workflows (by MALDI-ToF) reported a different pathogen (n=219) or a polyspecies infection (n=2), leaving 45 evaluable *E. coli* and 9 *K. pneumoniae* datasets from 54 bloodstream infections. Resistance profiles as part of routine diagnostic antimicrobial susceptibility testing using the BD Phoenix microbroth dilution (BD, USA) were

also available as part of the Govender *et al.* study.[136]

Constructing a Nextflow pipeline for processing sequencing reads

The demultiplexed and quality controlled reads were processed using a Nexflow pipeline that we constructed for this study (Figure 19). First, for each sample, we filtered out human reads using the human read removal tool Hostile (v0.1.0)[205]. We then performed strain-oblivious metagenome (contig) assembly of the reads using metaFlye (v2.9.5)[206] and mapped the reads with Minimap2 (v2.28)[171] back to the metagenome assembly as a reference, in order to capture variation present in the read pileups and not in the consensus assembly. Variants were called using the long read variant caller Longshot (v1.0.0).[207] This pipeline was constructed according to the recommended workflow specified in the metagenome strain haplotyping tool Floria by Shaw *et al.*[208]

Strain quantification in bloodstream infection samples

We employed the strain haplotyping tool Floria (v0.0.1) to estimate the strain count per sample. This software uses the nucleotide variation amongst reads mapped to a metagenome assembly to group reads by strain.[208] Specifically, Floria takes a set of mapped reads (in bam format) and SNPs (in vcf format) and outputs read clusters as strain-level sets.

Based on the strain-level read sets, which it refers to as 'haplosets', Floria also reports an average strain count per contig s_i, which is the average number of haplosets covering the SNPs in the contig. The average strain counts used here have been filtered using a minimum haplotype phasing quality (HAPQ)>15, which was the best-performing cut-off in the benchmarking performed by Shaw *et al.*[208] The HAPQ score is a heuristic score which quantifies the likelihood of the consensus haplotype of the haploset arising due to error, making it a spurious haploset. The score is defined between 0 and 60, with higher HAPQ indicating higher haploset goodness.

The average strain count per contig s_i can be used to calculate the species level strain count (SSN) as the size-adjusted mean of the strain counts of all contigs n with size c_i from a single species:

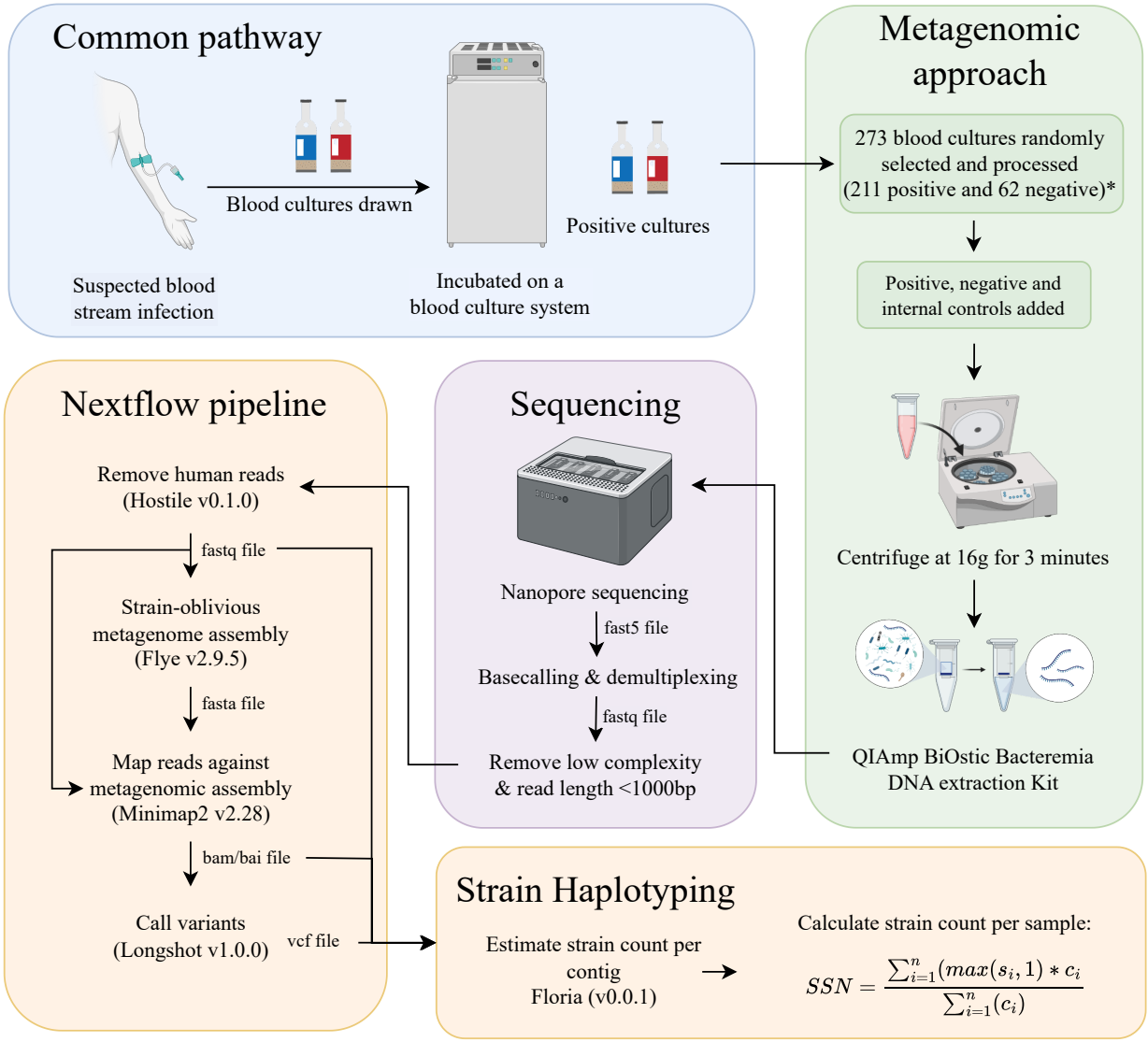


Figure 19: Blood culture sample processing for long-read metagenomic sequencing as described by Govender *et al.* [136] They randomly chose 273 samples for processing using a metagenomics approach. They then sequenced the samples using a Nanopore sequencer and demultiplexed and quality checked the reads. We obtained any fastq files from monomicrobial *E. coli* or *K. pneumoniae* blood stream infections from their dataset and processed them using our Nextflow pipeline (orange). Human reads were removed, and contigs assembled using metaFlye. Fastq reads were mapped back to the assembled contigs and variants were called based on the aligned reads. The variants in the vcf file, the aligned reads in the bam file and the raw reads in the fastq file were used as input for the Floria strain haplotyping tool. We obtained the strain count per contig for each sample, which was then used to calculate the strain count per sample (see equation provided). Figure adapted from Govender *et al.* [136]

$$SSN(Species\ A) = \frac{\sum_{i=1}^{n}(max\ s_i, 1) * c_i}{\sum_{i=1}^{n}(c_i)}$$

4.4 Results

4.4.1 Classifying subpopulations that contain rifampicin resistance associated variants as resistant improves sensitivity of resistance prediction in *M. tuberculosis*

Including resistant subpopulations has been shown to significantly improve the sensitivity of resistance prediction for fluoroquinolones in *M. tuberculosis*.[196] We aim to investigate if there is evidence supporting the same phenomenon in rifampicin resistance, which we will test in the following. The hypothesis is that bioinformatics pipelines with a high threshold for the fraction of read support (FRS) of a detected variant can mask resistant subpopulations for the downstream prediction method.

First, we checked the distribution of FRS for variants in the samples that show mutations in the RNA polymerase. The WHO catalogue version 1,[79] as well as the work of the CRyPTIC consortium[140] both used a very conservative FRS cut-off of 0.90. In our dataset of 35,538 samples, we will therefore describe samples with one or more rifampicin RAVs with a FRS$\geq$0.90 as *homogeneous*, and samples containing RAVs at a FRS$<$0.90 as *heterogeneous*. Mutations with an FRS$<$0.90 will be described as *heterogeneous* mutations. Out of the 10,568 samples containing a rifampicin RAV at any level of read support, 10,287 were homogeneous, 261 are heterogeneous and 20 are mixed, i.e. contain at least two rifampicin RAVs, one supported by FRS$\geq$0.90 and another with FRS$<$0.90 (Figure 20A). We also assessed which specific RAVs are found in the different types of samples (i.e. homogeneous vs heterogeneous vs mixed, Table 10). There are small differences in mutation distributions between homogeneous and heterogeneous samples, with heterogeneous samples showing a more balanced distribution of RAVs, with a lower proportion harbouring the canonical *rpoB* S450L mutation than homogeneous samples. This could be explained by some of the heterogeneous RAVs not yet reaching fixation, possibly because they carry a larger fitness cost than *rpoB* S450L. Still, both homogeneous and heterogeneous samples show *rpoB* S450L as the most prevalent mutation (66.6% and 40.6%, respectively), which is known to have the lowest fitness cost for *M. tuberculosis* out of the most common RAVs.[150] Interestingly, we see a striking difference in the samples that are mixed, although it should be noted that we have a limited number of mixed samples available (n=20). In these cases, the mutation *rpoB* L430P is the most prevalent RAV, which could be explained by the fact that L430P reportedly has to co-occur with other RAVs in order to cause clinical

Mutation	Homogeneous		Heterogeneous		Mixed	
	Count	%	Count	%	Count	%
rpoB S450L	7041	66.6	144	40.6	5	11.1
rpoB D435V	826	7.8	27	7.6	1	2.2
rpoB H445Y	374	3.5	28	7.9	2	4.4
rpoB H445D	368	3.5	16	4.5	0	0
rpoB D435Y	217	2.1	17	4.8	3	6.7
rpoB L430P	196	1.9	19	5.4	7	15.6
rpoB L452P	174	1.6	20	5.6	0	0
rpoB S450W	142	1.3	3	0.8	0	0
rpoB H445L	131	1.2	9	2.5	0	0
rpoB I491F	105	1.0	3	0.8	0	0
rpoB H445R	92	0.9	5	1.4	2	4.4
rpoB D435G	85	0.8	4	1.1	1	2.2
rpoB H445N	76	0.7	8	2.3	4	8.9
rpoB V170F	73	0.7	3	0.8	1	2.2
rpoB D435F	43	0.4	2	0.6	0	0
rpoB H445C	42	0.4	5	1.4	0	0
rpoB S441L	30	0.3	3	0.8	1	2.2
rpoB L430R	24	0.2	1	0.3	0	0
rpoB Q432P	24	0.2	8	2.3	0	0
rpoB Q432L	21	0.2	1	0.3	2	4.4

Table 10: Distribution of the top 20 resistance-associated variants (RAVs) by sample type (homogeneous, heterogeneous or mixed). Note that it is possible for multiple resistance mutations to co-occurr in one sample. We see a more balanced distribution of mutation frequencies in heterogeneous samples compared to homogeneous samples, but they still show *rpoB* S450L as their top-scoring RAV. Mixed samples however show a very unusual top-scoring mutation: *rpoB* L430P.

resistance.[209] In line with this finding, *rpoB* L430P is also one of the resistance mutations that occurrs in both phenotypically resistant and susceptible samples and is hence a 'disputed' resistance mutation, as discussed in Chapter 3.

Next, examining the distribution of heterogeneous rifampicin RAVs, we observe that these are seen at all levels of FRS, possibly increasing in frequency amongst samples with higher FRS values (Figure 20B). This indicates that the full range of values should be considered when assessing the impact of the FRS threshold on the performance of resistance prediction.

We then used v3.0.0 of the matched dataset of WGS and phenotypic AST data by the CRyPTIC consortium[143] to inspect the influence of the FRS cut-off on the performance of genotypic resistance calls. To reiterate, the genotypic resistance calls are based on whether the detected variants in the samples are catalogued RAVs, as defined by the WHO. The sensitivity and specificity of the genotypic resistance calls are calculated with respect to the phenotypic AST results.

We calculated the sensitivity and specificity of the rifampicin resistance prediction using different FRS cut-offs. As the FRS cut-off increases, the specificity rises proportionally, while sensitivity decreases and hence behaves inversely proportional to the cut-off value. Importantly, sensitivity varies more with FRS cut-off changes than the specificity (Figure 20C). Indeed, sensitivity increases from 94.3% to 96.4% as the minimum FRS required to call a RAV is decreased from 0.90 to 0.05; over the same FRS range the specificity falls slightly from 98.1% to 97.9% (Figure 20C, Table 11). The increase in sensitivity when the FRS is dropped from 0.90 to 0.05 is statistically significant (z-test, p-value = 8e-13), whilst the decrease in specificity is not (p-value = 2.05e-01, Figure 20D). These findings indicate that sensitivity can be increased by lowering the FRS threshold, with a minimal cost for specificity.

Extrapolating from this, applying a conservative FRS threshold appears to mask resistant subpopulations, resulting in falsely predicting some resistant samples as susceptible to rifampicin.

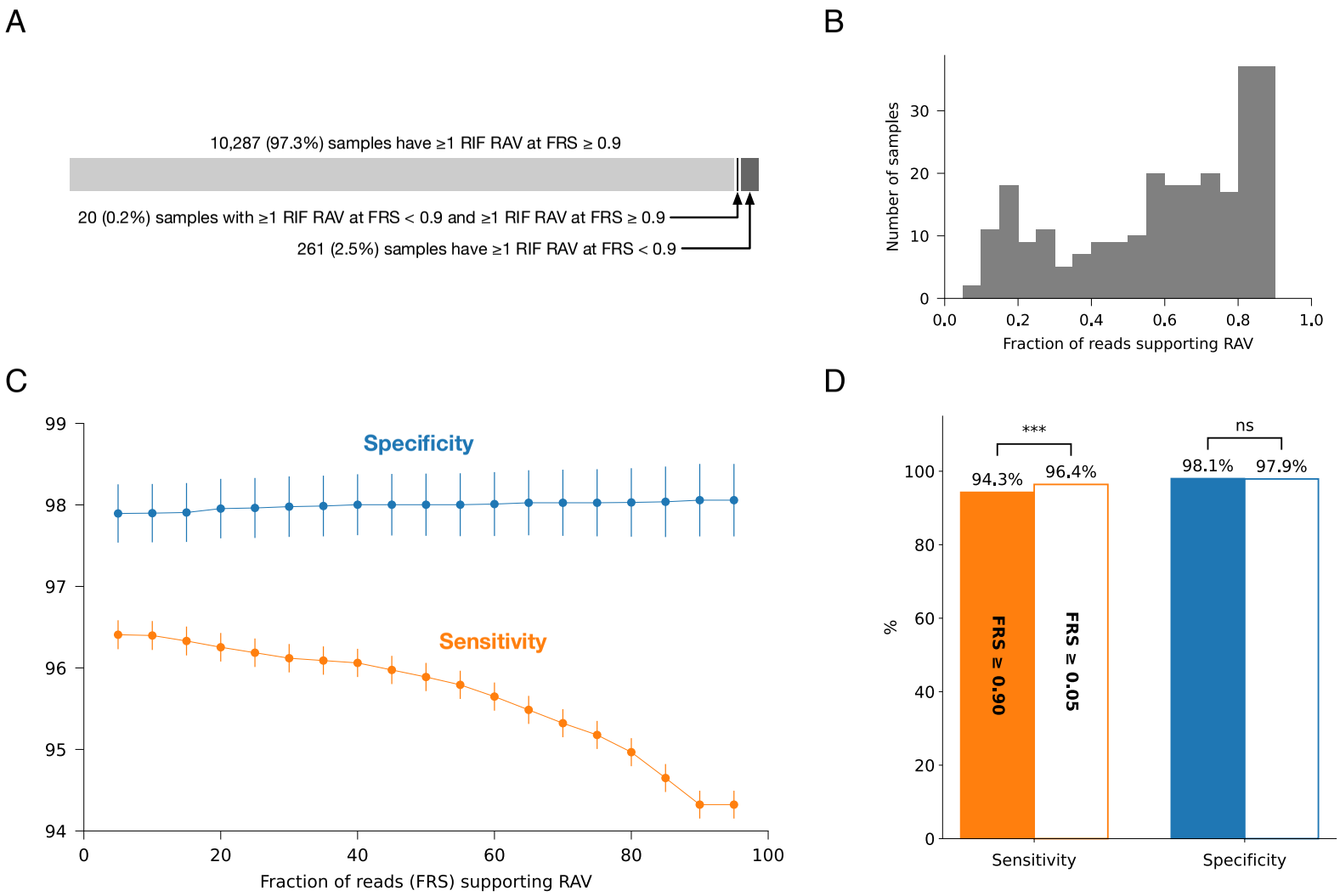


Figure 20: Lowering the fraction of read support (FRS) threshold for calling rifampicin resistance-associated variants (RAVs) increases sensitivity with no significant effect on specificity. **(A)** The majority of samples containing a RAV are homogeneous, meaning they show a rifampicin RAV at≥0.90 FRS. The heterogeneous samples make up for 2.5% of samples, but we rarely see samples with both a RAV at≥0.90 FRS and a RAV with lower read support. **(B)** The distribution of FRS for the RAVs in the 258 heterogeneous samples with $0.05 \leq \text{FRS} < 0.90$. **(C)** Decreasing the FRS threshold required to support a variant call that is a known RAV increases sensitivity of the prediction with little effect on specificity. The slopes of a linear regression for sensitivity and specificity are -0.02 and 0.002, respectively. Error bars (95% confidence limits) are plotted as calculated via the binomial proportion. **(D)** Sensitivity is significantly improved if the FRS threshold is lowered from 0.90 to 0.05 (z-test, p-value = 8e-13). There is no significant change in specificity. For clarity error bars are not plotted.

4.4.2 Combining compensatory mutations as resistance markers with varied fraction of read support thresholds reveals non-additive effect on resistance prediction

In Chapter 3, we evaluated the effect of adding CMs to the list of RAVs for resistance prediction. We now want to quantify the combined effect of lowering the FRS cut-off and adding CMs to our resistance catalogue, since their contribution to improving sensitivity and lowering the false negative rate (FNR) may not be purely additive.

First, we checked how many samples show compensation and at what FRS. We found that 4,678/35,538 (13.2%) samples contained at least one CM, with 4,289/4,678 supported by an FRS$\geq$0.90 and 384/4,678 having a CM with an FRS$<$0.90. As described in Chapter 3, including CMs in a list of RAVs reduced the number of false negative calls from 591 to 568 when applying an FRS threshold of 0.90 to call variants (Table 5), corresponding to an absolute decrease of 0.22% in FNR. If we repeat this evaluation with a FRS threshold of 0.05, the number of false negative calls decreases from 374 to 363 upon including CMs in the resistance prediction (Table 11), corresponding to an absolute decrease of 0.11% in FNR. This indicates that the effect of lowering the FRS threshold and including CMs in the list of RAVs are not entirely additive.

When allowing CMs with a FRS$\geq$0.05 to predict resistance in a sample, the number of samples falsely predicted to be resistant (false positives) surprisingly increases from 529 to 663. Most (99%) of these erroneous calls are due to samples which, despite containing the putative CM *rpoC* F452L, are not actually resistant. If we exclude this CM *post hoc*, the specificity does not decrease when allowing CMs to predict resistance at low FRS (Table 11).

The combined effect of lowering the FRS threshold to 0.05 and permitting CMs to predict resistance is a significant reduction of the FNR from 5.68% to 3.49% (Figure 21). Hence more than a third of the resistant samples incorrectly classified as susceptible are hereby explained and corrected (Table 11), a handy improvement.

RAVs	CMs	*rpoC* F452L	FRS_{min}	TP	FP	TN	FN	SE	SP
✓		✓	0.90	9819	488	24640	591	94.3%	98.1%
✓		✓	0.05	10036	529	24599	374	96.4%	97.9%
✓	✓	✓	0.90	9842	490	24638	568	94.5%	98.0%
✓	✓	✓	0.05	10047	663	24465	363	96.5%	97.4%
✓	✓		0.05	10047	531	24597	363	96.5%	97.9%
	✓	✓	0.05	4452	225	24903	5958	42.8%	99.1%
	✓	✓	0.90	4221	73	25055	6189	40.5%	99.7%

Table 11: Fraction of read support (FRS) and corresponding contingency table values and performance metrics for different scenarios of the catalogue-based predictions for rifampicin resistance. Scenarios are shown for different FRS thresholds, and are either using RAVs and/or CMs for prediction. In one scenario, the CM *rpoC* F452L was removed from the analysis. "SE" shows sensitivity and "SP" the specificity of the resistance prediction.

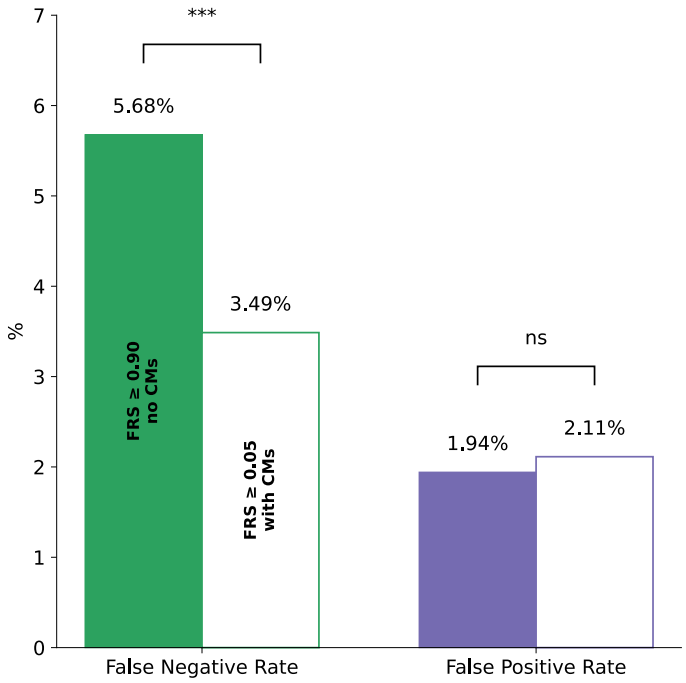


Figure 21: Overall prediction improvement when using both compensatory mutations (CMs) and a lower fraction of read support (FRS) threshold. The FNR is significantly improved if the FRS threshold is lowered from 0.90 to 0.05 and CMs are used for predicting resistance (Fisher's exact p-value = 4e-14). There is no significant change in FPR after removing the putative CM *rpoC* F452L from the list of CMs.

4.4.3 Heterogeneous resistant samples are less likely to show compensation

The proportion of heterogeneous samples with a CM is significantly lower than the fraction of homogeneous samples which are compensated (43.5% v 34.2%, Fishers exact test, p-value = 3.63e-4, Figure 22A). Furthermore, in the compensated heterogeneous samples, the CM FRS and RAV FRS values are correlated (Figure 22B, Pearson's p-value = 1.69e-22). This is consistent with these samples being composed of a susceptible strain and a rifampicin resistant strain that has acquired a CM.

While initially suggesting that this might be related to the timing of resistance emergence, since resistant heterogeneous samples are less likely to have had the time to evolve a CM, this observation could also be related to the type of resistance mutation that is present in each sample. When evaluating the relative frequency of resistance mutations in homogeneous vs heterogeneous samples with exactly one resistance mutation, we found that the main resistance mutation *rpoB* S450L accounted for 68.6% of homogeneous samples, but only for 60.5% of heterogeneous samples. The difference in relative frequencies to those observed in Table 10 is due to restricting our current analysis to samples with exactly one RAV, and hence looking specifically at the sample level and not the mutation level. This is to ensure that there is only ever one resistant subpopulation in our sample. Table 10 on the other hand includes samples with multiple RAVs, which explains why the ratio of samples with *rpoB* S450L for both sample types is slightly lower than the values reported here. This is because RAVs other than *rpoB* S450L are prone to appear together with other resistance mutations.

Those samples with *rpoB* S450L also contain almost all CMs in both heterogeneous (96%) and homogeneous samples (96.7%), showing the high association of CMs with this specific mutation. This also suggests that the lower level of compensation in heterogeneous samples is at least partially related to the lower proportion of *rpoB* S450L mutations in these samples. We can calculate the thereby explained difference in CM prevalence between the sample types by considering the ratio of compensated samples with S450L. The ratio of homogeneous S450L samples that are compensated is

$$\frac{43.5\% * 96.7\%}{68.6\%} = 61.3\%$$

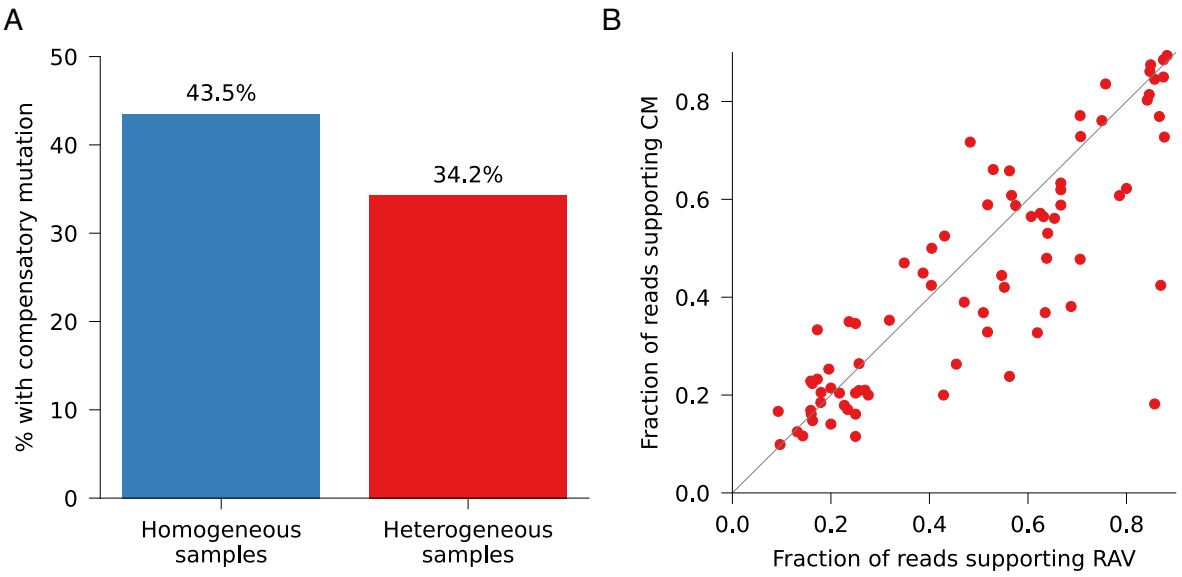


Figure 22: Heterogeneous resistant samples are less likely to also have a compensatory mutation (CM) than homogeneous resistant samples. **(A)** Percentage of samples showing CMs at any FRS in homogeneous vs heterogeneous resistant samples (Fishers exact p-value = 3.63e-4). **(B)** For the heterogeneous samples that are compensated, this plot shows the significant correlation between the FRS of the respective resistance and compensatory mutation (Pearson's p-value = 1.69e-22). The grey line indicates the hypothetical line of perfect correlation.

Assuming that this ratio is the same in both heterogeneous and homogeneous samples, the expected compensation ratio in heterogeneous resistant samples would be

$$\frac{61.3\% * 60.5\%}{96\%} = 38.6\%$$

This is still above the observed compensation ratio in heterogeneous samples of 34.2% (Figure 22A) and therefore implies that there are additional variables causing the lower compensation rate in heterogeneous samples.

4.4.4 At least 28% of heterogeneous resistant samples are the result of secondary infections

We can gain some insight into the source of resistance by examining the overall genetic diversity in the heterogeneous samples. If the resistant subpopulation arose by within-host evolution (Figure 17A), we would expect to see relatively few, if any, other heterogeneous mutations in the respective genome of the sample, given the low mutation rate of *M. tuberculosis*. A similar logic applies when a sample is taken midway through reversion of resistance following cessation of treatment (Figure 17B); again relatively few other heterogeneous mutations would be expected. If, however, the patient has been infected more than once (Figure 17C), the sample will most likely contain different strains and therefore we would expect a much greater number of heterogeneous mutations. This will not necessarily always be true: if the co-infecting strain is part of the same outbreak, the number of heterogeneous mutations will again be small. Notably, studies have also shown that the amount of variation accumulated within a patient can be as high as that observed between patients along a chain of transmission.[210] Altogether, we assume that only when the number of heterogeneous mutations is very high, e.g. if the sample contains multiple (sub)lineages, can we draw a tentative conclusion since in this case secondary infection is the only viable scenario given the data is correct. The proportion of heterogeneous samples estimated to be due to a secondary infection using this approach will therefore be a lower limit.

To determine this lower limit, we measured the number of heterogeneous mutations in each of the 248 heterogeneous resistant samples with a singular heterogeneous RAV. The latter condition is required to ensure that only a single resistant subpopulation is present. The resulting distribution is very broad; some samples have no or relatively few heterogeneous mutations, whilst others have several thousand (Figure 23A).

Mapping the lineages identified by mykrobe in each sample (single lineage, multiple sublineages, or multiple lineages) onto the distribution neatly segregates the data into three partially overlapping subgroups (Figure 23B). The first is the 163 samples which, according to mykrobe, consist of one lineage only (labelled ‘single’ in Figure 23B). The remaining 85 samples (34%) are assessed as either containing multiple sublineages belonging to the same lineage, or different lineages. These are labelled ‘sublineages’ and ‘lineages’, respectively, in Figure 23B. The only plausible explanation for the latter two cases is if the

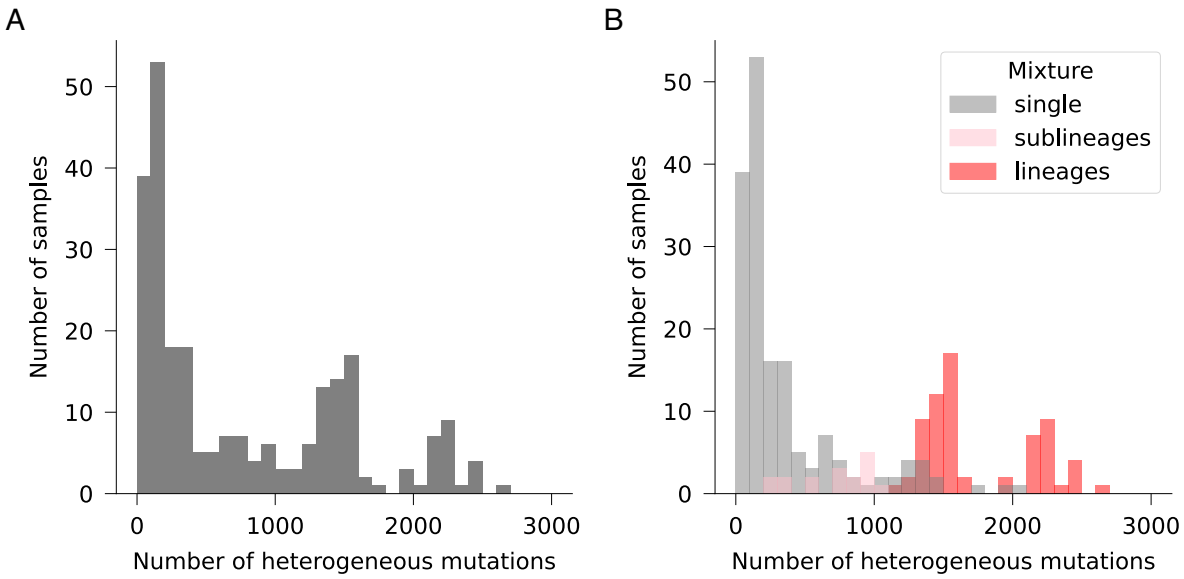


Figure 23: Quantifying genetic diversity within heterogeneous resistant samples using the amount of heterogeneous mutations detected. (A) Genetic diversity within the heterogeneous samples with resistant subpopulations, as measured by the amount of heterogeneous mutations per sample. Heterogeneous mutations are all variants in these samples which display an FRS below 0.90. **(B)** Same as in A, but the data is divided into three different batches based on the sample mixture type assigned by mykrobe (v0.12.1): A single lineage, multiple sublineages or multiple lineages per sample.

patient picked up a secondary infection. We also need to demonstrate that the secondary infection was already resistant i.e. the resistance did not evolve in the host. To establish this, we considered the phasing of the RAV FRS and the FRS distribution in each relevant sample.

The FRS distribution of samples with more than one (sub)lineage will show multiple distributions, whose modes should add up to 100%, representing the different subpopulations in the sample (e.g. Figure 24A). For some samples it is difficult to distinguish the two (or more) subpopulations, since the distributions overlap (Figure 24B). When examining the FRS plots of the 85 samples with multiple (sub)lineages, we see that in some plots the RAV FRS is very clearly in phase with one of the subpopulations (Figure 24C). For a few it is clearly out of phase, suggesting that resistance has evolved after the secondary infection event (Figure 24D) and some cases are difficult to call. The plots for all 85 samples are in the Appendix.

Based on our cases with a clear call (similar to Figure 24C), we find that at least 70/85 multiple (sub)lineage samples have a RAV in phase with one of the subpopulations. This strongly suggests that resistance was introduced directly through one of the subpopulations. We hence assume that at least 28.2% of heteroge-

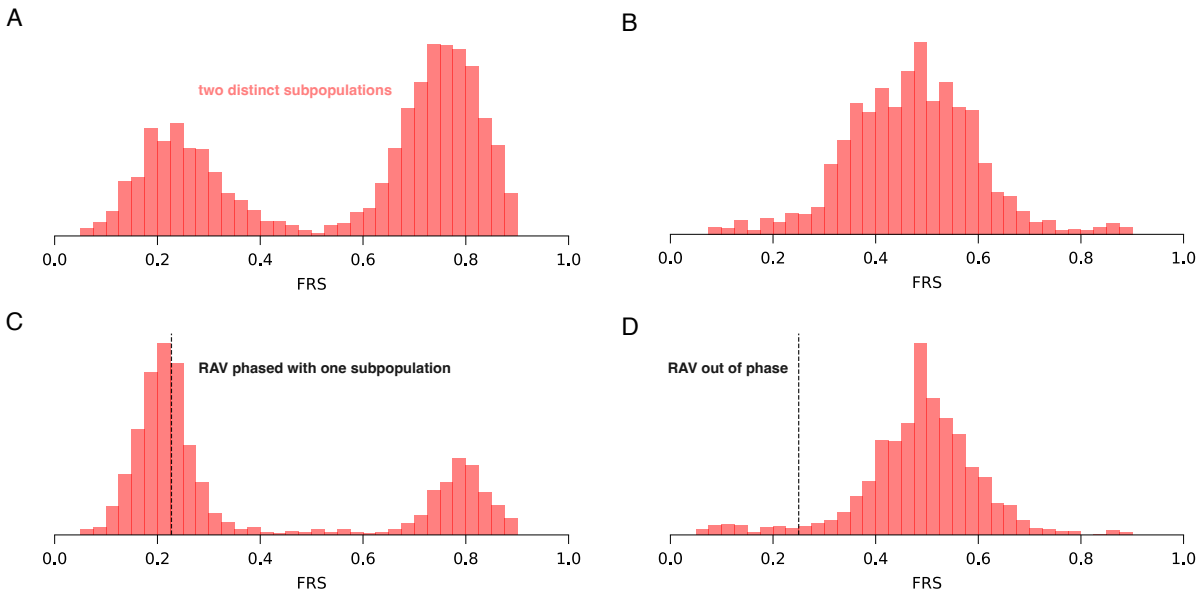


Figure 24: Fraction of read support (FRS) distribution of mutations in heterogeneous samples with multiple (sub)lineages can show if resistance-associated variants (RAVs) are in phase with one subpopulation. (A) FRS distribution in a single sample with multiple (sub)lineages of *M. tuberculosis*. The number of (heterogeneous) mutations at each FRS are shown in histogram representation. The two subpopulations can be clearly distinguished and peak at around 0.2 and 0.8 in this sample. **(B)** Plot structure identical to A, but in this sample we cannot distinguish between the subpopulations due to overlapping FRS counts. **(C)** Plot structure similar to A, but additionally a black dashed line indicates the FRS of the RAV in the sample. Here it is in phase with one of the distribution maxima. **(D)** Plot structure identical to C, but the dashed line indicating the RAV is out of phase with the distribution maxima, pointing towards recent evolution of resistance.

neous resistant samples in our dataset obtained resistance through a secondary infection event. It should be noted that this is a lower limit, so the actual percentage is likely to be higher.

Interestingly, there is also a significant association between a heterogeneous resistant sample having a CM and showing more than one (sub)lineage (Fishers exact test, p-value = 2.38e-04). Furthermore, within the heterogeneous samples which contain multiple *M. tuberculosis* lineages, we can clearly see two different clusters (Figure 23B). When evaluating the mix of lineages contained in these samples, we found that samples with a greater number of heterogeneous mutations are likely to contain Lineage 1, one of the ancient strains of *M. tuberculosis*. The modern lineages (Lineages 2-4) are more similar to the reference strain (H37Rv, Lineage 4) from which mutations are defined. Accordingly, the cluster with only modern lineages shows on average 358 heterogeneous mutations per sample fewer than the samples containing Lineage 1 (Figure 25).

We also found a significant association of heterogeneous samples that were assigned more than one

(sub)lineage with the CRyPTIC participating laboratory in Mumbai, India (Bonferroni-adjusted Fishers exact p-value = 1.79e-05). This might be due to the fact that in this geographical location, *M. tuberculosis* Lineages 1 and 3 are endemic, in contrast to the rest of the world, where 2 and 4 are the most common.[211] But Lineages 2 and 4 are still present, probably also due to spillover from other geographic locations. Hence all main *M. tuberculosis* lineages are in circulation in India, making it much more likely to acquire a polyclonal infection. Mumbai is also a very high-burden setting for *M. tuberculosis* in general, with an elevated force of infection, which is known to make mixed infections with two distinct strains likely.[119]

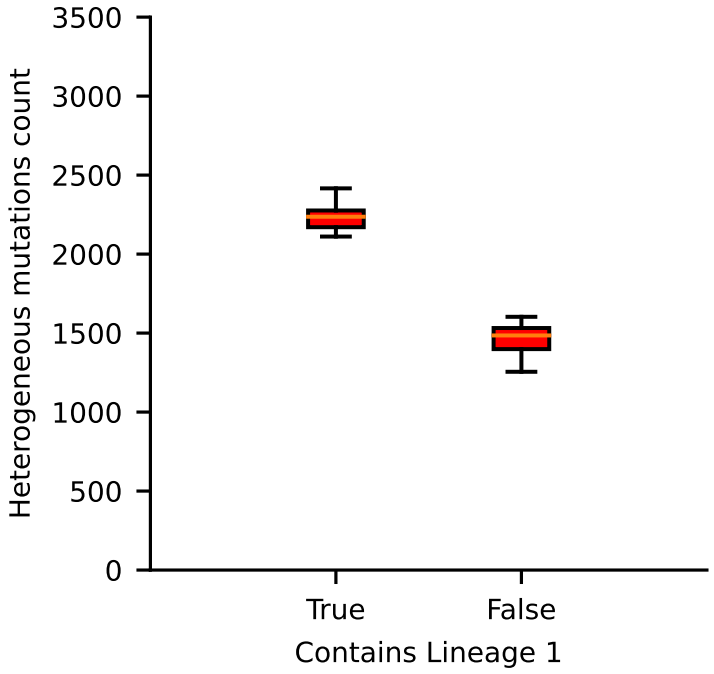


Figure 25: Samples that show multiple lineages and contain at least one ancient *M. tuberculosis* lineage (Lineages 1, 5-7) show more heterogeneous mutations. The average heterogeneous mutation count for samples with Lineage 1 is 2598, while for exclusively modern lineage (Lineages 2-4) samples it is 2240.

4.4.5 There is evidence of within-host genetic diversity in *E. coli* and *K. pneumoniae* metagenomic bloodstream samples

To establish whether within-host genetic variation is relevant in other pathogens, we investigated the example of metagenomic datasets derived from *E. coli* and *K. pneumoniae* bloodstream infections. This implicitly tests the validity of using single colony picks for WGS-based diagnostics for these pathogens.

Our sample set was composed of material directly sequenced from 52 blood culture bottles from patients with either confirmed *E. coli* or *K. pneumoniae* bloodstream infections. The blood culture bottles had been collected and the culture material sequenced using a metagenomics approach by Govender *et al.*[136] By using a metagenomics approach, they were able to capture genetic diversity at the sample- rather than the colony-level. This is in contrast to the standardised protocol of picking only a single colony for sequencing, which many laboratories follow.

To quantify the genetic diversity within individual samples, we used the off-the-shelf strain haplotyping tool called Floria.[208] To prepare the sequencing reads for consumption by the downstream tool, we built a Nextflow pipeline to process the fastq files. The pipeline performs strain-agnostic assembly of the reads, then aligns the reads and calls variants against the assembly as a reference. This allowed us to supply Floria with an indexed bam file (containing the aligned reads), a vcf file (containing the detected variant sites), and the fastq file (containing the raw reads). Floria then uses a strain haplotyping algorithm to cluster the reads into strain-level sets of haplotypes. Based on this it calculates an average strain count per contig, which we can transform into a strain count per sample using their supplied formula (Methods).

Half of the samples in our dataset (26/52) show a species level strain count (SSN) of one (Figure 26A), consistent with a clonal infection. This indicates that in those samples there is no evidence of within-sample diversity and hence within-host diversity of the infection. We see a decreasing number of samples as the SSN increases towards 1.4, with one singular outlier above a SSN of 1.8. Plotting this distribution as a scatter plot and by species, we observe that most of the samples with an elevated SSN cluster around a SSN of 1.2 (Figure 26B). The mean SSN is 1.08 for *E. coli* and 1.05 for *K. pneumoniae*, with no evidence of a difference between the two species (Mann-Whitney U test, p-value = 1.7e-01).

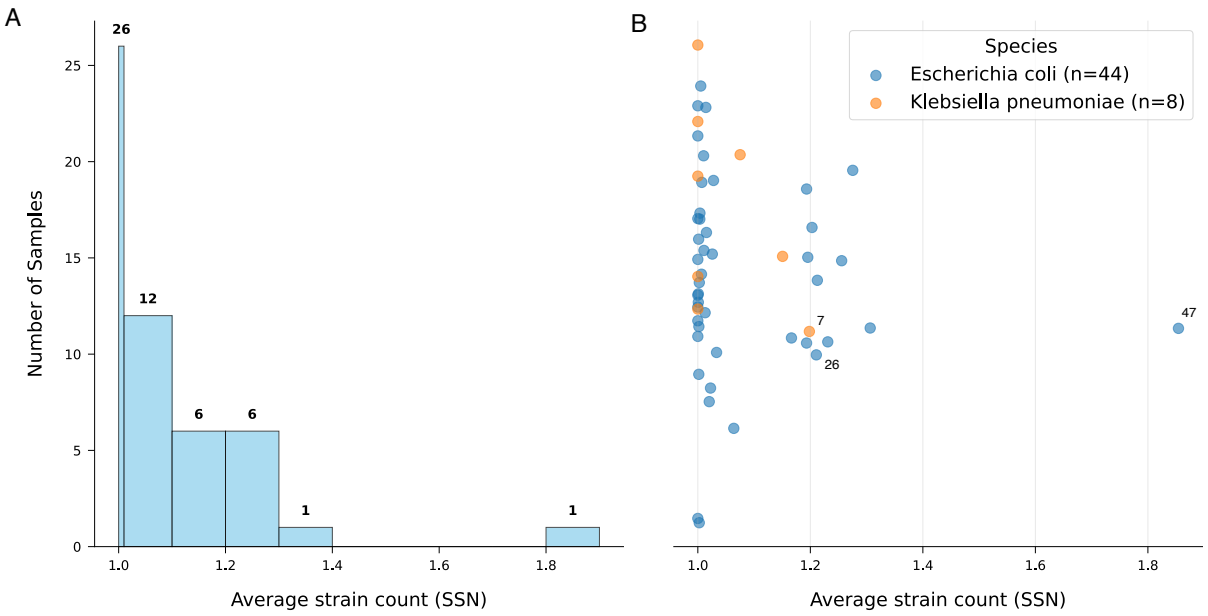


Figure 26: Species level strain count (SSN) as detected by Floria[208] indicates presence of genetic diversity in some samples. **(A)** The number of samples per SSN were plotted as a histogram, with the first bin covering the range between 1 and 1.01. The following bins cover the range until a SSN of 1.9 in steps of 0.1 at a time. Most samples fall within a SSN of 1 or slightly higher. **(B)** The SSN per sample is visualised as a jittered scatter plot. The y axis position hence does not have a quantitative meaning. *E. coli* samples are shown in blue and *K. pneumoniae* samples in orange. This plot allows to examine the exact SSN per sample, revealing clustering around a SSN of 1.2. Annotated are samples 7, 26 and 47, which are investigated in more detail in Figure 27.

Sample 47 (derived from a *E. coli* bloodstream infection) was a clear outlier, with a SSN of 1.86 (Figure 26B), hence we decided to examine in more detail the respective contig coverage detected by Floria. We should be able to see multiple different strain level read clusters (haplosets) covering the same location if we are dealing with a sample with multiple strains. Examining the largest contig (5Mb) from the assembly, we can see that Floria has detected two different haplosets covering the contig (Figure 27A, Contig 1). The coverage for one of the haplosets is higher (ca 300 vs ca 50), indicating it probably belongs to the dominant strain. The fact that one of the two haplosets contains many more reference SNPs than the other (colour scale in Figure 27A) indicates that they are genetically distinct. Also, the coverage of the minor haploset is still reasonable and the confidence score is high (HAPQ score, grey scale in Figure 27A).

The HAPQ or confidence score represents the likelihood that this haploset is not a duplicate or spurious haploset. An example of a low confidence score region by HAPQ can be seen in contig 6 of the same sample, where we see many low confidence HAPQ scores associated with the reads (27B). This contig is

therefore not considered for calculating the SSN, since we chose a HAPQ cut-off of 15 on a scale of 0-60. Other contigs for this sample were much shorter, and showed e.g. evidence of only one strain and/or low HAPQ scores. Overall, these analyses and the SSN of 1.86 indicate that sample 47 contains at least two strains of *E. coli*.

For samples clustering around the SSN of 1.2 (Figure 26B) it is more challenging to conclude how many strains are present. But since the SSN is a contig length-normalised measure, it is likely that these samples still show diversity in some regions of the genome, such as the accessory. We can see this for instance in contig 2 in sample 7 (*K. pneumoniae*, Figure 27C) and contig 5 in sample 26 (*E. coli*, Figure 27D). While these plots do show multiple haplosets, they either exhibit lower coverage (Figure 27C), or the length of the contig is comparatively short (Figure 27D). Both of these observations also explain the lower SSN compared to sample 47. For the quarter of samples with a SSN around 1.2 (13/52) it is hence likely that significant genetic diversity is present within the infecting bacterial population, although probably not at a level that suggests the presence of more than one strain.

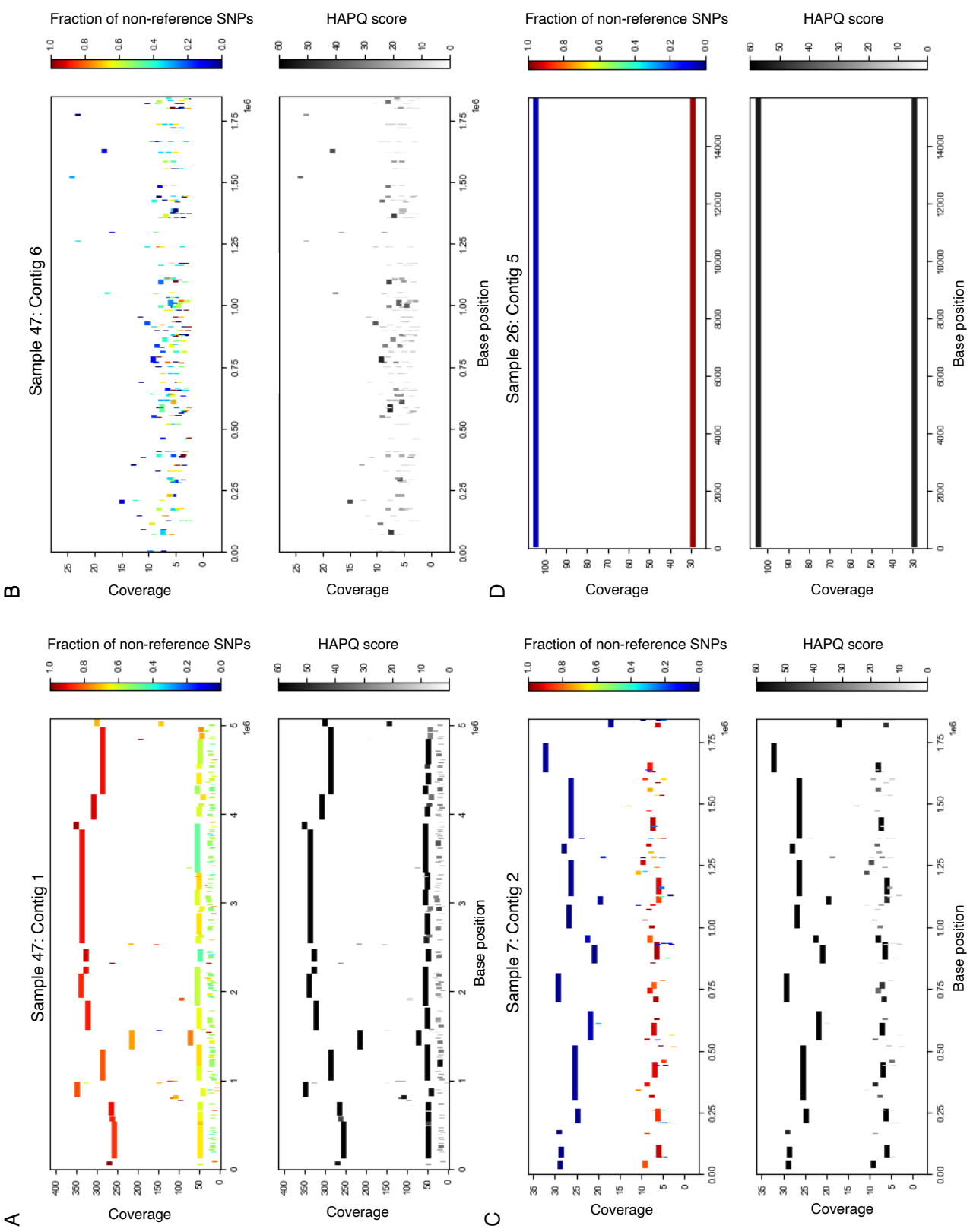


Figure 27: Presence of multiple strain level read clusters suggest high within-sample genetic diversity in samples with elevated species level strain count (SSN). (**A**) Coverage, reference SNP fraction and HAPQ scores are visualised for contig 1 in sample 47. The bars represent continuous stretches of reads mapped to the assembly along the base range on the x axis, with the read coverage shown on the y axis. In the upper plot, the colour scale indicates the fraction of non-reference SNPs for each read stretch and in the lower plot the gray scale indicates the haploset confidence score by HAPQ. **(B-D)** Plot layout identical to (A), but inspecting different contigs. This figure was generated using a script provided with the Floria software repository.[208]

4.5 Discussion

In this Chapter, we have demonstrated that it is important to consider within-sample diversity when predicting drug resistance using WGS in *M. tuberculosis* and have made a start at showing its relevance in other pathogens.

4.5.1 Main conclusions

In *M. tuberculosis*, although heteroresistance is usually detected by phenotypic AST, targeted sequencing approaches and rapid molecular tests (e.g. Gene Xpert MTB/RIF), WGS approaches often fail to resolve subpopulations. But by lowering the minimum FRS required to call a RAV from 0.90 to 0.05, we can significantly improve the sensitivity of WGS-based genotypic resistance classification by +2.1% to 96.4%, with no detectable effect on specificity. A higher sensitivity for resistance detection is likely to improve treatment decisions and lead to more effective surveillance.

Importantly, this adjustment also takes rifampicin prediction on this dataset above the required minimum threshold of 95% with respect to both sensitivity and specificity to pass the ISO standard for antimicrobial susceptibility test devices.[137] It would therefore seem sensible for WGS workflows that process *M. tuberculosis* samples to identify and call rifampicin RAVs if their presence is supported by just a few reads, regardless of the fraction of read support. Resistant subpopulations are likely even more relevant in second line drugs, where larger gaps exist in sensitivity performance. It has for instance been shown that including subpopulations in moxifloxacin resistance can improve sensitivity from 85.4% to 94.0%.[196]

Additionally allowing the presence of high-confidence CMs in the RNA polymerase to predict rifampicin resistance by association significantly reduces the FNR by 2.19% absolute percentage points, to 3.49%. This corresponds to 39% of very major errors, i.e. rifampicin-resistant samples classified as susceptible now being correctly classified by the WGS approach. In terms of the sensitivity of rifampicin resistance prediction, reducing the minimum FRS threshold to 0.05 and permitting the presence of CMs to predict rifampicin resistance in *M. tuberculosis* raises sensitivity further to 96.5%.

As noted above, it is vital that any list of CMs is accurate if they are to be used to infer resistance.[138]

The mutation *rpoC* F452L appears to not be associated with resistance when present at lower FRS, which could either be an error or point to an inverse causal relationship between this CM and the presence of RAVs. The mutation *rpoC* F452L is also homoplasic as per the investigation in Chapter 3, and is hence not associated with a specific sublineage only. This raises the question of whether F452L should be considered a CM or rather provides an advantageous genetic background for the acquisition of rifampicin resistance, by raising fitness *prior* to resistance emergence. Whilst speculation, this is supported by the observation that *rpoC* F452L was shown to stabilise the open promoter complex and increase the elongation rate of the *M. tuberculosis* RNA polymerase not only in the resistant *rpoB* S450L mutant, but also in the wild type.[151]

We can also gain insight into the timing and origin of resistance to rifampicin by examining the samples with resistant subpopulations. A fraction of the significantly lower prevalence of compensation in heterogeneous samples can be explained by the lower prevalence of the mutation *rpoB* S450L, which is known to acquire compensation and go to fixation very quickly. However, it is likely that other factors are at play too, such as that resistant heterogeneous samples have acquired resistance recently and are hence not compensated.

The number of other heterogeneous mutations in the heterogeneous samples hints at the probable sources of resistance. The striking level of genetic diversity detected within some of these samples is only achievable if the patient was subject to a secondary infection, which likely introduced the resistant subpopulation. The significant association of CMs with samples containing multiple (sub)lineages is most consistent with the secondary infection scenario, since the secondary infecting strain could already have acquired CMs before being transmitted.

We additionally made sure that the RAVs in the respective samples are in phase with one of the subpopulations and hence (sub)lineages in the sample, which ensures resistance was not acquired post secondary infection. With this in mind, our analysis is consistent with secondary infection, as opposed to within-host evolution, being the root cause of rifampicin resistance in *at least* 28% of the heterogeneous samples. This agrees with a longitudinal study which found that out of 107 patients with recurrent tuberculosis,

54 (50.5%) had become resistant to rifampicin via a secondary infection[212]. The finding is also relevant in light of increasing evidence that primary transmission, rather than the within-host *de novo* evolution of resistance is the main driver of MDR-TB.[164;213;214] Studying subpopulations can hence give valuable insights into resistance spread.

Stepping back, it is not surprising that our dataset contains genetically heterogeneous *M. tuberculosis* samples due to how clinical samples are grown and colonies extracted. For example, for *M. tuberculosis* WGS, aliquots taken from positive MGIT tubes usually contain multiple 'crumbs' and therefore multiple colonies. This makes it more likely that any subpopulations present are sequenced. One should therefore, perhaps, consider current mycobacterial WGS workflows as inherently metagenomic.

In contrast, the routine laboratory protocol for sequencing most other pathogens will not identify resistant subpopulations since single colony picks from solid media are used as a main basis for the diagnostics workflow. By analysing the metagenomic bloodstream infection dataset by Govender *et al.*[136] we found that one of the 52 samples from *E. coli*/ *K. pneumoniae* infections showed multiple strains within the same sample. In an additional quarter of samples (13/52), we observed substantial within-sample genetic diversity, although not necessarily sufficient to indicate the presence of multiple different strains.

Increased within-sample diversity is suggested by the presence of multiple strain level read clusters in the sample contigs, as detected by the haplotyping tool Floria. These read clusters lead to an elevated species level strain count (SSN), a metric used to quantify the number of strains per species in a metagenomics sample. Overall, the evidence suggests that significant diversity is present in over 25% of samples. This agrees with previous evidence for within-host genotypic and phenotypic diversity in both *E. coli* and *K. pneumoniae* infections.[194;195] Single colonies will not pick up on within-host diversity, hence clinically relevant resistance markers might readily escape routine detection.

In conclusion, this Chapter has demonstrated how to improve the sensitivity of WGS-based rifampicin resistance prediction in *M. tuberculosis* by including subpopulations, CMs, and population diversity, while also providing useful information on how resistance emerges and spreads clinically. We also provide preliminary evidence for within-host diversity in other types of infections, specifically *E. coli* and *K.*

pneumoniae infections, and hence challenges the single colony pick approach in current routine diagnostics. This might have implications for the development of future WGS diagnostics workflows involving these pathogens. In general, these findings will facilitate improved diagnostic strategies and more effective detection and management of drug-resistant infections.

4.5.2 Limitations

Despite the increase in sensitivity brought about by including CMs and lowering the FRS threshold used to identify RAVs, there remains discordance between genotypic and phenotypic AST in *M. tuberculosis*. The remaining discordance can likely be explained by several factors: there may be additional, unknown and most likely rare resistance mutations in the RNA polymerase that have not been classified as such by the second edition of the WHO resistance catalogue. Tackling this problem is non-trivial but could involve the use of machine learning models trained to predict the effect of individual *rpoB* mutations.[215] There are almost certainly errors in the laboratory processes, such as mislabelling or measurement error, that we can minimise but not eradicate. In addition, there is evidence that microheteroresistance (resistant subpopulations with an FRS$<$0.05), which is not captured here, could be indicative of the subsequent development of phenotypic drug resistance in clinical practice.[216]

We found evidence of elevated within-sample diversity, as measured by SSN, in more than a quarter of blood culture samples obtained from patients with *E. coli* and *K. pneumoniae* bloodstream infections. However, the exact magnitude of this effect should not be used to infer the fraction of any future bloodstream samples likely to show increased SSN. This is mainly because our sample set was very small, especially for *K. pneumoniae*, where we only had eight samples with pure infections available. To make any predictions of the likelihood of encountering high within-sample diversity in any additional samples we would need to test a larger sample set, ideally gathered by more than one hospital.

We also cannot make exact predictions about the number of strains present, since detecting multiple clusters of reads could readily occur within one strain after some generations of within-host evolution. However, we are not interested in the number of present strains per se, but rather in the clinically relevant within-sample diversity, such as resistance mutations present in only one of the haplotypes.

WGS-based resistance prediction resulting from single colony picks might not pick up on these, hence leading to treatment failure. The metagenomics approach on the other hand can probably detect clinically relevant within-host diversity due to taking samples directly from blood culture bottles, without pre-selection. But to evaluate this, we need both paired WGS-based resistance prediction results from single colony picks and high quality phenotypic AST data that is able to pick up on subpopulations, as a gold standard.

It should also be noted that the within-sample diversity detected through 'metagenomic' approaches in both *M. tuberculosis* and *Enterobacteriaceae* samples does not necessarily represent the entire within-host diversity. This is due to selection bias, which is introduced through the sampling process, possible pre-culturing steps, and even the sequencing process itself, which may favour certain sequences for amplification.[112] In addition, contamination can introduce traces of other pathogens that map to our assembled reference, especially in regions that are highly conserved between species. This in turn would increase the detected strain count in the affected contigs. Removing contaminating reads or masking highly conserved regions could be a way to prevent this.

4.5.3 Outlook and future work

We have shown that in *M. tuberculosis* infections, lowering the FRS threshold required to call variants significantly improves sensitivity of detecting rifampicin resistance. Since we know that lowering the FRS threshold also has a significant effect in fluoroquinolone antibiotics resistance prediction, it would be promising to quantify this for other drugs used in *M. tuberculosis* infection treatment. Additionally, as mentioned in Chapter 3, there is also evidence of compensation in fluoroquinolone and isoniazid resistant samples. Investigating the combined effect of lower FRS thresholds and allowing CMs to predict resistance could hence be considered for other drugs as well, potentially decreasing the number of very major errors and hence improving treatment outcome and decreasing resistance spread through better surveillance.

The phasing of RAVs with the FRS of subpopulations in heterogeneous *M. tuberculosis* infections allowed us to deduce which subpopulation introduced resistance (Figure 24). It might be possible to assign

other heterogeneous mutations to the two or more co-infecting populations by using the same approach. Mutations from the same subpopulations would be expected to cluster together, and hence should be identifiable in those plots (e.g. Figure 24A). Attributing mutations to one of the clusters should be possible as long as the subpopulations are separable and not overlapping (e.g. Figure 24B).

To evaluate whether resistance present in subpopulations is relevant in other pathogens, we plan to build further on our work on elevated within-infection diversity in samples from *E. coli* and *K. pneumoniae* bloodstream infections. If we could show that the diversity in these samples affects regions involved in resistance, it is likely that the diversity is not always detected in single colony pick based WGS workflows.

Initially, we could check the contig regions covered by more than one haplotype in the metagenomic assemblies. If these are flagged by ResFinder or other resistance detection software tools, it remains to be tested whether single colony picks can accurately detect resistance in those regions. Testing this could involve comparing the performance of genotypic resistance prediction made based on the metagenomics workflow to the prediction made based on the single colony pick workflow, which was unfortunately not performed as part of the study by Govender *et al.* To accurately evaluate resistance prediction performance, we would additionally have to collect phenotypic AST data that reflects the genetic diversity in the entire sample as the gold standard.

Unfortunately, we encounter the same problem as in genotypic AST: The routine workflows for phenotypic AST mainly use single colony picks. It would hence be interesting to explore the option of using the phenotypic AST equivalent of metagenomics in genotypic AST, such as direct rapid antimicrobial susceptibility testing (RAST) approaches.[132] Direct RAST, as proposed by EUCAST, is a phenotypic AST approach amenable to blood stream infection samples. It can be applied directly to the positive blood culture and uses disk diffusion. Since it uses the blood culture directly, it should be able to capture within-infection diversity. Comparing these AST results to the routine phenotypic AST based on single colony picks could be part of future work.

5 Machine Learning for structure-based resistance prediction

This chapter evaluates the usefulness of machine learning-based antibiotic resistance predictions for complementing existing rules-based approaches. The model-based approaches here will attempt to learn the effects of resistance on the structure and physicochemical properties of the target protein, and implicitly learn to predict the effect of novel mutations. The working example will be fluoroquinolone resistance in *M. tuberculosis* and *E. coli* infections.

5.1 Introduction

With the availability of large whole-genome sequencing (WGS) datasets paired with high-quality phenotypic data, 'rules-based' classification for resistance prediction based on current, curated knowledge (often captured in catalogues of known resistance-associated mutations) became possible. This is routinely used in clinical settings for *M. tuberculosis* infections[217–219] and for surveillance and research into Gram-negative infections.[133;134] Yet, even for *M. tuberculosis*, where this approach is more established, whole genome sequencing antimicrobial susceptibility testing (WGS-AST) paired with a catalogue is unable to return a definitive result if the WGS data contain novel mutations[172] and underperforms in sensitivity for new and repurposed drugs.[220]

Model-based antibiotic resistance prediction

While rules-based WGS-AST performs well for resistance prediction in a range of pathogen-drug combinations, its predictive power is based entirely on statistical association between resistance alleles and phenotype. It hence assumes that the entire genetic basis of the resistance phenotype is known, i.e. that all resistance markers have been identified. In addition, the rules-based approach usually ignores the possible interaction of multiple resistance loci, such as through epistasis. This is because the interaction of resistance alleles becomes a combinatorial problem, necessitating the analysis of large and diverse datasets to detect the relationship of different mutations and the phenotype. In the same vein, strain background can influence how the resistance allele affects the phenotype, which is also usually not considered.[75] More sophisticated statistical methods can be applied to increase the performance of rules-based approaches, but it will always remain an inferential method and therefore not be truly predictive.

To mitigate this problem, model-based approaches in the form of traditional machine learning (ML) classifiers, such as support vector machines, logistic regression and random forest have been applied to sequencing data to predict resistance in *M. tuberculosis* and Gram-negatives.[220–222] These are trained on paired sequencing and phenotypic data. The sequencing data can be raw reads or detected variants and the phenotypic data can be categorial or minimum inhibitory concentrations (MICs). If given the complete genetic information as input, model-based approaches can learn relationships between genetic features and phenotypes that might not be captured using the rules-based approach. Deep learning approaches are especially well equipped to learn complex association patterns. They have therefore been used to predict resistance and show good performance, complementing existing rules-based methods.[223–225]

However, the availability of diverse and high quality data for model training is a necessary condition when employing ML algorithms. In particular, low bias and high variance in the dataset are essential for producing generalisable models,[226] hence this is also required for predicting resistance with high accuracy. In addition, the majority of model-based approaches to WGS-AST only use the genetic sequence information of resistance-associated alleles and do not make direct use of the structural and physicochemical information available for the affected amino acids. However, this information might help in predicting the effect of novel mutations.

Using protein structure to infer resistance

Protein structural information is especially relevant for antibiotics whose mechanism of action relies on inhibiting specific bacterial protein targets by binding to them, such as rifampicin and the fluoroquinolones. These bind to the RNA polymerase and the DNA gyrase, respectively, and inhibit their essential function.[145;227] Consequently, the resistance mechanisms that arise in bacteria treated with these antibiotics, such as *M. tuberculosis* and *E. coli*, often involve structural changes to the antibiotic target protein to prevent binding of the drug.[227;228] How spatially close certain protein moieties are to the bound drug can hence be crucial for resistance prediction. Even amino acids that are distant in sequence can be spatially close to the drug binding site after the nascent amino acid chain reaches the equilibrium folding state of the protein. Unfortunately, the 3D structure of a protein cannot be readily deduced from the

nucleotide sequence without structural biology experiments or the use of sophisticated protein folding prediction tools like Alphafold.[229] But since highly essential proteins like the RNA polymerase and the DNA gyrase do not tend to change their structure much even when acquiring resistance, the wild type protein structure might already contain enough structural information to predict which amino acid mutations will cause resistance.

This assumption was made in the first study that used this structure-based approach to predict resistance to the drug pyrazinamide. This pro-drug binds to a bacterial protein in order to be converted to its active form, pyrazinoic acid. Hence resistance often arises in the gene coding for the protein required to catalyse this transition, the pyrazinamidase.[230] Carter *et al.* showed that the structural and physicochemical features of mutated amino acids in the resistance-associated non-essential gene *pncA* (in addition to sequence-based features) can predict pyrazinamide resistance with relatively high sensitivity and specificity, as well as potentially highlight novel mutations.[231] The model has been applied similarly to rifampicin resistance prediction, although with a lower payoff because rifampicin resistance is mainly caused by only a few different resistance mutations. In fact, they used a dataset of 219 susceptible mutations and only 46 resistance-associated variants (RAVs).[232] This lack of diversity is probably due to the fact that the protein target, the RNA polymerase, is an essential protein. Because of this, the rules-based approach performs really well for rifampicin resistance prediction already.[79] Applying a ML model is hence not as useful for rifampicin resistance as for pyrazinamide resistance, where resistance mutations are found at almost every codon position, making this a much more complex prediction problem and more diverse dataset (664 non-redundant missense mutations, 349 of which RAVs).[231]

However, these simple structure-based ML models are specifically trained to infer the effect of a single mutation on the phenotype. Since the assumption in this model is a causal relationship between a single nucleotide polymorphism (SNP) and the phenotype, with highly penetrant genetic variants, the approach can only consider samples with a single SNP in a resistance-associated gene. Therefore, this model cannot resolve any possible additive or epistatic effects of multiple resistance-associated alleles in a sample, which is similar to one of the issues affecting rules-based approaches to WGS-AMR.

But in contrast to model-based resistance prediction based on sequencing data alone, the approach allows protein structural and chemical information to be integrated and therefore uses information beyond the presence (or absence) of catalogued resistance mutations. Applying this to other resistance-associated genes could give insight into the importance of structural and chemical information for resistance prediction based on WGS data. The limitation to single missense mutations warrants applying this approach to organisms with limited variability in their genome and low mutation rate, such as *M. tuberculosis*.

Representing target proteins as graphs in graph convolutional network classification

If we want to consider the structural effect of multiple mutations per gene, or even a different allele, on the phenotype of a sample, a model must be trained using the structure of the entire target protein and the associated resistance phenotype. This captures all possible interactions between different resistance loci in the same gene, while still integrating both structural and sequencing information in a predictive model.

We can theoretically achieve this using a deep learning approach, graph convolutional networks (GCNs). GCNs use the classic convolutional neural network (CNN) architecture for processing data, where complex association patterns are learned by propagating data from an input layer through several hidden layers of a neural network (Figure 28). The layers apply convolution and pooling operations, which transform and sub-sample the original data. A classifier is then applied at the end to predict a binary or multi-class outcome.[233] By training the network on labelled data, the CNN can learn to predict sparse outcomes, such as resistance or susceptibility, based on complex input data, represented as large numerical matrices. The ability to use complex input data is ideal for processing images, because these are inherently represented as numerical matrices denoting the intensity of the RGB colours. This is why one of the first larger applications of CNNs was image classification (Figure 28).[234]

When graph-structured data is used as input for a deep neural network, we move from CNNs to GCNs.[235;236] Graphs can represent any connected data, such as social, transportation, and computer networks.[237] Even proteins can be represented as graphs and then input into a neural network to predict a specific outcome. GCNs have been applied to predicting protein binding and protein function, using both the protein structure and sequence as input.[238;239]

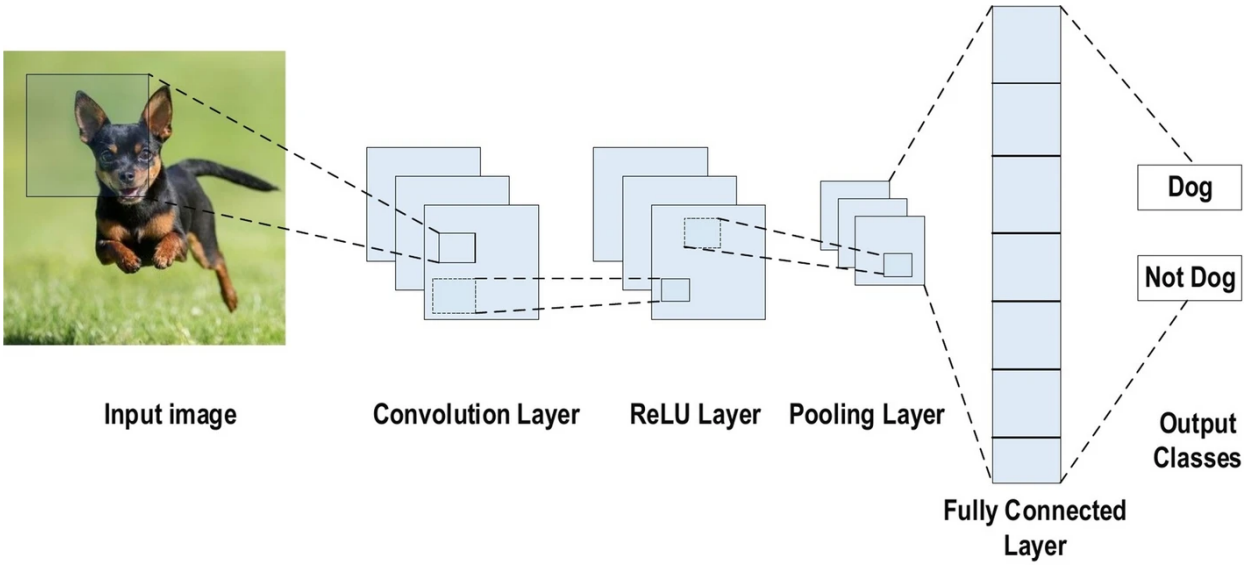


Figure 28: Exemplary overview of a convolutional neural network (CNN) architecture for image classification. The input image is treated as three matrices, representing the RGB colours. Several layers of convolution, activation (ReLU) and pooling are applied before the label is predicted from a fully connected network layer. This figure was reproduced from Alzubaidi *et al.* [233]

The most straightforward approach to encode a protein from the protein data bank (PDB) as a graph is for the graph nodes to represent single amino acids, which have an associated node vector listing specific physicochemical amino acid characteristics (Figure 29). Nodes are connected by edges if the amino acids are within a specified distance in the protein structure. The protein graph is then transformed into two matrices, one containing edge connectivity information (adjacency matrix) and one containing the feature information of each node (feature matrix). Similar to image classification, the resulting matrices can be used as input to a CNN.

In our case, we can try applying this to proteins that are targeted by an antibiotic of interest and harbour one or more mutations. The protein graphs will then be fed to a neural network to train a classifier for predicting resistance, taking into account the entire protein structure, hence the approach is applicable to alleles with more than one SNP. This will potentially enable us to train ML models capable of predicting

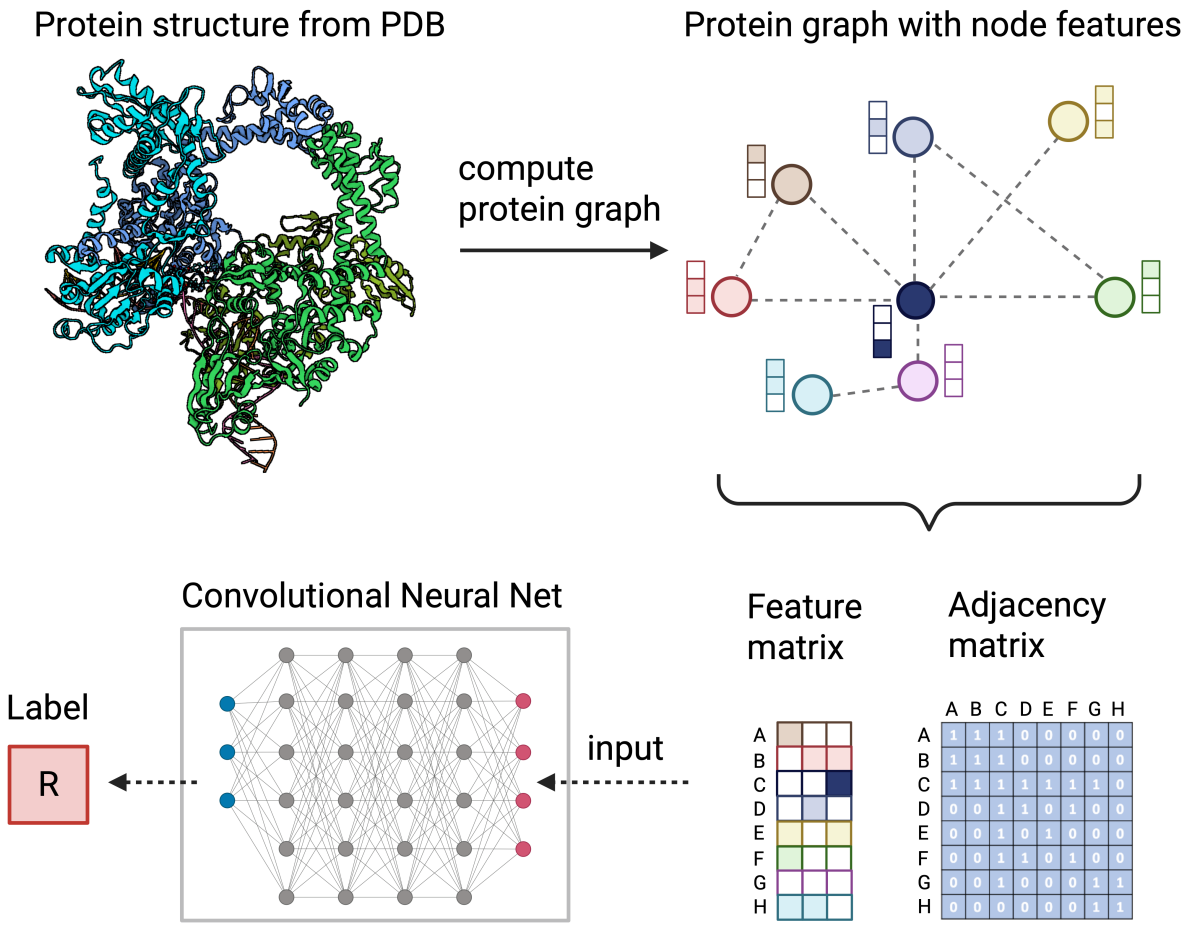


Figure 29: Exemplary illustration of a protein encoded as a graph processed by a graph convolutional network. This example is based on the DNA gyrase structure, but can be implemented similarly for many other proteins that show SNPs conferring resistance to antibiotics, including among others the RNA polymerase, pyrazinamidase and katG catalase-peroxidase.

resistance in organisms with higher genetic variability, e.g. *E. coli*.

Using this architecture, we can move away from using merely genetic sequencing data and also from using structure-based approaches that can only consider single mutations, towards an integrated approach that considers alleles and the structure of the proteins they encode for resistance prediction. Furthermore, the weights assigned within the GCN might reveal sub-structures of the protein that are most predictive of resistance, thereby improving our understanding of resistance mechanisms.

Fluoroquinolone resistance

To evaluate the two structure-based approaches to resistance prediction, we will use the example of fluoroquinolone resistance. Fluoroquinolones target a specific bacterial protein, the DNA gyrase, in many pathogens including *M. tuberculosis* and *E. coli*, and resistance arises mainly through mutations in this protein.[227] While *M. tuberculosis* has relatively few RAVs in the DNA gyrase, *E. coli* may show a higher variability, making these two interesting cases to test out first a simple machine learning and then a deep learning approach, the latter being capable of capturing more complex patterns.

The enzyme DNA gyrase is a type IIA topoisomerase, an essential protein that only exists in bacteria, making it an ideal drug target.[240] Topoisomerases modulate the topology of DNA and as such are involved in several essential processes in bacterial cells, among them DNA replication and transcription.[241] The DNA gyrase introduces topological changes that allow supercoiling or relaxation of the DNA molecule.[242] The protein complex consists of two subunits each of GyrA (*gyrA*) and GyrB (*gyrB*) (Figure 30A), which form a heterotetramer.[243] The fluoroquinolones bind to the protein close to the active site. This is where DNA is bound and cleaved in a functioning protein (Figure 30B). The drug binding traps the protein in an intermediate state and hence disrupts its function.

The majority of fluoroquinolone resistance mutations in *M. tuberculosis* are located in the 'quinolone resistance-determining region' (QRDR) on GyrA near the active site. In *E. coli*, there are two types of topoisomerases with redundant functions: The DNA gyrase and the topoisomerase IV, encoded by the genes *parC* and *parD*. Fluoroquinolone resistance in *E. coli* can hence arise through mutations in these genes as well. And while chromosomal mutations in the genes *gyrA/B* and *parC/D* are a common cause of resistance, it can also be caused by decreased uptake of the antibiotic or increased production of efflux pumps.[227] In addition, plasmid-mediated resistance to quinolones has been described in *E. coli* as well,[126] e.g. through the *qnr* gene, which encodes a pentapeptide. This molecule in turn blocks the action of quinolones on the DNA gyrase and topoisomerase IV.[227] In some cases, and especially for *E. coli*, we will therefore not be able to explain all resistance by just considering the DNA gyrase, again suggesting that these model-based approaches should be used to complement existing resistance prediction methods, rather than replace them.

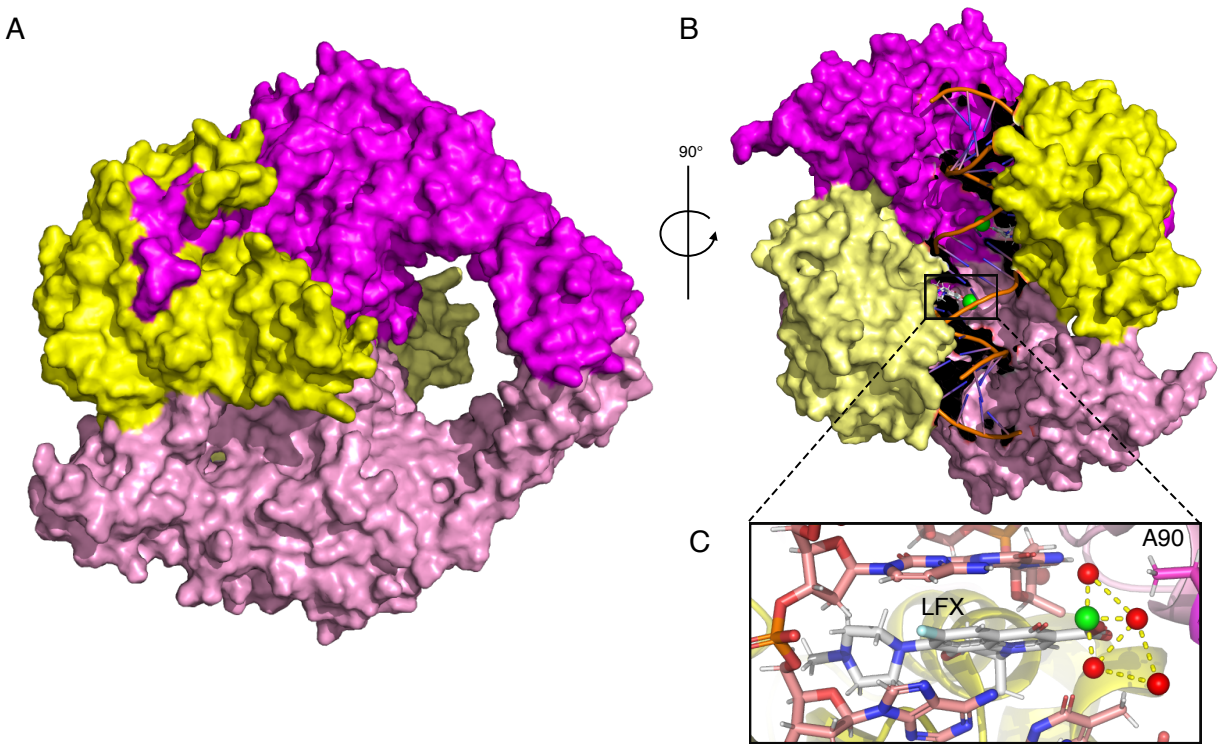


Figure 30: DNA gyrase structure with bound levofloxacin in *M. tuberculosis* as determined through crystallography by Blower *et al.* (protein data bank (PDB): 5BTG).[243] (A) The DNA gyrase is made up of four subunits, two of each GyrA and GyrB, forming a A2B2 heterotetramer. The GyrA subunits are shown in different shades of magenta, the GyrB subunits in different shades of yellow. **(B)** The DNA gyrase forms a clamp around the double stranded DNA (dsDNA) when engaged. The protein complex crystallised by Blower *et al.* has additionally bound levofloxacin (LFX, white) close to the active centre magnesium ion (green). Since there are two identical subunits each, the drug can bind in two congruent locations on the protein. **(C)** LFX binds close to the magnesium ion in the active site, by wedging between the bases of the dsDNA, inhibiting the protein function. The DNA chain is shown in light pink, the active site magnesium in green, and the water molecules in red, with the coordinating water network shown through yellow dashed lines. In *M. tuberculosis*, the engagement of the Alanine 90 (A90) amino acid with the coordinating water network around the active site magnesium is weak, whereas in *E. coli* a serine establishes close contact.[243]

As the function of the DNA gyrase is essential, its structure remains highly conserved throughout different bacterial species, although there are small organism- or clade-specific differences in both subunits.[244] These small differences however, such as the replacement of a serine for an alanine in the QRDR of *M. tuberculosis* DNA gyrase (Figure 30C), cause the *M. tuberculosis* DNA gyrase to be less susceptible to fluoroquinolone drugs than the DNA gyrase in *E. coli*.[245] Despite this, fluoroquinolones remain an important second-line treatment option for *M. tuberculosis* infections.

Determining the structure of the drug-bound DNA gyrase complex experimentally is difficult because of the flexibility of the protein. It has so far only been resolved for *M. tuberculosis*, and not for *E. coli*.

Blower *et al.* used a fused version of the *M. tuberculosis* subunits (GyrBA) to crystallise the protein-DNA-fluoroquinolone complex for the first time, permitting the structure to be elucidated using X-ray crystallography.[243] Their work showed that the fluoroquinolone molecules are wedged between two bases in the cleaved double-stranded DNA (Figure 30C). Interestingly, they found no substantial differences in the contacts formed between the complex and different fluoroquinolones, despite observed differences in MIC.[243]

5.2 Aims of this chapter

In this chapter, we will apply two different ML algorithms to resistance prediction using the example of fluoroquinolone resistance in *M. tuberculosis* and *E. coli*. Both approaches will attempt to learn which structural and physicochemical aspects of fluoroquinolone resistant DNA gyrase structures are indicative of resistance. The first approach is a simple ML classifier to predict resistance based on single mutations in the DNA gyrase of *M. tuberculosis* and the second uses deep learning through GCNs to predict the effect of multiple mutations in the *E. coli* DNA gyrase on the resistance classification.

We will evaluate this by looking at levofloxacin and moxifloxacin resistance for *M. tuberculosis*, since the CRyPTIC dataset has measured MICs for those two fluoroquinolones. In *E. coli*, we will only consider levofloxacin resistance, since the clinical data that we eventually want to use to evaluate performance only measured levofloxacin MICs.

The first approach will use the CRyPTIC dataset for training and testing the simple ML approach. A limited number of samples will be eligible for processing, because we can only use samples with one and only one mutation in either *gyrA* or *gyrB*. Hence, we will use the WHO catalogue version 1 to add additional known susceptible and resistance mutations in an attempt to inflate our dataset.[202]

The deep learning approach will be entirely based on *synthetic* datasets, with resistance mutations from the literature introduced into *E. coli* DNA gyrase alleles. This will therefore not yet be directly applicable to the real world, but does allow us to evaluate whether a GCN model can predict resistance based on the protein sequence and structural data alone.

The aims of this chapter are hence two-fold:

1. Test whether simple ML classifiers can predict fluoroquinolone resistance based on single mutations in the DNA gyrase of *M. tuberculosis*

2. Investigate whether a deep learning approach (GCN) can learn to predict fluoroquinolone resistance using the entire DNA gyrase structure in *E. coli*

5.3 Methods

The simple machine learning analysis in this Chapter can be reproduced using a GitHub repository and attendant Python3 Jupyter notebook available online.[168] The relevant analysis can be rerun locally on a user's computer or in the cloud using the corresponding google colab button in the README. Due to the heavy computation required to run some of the models, the latter might take a long time to run. The GCN prediction model analysis is even more heavyweight and was additionally documented using the Weights & Biases API and web interface[246] and hence cannot be rerun in the cloud using Google colab.

Simple ML for *M. tuberculosis* fluoroquinolone resistance prediction

Dataset sources

The dataset for training the simple ML models is derived from two published datasets for *M. tuberculosis* variants with associated phenotypes: the CRyPTIC dataset version 1[140;141] and the WHO catalogue of *M. tuberculosis* mutations version 1.[202] The CRyPTIC dataset version 1.1.1[142] contains matching genotypic-phenotypic data for 41,130 samples, whereas the WHO catalogue lists the associated phenotype for a number of genetic variants in *M. tuberculosis*, according to statistical evaluation and/or expert rules.

We created the overall dataset by filtering the CRyPTIC catalogue for samples with exactly one missense SNP mutation in either the *gyrA* or *gyrB* gene, excluding lineage markers. Out of the 3,974 samples that remained, 2,441 had high-quality phenotypic data available. We then split the dataset into a moxifloxacin and a levofloxacin dataset, according to the drug that the sample was tested for. We calculated the label support for each mutation seen in the datasets as the percentage agreement of the phenotypic label in all samples with the respective mutation. We then assigned the phenotype of the samples with this mutation to be the label with the higher label support, and dropped any samples with a label support of 50%, since we cannot determine the majority phenotype. Assuming that the higher frequency label portrays the true phenotype is a simplification, but we need to eliminate any discrepancy in the predicted phenotype in order to obtain a direct mutation - phenotype association for the overall datasets. After filtering, the resulting datasets contained samples with only one single non-lineage mutation in the DNA gyrase and a binary phenotypic label.

Lastly, we merged the WHO catalogue data with the curated CRyPTIC datasets, by adding any DNA gyrase SNP mutations that are not already present in the CRyPTIC data. We end up with a moxifloxacin dataset consisting of 1700 rows, of which 1665 are derived from CRyPTIC samples and 35 are reported WHO mutations. For levofloxacin, we obtain a dataset consisting of 1759 rows, of which 1723 are CRyPTIC samples and 36 are WHO mutations. These are our datasets for the sample-based approach. To obtain the datasets for the mutation-based approach, we aggregate the datasets by mutation. This will deduplicate the CRyPTIC dataset and leave us with one single row per mutation for the entire dataset.

Feature construction

Structural and physicochemical features were calculated for each mutation seen in the datasets. We used the Python package 'sbmlcore' to calculate these features.[247] It requires the gene mutation and the protein structure as input to calculate a number of mutation-specific features. We used DNA gyrase protein structures from the PDB, determined by crystallography by Blower *et al.* (PDBs: 5BS8 for moxifloxacin and 5BTG for levofloxacin-bound DNA gyrase).[243]

The Python package sbmlcore calculates a range of features, some of which are implemented as the absolute difference in value upon change of the amino acid in the protein. These were amino acid volume, hydropathy, molecular weight and isoelectric point. Another set of features involves the distance to important moieties in the PDB file. These include distance to the bound drug, and distances to the two active site magnesium ions, both measured as distance from the C_α of the mutated amino acid to the centre of mass of the relevant structure. Lastly, a few additional external tools were automatically run to calculate features based on the protein PDB. Stride allows calculation of the secondary structure of the protein, including ϕ and ψ protein backbone angles,[248] FreeSASA calculates the change in solvent accessible surface area,[249] SNAP2 is itself an ML model and predicts whether the mutation has a neutral or deleterious effect on the protein function,[250] and DeepDDG predicts the overall change in protein stability.[251] Additionally, the average distance of the centre of mass of the mutated amino acid from the solvent accessible surface was calculated, as well as a temperature factor-based indicator for how flexible or rigid the particular amino acid position is in the protein structure.

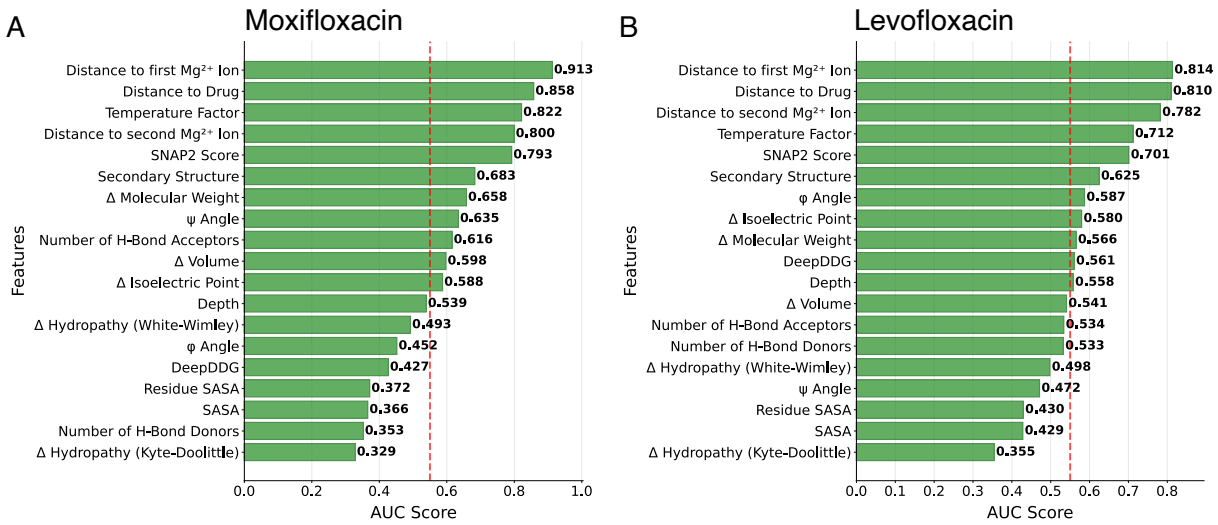


Figure 31: Features calculated for each amino acid mutation in the *M. tuberculosis* DNA gyrase. (A) The features calculated for each mutation in the moxifloxacin dataset. They are shown in order of their area under the receiver operator curve (AUC). This represents their ability to predict moxifloxacin resistance on their own in a univariate logistic regression. The red dashed line indicates the arbitrary cut-off at AUC=0.55, to keep the feature space sparse for model training. **(B)** Plot structure similar to A, but the features here were calculated for each mutation in the levofloxacin dataset and evaluated for levofloxacin resistance prediction.

Some features could not be calculated for all mutations, leading to a reduction in sample size to 1695 samples containing 153 unique mutations for the moxifloxacin dataset and 1749 samples containing 172 unique mutations for the levofloxacin dataset.

Feature selection and model training

We calculated a suite of 19 features for each missense mutation seen in the dataset. After scaling the features, we performed feature selection to assess which are most predictive of the phenotypic label, to help make the model as sparse as possible. This was done using a univariate logistic regression with a pooled 3-fold cross-validation using each individual feature for resistance prediction. The area under the receiver operator curve (AUC) of each feature was calculated. An AUC of 0.5 indicates that classification using the respective feature is no better than random chance, so we set an arbitrary cut-off at 0.55 to keep the feature space sparse (Figure 31), leaving us with 11 features per model. The distances to the active site magnesium ions and the bound drug were the most predictive of resistance on their own.

We evaluated two different ML models for resistance prediction using these features: a simple logistic regression (LR) and Extreme Gradient Boosting decision trees (XGBoost). These ML models showed the

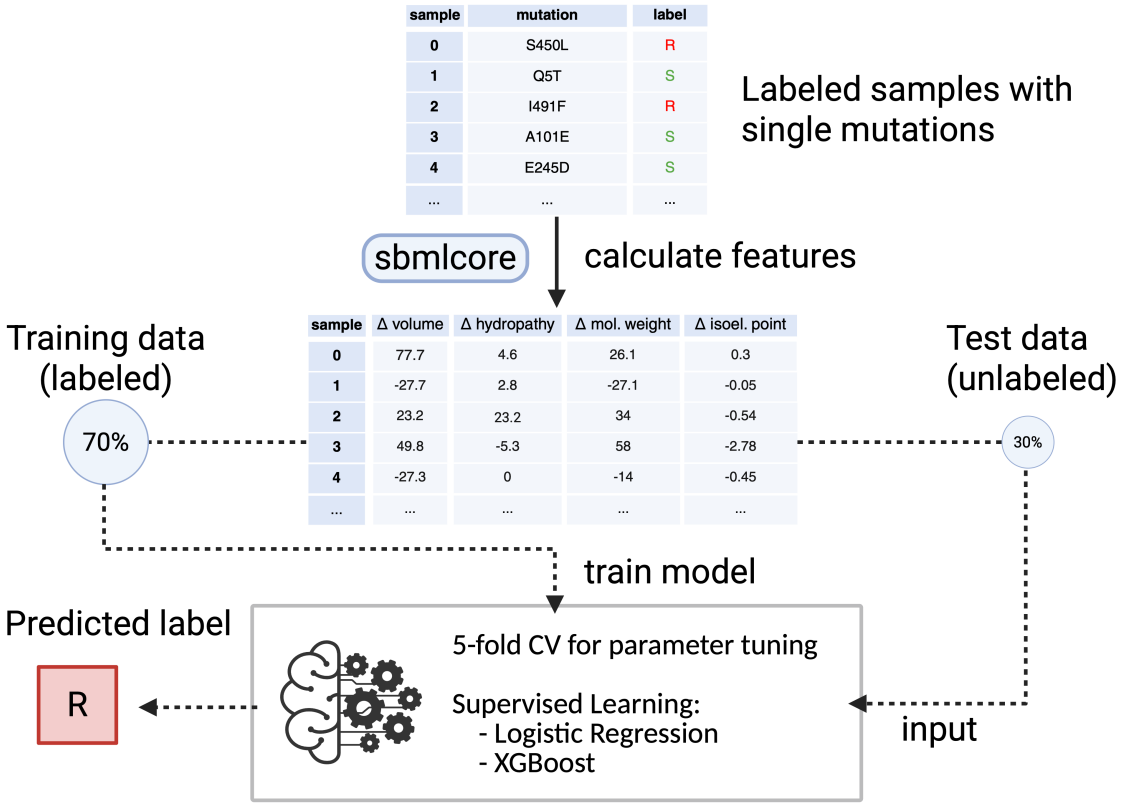


Figure 32: Simple machine learning approach to resistance prediction based on the structural impact of single mutations. Features are calculated for all missense mutations seen in the dataset of labelled samples with a single mutation in the DNA gyrase protein. The Python package sbmlcore calculates features based on structural and physicochemical characteristics of the amino acid affected by the mutation. The resulting dataset of sample IDs associated with a suite of 11 features is split into training and test sets and the training set is subjected to a 5-fold cross validation for parameter tuning for either logistic regression or XGBoost. The best performing model is selected for evaluation on the test set. If the mutation-based approach is used, the train - test split is performed 3 times to use the entire data for training and obtain an average performance over the 3 folds.

best performance in previous studies of rifampicin and pyrazinamide resistance prediction.[231;232] Since we want to minimise the number of false negatives, i.e. the number of very major errors (VMEs), we optimised all models for recall (sensitivity). Sensitivity and specificity were used as the readout for performance on the test set.

The data were split into training and test sets in a 70:30 ratio. For the sample-based approach, we performed a 5-fold cross-validation to tune the model hyperparameters via a grid search on the training set, and evaluated the performance on the held-out test set (Figure 32). This is best practice since we have sufficient samples available to keep the test set completely separate.

For the mutation-based approach, we used a nested cross-fold validation that performs a 3-fold cross-validation on the outer fold (the train - test split) and a 5-fold cross-validation on the inner fold (for hyperparameter tuning). The inner fold operation is identical to the sample-based approach, and the outer fold allows to run the entire model training 3 times, and evaluate it on a different test set each iteration. We hence use the entire dataset for training, which is necessary since we have very few samples for the mutation-based approach. We also obtain a mean and standard deviation for the test set by averaging over the three runs.

GCNs for *E. coli* levofloxacin resistance prediction

Dataset simulation

We used the Python package package 'sbmlsim' to construct a *synthetic* dataset of levofloxacin resistant and susceptible gene sequences for the *E. coli* DNA gyrase.[252] This package was written as a joint effort by Prof. Philip W Fowler, Dylan Dissanayake and myself.

When supplied with a GenBank file of a haploid organism, it constructs a specified number of alleles with a pre-defined average number of resistant and susceptible mutations per sample for any desired gene in the GenBank file. The number of mutations per sample is drawn from a Poisson distribution with the mean defined by the user. One can supply a list of both resistant and susceptible mutations, or assume that any non-resistant mutations are susceptible. For the latter, the susceptible mutation is drawn randomly from all remaining amino acid positions after resistance mutations have been introduced in the protein. In the randomly chosen wild type codon, any SNPs that introduce a stop codon or a resistance mutation as per the provided list are excluded. The susceptible mutation is then determined by a random one SNP change to the codon and the corresponding change is introduced to the allele.

We used sbmlsim to construct a dataset with levofloxacin resistant and susceptible *gyrA* and *gyrB* alleles. By simulating a dataset for model training rather than using a real-world dataset, we have exact control over class balance and possible resistance mutations (see simulated alleles in Figure 34).

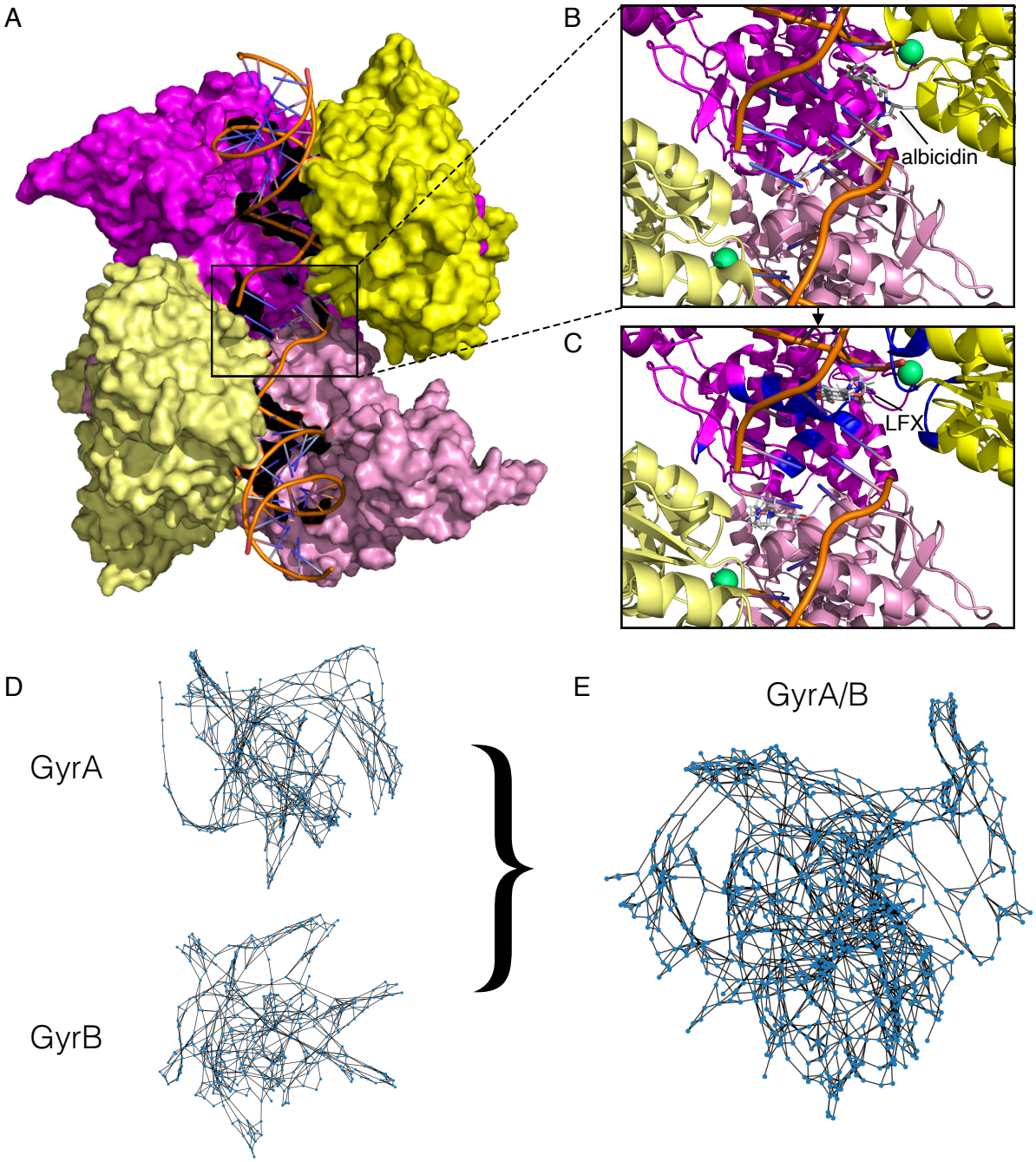


Figure 33: The *E. coli* DNA gyrase structure can be docked with the *M. tuberculosis* levofloxacin molecule and represented as a graph. (A) The DNA gyrase structure with bound albicidin as determined through cryo-electron microscopy by Michalczyk *et al.*[253] Just like in *M. tuberculosis*, the *E. coli* DNA gyrase is made up of four subunits, two of each GyrA and GyrB, forming a A2B2 heterotetramer. The GyrA subunits are shown in different shades of magenta, the GyrB subunits in different shades of yellow. **(B)** Albicidin is stabilising the cleavage complex by engaging with the protein subunits and the bound dsDNA.[253] Albicidin is bound in a similar location near one of the active centre magnesium ions as the fluoroquinolone drugs, but it is a much longer molecule and hence the binding has displaced one of the magnesium ions. **(C)** The albicidin was replaced with the levofloxacin molecule from the structural alignment with the GyrA subunit of the *M. tuberculosis* protein (PDB: 5BS8). The levofloxacin molecule is lodged between two DNA bases just like in the *M. tuberculosis* protein, and is located close to known resistance mutations in *E. coli* (coloured in blue, including known and inferred resistance mutations), indicating a realistic mapping of the drug. **(D)** The protein graph created based on the PDB of the *E. coli* DNA gyrase GyrA and GyrB subunits. Nodes represent the amino acids in the protein, and edges are drawn between nodes whose centres of mass are less than 6.3 Å apart. **(E)** The protein graph for two subunits of the *E. coli* DNA gyrase together. This constitutes only half of the heterotetramer protein, but the other half is identical.

Levofloxacin mapping and graph construction for the *E. coli* DNA gyrase structure

In order to build a GCN model for the DNA gyrase in *E. coli*, we first need to obtain a PDB file of the protein. We would ideally use a structure with a bound fluoroquinolone drug. Unfortunately, there is no *E. coli* DNA gyrase structure bound to any of the fluoroquinolone drugs available through the PDB. We hence decided to perform a structural alignment of the GyrA subunit in the *M. tuberculosis* DNA gyrase with bound levofloxacin (PDB:5BS8, Figure 30) and the *E. coli* DNA gyrase (PDB:7Z9C, Figure 33A). The *E. coli* DNA gyrase has albicidin and dsDNA bound, which are essential for stabilising the cleavage complex for cryo-electron microscopy (Figure 33B).[253] One magnesium ion appears to have been displaced by the bound albicidin drug, hence we only have one active site magnesium in *E. coli*. The alignment was performed using the MultiSeq analysis tool in VMD-1.9.4a57, which uses STAMP.[254;255] We decided to use only the GyrA subunit for alignment, as this gave the best root mean square deviation (RMSD) of 1.53 Å.

After structural alignment, we removed the *M. tuberculosis* protein and the albicidin structure and obtained a levofloxacin bound structure of the *E. coli* DNA gyrase (Figure 33C). The mapped drug molecules are near the most commonly observed resistance mutations in *E. coli* fluoroquinolone resistance (*gyrA* S83 and *gyrA* D87)[256] and are wedged between two bases of the dsDNA, just like in the *M. tuberculosis* complex (Figure 30C). The bound drug is required to measure the distance to the drug, which is calculated for each amino acid in the protein graph automatically by sbmlcore.

To obtain the graph that can be used as input for our GCN, we converted the protein subunits to undirected graphs, where the nodes represent amino acids, as described in detail below. We decided to train a model each for the subunits GyrA and GyrB (Figure 33D), and also a combined model for a graph of both subunits, forming a heterodimer (Figure 33E).

Feature construction

The GCN takes a feature matrix and an adjacency matrix as input (Figure 29). The adjacency matrix can easily be constructed from the protein graph, by computing the pairwise centre-of-mass distances between all residue pairs using MDAnalysis.[257;258] We then applied a distance cut-off of 6.3 Å to define

edges, meaning residues with centre-of-mass distances below this threshold are considered connected. We hence obtain a boolean distance matrix, representing edges between nodes within 6.3 Å of each other. This approach captures the three-dimensional structural context of the protein, allowing the GCN to learn from spatial relationships between residues (Figure 33D,E). The 6.3 Å cut-off was chosen to capture both direct contacts and near-neighbour interactions while maintaining computational efficiency.

The feature matrix contains features associated with each node in the graph, hence also called a node feature matrix. Each amino acid residue in the protein graphs is represented as a node with a nine-dimensional feature vector which contains physicochemical properties and structural information (Figure 34). The node features were computed using the sbmlcore framework and include amino acid volume, molecular weight, hydropathy indices, isoelectric point, the number of hydrogen bond acceptors and donors, the presence of side chain rings and distance to the bound drug. In contrast to the simple ML approach, we hence do not calculate mutation-specific features, but rather amino acid-specific features. This implies that they contain absolute values, not relative differences between the wild type amino acid and the mutated one (Figure 34).

The distance to the bound drug was calculated using the aligned *E. coli* DNA gyrase structure with transferred levofloxacin coordinates from the *M. tuberculosis* structure. This feature provides the crucial spatial context for understanding how mutations at different positions might affect drug binding. All features were normalized using MinMaxScaler to ensure comparable scales across different physicochemical properties and prevent any single feature from dominating the learning process.

Importantly, while node features varied between samples due to different mutation patterns, the edge connectivity remained constant across all samples. We are therefore assuming that the protein backbone does not change between alleles. This simplification allows the model to focus on learning how amino acid substitutions affect resistance while maintaining a consistent spatial context. One could predict the protein structure for each allele in the future by using Alphafold.[229]

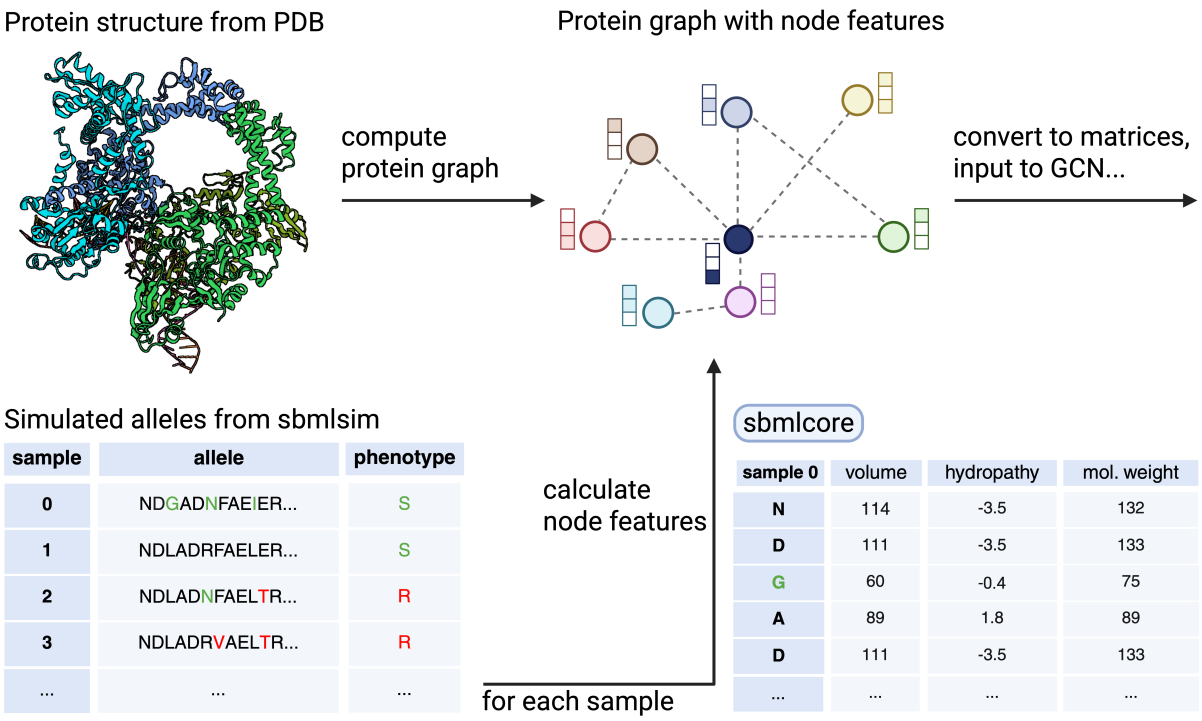


Figure 34: Constructing the feature and adjacency matrix for the graph convolutional network based on the *E. coli* DNA gyrase gene sequences and structures. The adjacency matrix is calculated based on distances between the amino acids in the protein structure from PDB. The feature matrix contains a nine-dimensional vector for each amino acid in the protein. It contains physicochemical properties of the amino acid, as well as structural information based on distance to the bound drug. These features are calculated with sbmlcore, using the simulated alleles from sbmlsim as input. The feature vector can hence be different for each sample, depending on the mutations introduced in the allele.

Model architecture and training

The Graph Convolutional Network was implemented using PyTorch Geometric with a standard architecture for graph classification (Table 12). Similar model architectures are being used across disciplines for different graph classification problems.

The synthetic dataset was split into training and test set using an 80:20 ratio, and model training was performed using a batch size of 64 samples with shuffling on the training set. We used cross-entropy loss as a measure of model training progression, since this is the standard for binary classification. Optimisation was performed using the Adam optimiser with configurable learning rate and weight decay.

The model trains via gradient-based optimization, where the Adam optimizer updates parameters based on the learning rate and the computed gradient from the cross-entropy loss function. A OneCycleLR

layer	dimensions
Input layer	9-dimensional node features
Hidden layers	Three sequential GraphConv layers with configurable hidden dimensions
Normalisation	Batch normalisation applied between the GraphConv layers
Activation	ReLU activation functions after normalisation
Global pooling	Mean pooling to aggregate node-level into graph-level embeddings
Output layer	Linear, binary classifier with dropout regularization

Table 12: Graph convolutional network architecture for levofloxacin resistance prediction. The layers are listed in order of their appearance in the network. The number of hidden channels is a hyperparameter and hence configurable. The last, binary classifier maps to two classes (resistant/susceptible).

hyperparameter	range
Learning rate	1e-7 to 1e-4
Hidden channels	[16, 32, 64, 128, 256]
Dropout rate	0.0 to 0.5
Weight decay	1e-5 to 1e-3
Optimisation metric	F1-score on the test set

Table 13: Hyperparameter space of the graph convolutional network for levofloxacin resistance prediction. The parameter sweeps choose randomly from the available parameter space in each run. The optimisation metric is the F1-score on the test set, to give a balanced score for both sensitivity and specificity. Each model was run 10 times, and the model with the best F1-score was chosen for evaluation.

learning rate scheduler was employed to adjust the learning rate during training. Model performance was evaluated using F1-score as the primary metric for hyperparameter optimization, while also tracking accuracy, sensitivity, and specificity.

The hyperparameters of the model were optimised using Weights & Biases (wandb)[246] random optimisation sweeps with the configuration specified in Table 13.

Three separate GCN models were trained following the above methodology:

1. A GyrA model, using only the GyrA subunit structure and mutations.

2. A GyrB model, using only the GyrB subunit structure and mutations

3. A combined model, using the complete heterodimeric gyrase structure with mutations from both subunits

We tested 10 different, random hyperparameter combinations for each model and chose the best-performing model in terms of F1-score for evaluation.

5.4 Results

5.4.1 The fluoroquinolone datasets show strong class imbalance towards resistant samples but with very few resistance associated variants

We wanted to test whether resistance prediction based on physicochemical and structural features in addition to sequence-based features can complement catalogue-based fluoroquinolone resistance prediction. We have a dataset of paired high-confidence phenotypic and genotypic data for moxifloxacin and levofloxacin resistance in *M. tuberculosis* from the CRyPTIC dataset[140;142;169] merged with additional data from version 1 of the WHO catalogue of *M. tuberculosis* SNP-based mutations.[202] The data was filtered to only include samples with a single mutation in either *gyrA* or *gyrB*, to establish the direct mutation - phenotype association. For the same reason, samples with mutations that have label support (phenotypic resistance call agreement in the aggregated dataset) at 50% were discarded, and the majority label was assigned to all other mutations.

The final moxifloxacin dataset contained 1695 samples, 1420 (84%) of which resistant to the drug. The levofloxacin dataset had a similar proportion, with 1429/1749 (82%) samples resistant. Both drugs hence show a class imbalance towards resistant samples. When examining the distribution of mutations however, we observe the opposite: with only 23/153 (15%) unique moxifloxacin resistance mutations and 29/172 (17%) unique levofloxacin resistance mutations, there are relatively few resistance mutations responsible for many resistant samples. This makes our fluoroquinolone dataset very similar to the rifampicin resistance dataset used by Lynch *et al.*[232], where 46 RAVs out of 265 unique mutations (17%) are responsible for causing resistance. In fact, in our dataset, the two most common resistance mutations, D94G and A90V on *gyrA* are responsible for over 50% of resistant samples (Table 14). These two amino acids are also very close to the active site of the protein, as shown in Figure 30C. Among the most common mutations we also see many that affect the same codon, which additionally makes the dataset more sparse for some features. Again, the paucity of observed RAVs is probably due to both the RNA polymerase and DNA gyrase being essential proteins, allowing very limited scope for evolution to introduce resistance-inducing mutations that do not affect protein function.

Mutation	Levofloxacin	Moxifloxacin	total %
gyrA D94G	632	621	36.12
gyrA A90V	340	338	19.54
gyrA D94N	137	138	7.93
gyrA D94A	85	88	4.99
gyrA S91P	63	67	3.75
gyrA D94Y	60	62	3.52
gyrA D94H	38	35	2.10
gyrA N193S	27	20	1.35
gyrA P472S	23	14	1.07
gyrA G88C	22	24	1.33
gyrB E501D	18	17	1.01
gyrA T267I	12	12	0.69
gyrB D461N	10	13	0.66
gyrA P154R	9	9	0.52
gyrB R446C	9	10	0.55

Table 14: Number of samples per mutation for levofloxacin and moxifloxacin resistant samples. Shown are the 15 most commonly observed mutations in the levofloxacin and moxifloxacin dataset. The total percentage adds up all samples for both drugs and is hence a weighted average percentage for both.

5.4.2 Low diversity of fluoroquinolone resistance mutations in *M. tuberculosis* prevents training machine learning models for resistance prediction based on structural data

The strong class imbalance and the paucity of RAVs can make resistance prediction very challenging in practice, as is evident in the example of rifampicin resistance.[232] To exemplify the consequences of naively applying ML models to clinical datasets without accounting for these factors, we analysed the performance of the simple ML approach to fluoroquinolone resistance prediction using both a sample-based and a mutation-based dataset. The sample-based dataset will use the samples as they are, hence including many duplicated samples with identical mutation - phenotype associations. The mutation-based dataset will be aggregated by mutation, which deduplicates the dataset to show one sample per mutation - phenotype association. The sample-based dataset is hence essentially an upsampled version of the mutation-based dataset.

The features for each sample were calculated based on the physicochemical and structural features of the respective mutated amino acid as described in the methods. We ended up with 11 features for both drugs. Feature importance analysis yielded distance to drug binding site and distance to the two active site magnesium ions as the most important features, which indicates that structure does play an important but straightforward role in prediction. Two different ML models were trained, a simple LR and XGBoost,

since they were shown to perform the best for previous attempts to predict resistance using simple machine learning.[231;232] Models were trained by first performing hyperparameter tuning via cross-validation on the training set. Using the best parameters for each model (LR or XGBoost), we then predicted the labels of samples in the test set and reported performance through sensitivity and specificity.

Naively applying the models to the sample-based dataset for moxifloxacin gave very good performance on the test set, with sensitivity and specificity above 95% for XGBoost, and logistic regression performing similarly for sensitivity but showing a lower specificity of 87.8% (Figure 35A). These performance metrics are unexpectedly high, and we observe similarly high performance for levofloxacin resistance prediction (Figure 36A). Since so many samples are duplicates of each other, especially resistant samples, the main concern of this sample-based approach is data leakage. Indeed, we observed that with our train - test split of 70:30, 25/55 (45%) susceptible and 12/17 (71%) resistant mutations in the moxifloxacin test set have already been seen in the training set (Figure 35B). For levofloxacin, the numbers are similar with 28/66 (42%) susceptible and 13/16 (81%) resistant mutations being leaked from the training to the test dataset (Figure 36B). This is problematic, since it will reward memorisation over mechanistic understanding when training our ML models. And simple memorisation will fail to achieve our goal of predicting the effect of novel mutations. In fact, using the XGBoost approach, a combined 11/15 (73%) misclassified samples in the test sets of the moxifloxacin and levofloxacin datasets contained mutations not seen in the training set, indicating limited learning of unseen mutations.

With the mutation-based approach we can eliminate the data leakage problem almost entirely, since each mutation - phenotype association is only present once in the overall dataset. However, the problem of multiple mutations in the same codon location persists and might lead to residual memorisation based on codon identity. In addition, the aggregation of data naturally comes at the cost of reducing the size of the dataset. We end up with a dataset of 153 unique mutations for moxifloxacin and 172 for levofloxacin, reducing the size of the datasets by a factor of 10. However, the mutation-based model is much more likely to learn to predict resistance based on structural and physicochemical features, and is therefore more likely to generalise to novel mutations.

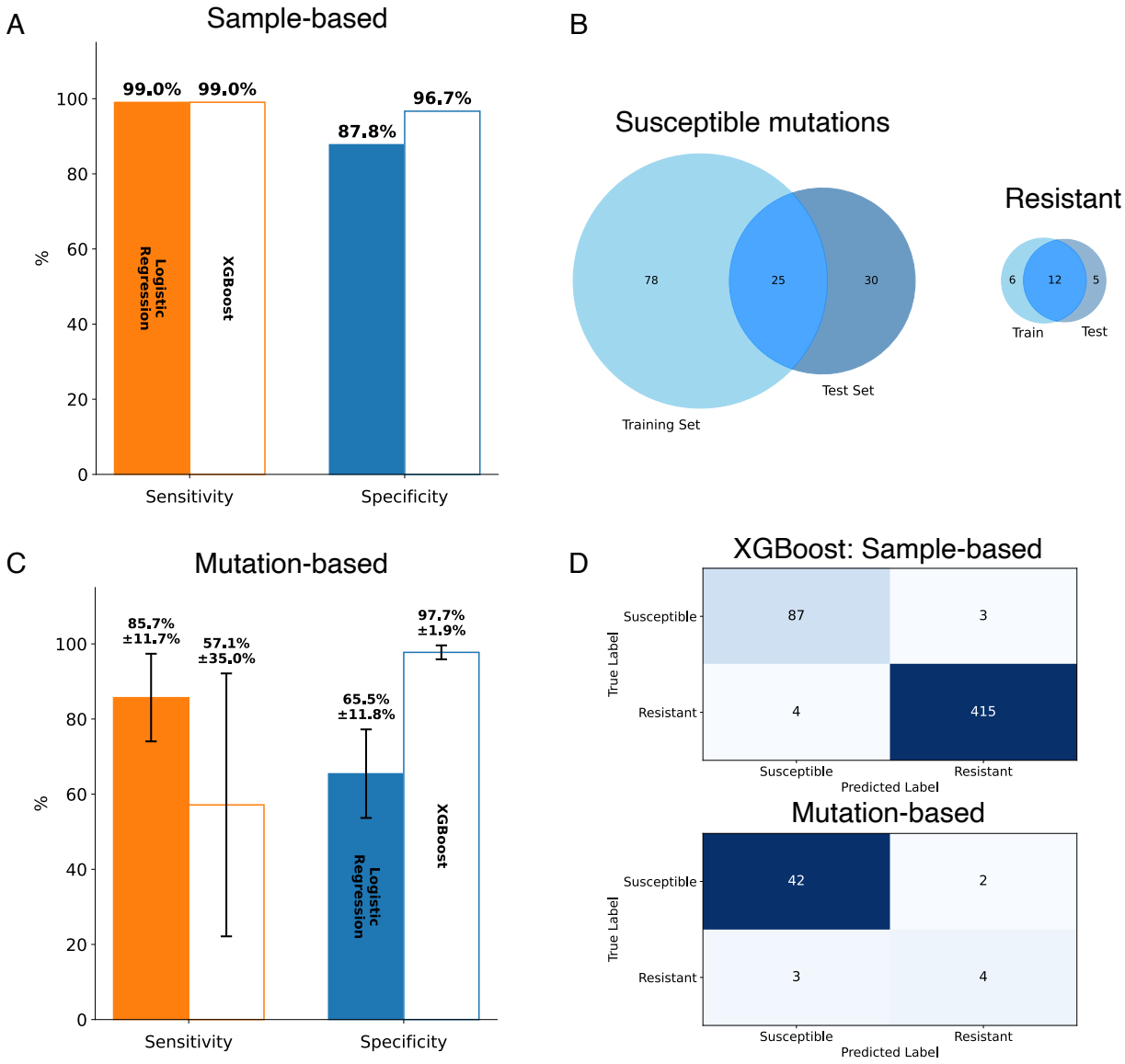


Figure 35: Performance of simple machine learning approach for moxifloxacin resistance prediction is inflated by data leakage in sample-based approach. **(A)** The performance of the sample-based approach based on the entire dataset of moxifloxacin samples with either a single resistance or susceptible mutation. Sensitivity and specificity are shown for both the logistic regression and XGBoost models. **(B)** Number of susceptible mutations and resistance mutations, as well as their randomised distribution into training and test set using a 70:30 overall split. The overlap highlights the mutations in the test set that have already been seen in the training set and hence cause a data leakage. **(C)** The performance of the mutation-based approach on the aggregated dataset of moxifloxacin samples (samples deduplicated by mutation). Sensitivity and specificity are shown as the mean and standard deviation of the 3-fold cross validation for both the logistic regression and XGBoost model. **(D)** Confusion matrices of the XGBoost model for both the sample-based and a random fold of the mutation-based approach. The class imbalance is clearly on the side of resistant samples for the sample-based approach, and only 4/419 (1%) resistant samples are misclassified. The class imbalance for the mutation-based approach is towards susceptible mutations, leaving only 7 resistance mutations in the test set for moxifloxacin. Only 3/7 (43%) resistance mutations are misclassified, but the impact on the sensitivity is much higher due to the small sample size.

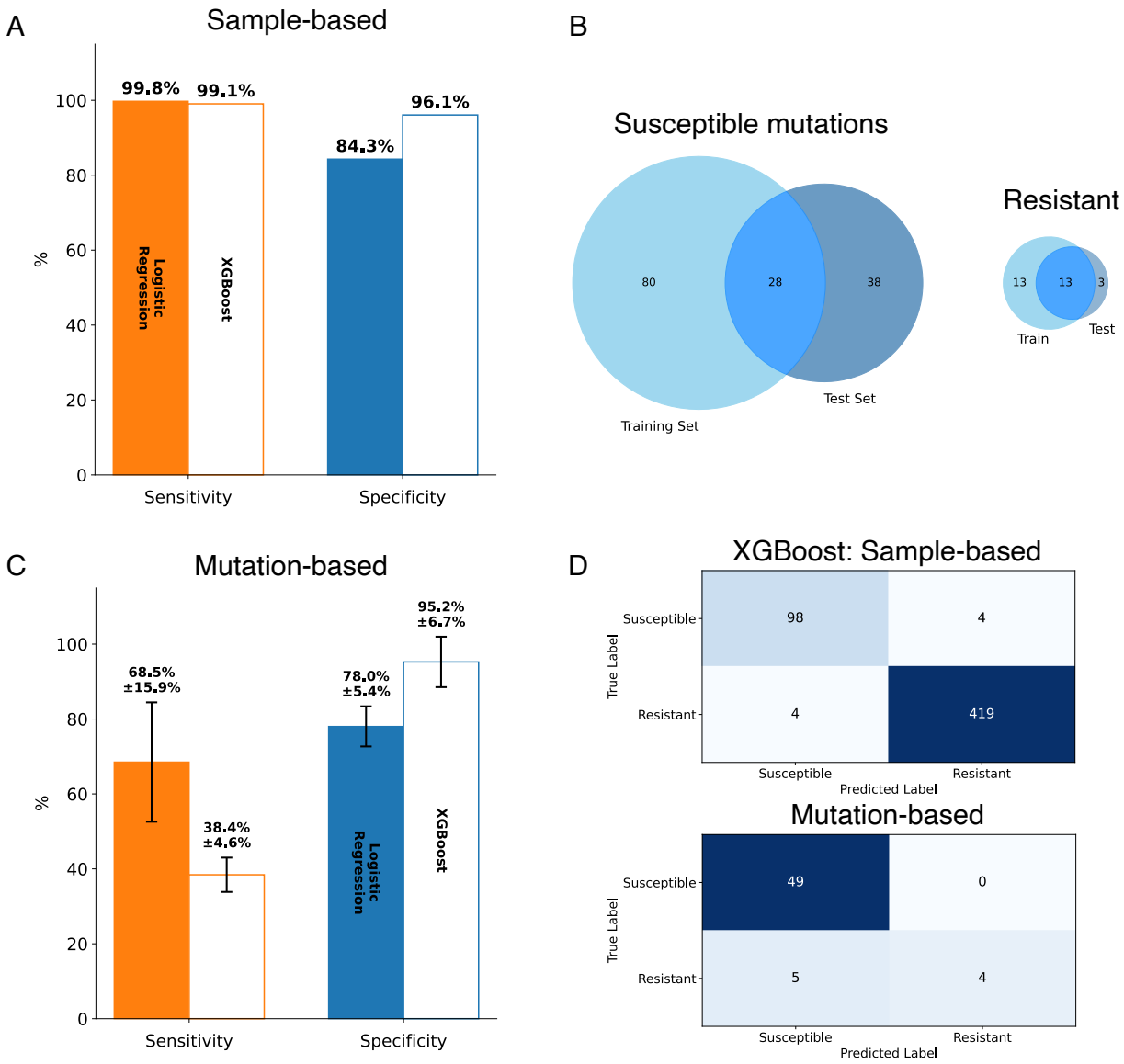


Figure 36: Performance of simple machine learning model for levofloxacin resistance prediction is inflated by data leakage in sample-based approach. **(A)** The performance of the sample-based approach based on the entire dataset of levofloxacin samples with either a single resistance or susceptible mutation. Sensitivity and specificity are shown for both the logistic regression and XGBoost models. **(B)** Number of susceptible mutations and resistance mutations, as well as their randomised distribution into training and test set using a 70:30 overall split. The overlap highlights the mutations in the test set that have already been seen in the training set and hence cause a data leakage. **(C)** The performance of the mutation-based approach on the aggregated dataset of levofloxacin samples (samples deduplicated by mutation). Sensitivity and specificity are shown as the mean and standard deviation of the 3-fold cross validation for both the logistic regression and XGBoost model. **(D)** Confusion matrices of the XGBoost model for both the sample-based and a random fold of the mutation-based approach. The class imbalance is clearly on the side of resistant samples for the sample-based approach, and only 4/423 (1%) resistant samples are misclassified. The class imbalance for the mutation-based approach is towards susceptible mutations, leaving only 9 resistance mutations in the test set for levofloxacin. Only 5/9 (56%) resistance mutations are misclassified, but the impact on the sensitivity is much higher due to the small sample size.

Since we have a much smaller dataset for the mutation-based approach, we expect the outcome of the resistance prediction to be highly initialisation dependent, i.e. different runs will lead to vastly different performances depending on the random train - test split. We decided to run the mutation-based models with a nested cross-validation, since this will give us a mean and standard deviation over multiple model runs. The mutation-based approach using XGBoost shows similar specificity in the test set as the sample-based approach in both moxifloxacin and levofloxacin resistance prediction (Figure 35C, 36C). However, sensitivity is significantly decreased, achieving mean values of 57% and 38%, respectively. This is surprising, since XGBoost was the better performing model previously, and is generally one of the most accurate and fast algorithms for classification problems. Although LR performs slightly better in terms of sensitivity, the specificity has decreased here as well and additionally the standard deviation is very high for all metrics in both ML models for the mutation-based approach (Figure 35C, 36C). This indicates that as expected, the prediction performance is not reproducible and is highly dependent on the current training and test set.

We also observe that susceptibility is predicted reasonably well by the mutation-based approach, as measured by specificity (Figure 35D, 36D). On the other hand, resistance mutations are misclassified much more frequently, which constitutes very major errors (VMEs). The reason for this becomes apparent when examining the confusion matrix for the XGBoost model of the sample-based approach and comparing it with the confusion matrix of a randomly chosen fold for the mutation-based model. The class imbalance in the mutation-based approach has clearly shifted towards susceptibility (Figure 35D, 36D). This makes it inherently difficult to learn resistance prediction based on the tiny test data set with merely seven or nine resistance mutations, respectively, for moxifloxacin and levofloxacin. We are probably underpowered for learning structure-based resistance prediction based on these datasets, especially given the number of features we are using.

5.4.3 Graph convolutional network models can learn to predict fluoroquinolone resistance in *E. coli in silico* but require further testing

The drastic reduction in sample size when using the simple ML approach stems mainly from restricting the dataset to one single missense mutation per sample. To mitigate this issue, we therefore need a model capable of learning to predict resistance based on the entire protein structure. This in turn will allow for an unlimited number of mutations per sample, as long as they are in the same target protein. We will hence evaluate whether a GCN model, which can use the entire protein graph as input, can predict resistance based solely on the physicochemical and structural features of the amino acids in the protein. We will also use a different organism that shows more genetic variation.

The input data for this model will be from a simulated dataset, as described in the Methods. We will have one model each for *gyrA* and *gyrB*, and a joint model for the GyrA/GyrB heterodimer. For each model, we simulated 1000 alleles (or allele pairs for the heterodimer) with half of the samples labelled resistant. The resistant samples show at least one resistance mutation, with the exact number of mutations randomly chosen from a Poisson distribution with mean one. The mutations are sampled from a list of 25 known and inferred *E. coli* resistance mutations in the DNA gyrase (Table 15). These resistance mutations are also shown in blue in Figure 33C, with one *gyrA* mutation being out of frame due to being located far from the drug binding site (*gyrA* A196E).

The other half of samples is labelled susceptible and contains no resistance mutations from our list (Table 15). We introduced on average eight random susceptible mutations per sample (also in resistant samples), with susceptible mutations being all possible one SNP mutations that are not resistant as per our list. The data were split into training and test set using an 80:20 ratio and the GCN was trained for 500 epochs. The sensitivity and specificity were tracked as performance metrics during model training, and the loss function is recorded to monitor model convergence (Figure 37).

For the *gyrA* subunit, the performance on the test set increases over time as model training proceeds, with an average test sensitivity over the last 10 iterations of 89.9% and test specificity of 90.1%(Figure 37A,B). The consistently low loss function after about 350 iterations of model training (steps) indicates

mutation	NCBI	proximity to drug
gyrA A51V	✓	
gyrA N57K	✓	
gyrA A67S	✓	
gyrA V70Y		✓
gyrA I74W		✓
gyrA H80P	✓	
gyrA G81C	✓	
gyrA D82A	✓	
gyrA S83A	✓	
gyrA A84P	✓	
gyrA V85P		✓
gyrA Y86E	✓	
gyrA D87G	✓	
gyrA V90Y		✓
gyrA Q106H	✓	
gyrA A196E	✓	
gyrB E424P		✓
gyrB G425P		✓
gyrB D426I	✓	
gyrB Q465P		✓
gyrB K447E	✓	
gyrB G448P		✓
gyrB K449E		✓
gyrB S464D		✓
gyrB E466C		✓

Table 15: Mutations in GyrA and GyrB of *E. coli* that were employed as resistance markers in the dataset simulation. The checkmarks indicate whether the mutations were taken from the NCBI Reference Gene Catalog (taken from AMRFinderPlus[259] version 3.12.8 with database version 2024-01-31.1) or chosen due to their proximity to the fluoroquinolone drug. We have 16 resistance mutations in *gyrA* and 9 mutations in *gyrB*. We emphasise that these mutations have been chosen to test if a GCN model can learn to predict resistance in general, but those not based on the NCBI database are somewhat arbitrary and unlikely to consistently confer resistance in practise.

that we will not see a substantial increase in performance past this point since the model has converged (Figure 37A).

When evaluating performance for the GyrB subunit, we obtain even better sensitivity towards the end with 98% average test sensitivity over the last 10 iterations and a specificity of 99% (Figure 37C,D). The loss function converges after about the same number of training iteration, but on a lower level than for GyrA.

Based on this, it seems reasonable to assume that performance for the heterodimer GyrA/B will be lower than for GyrB, but higher than for GyrA. Indeed, we see that sensitivity after convergence is at 95.1% and specificity is at 98.2% (Figure 37E,F), which is exactly between the performance of the two subunits alone. The model converges later than for the single subunit models, after around 450 iterations of model training, likely due to the larger graph structure making the model tuning more elaborate.

Overall, performance on the training set is better than on the test set in every approach and especially for specificity in the GyrA model and sensitivity in the joint model (Figure 37B,E). This indicates some overfitting on the training data, which is commonly observed in ML models but needs to be monitored closely. In our application, it could once again prompt concerns about the model learning to predict resistance based on mutation identity instead of structural and physicochemical consequences for amino acids in the protein.

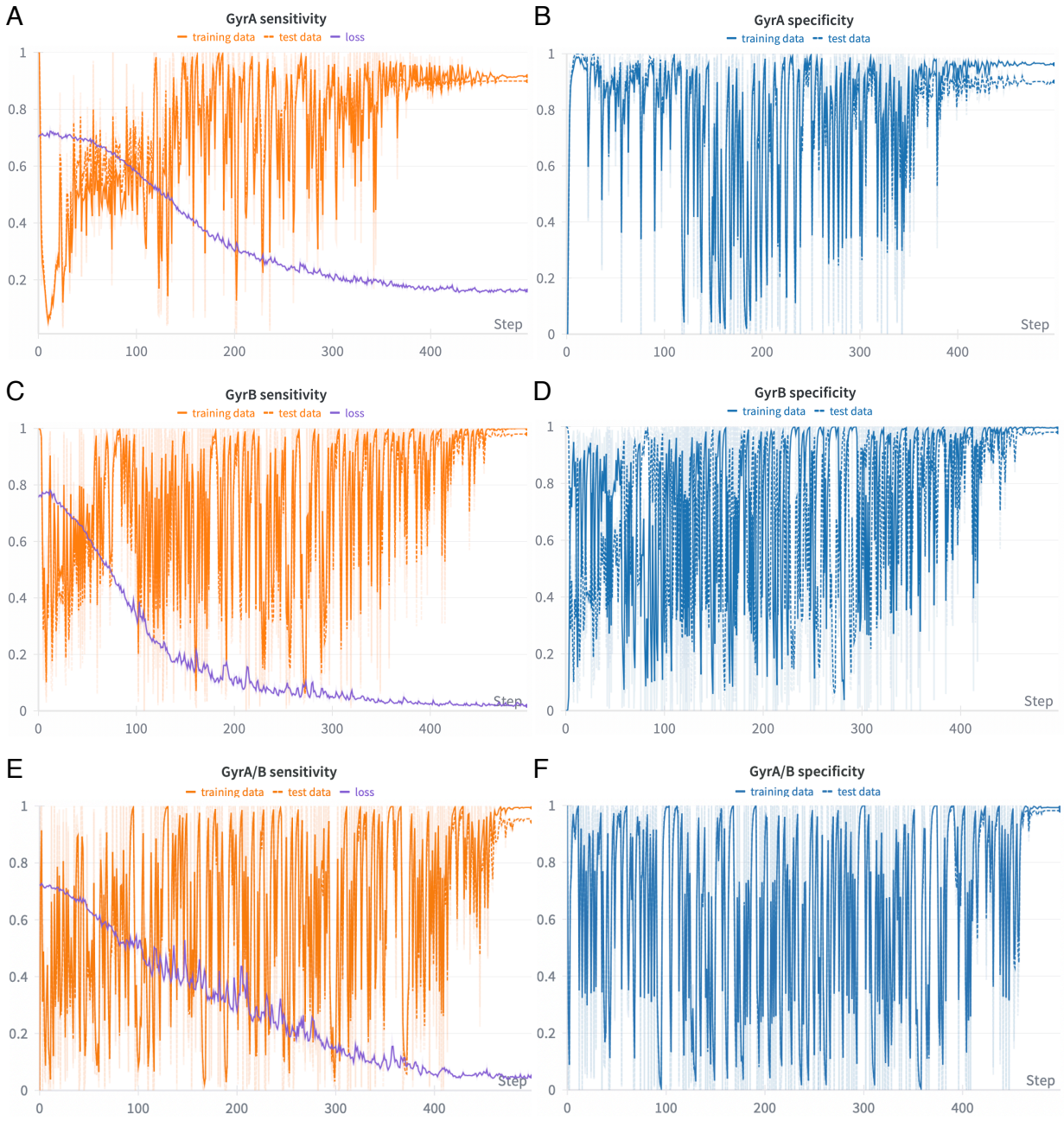


Figure 37: Sensitivity and specificity of graph convolutional network predictions for the GyrA and GyrB subunits of the DNA gyrase show overfitting in most cases. **(A)** Sensitivity and training loss function of levofloxacin resistance prediction when using the GyrA subunit dataset. Performance on the training data is shown as a solid orange line, while test data performance is shown as a dashed orange line. The training loss is shown in purple. The plots have been smoothed using time-weighted exponential moving average with a smoothing factor of 0.5 to allow a cleaner representation of the data. **(B)** Specificity of levofloxacin resistance prediction for the GyrA subunit dataset. Performance on the training data is shown as a solid blue line, while test data performance is shown as a dashed blue line. We see some overfitting on the training data. **(C)** Plot structure identical to A, but for the GyrB subunit dataset. **(D)** Plot structure identical to B, but for the GyrB subunit dataset. **(E)** Plot structure identical to A, but for the joint GyrA/B dataset. We see some overfitting on the training data. **(F)** Plot structure identical to B, but for the joint GyrA/B dataset. This figure was put together using the wandb web UI, where performance metrics were tracked as well.[246]

5.5 Discussion

In this Chapter, we have evaluated two approaches to model-based antibiotic resistance prediction using structural and physicochemical features of the target proteins. We have shown that the model design and prevention of data leakage are integral for ensuring proper learning, as opposed to simple label memorisation. More work is needed to tap the full potential of these machine learning models for resistance prediction.

5.5.1 Main conclusions

We have learned that simple ML models for resistance prediction based on sequence, structural and physicochemical characteristics of the antibiotic target protein of fluoroquinolones show moderate performance on the test sets when aggregated by mutation. The sample-based approach on the other hand shows good performance, likely due to a combination of the data leakage and the class imbalance of mutations (towards susceptibility) and class imbalance of samples (towards resistance), which artificially inflates the performance.

There are ways to address class imbalance in the mutation-based dataset, however these can only lead to limited improvement. One approach would be upsampling, but this has already been shown to lead to data leakage, as exemplified by the sample-based approach. In the latter approach, the data leakage from training to test set leads to the model learning labels only, and hence being unable to infer resistance based on structural and physicochemical features. Since our overall goal is to predict the effect of unseen mutations, this defeats the purpose, as it offers no advantage over rules-based approaches to resistance prediction.

The overall small size of the dataset makes training the mutation-based model difficult, since we have 11 features for model training, but only 153 unique mutations for moxifloxacin, 23 of which resistance mutations and 172 for levofloxacin, 29 of which resistance mutations. We can quickly analyse the degrees of freedom (DOF) of the LR model to estimate how many independent samples would be needed for effective model training in the mutation-based approach. In general, LR needs fewer samples for training than XGBoost, since fewer parameters are involved in the classification process, which is why we can use

this as a lower bound of how many samples we need. If we were to use a non-regularised LR, the DOF would be equal to the number of features plus the intercept, leading to 12 degrees of freedom. But since we are using L2 regularisation to prevent overfitting, we need to use an approximation to calculate the DOF for LR using singular value decomposition, as described by Hastie *et al.*[260] Because we are using a 3-fold cross-validation for the outer fold, we then calculate the average DOF over the 3 folds.

Using the example of LR for levofloxacin resistance, we obtain an approximate mean DOF of 8.02 ± 4.98. A common rule of thumb is the 'rule of 10', which suggests that at least 10 events per DOF are needed for effective model learning.[261] According to this we would need at least 80 ± 49 samples, i.e. mutations, to successfully train the LR model. We have 172 unique mutations for levofloxacin and are hence above the total needed mutation count, but the strong class imbalance toward susceptible mutations likely increases the number of mutations needed. Since we want to predict resistance, it seems appropriate to only call resistance mutations an 'event' as described in the rule of 10. With this constraint, we have only 29 events for levofloxacin, hence being below the lower boundary for the theoretically needed amount of mutations. Since we have even fewer mutations available for moxifloxacin resistance (153 unique mutations, 23 of which resistant), we likely fail to meet the threshold here as well.

For XGBoost, we cannot directly approximate the DOF, since it is a tree-based method that does not have a fixed number of parameters. However, it is generally accepted that XGBoost is a more sophisticated algorithm than LR and as such is able to capture more complex patterns, but in turn requires more data for model training. This might also explain why it breaks down when applied to a low complexity datasets like our mutation-based datasets with very few resistance mutations, while LR performs slightly better.

That the number of known resistance mutations is a limiting factor is supported by the fact that in the ML studies by Carter *et al.*[231] and Lynch *et al.*,[232] the sensitivity and specificity of resistance prediction were higher, with less variability. For the pyrazinamide dataset, 664 SNP missense mutations were available, out of which 349 were associated with resistance. This easily exceeds the threshold needed for proper learning of LR in a mutation-based approach, assuming that the threshold will be the same as in our model.

For rifampicin resistance prediction, 219 susceptible mutations and 46 resistance mutations were available. This is a rather low number of resistance mutations, but this model achieved better performance compared to fluoroquinolone resistance.[232] Still, it is likely that the actual performance was lower since overfitting was detected, probably due to the strong class imbalance in the rifampicin resistance dataset. As eluded to by Carter *et al.*, the most promising targets for this simple machine learning approach in *M. tuberculosis* are hence non-essential proteins, similar to the pro-drug system in pyrazinamide resistance.[231] Since disrupting the protein function is not off limits, these systems show both a higher variety of mutation, as well as a more even spread of resistance mutations throughout the protein, leading to an even ratio of susceptible to resistance mutations. Another option would be to look at resistance in microbes with more reported genetic variation, such as *E. coli*, and to allow multiple mutations to be present in a sample. This would vastly increase the number of eligible samples and mostly prevent data leakage based on single mutations.

For the GCN approach, we simulated a dataset of 1000 samples with arbitrary chosen 'resistance' mutations in 50% of the simulated *E. coli* alleles or allele pairs. We hence had a balanced dataset with no class imbalance, and also had the option to introduce more than one mutation per allele, making the dataset much more genetically diverse. Sensitivity and specificity values of levofloxacin resistance prediction were higher for the simulated dataset of the GyrB subunit than for the GyrA subunit. The model might be able to learn to predict resistance more easily for GyrB because we have fewer resistance mutations in the simulated dataset compared to GyrA (9 vs 16, with comparable gene length). In addition, the resistance mutations in GyrB are all in close proximity to the drug binding site. The model hence has an easy way to learn resistance prediction solely based on distance to the drug, while the GyrA dataset contains mutations further away from the binding site (e.g. *gyrA* A196E), hence the model would have to learn a combination of features.

When using the entire graph structure of the protein as input instead of singular mutations, as in the mutation-based simple ML approach, we have taken away the model's ability to memorise labels directly. We also introduce additional noise through the presence of susceptible mutations in both resistant and susceptible samples. However, the overfitting on the training data observed in the GCN model still

indicates that the model might not only be learning generalisable properties of the resistance phenotype. It is hence still possible that the model has learned a mutation - resistance association, even if this is more difficult than in the simple ML approach, instead of learning the structural and physicochemical causes of resistance.

5.5.2 Limitations

Apart from the need to improve the performance of the simple ML approach, there are several limitations owed to the assumptions made when building the models. First of all, we restricted ourselves to the resolved crystallised regions of a single protein for predicting resistance. Promoter regions and any effects through transcription factors or mutations in other proteins are hence not included. These could include upregulation of efflux pump activity and/or decreased uptake of the drug. In addition, we have assumed that resistance is always caused by a singular, highly penetrant genetic variant. This assumption does not always hold, especially in cases where the MIC is incrementally increased by the presence of multiple resistance-associated mutations. In *E. coli*, it is known that the presence of multiple resistance mutations in the DNA gyrase (*gyrA/B*) or topoisomerase IV *parC/D* can lead to successive increases in MIC.[262]

The genetic variant is also required to be a missense mutation, as insertions, deletions, and nonsense mutations cannot be considered using this paradigm. Missense mutations are additionally assumed to not change the overall structure of the protein, which is reasonable for essential proteins, but might not always be the case.

Some of these limitations can be addressed using the GCN approach: it allows multiple mutations per protein to be considered, and it can, in theory, integrate Alphafold[229] for structure prediction for each allele. Even without using Alphafold, the GCN was able to predict resistance to levofloxacin with high sensitivity and specificity when only the spatial connections and physicochemical features of the amino acids in the protein were given as input. Since the DNA gyrase is an essential protein, it is unlikely that large changes in protein structure would be incurred by resistance mutations as that would likely impair protein function. The use of structure prediction tools would hence be more useful in the prediction

of resistance in non-essential proteins, such as the pyrazinamidase. The GCN approach paired with our synthetic datasets could also allow us to test if the model could be trained to predict incremental increases in MIC, which might be especially useful when individual genetic variants are not highly penetrant.

However, the GCN approach still limits us to the protein coding regions of genes. And while it might be possible to test GCNs using multiple protein graphs as input, promoter regions and inter-genic regions of the DNA will still not be considered. It is also difficult to process insertions, deletions and nonsense mutations (leading to chain-terminating stop codons) in a GCN approach, since the protein graphs would end up being of different lengths. This could strongly bias the learning process. An additional problem arises from the relative lack of known *E. coli* fluoroquinolone resistance and susceptible mutations when compared to *M. tuberculosis*, which restricted the diversity of our simulated dataset and hence made it more difficult to train a generalisable ML model.

5.5.3 Outlook and future work

We have established that the simple ML approach is not amenable to fluoroquinolone resistance prediction in *M. tuberculosis* due to a lack of mutation diversity associated with resistance. Our GCN model for *E. coli* allowed us to capture more complex relationships using entire protein graphs, however we observe overfitting on the training data and have a low number of known resistance mutations, especially for the GyrB subunit. This once again prompts concerns about data leakage, even if multiple mutations are present per sample.

To address this problem, we could segregate resistance mutations between a training and test set list and simulate gene sequences based on the respective lists only, similar to the mutation-based approach in the simple ML model. The problem arising here is, aside from the pool of resistance mutations being very small, that we do not have a list of susceptible mutations as we did for *M. tuberculosis*. For the GCN approach in *E. coli* so far, we have assumed that all mutations that are not resistance mutations are susceptible to levofloxacin. However, this will not work if we segregate the mutations into a test and training dataset, since we will end up implicitly calling the resistance mutations from the training set susceptible in the test set. And if we instruct the algorithm to exclude these mutations in the test

set, we will again introduce a potential for data leakage, since the absence of a mutation can also carry information. We will therefore need either i) a list of missense mutations associated with susceptibility or ii) sufficient alleles with known resistance labels, making the use of sbmlsim obsolete.

A more straightforward test would be to introduce more than one resistance mutation in all resistant samples in our synthetic dataset, which would make it harder for the model to learn a mutation - phenotype association directly. In the current approach, we introduce at least one resistance mutation per resistant sample. Although most samples only have one resistance mutation, a Poisson distribution is applied to allow the occasional sample with two or even three resistance mutations. This could easily be increased to a Poisson distribution with a mean of two or three, which would increase the average number of resistance mutations per sample.

For fluoroquinolone resistance in *E. coli*, simulating multiple resistance mutations per sample could also be more representative of what is seen in clinical samples.[262] The previously mentioned incremental increase in MIC could also be reflected by redefining how we generate our synthetic datasets. We could e.g. only assign a phenotypic resistance label to a sample if multiple resistance mutations are present, or if predefined combinations of MIC-increasing mutations are in the same sample. One could also define epistatic relationships between different mutations. This of course increases the difficulty of generating the synthetic dataset, but in turn makes it more likely to reflect the biological reality of fluoroquinolone resistance in *E. coli* and therefore becomes a better test of the method. A model trained on a dataset where the phenotypic label depends on the combination of different mutations would be much more likely to be applicable to a real world dataset of clinical *E. coli* samples. In the end, the aim of this model-based approach to resistance prediction is to predict the effect of rare mutations in clinical samples of both *M. tuberculosis* and *E. coli* infections. While the GCN approach is promising, it needs further development to achieve this goal. Once the performance on synthetic data is improved, we will need to collate a real-world dataset for training and testing the model on real clinical samples.

To illustrate the difficulty in designing synthetic datasets for ML purposes I would like to draw a comparison to the use of CNNs. There is another reason why image classification is one of the most sophisticated

and progressive fields of AI use. The classification of visual features is something that the human brain is naturally very good at, since it is integral for our survival. Designing diverse datasets for training a CNN in image classification is hence much more trivial than for a problem that we ourselves have not quite understood, such as the underlying cause of resistance to many drugs. Therefore, designing a dataset that is able to learn the real underlying cause of resistance is non-trivial when we struggle to understand which type of diversity is important. Our ability to use ML methods for predicting resistance for different drug-pathogen combinations hence also depends on our previous knowledge of the resistance mechanisms.

Moving forward, the limited genetic variability in essential genes involved in drug resistance probably complicates the use of ML models for resistance prediction in all pathogens. The more straightforward targets for this type of approach are hence non-essential genes, such as the *pncA* gene involved in pyrazinamide resistance. Both the simple ML approach based on singleton data and the GCN approach have been shown to perform better on this drug target in *M. tuberculosis* than on any of the other drug target proteins tested.[231;263] For resistance prediction in essential genes however it is unlikely that the use of ML methods will significantly increase the performance over the levels achieved using catalogued resistance mutations.

6 Conclusions and future work

In this thesis, I have explored several avenues to improve predicting antibiotic resistance through whole genome sequencing antimicrobial susceptibility testing (WGS-AST) in *M. tuberculosis* and *E. coli*. This involved examining the genetic and evolutionary landscape of antimicrobial resistance (AMR) from multiple perspectives, spanning molecular mechanisms, within-host diversity, and predictive modelling. In doing so, this thesis contributes to a more nuanced view of how resistance prediction can be advanced in a way that is both biologically grounded and clinically actionable.

In Chapter 3, I identified a comprehensive list of 51 compensatory mutations (CMs) through their strong association with known rifampicin resistance mutations, which was complicated by the strong linkage disequilibrium in the highly clonal organism *M. tuberculosis*. I showed that the fitness cost associated with rifampicin resistance, as well as the compensatory effect of CMs in Lineage 2 can be seen in the result of *in vitro* growth assays. During this *in vitro* growth investigation, I also found that the increased growth phenotype is strongly confounded with high-fitness clades within the phylogenetic tree of *M. tuberculosis*. The retrospective finding that some putative CMs, such as *rpoC* F452L, may act as predisposing fitness mutations rather than true compensatory changes, underscores the complexity of disentangling fitness effects from resistance-associated effects.

This is in line with new interest in the interactions of lineage background and different resistance-associated mutations in *M. tuberculosis*, which potentially form epistatic relationships that influence fitness *in vivo*. One example of this is the proposed synergistic action of mutations in *rpoB* and *gyrA*, potentially affecting rifampicin and fluoroquinolone resistance.[264] The investigation of CMs is hence intertwined with the investigation of fitness-related mutations throughout the genome. In terms of WGS-AST for predicting rifampicin resistance, I showed that the presence of CMs is a highly specific indicator for resistance. On these grounds, it might be worth starting a discussion around the definition of legitimate resistance markers in WGS-AST data. Hitherto, only (assumed) directly resistance-associated mutations are considered, and the causality between a mutation and the resistance phenotype remains difficult to prove *in vivo*.

In Chapter 4, I demonstrated that the commonly applied fraction-of-read-support (FRS) threshold can obscure rifampicin resistant subpopulations in *M. tuberculosis*, such that lowering the threshold improves sensitivity of resistance prediction. Overall, we find that detecting resistant subpopulations is not only useful for resistance prediction, but also provides evidence for different infection scenarios through considering within-sample diversity. This prompted me to investigate possible heterogeneity in samples from *Enterobacteriaceae* infections, where single-colony picks are the prevalent starting point for routine diagnostics. Analysis of a metagenomics dataset of blood stream infections with *E. coli* and *K. pneumoniae* produced some evidence of elevated within-sample diversity, which lays the groundwork for investigating clinically relevant diversity.

In Chapter 5, I explored the possibility of using machine learning (ML) approaches based on structural and physicochemical features derived from the fluoroquinolone antibiotic target protein, with the main goal of predicting the effect of novel mutations. The simple classifiers targeting fluoroquinolone resistance in *M. tuberculosis* via GyrA/GyrB were limited by data leakage, class imbalance and low genetic diversity in essential genes. The GCN model approach for fluoroquinolone resistance in *E. coli* is more promising, however it currently relies on simulated data of the DNA gyrase subunit genes and also exhibited overfitting on the training data. To make sure that the model is generalisable and capable of predicting the effect of novel mutations, a more realistic dataset is required initially, followed by clinical validation.

In conclusion, integrating WGS-AST into resistance prediction comes with both opportunities and challenges. The observation that CMs in *M. tuberculosis* have predictive value for resistance highlights the tension between curating catalogues to achieve the best performance in clinical use versus the need for mechanistic clarity. The demonstration that resistant subpopulations are readily detectable when thresholds are relaxed calls into question the reliance on rigid bioinformatic pipelines, particularly in diseases such as tuberculosis (TB) where within-host heterogeneity has been shown to shape treatment outcomes.[210;216;265] Similarly, evidence of high within-sample diversity in *E. coli* and *K. pneumoniae* emphasises that single-colony picks as a basis for AST may underestimate clinically important heterogeneity in bloodstream infections. Finally, the contrasting performance of ML models for different drug-species combinations underscores that while ML approaches hold real promise, their success is constrained by

the underlying genetic architecture of resistance and the availability of balanced, diverse training data.

The most promising way to build on the findings from Chapter 3 is to prove the predictive value of CMs by simulating the scenario of low sequencing depth by downsampling the reads. It would also be promising to examine the few samples with CMs and no resistance-associated mutations (according to the WHO catalogue) for other, potentially novel resistance-associated mutations. Chapter 4 started investigating the relevance of subpopulations in *E. coli* and *K. pneumoniae*, with the low hanging fruit of proving that clinically relevant heterogeneity is present in those samples with elevated within-sample diversity. A more thorough investigation of this would require larger datasets from bloodstream infections. The metagenomic approach to WGS-AST employed for these *Enterobacteriaceae* samples is especially interesting since it allows sequencing directly from positively flagged blood culture bottle, bringing us one step closer to realising direct-from-sample sequencing and avoiding pre-culturing steps for sequencing. Chapter 5 leads into refining the GCN by optimising the dataset simulation, to more accurately represent real-world samples. Then one could train the model on available in-house datasets of *E. coli* infections. However, it is more promising to focus efforts on non-essential proteins with more unexplained resistance.

The comparison of diagnostics for *M. tuberculosis* and *Enterobacteriaceae* illustrates nicely that the combined characteristics of bacteria and drug dictate how we perform AST. A basic difference that has not been a subject of discussion in this thesis yet is the difference in pathogenicity. While many *Enterobacteriaceae* are commensals or opportunistic pathogens, which frequently colonise the gut of their hosts and replicate without inducing an infection,[266] *M. tuberculosis* can be in a latent state but will always display pathogenicity once actively replicating. This makes diagnostics for *Enterobacteriaceae* difficult, if not based on sterile samples like blood cultures, since commensals must be differentiated from pathogens. On the other hand, the difficulties in culturing *M. tuberculosis* bacteria have led to the development of sophisticated methods for genotypic AST, which has been less of a focus in *E. coli* and *K. pneumoniae* due to their fast growth and non-fastidious nature. But even though WGS-AST for *Enterobacteriaceae* is more challenging due to the presence of large accessory genomes, which makes resistance arising through chromosomal mutations less relevant,[124;125] there are lessons to be learned from the WGS-AST methods developed for *M. tuberculosis*. The necessity to sequence 'crumbs' of multiple colonies instead of single-

colony picks,[111] for example, comes with the helpful side effect of capturing within-sample diversity. The efforts invested in establishing WGS-AST as a routine tool in *M. tuberculosis* diagnostics can hence pave the way for setting this up for other bacteria like members of the *Enterobacteriaceae* family. To get there, the use of ML methods will likely be superior compared to rules-based approaches, since resistance patterns in *Enterobacteriaceae* are much more complex and difficult to formulate in consistent rules or in catalogues using conventional statistics. The presence of multiple resistance genes and even resistance mechanisms in a single sample is common, and the therefore complex relationship with the phenotype is probably easier to capture using ML.[267] Training these models will necessitate the availability of large and sufficiently diverse clinical datasets with paired phenotypic data.

Despite the notable advances in the field of WGS-AST, it is clear that phenotypic AST will still be needed in the future. It is the only available approach to reliably correlate the presence of mutations and their phenotypic expression, which is especially important for new antimicrobial drugs with novel mechanisms and hence a high number of unknown mutations. This is e.g. essential for validating and updating catalogues of resistance-associated and susceptible mutations in *M. tuberculosis*. As we have seen, even ML approaches will not be able to fill in this gap without some prior knowledge of the resistance mechanism, if there are insufficiently diverse samples with phenotypes available.

Looking ahead, the integration of genomic surveillance, quantitative measures of within-host diversity, and advanced ML offers a powerful route toward more reliable and responsive resistance prediction. The work presented here shows that compensatory evolution, resistant subpopulations, and species-specific genetic architectures all complicate the simplistic "one variant–one phenotype" model that still underpins most clinical catalogues. By systematically testing these complexities, this thesis helps to define the limits of current approaches while pointing toward more flexible, data-driven solutions. Continued progress will depend on large, well-annotated datasets, careful validation of predictive markers, and the translation of algorithmic advances into tools that are interpretable and robust enough for routine clinical use. As global AMR pressures intensify, there is a clear opportunity for genomics-informed diagnostics and predictive models to support earlier, more targeted treatment decisions, ultimately improving patient outcomes and slowing the spread of resistance.

7 References

[1] Sakai, T. & Morimoto, Y. The History of Infectious Diseases and Medicine. *Pathogens* **11**, 1147 (2022).

[2] Seersholm, F. V. *et al.* Repeated Plague Infections across Six Generations of Neolithic Farmers. *Nature* **632**, 114–121 (2024).

[3] Lewis, C. M., Akinyi, M. Y., DeWitte, S. N. & Stone, A. C. Ancient Pathogens provide a Window into Health and Well-being. *Proceedings of the National Academy of Sciences* **120**, e2209476119 (2023).

[4] Polgreen, P. M. & Polgreen, E. L. Emerging and Re-emerging Pathogens and Diseases, and Health Consequences of a Changing Climate. *Infectious Diseases* 40–48.e2 (2017).

[5] Prentice, M. B. & Rahalison, L. Plague. *The Lancet* **369**, 1196–1207 (2007).

[6] Barberis, I., Bragazzi, N., Galluzzo, L. & Martini, M. The History of Tuberculosis: From the first Historical Records to the Isolation of Koch's Bacillus. *Journal of Preventive Medicine and Hygiene* **58**, E9–E12 (2017).

[7] Koch, R., Brock, T. D. & Fred, E. B. The Etiology of Tuberculosis. *Reviews of Infectious Diseases* **4**, 1270–1274 (1982).

[8] Sravanthi, K. *et al.* Robert Koch: From Anthrax to Tuberculosis – A Journey in Medical Science. *Cureus* **16**, e72955 (2024).

[9] Blevins, S. M. & Bronze, M. S. Robert Koch and the 'Golden Age' of Bacteriology. *International Journal of Infectious Diseases* **14**, e744–e751 (2010).

[10] King, L. S. Dr. Koch's Postulates. *Journal of the History of Medicine and Allied Sciences* **7**, 350–361 (1952).

[11] Zhang, Z. & Zhao, W. Koch's Postulates: From Classical Framework to Modern Applications in Medical Microbiology. *Global Medical Education* (2025).

[12] Murray, J. F. Tuberculosis and World War I. *American Journal of Respiratory and Critical Care Medicine* **192**, 411–414 (2015).

[13] Patterson, K. D. & Pyle, G. F. The Geography and Mortality of the 1918 Influenza Pandemic. *Bulletin of the History of Medicine* **65**, 4–21 (1991).

[14] Fleming, A. On the Antibacterial Action of Cultures of a Penicillium, with Special Reference to their Use in the Isolation of B. influenzæ. *British journal of experimental pathology* **10**, 226–236 (1929).

[15] Christensen, S. B. Drugs That Changed Society: History and Current Status of the Early Antibiotics: Salvarsan, Sulfonamides, and Beta-Lactams. *Molecules* **26**, 6057 (2021).

[16] Waksman, S. A., Schatz, A. & Reynolds, D. M. Production of Antibiotic Substances by Actinomycetes. *Annals of the New York Academy of Sciences* **1213**, 112–124 (2010).

[17] Hutchings, M. I., Truman, A. W. & Wilkinson, B. Antibiotics: Past, Present and Future. *Current Opinion in Microbiology* **51**, 72–80 (2019).

[18] Katz, L. & Baltz, R. H. Natural Product Discovery: Past, Present, and Future. *Journal of Industrial Microbiology and Biotechnology* **43**, 155–176 (2016).

[19] Kapoor, G., Saigal, S. & Elongavan, A. Action and Resistance Mechanisms of Antibiotics: A Guide for Clinicians. *Journal of Anaesthesiology, Clinical Pharmacology* **33**, 300–305 (2017).

[20] Sakenova, N. *et al.* Systematic Mapping of Antibiotic Cross-Resistance and Collateral Sensitivity with Chemical Genetics. *Nature Microbiology* **10**, 202–216 (2025).

[21] Davies, J. & Davies, D. Origins and Evolution of Antibiotic Resistance. *Microbiology and Molecular Biology Reviews : MMBR* **74**, 417–433 (2010).

[22] Wright, G. D. & Poinar, H. Antibiotic Resistance is Ancient: Implications for Drug Discovery. *Trends in Microbiology* **20**, 157–159 (2012).

[23] Abraham, E. P. & Chain, E. An Enzyme from Bacteria Able to Destroy Penicillin. *Reviews of Infectious Diseases* **10**, 677–678 (1988).

[24] Fleming, A. Nobel Lecture. 83–93 (1945).

[25] Jevons, M. P. "Celbenin" - resistant Staphylococci. *British Medical Journal* **1**, 124–125 (1961).

[26] Lobanovska, M. & Pilla, G. Penicillin's Discovery and Antibiotic Resistance: Lessons for the Future? *The Yale Journal of Biology and Medicine* **90**, 135–145 (2017).

[27] Davies, J. Vicious Circles: Looking Back on Resistance Plasmids. *Genetics* **139**, 1465–1468 (1995).

[28] Ventola, C. L. The Antibiotic Resistance Crisis. *Pharmacy and Therapeutics* **40**, 277–283 (2015).

[29] Smith, P. A. & Romesberg, F. E. Combating Bacteria and Drug Resistance by Inhibiting Mechanisms of Persistence and Adaptation. *Nature Chemical Biology* **3**, 549–556 (2007).

[30] Nikaido, H. Multidrug Resistance in Bacteria. *Annual Review of Biochemistry* **78**, 119–146 (2009).

[31] Iskandar, K. *et al.* Antibiotic Discovery and Resistance: The Chase and the Race. *Antibiotics* **11**, 182 (2022).

[32] Tacconelli, E. *et al.* Discovery, Research, and Development of New Antibiotics: The WHO Priority List of Antibiotic-Resistant Bacteria and Tuberculosis. *The Lancet Infectious Diseases* **18**, 318–327 (2018).

[33] Murray, C. J. L. *et al.* Global Burden of Bacterial Antimicrobial Resistance in 2019: A Systematic Analysis. *The Lancet* **399**, 629–655 (2022).

[34] Naghavi, M. *et al.* Global Burden of Bacterial Antimicrobial Resistance 1990–2021: A Systematic Analysis with Forecasts to 2050. *The Lancet* **404**, 1199–1226 (2024).

[35] O'Neill, J. Tackling Drug-Resistant Infections Globally: Final Report and Recommendations. Tech. Rep. (2016).

[36] Tripathi, N., Zubair, M. & Sapra, A. Gram Staining. In *StatPearls* (StatPearls Publishing, Treasure Island (FL), 2025).

[37] Salam, M. A., Al-Amin, M. Y., Pawar, J. S., Akhter, N. & Lucy, I. B. Conventional Methods and

Future Trends in Antimicrobial Susceptibility Testing. *Saudi Journal of Biological Sciences* **30**, 103582 (2023).

[38] Heatley, N. G. A Method for the Assay of Penicillin. *Biochemical Journal* **38**, 61–65 (1944).

[39] Schmith, K. & Reymann, F. Experimentelle og Kliniske Undersogelser over Gonococcers Folsomhed Overfor Sulfapyridin. *Nordisk Medicin* **8**, 2493–9 (1940).

[40] Smaill, F. Antibiotic Susceptibility and Resistance Testing: An Overview. *Canadian Journal of Gastroenterology and Hepatology* **14**, 382415 (2000).

[41] Kahlmeter, G. & Turnidge, J. How to: ECOFFs—the Why, the How, and the Don'ts of EUCAST Epidemiological Cutoff Values. *Clinical Microbiology and Infection* **28**, 952–954 (2022).

[42] Wheat, P. F. History and Development of Antimicrobial Susceptibility Testing Methodology. *Journal of Antimicrobial Chemotherapy* **48**, 1–4 (2001).

[43] UK Health Security Agency. Investigation of Specimens for Mycobacterium Species. *UK Standards for Microbiology Investigations.* **V 11 Issue 5** (2023).

[44] Clark, A. E., Kaleta, E. J., Arora, A. & Wolk, D. M. Matrix-Assisted Laser Desorption Ionization–Time of Flight Mass Spectrometry: A Fundamental Shift in the Routine Practice of Clinical Microbiology. *Clinical Microbiology Reviews* **26**, 547–603 (2013).

[45] Patel, R. Matrix-Assisted Laser Desorption Ionization–Time of Flight Mass Spectrometry in Clinical Microbiology. *Clinical Infectious Diseases* **57**, 564–572 (2013).

[46] Tenover, F. C. Diagnostic Deoxyribonucleic Acid Probes for Infectious Diseases. *Clinical Microbiology Reviews* **1**, 82–101 (1988).

[47] Schmitz, J. E., Stratton, C. W., Persing, D. H. & Tang, Y.-W. Forty Years of Molecular Diagnostics for Infectious Diseases. *Journal of Clinical Microbiology* **60**, e02446–21 (2022).

[48] Mullis, K. *et al.* Specific Enzymatic Amplification of DNA *In Vitro*: The Polymerase Chain Reaction. *Biotechnology (Reading, Mass.)* **24**, 17–27 (1992).

[49] Boom, R. *et al.* Rapid and Simple Method for Purification of Nucleic Acids. *Journal of Clinical Microbiology* **28**, 495–503 (1990).

[50] Roberts, G. A. & Dryden, D. T. F. DNA Electrophoresis: Historical and Theoretical Perspectives. In Makovets, S. (ed.) *DNA Electrophoresis: Methods and Protocols*, 1–9 (2013).

[51] Miller, M. B. & Tang, Y.-W. Basic Concepts of Microarrays and Potential Applications in Clinical Microbiology. *Clinical Microbiology Reviews* **22**, 611–633 (2009).

[52] Hodinka, R. L., Kaiser, L., Hodinka, R. L. & Kaiser, L. Point-Counterpoint: Is the Era of Viral Culture Over in the Clinical Microbiology Laboratory? *Journal of Clinical Microbiology* **51**, 2–8 (2013).

[53] Lawn, S. D. *et al.* Advances in Tuberculosis Diagnostics: The Xpert MTB/RIF Assay and Future Prospects for a Point-Of-Care Test. *The Lancet Infectious Diseases* **13**, 349–361 (2013).

[54] Larsen, M. V. *et al.* Multilocus Sequence Typing of Total-Genome-Sequenced Bacteria. *Journal of Clinical Microbiology* **50**, 1355–1361 (2012).

[55] Walker, T. M. *et al.* Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: A retrospective observational study. *The Lancet Infectious Diseases* **13**, 137–146 (2013).

[56] Roetzer, A. *et al.* Whole Genome Sequencing versus Traditional Genotyping for Investigation of a *Mycobacterium tuberculosis* Outbreak: A Longitudinal Molecular Epidemiological Study. *PLOS Medicine* **10**, e1001387 (2013).

[57] Köser, C. U., Ellington, M. J. & Peacock, S. J. Whole-Genome Sequencing to Control Antimicrobial Resistance. *Trends in Genetics* **30**, 401–407 (2014).

[58] Gilbert, W. & Maxam, A. The Nucleotide Sequence of the lac Operator. *Proceedings of the National Academy of Sciences* **70**, 3581–3584 (1973).

[59] Sanger, F., Nicklen, S. & Coulson, A. R. DNA Sequencing with Chain-Terminating Inhibitors. *Proceedings of the National Academy of Sciences* **74**, 5463–5467 (1977).

[60] Maniatis, T., Jeffrey, A. & Van deSande, H. Chain Length Determination of Small Double- and

Single-Stranded DNA Molecules by Polyacrylamide Gel Electrophoresis. *Biochemistry* **14**, 3787–3794 (1975).

[61] Smith, L. M. *et al.* Fluorescence Detection in Automated DNA Sequence Analysis. *Nature* **321**, 674–679 (1986).

[62] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).

[63] Shendure, J. *et al.* DNA Sequencing at 40: Past, Present and Future. *Nature* **550**, 345–353 (2017).

[64] Wetterstrand, K. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP), National Human Genome Research Institute.

[65] Slatko, B. E., Gardner, A. F. & Ausubel, F. M. Overview of Next Generation Sequencing Technologies. *Current protocols in molecular biology* **122**, e59 (2018).

[66] Greenleaf, W. J. & Sidow, A. The Future of Sequencing: Convergence of Intelligent Design and Market Darwinism. *Genome Biology* **15**, 303 (2014).

[67] Heather, J. M. & Chain, B. The Sequence of Sequencers. *Genomics* **107**, 1–8 (2016).

[68] Eid, J. *et al.* Real-Time DNA Sequencing from Single Polymerase Molecules. *Science* **323**, 133–138 (2009).

[69] Bayley, H. Nanopore Sequencing: From Imagination to Reality. *Clinical Chemistry* **61**, 25–31 (2015).

[70] Check Hayden, E. Pint-sized DNA Sequencer impresses First Users. *Nature* **521**, 15–16 (2015).

[71] Utturkar, S. M. *et al.* Evaluation and Validation of *de novo* and Hybrid Assembly Techniques to Derive High-Quality Genome Sequences. *Bioinformatics* **30**, 2709–2716 (2014).

[72] Boolchandani, M., D'Souza, A. W. & Dantas, G. Sequencing-Based Methods and Resources to study Antimicrobial Resistance. *Nature Reviews Genetics* **20**, 356–370 (2019).

[73] Loman, N. J., Quick, J. & Simpson, J. T. A Complete Bacterial Genome Assembled *de novo* using only Nanopore Sequencing Data. *Nature Methods* **12**, 733–735 (2015).

[74] Bejaoui, S. *et al.* Comparison of Illumina and Oxford Nanopore Sequencing Data Quality for *Clostridioides difficile* Genome Analysis and their Application for Epidemiological Surveillance. *BMC Genomics* **26**, 92 (2025).

[75] Su, M., Satola, S. W. & Read, T. D. Genome-Based Prediction of Bacterial Antibiotic Resistance. *Journal of Clinical Microbiology* **57**, e01405–18 (2019).

[76] McArthur, A. G. *et al.* The Comprehensive Antibiotic Resistance Database. *Antimicrobial Agents and Chemotherapy* **57**, 3348–3357 (2013).

[77] Alcock, B. P. *et al.* CARD 2023: Expanded Curation, Support for Machine Learning, and Resistome Prediction at the Comprehensive Antibiotic Resistance Database. *Nucleic Acids Research* **51**, D690–D699 (2023).

[78] Zankari, E. *et al.* Identification of Acquired Antimicrobial Resistance Genes. *Journal of Antimicrobial Chemotherapy* **67**, 2640–2644 (2012).

[79] World Health Organization. *Catalogue of mutations in Mycobacterium tuberculosis complex and their association with drug resistance, 2nd edition* (2023).

[80] Doyle, R. M. *et al.* Direct Whole-Genome Sequencing of Sputum Accurately Identifies Drug-Resistant *Mycobacterium tuberculosis* Faster than MGIT Culture Sequencing. *Journal of Clinical Microbiology* **56**, 10.1128/jcm.00666–18 (2018).

[81] Votintseva, A. A. *et al.* Same-Day Diagnostic and Surveillance Data for Tuberculosis via Whole-Genome Sequencing of Direct Respiratory Samples. *Journal of Clinical Microbiology* **55**, 1285–1298 (2017).

[82] Ellington, M. J. *et al.* The Role of Whole Genome Sequencing in Antimicrobial Susceptibility Testing of Bacteria: Report from the EUCAST Subcommittee. *Clinical Microbiology and Infection* **23**, 2–22 (2017).

[83] Gagneux, S. Host–Pathogen Coevolution in Human Tuberculosis. *Philosophical Transactions of the Royal Society B: Biological Sciences* **367**, 850–859 (2012).

[84] World Health Organisation. *Global tuberculosis report* (2024).

[85] Bagcchi, S. WHO's Global Tuberculosis Report 2022. *The Lancet Microbe* **4**, e20 (2023).

[86] Delogu, G., Sali, M. & Fadda, G. The Biology of *Mycobacterium tuberculosis* Infection. *Mediterranean Journal of Hematology and Infectious Diseases* **5**, e2013070 (2013).

[87] Lin, P. L. & Flynn, J. L. Understanding Latent Tuberculosis: A Moving Target. *Journal of Immunology* **185**, 15–22 (2010).

[88] Muñoz, L., Stagg, H. R. & Abubakar, I. Diagnosis and Management of Latent Tuberculosis Infection. *Cold Spring Harbor Perspectives in Medicine* **5**, a017830 (2015).

[89] Esmail, H., C. E. Barry, r., Young, D. B. & Wilkinson, R. J. The Ongoing Challenge of Latent Tuberculosis. *Philosophical Transactions of the Royal Society B: Biological Sciences* (2014).

[90] Houben, R. M. G. J. & Dodd, P. J. The Global Burden of Latent Tuberculosis Infection: A Re-estimation Using Mathematical Modelling. *PLoS Medicine* **13**, e1002152 (2016).

[91] Rocha, D. M. G. C., Viveiros, M., Saraiva, M. & Osório, N. S. The Neglected Contribution of Streptomycin to the Tuberculosis Drug Resistance Problem. *Genes* **12**, 2003 (2021).

[92] Murray, J. F., Schraufnagel, D. E. & Hopewell, P. C. Treatment of Tuberculosis. A Historical Perspective. *Annals of the American Thoracic Society* **12**, 1749–1759 (2015).

[93] Cohen, K. A. *et al.* Evidence for Expanding the Role of Streptomycin in the Management of Drug-Resistant *Mycobacterium tuberculosis*. *Antimicrobial Agents and Chemotherapy* **64**, 10.1128/aac.00860–20 (2020).

[94] World Health Organization. *WHO Consolidated Guidelines on Drug-Resistant Tuberculosis Treatment*. WHO Guidelines Approved by the Guidelines Review Committee (2019).

[95] World Health Organization. Rapid Communication: Key Changes to Treatment of Multidrug- and Rifampicin-Resistant Tuberculosis (MDR/RR-TB) (2018).

[96] Gygli, S. M., Borrell, S., Trauner, A. & Gagneux, S. Antimicrobial Resistance in *Mycobacterium tuberculosis*: Mechanistic and Evolutionary Perspectives. *FEMS Microbiology Reviews* **41**, 354–373 (2017).

[97] Almeida Da Silva, P. E. & Palomino, J. C. Molecular Basis and Mechanisms of Drug Resistance in *Mycobacterium tuberculosis*: Classical and New Drugs. *Journal of Antimicrobial Chemotherapy* **66**, 1417–1430 (2011).

[98] Brandis, G., Wrande, M., Liljas, L. & Hughes, D. Fitness-Compensatory Mutations in Rifampicin-Resistant RNA Polymerase. *Molecular Microbiology* **85**, 142–151 (2012).

[99] Chan, E. D. & Iseman, M. D. Multidrug-Resistant and Extensively Drug-Resistant Tuberculosis: A Review. *Current Opinion in Infectious Diseases* **21**, 587–595 (2008).

[100] Merker, M. *et al.* Evolutionary History and Global Spread of the *Mycobacterium tuberculosis* Beijing Lineage. *Nature Genetics* **47**, 242–249 (2015).

[101] World Health Organization. *WHO Bacterial Priority Pathogens List 2024: Bacterial Pathogens of Public Health Importance, to Guide Research, Development, and Strategies to Prevent and Control Antimicrobial Resistance, 1st ed* (2024).

[102] World Health Organization. Technical Manual for Drug Susceptibility Testing of Medicines Used in the Treatment of Tuberculosis. Tech. Rep. 9789241514842, World Health Organization, Geneva (2018).

[103] Folkvardsen, D. B. *et al.* Rifampin Heteroresistance in *Mycobacterium tuberculosis* Cultures as Detected by Phenotypic and Genotypic Drug Susceptibility Test Methods. *Journal of Clinical Microbiology* **51**, 4220–4222 (2013).

[104] Ghodbane, R., Raoult, D. & Drancourt, M. Dramatic Reduction of Culture Time of *Mycobacterium tuberculosis*. *Scientific Reports* **4**, 4236 (2014).

[105] Pfyffer, G. E. *et al.* Comparison of the Mycobacteria Growth Indicator Tube (MGIT) with Radiometric and Solid Culture for Recovery of Acid-Fast Bacilli. *Journal of Clinical Microbiology* **35**, 364–368 (1997).

[106] Yusoof, K. A. *et al.* Tuberculosis Phenotypic and Genotypic Drug Susceptibility Testing and Immunodiagnostics: A Review. *Frontiers in Immunology* **13** (2022).

[107] Pankhurst, L. J. *et al.* Rapid, Comprehensive, and Affordable Mycobacterial Diagnosis with Whole-Genome Sequencing: A Prospective Study. *The Lancet Respiratory Medicine* **4**, 49–58 (2016).

[108] Hillemann, D., Rüsch-Gerdes, S. & Richter, E. Evaluation of the GenoType MTBDRplus Assay for Rifampin and Isoniazid Susceptibility Testing of *Mycobacterium tuberculosis* Strains and Clinical Specimens. *Journal of Clinical Microbiology* **45**, 2635–2640 (2007).

[109] World Health Organization. *Rapid Implementation of the Xpert MTB/RIF Diagnostic Test: Technical and Operational 'How-to'; Practical Considerations* (2011).

[110] Papaventsis, D. *et al.* Whole Genome Sequencing of *Mycobacterium tuberculosis* for Detection of Drug Resistance: A Systematic Review. *Clinical Microbiology and Infection* **23**, 61–68 (2017).

[111] Votintseva, A. A. *et al.* Mycobacterial DNA Extraction for Whole-Genome Sequencing from Early Positive Liquid (MGIT) Cultures. *Journal of Clinical Microbiology* **53**, 1137–1143 (2015).

[112] Meehan, C. J. *et al.* Whole Genome Sequencing of *Mycobacterium tuberculosis*: Current Standards and Open Issues. *Nature Reviews Microbiology* **17**, 533–545 (2019).

[113] Cole, S. T. *et al.* Deciphering the Biology of *Mycobacterium tuberculosis* from the Complete Genome Sequence. *Nature* **393**, 537–544 (1998).

[114] Zignol, M. *et al.* Genetic Sequencing for Surveillance of Drug Resistance in Tuberculosis in Highly Endemic Countries: A Multi-Country Population-Based Surveillance Study. *The Lancet Infectious Diseases* **18**, 675–683 (2018).

[115] Walker, T. M. *et al.* Tuberculosis is Changing. *The Lancet Infectious Diseases* **17**, 359–361 (2017).

[116] World Health Organization. *WHO Consolidated Guidelines on Tuberculosis: Rapid Diagnostics for Tuberculosis Detection; Module 3: Diagnosis, Third Edition* (2024).

[117] Goulooze, S. C., Cohen, A. F. & Rissmann, R. Bedaquiline. *British Journal of Clinical Pharmacology* **80**, 182–184 (2015).

[118] Walker, T. M. *et al.* The 2021 WHO Catalogue of *Mycobacterium tuberculosis* Complex Mutations Associated with Drug Resistance: A Genotypic Analysis. *The Lancet Microbe* **3**, e265–e273 (2022).

[119] Warner, D. F., Koch, A. & Mizrahi, V. Diversity and disease pathogenesis in *Mycobacterium tuberculosis*. *Trends in Microbiology* **23**, 14–21 (2015).

[120] Carlet, J. *et al.* Society's Failure to Protect a Precious Resource: Antibiotics. *The Lancet* **378**, 369–371 (2011).

[121] Rodríguez-Baño, J., Gutiérrez-Gutiérrez, B., Machuca, I. & Pascual, A. Treatment of Infections Caused by Extended-Spectrum-Beta-Lactamase-, AmpC-, and Carbapenemase-Producing Enterobacteriaceae. *Clinical Microbiology Reviews* **31**, 10.1128/cmr.00079–17 (2018).

[122] Anderson, M. T. *et al.* Replication Dynamics for Six Gram-Negative Bacterial Species during Bloodstream Infection. *mBio* **12**, e01114–21 (2021).

[123] Zhu, M. & Dai, X. On the Intrinsic Constraint of Bacterial Growth Rate: *M. tuberculosis*'s View of the Protein Translation Capacity. *Critical Reviews in Microbiology* **44**, 455–464 (2018).

[124] Miller, W. R. & Arias, C. A. ESKAPE Pathogens: Antimicrobial Resistance, Epidemiology, Clinical Impact and Therapeutics. *Nature Reviews Microbiology* **22**, 598–616 (2024).

[125] Doi, Y., Adams-Haduch, J. M., Peleg, A. Y. & D'Agata, E. M. The Role of Horizontal Gene Transfer in the Dissemination of Extended-Spectrum Beta-Lactamase-Producing *Escherichia coli* and *Klebsiella pneumoniae* Isolates in an Endemic Setting. *Diagnostic microbiology and infectious disease* **74**, 34–38 (2012).

[126] Paterson, D. L. Resistance in Gram-Negative Bacteria: Enterobacteriaceae. *The American Journal of Medicine* **119**, S20–S28 (2006).

[127] Meletis, G. Carbapenem Resistance: Overview of the Problem and Future Perspectives. *Therapeutic Advances in Infectious Disease* **3**, 15–21 (2016).

[128] EUCAST. European Committee on Antimicrobial Susceptibility Testing Reading Guide for Broth Microdilution, Version 5.0 (2024).

[129] Carroll, K. C. *et al.* Evaluation of the BD Phoenix Automated Microbiology System for Identification and Antimicrobial Susceptibility Testing of *Enterobacteriaceae*. *Journal of Clinical Microbiology* **44**, 3506–3509 (2006).

[130] BD Becton Dickinson. Phoenix™ M50 Automated Microbiology System User's Manual (2025).

[131] Jonasson, E., Matuschek, E. & Kahlmeter, G. The EUCAST Rapid Disc Diffusion Method for Antimicrobial Susceptibility Testing Directly from Positive Blood Culture Bottles. *Journal of Antimicrobial Chemotherapy* **75**, 968–978 (2020).

[132] Åkerlund, A. *et al.* EUCAST Rapid Antimicrobial Susceptibility Testing (RAST) in Blood Cultures: Validation in 55 European Laboratories. *Journal of Antimicrobial Chemotherapy* **75**, 3230–3238 (2020).

[133] Vanstokstraeten, R. *et al.* Genotypic Resistance Determined by Whole Genome Sequencing versus Phenotypic Resistance in 234 *Escherichia coli* Isolates. *Scientific Reports* **13**, 449 (2023).

[134] Shelburne, S. A. *et al.* Whole-Genome Sequencing Accurately Identifies Resistance to Extended-Spectrum Beta-Lactams for Major Gram-Negative Bacterial Pathogens. *Clinical Infectious Diseases* **65**, 738–745 (2017).

[135] Stoesser, N. *et al.* Predicting Antimicrobial Susceptibilities for *Escherichia coli* and *Klebsiella pneumoniae* Isolates Using Whole Genomic Sequence Data. *Journal of Antimicrobial Chemotherapy* **68**, 2234–2244 (2013).

[136] Govender, K. N. *et al.* Rapid Clinical Diagnosis and Treatment of Common, Undetected, and Uncultivable Bloodstream Infections Using Metagenomic Sequencing from Routine Blood Cultures with Oxford Nanopore. *medRxiv* (2025).

[137] International Organization for Standardization. ISO 20776-2: Clinical Laboratory Testing and *in vitro* Diagnostic Test Systems - Susceptibility Testing of Infectious Agents and Evaluation of Performance of Antimicrobial Susceptibility Test Devices (2021).

[138] Brunner, V. M. & Fowler, P. W. Compensatory Mutations are Associated with Increased *In Vitro* Growth in Resistant Clinical Samples of *Mycobacterium tuberculosis*. *Microbial Genomics* **10**, 001187 (2024).

[139] Brunner, V. M. & Fowler, P. W. Subpopulations in Clinical Samples of *M. tuberculosis* Can Give Rise to Rifampicin Resistance and Shed Light on How Resistance Is Acquired. *JAC-Antimicrobial Resistance* **7**, dlaf175 (2025).

[140] The CRyPTIC Consortium. Epidemiological Cutoff Values for a 96-Well Broth Microdilution Plate for High-Throughput Research Antibiotic Susceptibility Testing of *M . tuberculosis*. *European Respiratory Journal* **60**, 2200239 (2022).

[141] The CRyPTIC Consortium. A Data Compendium Associating the Genomes of 12,289 *Mycobacterium tuberculosis* Isolates with Quantitative Resistance Phenotypes to 13 Antibiotics. *PLoS Biology* **20**, e3001721 (2022).

[142] The CRyPTIC Consortium & Fowler, P. The CRyPTIC Consortium Dataset (v1.1.1), [Dataset]. Zenodo. https://doi.org/10.5281/zenodo.15679731 (2021).

[143] The CRyPTIC Consortium & Fowler, P. The CRyPTIC Consortium Dataset (v3.0.0), [Dataset]. Zenodo. https://doi.org/10.5281/zenodo.16041005 (2025).

[144] Bortoluzzi, A. *et al.* *Mycobacterium tuberculosis* RNA Polymerase-Binding Protein A (RbpA) and its Interactions with Sigma Factors. *Journal of Biological Chemistry* **288**, 14438–14450 (2013).

[145] Lin, W. *et al.* Structural Basis of *Mycobacterium tuberculosis* Transcription and Transcription Inhibition. *Molecular Cell* **66**, 169–179.e8 (2017).

[146] Yuen, L. K. W., Leslie, D. & Coloe, P. J. Bacteriological and Molecular Analysis of Rifampin-Resistant *Mycobacterium tuberculosis* Strains Isolated in Australia. *Journal of Clinical Microbiology* **37**, 3844–3850 (1999).

[147] Aristoff, P. A., Garcia, G. A., Kirchhoff, P. D. & Hollis Showalter, H. D. Rifamycins – Obstacles and Opportunities. *Tuberculosis* **90**, 94–118 (2010).

[148] Rothstein, D. M. Rifamycins, Alone and in Combination. *Cold Spring Harbor Perspectives in Medicine* **6**, a027011 (2016).

[149] Billington, O. J., McHugh, T. D. & Gillespie, S. H. Physiological Cost of Rifampin Resistance Induced In Vitro in *Mycobacterium tuberculosis*. *Antimicrobial Agents and Chemotherapy* **43**, 1866–1869 (1999).

[150] Gagneux, S. *et al.* The Competitive Cost of Antibiotic Resistance in *Mycobacterium tuberculosis*. *Science* **312**, 1944–1946 (2006).

[151] Stefan, M. A., Ugur, F. S. & Garcia, G. A. Source of the Fitness Defect in Rifamycin-Resistant Mycobacterium tuberculosis RNA Polymerase and the Mechanism of Compensation by Mutations in the Beta' Subunit. *Antimicrobial Agents and Chemotherapy* **62**, e00164–18 (2018).

[152] Maisnier-Patin, S. & Andersson, D. I. Adaptation to the Deleterious Effects of Antimicrobial Drug Resistance Mutations by Compensatory Evolution. *Research in Microbiology* **155**, 360–369 (2004).

[153] Alame Emane, A. K., Guo, X., Takiff, H. E. & Liu, S. Drug Resistance, Fitness and Compensatory Mutations in *Mycobacterium tuberculosis*. *Tuberculosis* **129**, 102091 (2021).

[154] Song, T. *et al.* Fitness Costs of Rifampicin Resistance in *Mycobacterium tuberculosis* are Amplified Under Conditions of Nutrient Starvation and Compensated by Mutation in the Beta' Subunit of RNA Polymerase. *Molecular Microbiology* **91**, 1106–1119 (2014).

[155] Comas, I. *et al.* Whole-Genome Sequencing of Rifampicin-Resistant *Mycobacterium tuberculosis* Strains Identifies Compensatory Mutations in RNA Polymerase Genes. *Nature Genetics* **44**, 106–110 (2012).

[156] de Vos, M. *et al.* Putative Compensatory Mutations in the *rpoC* Gene of Rifampin-Resistant *Mycobacterium tuberculosis* Are Associated with Ongoing Transmission. *Antimicrobial Agents and Chemotherapy* **57**, 827–832 (2013).

[157] Li, Q. *et al.* Compensatory Mutations of Rifampin Resistance Are Associated with Transmission

of Multidrug-Resistant *Mycobacterium tuberculosis* Beijing Genotype Strains in China. *Antimicrobial Agents and Chemotherapy* **60**, 2807–2812 (2016).

[158] Casali, N. *et al.* Microevolution of Extensively Drug-Resistant Tuberculosis in Russia. *Genome Research* **22**, 735–745 (2012).

[159] Ali, A. *et al.* Whole Genome Sequencing Based Characterization of Extensively Drug-Resistant *Mycobacterium tuberculosis* Isolates from Pakistan. *PLOS ONE* **10**, e0117771 (2015).

[160] Ma, P. *et al.* Compensatory Effects of *M. tuberculosis rpoB* Mutations Outside the Rifampicin Resistance-Determining Region. *Emerging Microbes & Infections* **10**, 743–752 (2021).

[161] Ruiz, V. & Paula, A. Determination of Potentially Novel Compensatory Mutations in *rpoC* Associated with Rifampin Resistance and *rpoB* Mutations in *Mycobacterium tuberculosis* Clinical Isolates from Peru. *Int J Mycobacteriol* **9**, 121–137 (2020).

[162] Gygli, S. M. *et al.* Prisons as Ecological Drivers of Fitness-Compensated Multidrug-Resistant *Mycobacterium tuberculosis*. *Nature Medicine* **27**, 1171–1177 (2021).

[163] Merker, M. *et al.* Compensatory Evolution Drives Multidrug-Resistant Tuberculosis in Central Asia. *eLife* **7**, e38200 (2018).

[164] Loiseau, C. *et al.* The Relative Transmission Fitness of Multidrug-Resistant *Mycobacterium tuberculosis* in a Drug Resistance Hotspot. *Nature Communications* **14**, 1988 (2023).

[165] Chen, Y., Liu, Q., Takiff, H. E. & Gao, Q. Comprehensive Genomic Analysis of *Mycobacterium tuberculosis* Reveals Limited Impact of High-Fitness Genotypes on MDR-TB Transmission. *Journal of Infection* **85**, 49–56 (2022).

[166] Goig, G. A. *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* **4**, e506–e515 (2023).

[167] Fowler, P. W. *et al.* Automated Detection of Bacterial Growth on 96-well Plates for High-Throughput Drug Susceptibility Testing of *Mycobacterium tuberculosis*. *Microbiology* **164**, 1522–1530 (2018).

[168] Brunner, V. Accompanying Code to Reproduce Analysis and Figures in this Thesis. https://github.com/viktoria023/thesis-repository.

[169] The CRyPTIC Consortium and the 100,000 Genomes Project *et al.* Prediction of Susceptibility to First-Line Tuberculosis Drugs by DNA Sequencing. *The New England Journal of Medicine* **379**, 1403–1415 (2018).

[170] Wood, D. E., Lu, J. & Langmead, B. Improved Metagenomic Analysis with Kraken 2. *Genome Biology* **20**, 257 (2019).

[171] Li, H. Minimap2: Pairwise Alignment for Nucleotide Sequences. *Bioinformatics* **34**, 3094–3100 (2018).

[172] Hunt, M. *et al.* Antibiotic Resistance Prediction for *Mycobacterium tuberculosis* from Genome Sequence Data with Mykrobe. *Wellcome Open Research* **4**, 191 (2019).

[173] Hunt, M. *et al.* Minos: Variant Adjudication and Joint Genotyping of Cohorts of Bacterial Genomes. *Genome Biology* **23**, 147 (2022).

[174] Westhead, J. & Fowler, P. W. gnomonicus: https://github.com/oxfordmmm/gnomonicus (2023).

[175] Rancoita, P. M. V. *et al.* Validating a 14-Drug Microtiter Plate Containing Bedaquiline and Delamanid for Large-Scale Research Susceptibility Testing of *Mycobacterium tuberculosis*. *Antimicrobial Agents and Chemotherapy* **62**, e00344–18 (2018).

[176] Earle, S. G. *et al.* Identifying Lineage Effects when Controlling for Population Structure Improves Power in Bacterial Association Studies. *Nature Microbiology* **1**, 1–8 (2016).

[177] Chen, P. E. & Shapiro, B. J. Classic Genome-Wide Association Methods are Unlikely to Identify Causal Variants in Strongly Clonal Microbial Populations. *bioRxiv* (2021).

[178] Malone, K. M. & Brankin, A. E. CRyPTIC phylogenetic tree: https://github.com/kerrimalone/Brankin_Malone_2022 (2022).

[179] Hunt, M. *et al.* mykrobe: https://github.com/mykrobe-tools/mykrobe (2024).

[180] Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: An Online Tool for Phylogenetic Tree Display and Annotation. *Nucleic Acids Research* **49**, W293–W296 (2021).

[181] Miotto, P., Cabibbe, A. M., Borroni, E., Degano, M. & Cirillo, D. M. Role of Disputed Mutations in the *rpoB* Gene in Interpretation of Automated Liquid MGIT Culture Results for Rifampin Susceptibility Testing of Mycobacterium tuberculosis. *Journal of Clinical Microbiology* **56**, e01599–17 (2018).

[182] Karmakar, M., Trauer, J. M., Ascher, D. B. & Denholm, J. T. Hyper Transmission of Beijing Lineage *Mycobacterium tuberculosis*: Systematic Review and Meta-Analysis. *Journal of Infection* **79**, 572–581 (2019).

[183] UK Health Security Agency. *Mycobacterium tuberculosis* Whole-Genome Sequencing and Cluster Investigation Handbook. *GOV.UK* (2022).

[184] Nickels, B. E. & Hochschild, A. Regulation of RNA Polymerase through the Secondary Channel. *Cell* **118**, 281–284 (2004).

[185] San, L. L. *et al.* Insight into Multidrug-Resistant Beijing Genotype *Mycobacterium tuberculosis* Isolates in Myanmar. *International Journal of Infectious Diseases* **76**, 109–119 (2018).

[186] Ribeiro, S. C. M. *et al.* *Mycobacterium tuberculosis* Strains of the Modern Sublineage of the Beijing Family Are More Likely To Display Increased Virulence than Strains of the Ancient Sublineage. *Journal of Clinical Microbiology* **52**, 2615–2624 (2014).

[187] Ford, C. B. *et al.* *Mycobacterium tuberculosis* Mutation Rate Estimates from Different Lineages Predict Substantial Differences in the Emergence of Drug-Resistant Tuberculosis. *Nat Genet* **45**, 784–790 (2013).

[188] van den Berg, S., Vandenplas, J., van Eeuwijk, F. A., Lopes, M. S. & Veerkamp, R. F. Significance Testing and Genomic Inflation Factor Using High-Density Genotypes or Whole-Genome Sequence Data. *Journal of Animal Breeding and Genetics* **136**, 418–429 (2019).

[189] Zhou, X. & Stephens, M. Genome-Wide Efficient Mixed-Model Analysis for Association Studies. *Nature Genetics* **44**, 821–824 (2012).

[190] The CRyPTIC Consortium. Genome-wide Association Studies of Global *Mycobacterium tuberculosis* Resistance to 13 Antimicrobials in 10,228 Genomes Identify New Resistance Mechanisms. *PLOS Biology* **20**, e3001755 (2022).

[191] Napier, G., Campino, S., Phelan, J. E. & Clark, T. G. Large-Scale Genomic Analysis of *Mycobacterium tuberculosis* Reveals Extent of Target and Compensatory Mutations Linked to Multi-Drug Resistant Tuberculosis. *Scientific Reports* **13**, 623 (2023).

[192] Bastolla, U. *et al.* Fitness Effect of the Isoniazid Resistance Mutation S315T of the Catalase-Peroxidase Enzyme KatG of *Mycobacterium tuberculosis*. *Genome Biology and Evolution* **17**, evaf120 (2025).

[193] Reynolds, M. G. Compensatory Evolution in Rifampin-Resistant *Escherichia coli*. *Genetics* **156**, 1471–1481 (2000).

[194] Cheng, S. *et al.* Within-Host Genotypic and Phenotypic Diversity of Contemporaneous Carbapenem-Resistant *Klebsiella pneumoniae* from Blood Cultures of Patients with Bacteremia. *mBio* **13**, e02906–22 (2022).

[195] Levert, M. *et al.* Molecular and Evolutionary Bases of Within-Patient Genotypic and Phenotypic Diversity in *Escherichia coli* Extraintestinal Infections. *PLOS Pathogens* **6**, e1001125 (2010).

[196] Brankin, A. E. & Fowler, P. W. Inclusion of minor alleles improves catalogue-based prediction of fluoroquinolone resistance in *Mycobacterium tuberculosis*. *JAC-Antimicrobial Resistance* **5**, dlad039 (2023).

[197] Andersson, D. I., Nicoloff, H. & Hjort, K. Mechanisms and Clinical Relevance of Bacterial Heteroresistance. *Nature Reviews Microbiology* **17**, 479–496 (2019).

[198] Faksri, K. *et al.* Comparisons of Whole-Genome Sequencing and Phenotypic Drug Susceptibility Testing for *Mycobacterium tuberculosis* Causing MDR-TB and XDR-TB in Thailand. *International Journal of Antimicrobial Agents* **54**, 109–116 (2019).

[199] Bryant, J. M. *et al.* Whole-Genome Sequencing to Establish Relapse or Re-Infection with *My-*

cobacterium tuberculosis: A Retrospective Observational Study. *The Lancet Respiratory Medicine* **1**, 786–792 (2013).

[200] Köser, C. U. *et al.* Whole-Genome Sequencing for Rapid Susceptibility Testing of *M. tuberculosis*. *New England Journal of Medicine* **369**, 290–292 (2013).

[201] Votintseva, A. A. *et al.* Mycobacterial DNA Extraction for Whole-Genome Sequencing from Early Positive Liquid (MGIT) Cultures. *Journal of Clinical Microbiology* **53**, 1137–1143 (2015).

[202] World Health Organization. *Catalogue of mutations in Mycobacterium tuberculosis complex and their association with drug resistance, 1st edition* (2021).

[203] Heyman, G. *et al.* Prevalence, Misclassification, and Clinical Consequences of the Heteroresistant Phenotype in *Escherichia coli* Bloodstream Infections in Patients in Uppsala, Sweden: A Retrospective Cohort Study. *The Lancet Microbe* **6**, 101010 (2025).

[204] Westhead, J. *et al.* Enhancement and Validation of the Antibiotic Resistance Prediction Performance of a Cloud-Based Genetics Processing Platform for Mycobacteria. *bioRxiv* 2024.11.08.622466 (2025).

[205] Constantinides, B., Hunt, M. & Crook, D. W. Hostile: Accurate Decontamination of Microbial Host Sequences. *Bioinformatics* **39**, btad728 (2023).

[206] Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of Long, Error-Prone Reads Using Repeat Graphs. *Nature Biotechnology* **37**, 540–546 (2019).

[207] Edge, P. & Bansal, V. Longshot Enables Accurate Variant Calling in Diploid Genomes from Single-Molecule Long Read Sequencing. *Nature Communications* **10**, 4660 (2019).

[208] Shaw, J., Gounot, J.-S., Chen, H., Nagarajan, N. & Yu, Y. W. Floria: Fast and Accurate Strain Haplotyping in Metagenomes. *Bioinformatics* **40**, i30–i38 (2024).

[209] Hameed, H. M. A. *et al.* Characterization of Genetic Variants Associated with Rifampicin Resistance Level in *Mycobacterium tuberculosis* Clinical Isolates Collected in Guangzhou Chest Hospital, China. *Infection and Drug Resistance* **15**, 5655–5666 (2022).

[210] Pérez-Lago, L. *et al.* Whole Genome Sequencing Analysis of Intrapatient Microevolution in *Mycobacterium tuberculosis*: Potential Impact on the Inference of Tuberculosis Transmission. *The Journal of Infectious Diseases* **209**, 98–108 (2014).

[211] Poonawala, H., Kumar, N. & Peacock, S. J. A Review of Published Spoligotype Data Indicates the Diversity of *Mycobacterium tuberculosis* from India is Under-Represented in Global Databases. *Infection, Genetics and Evolution* **78**, 104072 (2020).

[212] Van Deun, A. *et al.* Acquired Rifampicin Resistance During First TB Treatment: Magnitude, Relative Importance, Risk Factors and Keys to Control in Low-Income Settings. *JAC-Antimicrobial Resistance* **4**, dlac037 (2022).

[213] Kendall, E. A., Fofana, M. O. & Dowdy, D. W. Burden of Transmitted Multidrug Resistance in Epidemics of Tuberculosis: A Transmission Modelling Analysis. *The Lancet Respiratory Medicine* **3**, 963–972 (2015).

[214] Trauer, J. M., Denholm, J. T. & McBryde, E. S. Construction of a Mathematical Model for Tuberculosis Transmission in Highly Endemic Regions of the Asia-Pacific. *Journal of Theoretical Biology* **358**, 74–84 (2014).

[215] Lynch, C. I., Adlard, D. & Fowler, P. W. Predicting rifampicin resistance in M. tuberculosis using machine learning informed by protein structural and chemical features. *ERJ Open Research* (2025).

[216] Engelthaler, D. M. *et al.* Minority *Mycobacterium tuberculosis* Genotypic Populations as an Indicator of Subsequent Phenotypic Resistance. *American Journal of Respiratory Cell and Molecular Biology* **61**, 789–791 (2019).

[217] Walker, T. M. *et al.* Whole-Genome Sequencing for Prediction of *Mycobacterium tuberculosis* Drug Susceptibility and Resistance: A Retrospective Cohort Study. *The Lancet Infectious Diseases* **15**, 1193–1202 (2015).

[218] Coll, F. *et al.* Rapid Determination of Anti-Tuberculosis Drug Resistance from Whole-Genome Sequences. *Genome Medicine* **7**, 51 (2015).

[219] Georghiou, S. B. *et al.* Evaluation of Genetic Mutations Associated with *Mycobacterium tuberculosis* Resistance to Amikacin, Kanamycin and Capreomycin: A Systematic Review. *PLOS ONE* **7**, e33275 (2012).

[220] Kouchaki, S. *et al.* Application of Machine Learning Techniques to Tuberculosis Drug Resistance Analysis. *Bioinformatics* **35**, 2276–2282 (2019).

[221] Yang, Y. *et al.* Machine Learning for Classifying Tuberculosis Drug-Resistance from DNA Sequencing Data. *Bioinformatics* **34**, 1666–1671 (2018).

[222] Van Camp, P.-J., Haslam, D. B. & Porollo, A. Prediction of Antimicrobial Resistance in Gram-Negative Bacteria From Whole-Genome Sequencing Data. *Frontiers in Microbiology* **11** (2020).

[223] Kuang, X., Wang, F., Hernandez, K. M., Zhang, Z. & Grossman, R. L. Accurate and Rapid Prediction of Tuberculosis Drug Resistance from Genome Sequence Data Using Traditional Machine Learning Algorithms and CNN. *Scientific Reports* **12**, 2427 (2022).

[224] Zhang, A., Teng, L. & Alterovitz, G. An Explainable Machine Learning Platform for Pyrazinamide Resistance Prediction and Genetic Feature Identification of *Mycobacterium tuberculosis*. *Journal of the American Medical Informatics Association* **28**, 533–540 (2021).

[225] Jin, C. *et al.* Predicting Antimicrobial Resistance in *E. coli* with Discriminative Position Fused Deep Learning Classifier. *Computational and Structural Biotechnology Journal* **23**, 559–565 (2024).

[226] Gudivada, V., Apon, A. & Ding, J. Data Quality Considerations for Big Data and Machine Learning: Going Beyond Data Cleaning and Transformations. *International Journal on Advances in Software* **10**, 1–20 (2017).

[227] Fàbrega, A., Madurga, S., Giralt, E. & Vila, J. Mechanism of Action of and Resistance to Quinolones. *Microbial biotechnology* **2**, 40–61 (2009).

[228] Goldstein, B. P. Resistance to Rifampicin: A Review. *The Journal of Antibiotics* **67**, 625–630 (2014).

[229] Jumper, J. *et al.* Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

[230] Zhang, Y., Shi, W., Zhang, W. & Mitchison, D. Mechanisms of Pyrazinamide Action and Resistance. *Microbiology Spectrum* **2**, 10.1128/microbiolspec.mgm2–0023–2013 (2014).

[231] Carter, J. J. *et al.* Prediction of Pyrazinamide Resistance in *Mycobacterium tuberculosis* Using Structure-Based Machine-Learning Approaches. *JAC-Antimicrobial Resistance* **6**, dlae037 (2024).

[232] Lynch, C. I., Adlard, D. & Fowler, P. W. Predicting Rifampicin Resistance in *Mycobacterium tuberculosis* Using Machine Learning Informed by Protein Structural and Chemical Features. *ERJ Open Research* **11** (2025).

[233] Alzubaidi, L. *et al.* Review of Deep Learning: Concepts, CNN Architectures, Challenges, Applications, Future Directions. *Journal of Big Data* **8**, 53 (2021).

[234] Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. *Communications of the ACM* **60**, 84–90 (2017).

[235] Kipf, T. N. & Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. *5th International Conference on Learning Representations (ICLR)* (2017).

[236] Zhang, Z. *et al.* Hierarchical Multi-View Graph Pooling with Structure Learning. *IEEE Transactions on Knowledge and Data Engineering* 1–1 (2021).

[237] Defferrard, M., Bresson, X. & Vandergheynst, P. Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering. *Advances in Neural Information Processing Systems 29* 3844 – 3852 (2016).

[238] Jiménez, J., Doerr, S., Martínez-Rosell, G., Rose, A. S. & De Fabritiis, G. DeepSite: Protein-Binding Site Predictor Using 3D-Convolutional Neural Networks. *Bioinformatics* **33**, 3036–3042 (2017).

[239] Gligorijević, V. *et al.* Structure-Based Protein Function Prediction Using Graph Convolutional Networks. *Nature Communications* **12**, 3168 (2021).

[240] Rajakumari, K. *et al.* Comprehensive Review of DNA Gyrase as Enzymatic Target for Drug Discovery and Development. *European Journal of Medicinal Chemistry Reports* **12**, 100233 (2024).

[241] Vos, S. M., Tretter, E. M., Schmidt, B. H. & Berger, J. M. All Tangled up: How Cells Direct, Manage and Exploit Topoisomerase Function. *Nature Reviews Molecular Cell Biology* **12**, 827–841 (2011).

[242] Reece, R. J. & Maxwell, A. DNA Gyrase: Structure and Function. *Critical Reviews in Biochemistry and Molecular Biology* **26**, 335–375 (1991).

[243] Blower, T. R., Williamson, B. H., Kerns, R. J. & Berger, J. M. Crystal structure and stability of gyrase–fluoroquinolone cleaved complexes from *Mycobacterium tuberculosis*. *Proceedings of the National Academy of Sciences* **113**, 1706–1713 (2016).

[244] Weidlich, D. & Klostermeier, D. Functional Interactions between Gyrase Subunits Are Optimized in a Species-Specific Manner. *The Journal of Biological Chemistry* **295**, 2299–2312 (2020).

[245] Guillemin, I., Jarlier, V. & Cambau, E. Correlation between Quinolone Susceptibility Patterns and Sequences in the A and B Subunits of DNA Gyrase in Mycobacteria. *Antimicrobial Agents and Chemotherapy* **42**, 2084–2088 (1998).

[246] Biewald, L. Experiment tracking with weights and biases: https://www.wandb.com/ (2020).

[247] Fowler, P. W., Lynch, C. & Adlard, D. sbmlcore (v0.2.8). [Python Package]. Zenodo. https://github.com/fowler-lab/fowler-lab/sbmlcore (2024).

[248] Heinig, M. & Frishman, D. STRIDE: A Web Server for Secondary Structure Assignment from Known Atomic Coordinates of Proteins. *Nucleic Acids Research* **32**, W500–W502 (2004).

[249] Mitternacht, S. FreeSASA: An Open Source C Library for Solvent Accessible Surface Area Calculations. *F1000Research* **5**, 189 (2016).

[250] Hecht, M., Bromberg, Y. & Rost, B. Better Prediction of Functional Effects for Sequence Variants. *BMC Genomics* **16**, S1 (2015).

[251] Cao, H., Wang, J., He, L., Qi, Y. & Zhang, J. Z. DeepDDG: Predicting the Stability Change of

Protein Point Mutations Using Neural Networks. *Journal of Chemical Information and Modeling* **59**, 1508–1514 (2019).

[252] Fowler, P. W., Brunner, V. & Dissanayake, D. sbmlsim (v0.1.0). [Python Package]. Zenodo. https://github.com/fowler-lab/fowler-lab/sbmlsim (2025).

[253] Michalczyk, E. *et al.* Molecular Mechanism of Topoisomerase Poisoning by the Peptide Antibiotic Albicidin. *Nature Catalysis* **6**, 52–67 (2023).

[254] Russell, R. B. & Barton, G. J. Multiple Protein Sequence Alignment from Tertiary Structure Comparison: Assignment of Global and Residue Confidence Levels. *Proteins* **14**, 309–323 (1992).

[255] Russell, R. B., Walsh, T. & Barton, G. *Structural Alignment of Multiple Proteins Version 4.4 User Guide.* (Laboratory of Molecular Biophysics, University of Oxford.).

[256] Hooper, D. C. & Jacoby, G. A. Mechanisms of Drug Resistance: Quinolone Resistance. *Annals of the New York Academy of Sciences* **1354**, 12–31 (2015).

[257] Michaud-Agrawal, N., Denning, E. J., Woolf, T. B. & Beckstein, O. MDAnalysis: A Toolkit for the Analysis of Molecular Dynamics Simulations. *Journal of Computational Chemistry* **32**, 2319–2327 (2011).

[258] Gowers, R. J. *et al.* MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations. *scipy* (2016).

[259] Feldgarden, M. *et al.* AMRFinderPlus and the Reference Gene Catalog Facilitate Examination of the Genomic Links among Antimicrobial Resistance, Stress Response, and Virulence. *Scientific Reports* **11**, 12728 (2021).

[260] Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning*. Springer Series in Statistics (New York, NY, USA, 2001).

[261] Peduzzi, P., Concato, J., Kemper, E., Holford, T. R. & Feinstein, A. R. A Simulation Study of the Number of Events per Variable in Logistic Regression Analysis. *Journal of Clinical Epidemiology* **49**, 1373–1379 (1996).

[262] van der Putten, B. C. L. *et al.* Quantifying the Contribution of Four Resistance Mechanisms to Ciprofloxacin MIC in *Escherichia coli*: A Systematic Review. *Journal of Antimicrobial Chemotherapy* **74**, 298–310 (2019).

[263] Dissanayake, D., Morrone, J. A. & Fowler, P. W. Predicting Pyrazinamide Resistance in *Mycobacterium tuberculosis* Using a Graph Convolutional Network. *bioRxiv* (2025).

[264] Bouaouina, S. *et al.* Epistasis between Drug Resistance-Conferring Mutations in *Mycobacterium tuberculosis*. *bioRxiv* 2025.08.07.669101 (2025).

[265] Cohen, T. *et al.* Within-Host Heterogeneity of *Mycobacterium tuberculosis* Infection Is Associated With Poor Early Treatment Response: A Prospective Cohort Study. *The Journal of Infectious Diseases* **213**, 1796–1799 (2016).

[266] Moreira de Gouveia, M. I., Bernalier-Donadille, A. & Jubelin, G. *Enterobacteriaceae* in the Human Gut: Dynamics and Ecological Roles in Health and Disease. *Biology* **13**, 142 (2024).

[267] Pesesky, M. W. *et al.* Evaluation of Machine Learning and Rules-Based Approaches for Predicting Antimicrobial Resistance Profiles in Gram-Negative Bacilli from Whole Genome Sequence Data. *Frontiers in Microbiology* **7**, 1887 (2016).

8 Appendix

resistance	putative CM	only CM	both	-$\log_{10}$(p-value)
rpoB S450L	*rpoC* I491V	19	665	inf
rpoB S450L	*rpoC* E1092D	2012	1989	inf
rpoB S450L	*rpoC* V483A	33	586	inf
rpoB S450L	*rpoC* V483G	37	1206	inf
rpoB S450L	*rpoC* I491T	10	457	293.055074
rpoB D435G	*rpoB* I1106T	3	103	263.704207
rpoB S450L	*rpoC* F452S	2	345	230.575216
rpoB S450L	*rpoC* P1040R	32	396	225.093082
rpoB L452P	*rpoB* I1106T	3	103	212.491999
rpoB S450L	*rpoB* E761D	0	304	207.047893
rpoB S450L	*rpoB* L731P	1	226	151.442572
rpoB S450L	*rpoC* N698S	2	205	135.233842
rpoB S450L	*rpoC* C62C	85	313	133.388191
rpoB S450L	*rpoC* V517L	1	184	122.869655
rpoB S450L	*rpoC* D485Y	8	194	118.885081
rpoB S450L	*rpoC* V1252L	3	175	113.236538
rpoB S450L	*rpoA* T187A	3	171	110.539929
rpoB S450L	*rpoC* D485N	2	166	108.820765
rpoB S450L	*rpoC* G332S	16	179	100.192785
rpoB H445R	*rpoC* S561P	2	39	98.271832
rpoB I491F	*rpoC* E1033A	25	46	95.599395
rpoB S450L	*rpoC* L516P	3	144	92.368503
rpoB S450L	*rpoC* G433S	3	141	90.353083
rpoB S450L	*rpoB* A1075A	28594	8806	76.095081
rpoB S450L	*rpoC* P1040S	3	118	74.930148
rpoB H445Y	*rpoC* S548S	56	54	73.464142
rpoB S450L	*rpoC* P1040A	1	110	72.701835
rpoB S450L	*rpoB* R827C	8	123	72.056710
rpoB S450L	*rpoC* L527V	3	113	71.584879
rpoB S450L	*rpoC* G332R	5	113	68.947351
rpoB S450L	*rpoC* K445R	0	98	66.488253
rpoB V170F	*rpoB* V168A	2	25	64.982745
rpoB S450L	*rpoC* F452L	3	96	60.235512
rpoB S450L	*rpoC* L547V	2	80	50.937746
rpoB S450L	*rpoB* K891E	0	74	50.182923
rpoB S441A	*rpoB* S428S	20	15	49.332972
rpoB S450L	*rpoB* I480V	3	78	48.270034
rpoB D435V	*rpoC* A542A	10091	453	47.280576

rpoB L452P	*rpoB* H1028R	1	23	46.468449
rpoB S450W	*rpoA* P25R	1	20	45.519323
rpoB S450L	*rpoZ* S22S	1	68	44.370425
rpoB S450L	*rpoC* V1039A	1	68	44.370425
rpoB S450L	*rpoC* N416S	3	72	44.296722
rpoB S450L	*rpoC* N826T	0	64	43.393349
rpoB S450L	*rpoC* W484G	11	80	41.688865
rpoB S450L	*rpoC* L507V	1	64	41.680760
rpoB S450L	*rpoC* A521D	2	65	40.928287
rpoB S450L	*rpoB* V496A	0	60	40.678230
rpoB S450L	*rpoA* V183G	10	77	40.623323
rpoB S441A	*rpoB* L464M	3	11	39.118755
rpoB S441A	*rpoB* Q432Q	141	15	38.366789
rpoB S450L	*rpoB* P45S	5	65	37.490968
rpoB S450L	*rpoB* A692T	16	79	37.445343
rpoB S450L	*rpoC* V1252M	2	59	36.937475
rpoB S450L	*rpoB* I488V	2	59	36.937475
rpoB S450L	*rpoC* V431M	4	62	36.579062
rpoB S441A	*rpoC* L405M	28	12	36.045299
rpoB S450W	*sigA* A223T	0	15	35.074417
rpoB S441A	*rpoB* L443L	138	14	34.998919
rpoB S441A	*rpoB* A451A	4	10	34.646643
rpoB S450L	*rpoC* K1152Q	0	51	34.570692
rpoB S450L	*rpoC* L449V	0	50	33.892204
rpoB S450L	*rpoB* R827L	0	50	33.892204
rpoB D435G	*rpoB* I491L	18	17	32.890847
rpoB S450L	*rpoB* Q409R	14	69	32.770860
rpoB S450L	*rpoB* A286V	4	56	32.675893
rpoB S450L	*rpoA* A180V	0	48	32.535302
rpoB H445D	*rpoC* G388A	1	17	32.329453
rpoB S441A	*rpoB* G453G	128	13	32.105190
rpoB S441A	*sigA* V375V	31	11	32.049217
rpoB S441A	*sigA* I382V	3	9	31.338741
rpoB S441A	*sigA* G380A	3	9	31.338741
rpoB S450L	*rpoA* D190G	0	46	31.178502
rpoB S450L	*rpoC* F452C	1	48	30.944960
rpoB V170F	*rpoC* G571R	11	14	30.914415
rpoB S441A	*sigA* L386M	4	9	30.826900
rpoB S450L	*rpoC* T812I	3	51	30.476049
rpoB H445D	*rpoC* A542A	10307	237	30.336130
rpoB S441A	*rpoB* G426G	20	10	30.169890

rpoB Q432K	*rpoZ* T107I	1	10	30.130374
rpoB D435G	*rpoC* A542A	10437	107	29.915573
rpoB S450L	*rpoA* G31S	0	44	29.821803
rpoB D435Y	*rpoB* R167C	1	15	29.718602
rpoB H445Y	*rpoB* E207K	0	15	29.047955
rpoB Q432P	*rpoC* T853A	3	10	28.715938
rpoB S450L	*rpoB* Q975H	17	65	28.567694
rpoB S441A	*rpoB* P416P	75	11	28.293538
rpoB S441A	*rpoB* G442G	82	11	27.897222
rpoB S441A	*sigA* E385Q	14	9	27.769306
rpoB S441A	*rpoC* I128V	3	8	27.555788
rpoB S450L	*rpoC* N698K	4	48	27.502124
rpoB K446Q	*rpoB* T444T	5	8	27.295752
rpoB S450L	*rpoA* E184D	0	40	27.108709
rpoB S441A	*rpoC* D404D	101	11	26.962895
rpoB S441A	*rpoB* E460D	5	8	26.663789
rpoB S441A	*rpoB* S431S	113	11	26.455069
rpoB S441A	*rpoB* V469V	62	10	25.919864
rpoB S441A	*rpoB* R791T	0	7	25.916366
rpoB S450L	*rpoC* P434R	3	44	25.910863
rpoB S450W	*rpoC* G519D	52	18	25.858436
rpoB S450L	*rpoB* F503S	1	40	25.595375
rpoB S441A	*rpoC* R459R	76	10	25.100601
rpoB S450L	*rpoB* I491V	0	37	25.074154
rpoB S450W	*sigA* D146E	143	22	24.718437
rpoB S441A	*rpoB* G463G	88	10	24.503851
rpoB S441A	*sigA* G340G	176	11	24.426276
rpoB S450L	*rpoB* S874Y	0	36	24.396020
rpoB S450L	*rpoB* R552L	0	36	24.396020
rpoB S450L	*rpoA* G31A	0	35	23.717911
rpoB S450L	*rpoC* P1040L	0	35	23.717911
rpoB S450L	*rpoC* D747A	0	35	23.717911
rpoB D435Y	*rpoB* N487S	1	12	23.611072
rpoB S450L	*rpoB* T400A	1	37	23.593619
rpoB S441A	*rpoB* L796L	19	8	23.427668
rpoB S450L	*rpoC* H525Q	8	46	22.959081
rpoB S441A	*rpoB* L411L	135	10	22.735784
rpoB Q432P	*rpoB* K799Q	3	8	22.648819
rpoB H445L	*rpoB* P802P	18	13	22.548069
rpoB S450L	*rpoC* S428A	2	37	22.402814
rpoB S441A	*rpoB* P768P	7	7	22.381189

rpoB S450W	*rpoB* L862R	1	10	22.306854
rpoB S441A	*sigA* Q383Q	156	10	22.131372
rpoB S441A	*rpoC* V378V	75	9	22.118200
rpoB S441A	*rpoB* T482S	0	6	22.105354
rpoB S441A	*rpoC* G393G	77	9	22.021672
rpoB V359A	*rpoB* c-61t	8727	25	21.736503
rpoB S441A	*sigA* A347A	176	10	21.624922
rpoB S441A	*rpoC* V431V	35	8	21.613406
rpoB I491F	*rpoC* H748P	18	12	21.559555
rpoB D435Y	*rpoB* c-61t	8587	165	21.522392
rpoB S441A	*rpoB* D772E	1	6	21.260314
rpoB S441A	*rpoC* R427R	195	10	21.193143
rpoB I491F	*rpoC* G594E	10510	100	21.163429
rpoB V359A	*rpoB* G876G	9211	25	21.151306
rpoB S450L	*rpoC* E1033K	2	35	21.092603
rpoB S441A	*rpoB* R813R	43	8	20.971203
rpoB S450L	*rpoC* E1137G	1	33	20.929209
rpoB S450W	*rpoC* L873L	113	18	20.561958
rpoB S441A	*rpoB* L430L	16	7	20.527767
rpoB S441A	*rpoB* R476R	50	8	20.492939
rpoB S441A	*rpoB* V417V	118	9	20.437508
rpoB S441A	*rpoC* L402I	17	7	20.378058
rpoB S441A	*rpoC* R414R	123	9	20.281910
rpoB S441A	*rpoC* K437Q	3	6	20.181248
rpoB H445N	*rpoB* I491M	16	11	20.082604
rpoB S441A	*rpoB* R467R	58	8	20.016715
rpoB D435Y	*rpoB* G876G	9069	167	19.944520
rpoB S441A	*rpoB* E481E	136	9	19.904239
rpoB S441A	*rpoB* G793G	137	9	19.876649
rpoB S441A	*rpoB* S458S	138	9	19.849253
rpoB S441A	*rpoB* R448R	138	9	19.849253
rpoB S450L	*rpoB* P45L	10	43	19.849063
rpoB S441A	*rpoC* L446Q	21	7	19.844110
rpoB L430P	*rpoC* E1092D	3920	81	19.824687
rpoB S450L	*rpoC* T1230I	2	33	19.785067
rpoB S441A	*rpoC* G426G	4	6	19.783365
rpoB S441A	*rpoB* I751V	4	6	19.783365
rpoB S441A	*rpoC* Y116Y	142	9	19.741556
rpoB S450L	*rpoC* G519D	19	51	19.665774
rpoB S450L	*rpoC* N698H	0	29	19.649786
rpoB S441A	*rpoC* T55T	24	7	19.497749

rpoB S441A	*rpoB* V784I	5	6	19.441000
rpoB S441A	*sigA* F317F	155	9	19.410789
rpoB S441A	*rpoC* P454P	26	7	19.287104
rpoB S441A	*rpoB* M1025L	27	7	19.187042
rpoB S441A	*rpoC* A85S	28	7	19.090185
rpoB S441A	*rpoC* N384N	171	9	19.038845
rpoB L452P	*rpoC* A542A	10369	175	18.941368
rpoB S441A	*rpoB* A609A	7	6	18.871240
rpoB S441A	*rpoB* T1018S	7	6	18.871240
rpoB S441A	*rpoB* D792A	7	6	18.871240
rpoB H445N	*rpoC* Y61Y	164	17	18.826016
rpoB S441A	*rpoC* V456V	31	7	18.816937
rpoB H445D	*rpoZ* T107I	1	10	18.694010
rpoB S441A	*rpoB* T809T	87	8	18.692525
rpoB S450W	*rpoB* R273C	0	8	18.665573
rpoB S441A	*rpoC* L374L	92	8	18.507804
rpoB S441A	*rpoB* E465E	9	6	18.406468
rpoB S441A	*rpoC* T392A	0	5	18.335729
rpoB S441A	*rpoB* E773E	0	5	18.335729
rpoB S450L	*rpoA* R182L	0	27	18.293947
rpoB S441A	*rpoB* R415R	101	8	18.198272
rpoB H445R	*rpoC* L566V	0	7	17.873658
rpoB S441A	*rpoC* L381L	44	7	17.855077
rpoB S441A	*rpoC* R412R	44	7	17.855077
rpoB S450L	*rpoC* L449R	2	30	17.829461
rpoB S441A	*rpoB* A462A	13	6	17.672620
rpoB S441A	*rpoB* V774S	1	5	17.557639
rpoB S441A	*rpoC* F70Y	1	5	17.557639
rpoB S441A	*rpoB* R781R	123	8	17.541003
rpoB D435Y	*rpoB* R167H	1	9	17.534226
rpoB S441A	*sigA* Y341F	14	6	17.517776
rpoB L430P	*rpoB* F424L	3	9	17.477836
rpoB S441A	*rpoB* D779D	128	8	17.407555
rpoB S441A	*rpoB* R461R	131	8	17.329889
rpoB S441A	*rpoB* I770V	132	8	17.304382
rpoB S441A	*rpoC* A448A	54	7	17.279465
rpoB S441A	*rpoB* T407T	135	8	17.228972
rpoB Q432P	*rpoB* R827H	30	8	17.184375
rpoB I491F	*rpoB* S493T	0	7	17.158490
rpoB S441A	*rpoC* S428S	58	7	17.076704
rpoB S441A	*rpoC* D359E	2	5	17.013632

rpoB S441A	*sigA* L379L	60	7	16.980195
rpoB S450L	*rpoC* G332C	1	27	16.945026
rpoB S450L	*rpoC* E518D	0	25	16.938209
rpoB S450L	*rpoB* R552H	0	25	16.938209
rpoB S441A	*rpoC* L436L	19	6	16.858229
rpoB S441A	*rpoC* E364E	158	8	16.699716
rpoB S441A	*sigA* M403L	3	5	16.587725
rpoB S441A	*rpoC* V135M	3	5	16.587725
rpoB S441A	*sigA* R313R	164	8	16.574006
rpoB S441A	*rpoB* G475G	22	6	16.530578
rpoB S441A	*sigA* L318L	71	7	16.498203
rpoB S450L	*rpoB* V496M	1	26	16.282794
rpoB S450L	*rpoB* E550G	0	24	16.260378
rpoB S450L	*rpoC* P1040Q	0	24	16.260378
rpoB L430P	*rpoB* a-83g	33	12	16.059468
rpoB D545E	*rpoB* E821E	0	5	15.992566
rpoB D545E	*rpoB* V514I	0	5	15.992566
rpoB D545E	*rpoC* D747D	0	5	15.992566
rpoB D545E	*rpoB* V517L	0	5	15.992566
rpoB D545E	*rpoB* T806T	0	5	15.992566
rpoB D545E	*rpoC* Q687Q	0	5	15.992566
rpoB D545E	*rpoC* E665D	0	5	15.992566
rpoB D545E	*rpoB* E738E	0	5	15.992566
rpoB D545E	*rpoB* S14A	0	5	15.992566
rpoB D545E	*rpoC* P739P	0	5	15.992566
rpoB D545E	*rpoC* D796D	0	5	15.992566
rpoB D545E	*rpoC* E90E	0	5	15.992566
rpoB D545E	*rpoC* A719A	0	5	15.992566
rpoB D545E	*rpoA* 679$_i$*ndel*	0	5	15.992566
rpoB D545E	*rpoC* 2513$_i$*ndel*	0	5	15.992566
rpoB 1296$_i$*ndel*	*sigA* 247$_i$*ndel*	68	7	15.979535
rpoB S441A	*rpoC* I358L	5	5	15.934636
rpoB S450L	*rpoC* W484S	5	31	15.922741
rpoB D435Y	*rpoB* I491L	23	12	15.879803
rpoB S441A	*rpoB* R816R	88	7	15.877654
rpoB S441A	*rpoC* C441C	6	5	15.671456
rpoB S441A	*rpoB* G798G	6	5	15.671456
rpoB S450L	*rpoC* Q523E	1	25	15.621172
rpoB S450L	*rpoB* F971L	0	23	15.582572
rpoB L430P	*rpoB* A1075A	37084	316	15.532371
rpoB S450F	*sigA* t-97g	0	6	15.341942

rpoB S441A	*rpoC* S403S	106	7	15.335156
rpoB S441A	*sigA* L387L	37	6	15.322406
rpoB S450L	*rpoB* V534M	2	26	15.234722
rpoB D545E	*rpoB* L1156L	1	5	15.214628
rpoB D545E	*rpoB* A286A	1	5	15.214628
rpoB D545E	*rpoC* D660E	1	5	15.214628
rpoB D545E	*rpoB* G319N	1	5	15.214628
rpoB D545E	*rpoC* H1104H	1	5	15.214628
rpoB D545E	*sigA* Q353Q	1	5	15.214628
rpoB D545E	*sigA* L273L	1	5	15.214628
rpoB D545E	*rpoB* E320H	1	5	15.214628
rpoB D545E	*rpoC* T1050T	1	5	15.214628
rpoB D545E	*rpoC* H653T	1	5	15.214628
rpoB D545E	*rpoB* R273R	1	5	15.214628
rpoB S441A	*rpoB* A1099A	9	5	15.034818
rpoB S441A	*rpoC* R372R	42	6	15.018873
rpoB S450L	*rpoC* P434T	0	22	14.904791
rpoB S450L	*rpoC* V1039G	0	22	14.904791
rpoB S441A	*rpoB* G787G	124	7	14.875256
rpoB S441A	*rpoC* Y61A	10	5	14.858789
rpoB S441A	*rpoC* R421R	125	7	14.851645
rpoB S450W	*rpoC* D1218A	2	7	14.774517
rpoB S441A	*rpoB* A753A	130	7	14.736278
rpoB D545E	*rpoB* R213R	2	5	14.670772
rpoB D545E	*rpoB* N1155N	2	5	14.670772
rpoB D545E	*rpoC* Y793Y	2	5	14.670772
rpoB D545E	*rpoB* E396E	2	5	14.670772
rpoB D545E	*rpoB* E132E	2	5	14.670772
rpoB D545E	*rpoC* E769E	2	5	14.670772
rpoB D545E	*rpoB* !1173!	2	5	14.670772
rpoB D545E	*rpoB* S21S	2	5	14.670772
rpoB D545E	*rpoC* a-35g	2	5	14.670772
rpoB D545E	*sigA* K251K	2	5	14.670772
rpoB D545E	*rpoB* P15Q	2	5	14.670772
rpoB D545E	*rpoB* S201S	2	5	14.670772
rpoB D545E	*rpoC* c-36g	2	5	14.670772
rpoB D545E	*rpoB* V1129V	2	5	14.670772
rpoB D545E	*rpoB* A10D	2	5	14.670772
rpoB D545E	*rpoB* N24S	2	5	14.670772
rpoB D545E	*rpoB* S22S	2	5	14.670772
rpoB D545E	*rpoB* D150D	2	5	14.670772

rpoB D545E	*rpoC* g-40t	2	5	14.670772
rpoB D545E	*rpoC* Q657H	2	5	14.670772
rpoB D545E	*rpoB* P13P	2	5	14.670772
rpoB S441A	*rpoB* P471P	49	6	14.645930
rpoB S441A	*rpoC* L117V	0	4	14.603885
rpoB S441A	*rpoC* E96D	0	4	14.603885
rpoB S441A	*rpoB* G485G	137	7	14.581815
rpoB S450L	*rpoC* H525N	3	26	14.347464
rpoB S441A	*rpoC* G53G	150	7	14.314343
rpoB L452P	*rpoB* E481A	8	9	14.312102
rpoB S450L	*rpoC* P434Q	1	23	14.299943
rpoB S450L	*rpoC* R770H	1	23	14.299943
rpoB S441A	*rpoC* P363P	152	7	14.275215
rpoB D545E	*rpoZ* I55I	3	5	14.245016
rpoB D545E	*rpoC* L679P	3	5	14.245016
rpoB D545E	*rpoC* E477E	3	5	14.245016
rpoB D545E	*rpoC* P643S	3	5	14.245016
rpoB D545E	*rpoB* E257E	3	5	14.245016
rpoB D545E	*sigA* Q277Q	3	5	14.245016
rpoB D545E	*rpoC* K64K	3	5	14.245016
rpoB D545E	*rpoB* A1153A	3	5	14.245016
rpoB D545E	*rpoC* S1095S	3	5	14.245016
rpoB D545E	*rpoB* N31N	3	5	14.245016
rpoB D545E	*rpoC* A661P	3	5	14.245016
rpoB D545E	*rpoB* N23N	3	5	14.245016
rpoB D545E	*rpoB* L525L	3	5	14.245016
rpoB D545E	*rpoB* A420A	3	5	14.245016
rpoB D545E	*rpoC* A466A	3	5	14.245016
rpoB D545E	*sigA* -50$_i$*ndel*	3	5	14.245016
rpoB D545E	*rpoC* t-39c	3	5	14.245016
rpoB D545E	*rpoB* V129V	3	5	14.245016
rpoB S450L	*rpoC* K1152N	0	21	14.227036
rpoB S441A	*rpoC* K355K	158	7	14.160781
rpoB S441A	*rpoB* L758L	15	5	14.146208
rpoB S441A	*sigA* I351I	164	7	14.050523
rpoB S450W	*rpoB* Q19R	0	6	13.990529
rpoB D435V	*rpoB* H1028Y	1	9	13.920825
rpoB S450N	*rpoB* L430L	19	4	13.916768
rpoB S441A	*sigA* P396P	17	5	13.916258
rpoB S441A	*rpoC* R356R	66	6	13.915363
rpoB S441A	*rpoB* P810S	1	4	13.904979

rpoB S441A	*rpoB* G1020G	1	4	13.904979
rpoB S441A	*rpoB* N1105D	1	4	13.904979
rpoB M434I	*rpoB* c-61t	8726	26	13.898173
rpoB D545E	*rpoB* T526T	4	5	13.893046
rpoB D545E	*sigA* I284I	4	5	13.893046
rpoB D545E	*rpoC* E183Q	4	5	13.893046
rpoB D545E	*rpoB* Q241R	4	5	13.893046
rpoB D545E	*rpoB* R1151R	4	5	13.893046
rpoB D545E	*rpoC* R77R	4	5	13.893046
rpoB D545E	*rpoB* D688E	4	5	13.893046
rpoB D545E	*rpoC* Y1142Y	4	5	13.893046
rpoB D545E	*rpoC* P740P	4	5	13.893046
rpoB D545E	*rpoC* E250E	4	5	13.893046
rpoB D545E	*rpoB* P566A	4	5	13.893046
rpoB D545E	*rpoC* G669G	4	5	13.893046
rpoB D545E	*rpoC* E488E	4	5	13.893046
rpoB S441A	*sigA* G327G	177	7	13.824597
rpoB S441A	*rpoB* L440L	19	5	13.708468
rpoB S450L	*rpoA* V183A	1	22	13.640446
rpoB S441A	*sigA* A331A	20	5	13.611619
rpoB D545E	*rpoC* S1115S	5	5	13.592229
rpoB D545E	*rpoC* L438L	5	5	13.592229
rpoB D545E	*rpoB* D53D	5	5	13.592229
rpoB D545E	*rpoZ* 212$_i$*ndel*	5	5	13.592229
rpoB D545E	*rpoC* L2L	5	5	13.592229
rpoB D545E	*rpoB* A1152A	5	5	13.592229
rpoB D545E	*rpoB* D520D	5	5	13.592229
rpoB D545E	*rpoC* E171E	5	5	13.592229
rpoB D545E	*rpoC* L774L	5	5	13.592229
rpoB Q432E	*rpoC* P481T	167	6	13.554601
rpoB S450L	*rpoC* A734V	0	20	13.549305
rpoB S450L	*rpoC* G433A	0	20	13.549305
rpoB S450L	*rpoC* P1040T	0	20	13.549305
rpoB S441A	*rpoB* V613I	2	4	13.427922
rpoB S441A	*sigA* A316Q	2	4	13.427922
rpoB S441A	*rpoB* V418V	81	6	13.406741
rpoB S441A	*rpoC* R353R	83	6	13.345875
rpoB M434I	*rpoB* G876G	9210	26	13.332744
rpoB D545E	*rpoB* S324T	6	5	13.329200
rpoB D545E	*rpoC* A807A	6	5	13.329200
rpoB D545E	*rpoB* V26V	6	5	13.329200

rpoB D545E	*rpoB* D634D	6	5	13.329200
rpoB D545E	*rpoB* R824R	6	5	13.329200
rpoB D545E	*rpoC* L676L	6	5	13.329200
rpoB D545E	*rpoB* V740I	6	5	13.329200
rpoB S450L	*rpoB* T399I	2	23	13.300570
rpoB S441A	*rpoB* D752D	86	6	13.257173
rpoB S450W	*rpoB* P45L	43	10	13.133690
rpoB S441A	*rpoC* A362A	26	5	13.107107
rpoB D545E	*rpoC* Q1084Q	7	5	13.095329
rpoB D545E	*rpoC* Y134Y	7	5	13.095329
rpoB D545E	*rpoB* P30P	7	5	13.095329
rpoB D545E	*sigA* L262L	7	5	13.095329
rpoB D545E	*rpoB* G354G	7	5	13.095329
rpoB D545E	*rpoB* A559A	7	5	13.095329
rpoB D545E	*rpoC* E751D	7	5	13.095329
rpoB D545E	*rpoC* E1056E	7	5	13.095329
rpoB D545E	*rpoB* L42L	7	5	13.095329
rpoB D545E	*rpoC* G1126G	7	5	13.095329
rpoB D545E	*rpoB* S325T	7	5	13.095329
rpoB D545E	*rpoC* V1141V	7	5	13.095329
rpoB D545E	*rpoB* L372L	7	5	13.095329
rpoB D545E	*rpoC* Q435Q	7	5	13.095329
rpoB S441A	*rpoB* S576S	3	4	13.060010
rpoB S441A	*rpoC* G93G	3	4	13.060010
rpoB D545E	*rpoB* D755D	8	5	12.884688
rpoB D545E	*rpoB* A599A	8	5	12.884688
rpoB D545E	*rpoC* A773A	8	5	12.884688
rpoB D545E	*rpoB* S771S	8	5	12.884688
rpoB D545E	*rpoC* V1105V	8	5	12.884688
rpoB D545E	*rpoC* V785V	8	5	12.884688
rpoB D545E	*rpoB* E266E	8	5	12.884688
rpoB D545E	*rpoB* T347A	8	5	12.884688
rpoB D545E	*rpoB* L47L	8	5	12.884688
rpoB D545E	*rpoB* V33V	8	5	12.884688
rpoB S450L	*rpoB* L42V	0	19	12.871601
rpoB S450L	*rpoC* E750D	0	19	12.871601
rpoB S450L	*rpoC* Q1110H	0	19	12.871601
rpoB D435V	*rpoC* P481T	148	25	12.867860
rpoB S441A	*sigA* T348T	101	6	12.854150
rpoB S450L	*rpoC* F831L	6	27	12.838477
rpoB S441A	*rpoC* T33T	103	6	12.804853

rpoB D545E	*sigA* a-99t	0	4	12.772606
rpoB H445D	*rpoC* R741S	51	13	12.761295
rpoB S441A	*rpoB* M587L	4	4	12.759044
rpoB S441A	*rpoC* D462E	4	4	12.759044
rpoB S441A	*rpoB* L713L	4	4	12.759044
rpoB H445R	*rpoC* S727G	0	5	12.754901
rpoB D545E	*sigA* R278R	9	5	12.693015
rpoB D545E	*rpoB* P41P	9	5	12.693015
rpoB H445D	*sigA* G312G	28	11	12.601454
rpoB S441A	*rpoC* G411G	35	5	12.519651
rpoB D545E	*rpoB* V690V	10	5	12.517137
rpoB D545E	*rpoB* A36A	10	5	12.517137
rpoB D545E	*rpoC* L406L	10	5	12.517137
rpoB D545E	*rpoC* V1067V	10	5	12.517137
rpoB D545E	*rpoB* L38L	10	5	12.517137
rpoB D545E	*rpoB* E465E	10	5	12.517137
rpoB D545E	*rpoC* L1111L	10	5	12.517137
rpoB L452P	*rpoB* V168M	0	6	12.440683
rpoB S441A	*rpoC* G115G	120	6	12.419842
rpoB S450L	*rpoB* H835R	10	30	12.384351
rpoB D545E	*rpoB* S61S	11	5	12.354622
rpoB D545E	*rpoB* L289L	11	5	12.354622
rpoB D545E	*rpoC* V692V	11	5	12.354622
rpoB D545E	*rpoB* S672S	11	5	12.354622
rpoB H445N	*rpoB* G453A	0	5	12.323472
rpoB S441A	*rpoB* R397R	39	5	12.302295
rpoB S441A	*rpoB* I1035I	6	4	12.282052
rpoB D545E	*rpoC* V709V	12	5	12.203567
rpoB D545E	*sigA* A276A	12	5	12.203567
rpoB D545E	*rpoC* A1202A	12	5	12.203567
rpoB S450L	*rpoC* E1113G	0	18	12.193921
rpoB S450L	*rpoC* G433C	0	18	12.193921
rpoB S441A	*rpoB* I1046I	7	4	12.085822
rpoB S441A	*rpoB* P408P	7	4	12.085822
rpoB D545E	*sigA* t-91c	1	4	12.073845
rpoB D545E	*rpoB* A234A	13	5	12.062450
rpoB D545E	*rpoB* V179V	13	5	12.062450
rpoB D545E	*rpoC* A193A	13	5	12.062450
rpoB D545E	*rpoB* G46G	13	5	12.062450
rpoB D545E	*rpoB* V318T	13	5	12.062450
rpoB S441A	*rpoB* P486P	44	5	12.058112

rpoB S441A	*rpoC* L399L	139	6	12.047987
rpoB H445S	*rpoB* T444T	8	5	12.022144
rpoB S441A	*rpoB* R598R	45	5	12.012416
rpoB S441A	*rpoB* R1061R	141	6	12.011779
rpoB H445R	*rpoC* D869E	1	5	11.977752
rpoB D545E	*rpoC* D795S	14	5	11.930037
rpoB D545E	*rpoC* T666T	14	5	11.930037
rpoB S441A	*rpoC* V461V	146	6	11.923415
rpoB H445L	*rpoC* V1138G	0	5	11.827199
rpoB H445L	*rpoB* E156G	0	5	11.827199
rpoB D545E	*rpoC* L1055L	15	5	11.805311
rpoB D545E	*rpoC* G786G	15	5	11.805311
rpoB D545E	*rpoB* L626L	15	5	11.805311
rpoB D545E	*sigA* T264A	15	5	11.805311
rpoB S441A	*rpoC* L354L	153	6	11.804564
rpoB S441A	*sigA* S344S	51	5	11.756824
rpoB S441A	*rpoC* A380A	9	4	11.750159
rpoB S450L	*rpoC* Q1125H	4	23	11.728344
rpoB D545E	*rpoC* R345R	16	5	11.687424
rpoB S441A	*rpoC* R389R	161	6	11.675101
rpoB S450L	*rpoB* H723D	1	19	11.667228
rpoB S441A	*rpoC* R67R	165	6	11.612721
rpoB S441A	*rpoB* V715V	10	4	11.604095
rpoB H445Y	*rpoC* M515T	0	6	11.603718
rpoB D545E	*rpoB* P62Q	2	4	11.596933
rpoB D545E	*rpoC* P577P	2	4	11.596933
rpoB D545E	*rpoC* I557V	2	4	11.596933
rpoB D545E	*rpoB* S4F	2	4	11.596933
rpoB D545E	*rpoB* $2663_i ndel$	2	4	11.596933
rpoB D435V	*rpoB* N292D	0	7	11.596258
rpoB D545E	*rpoC* G385G	17	5	11.575663
rpoB D545E	*rpoB* T122T	17	5	11.575663
rpoB D435Y	*rpoB* Q172R	5	7	11.525361
rpoB S450L	*rpoC* Q479R	0	17	11.516266
rpoB S441A	*rpoB* N733Q	11	4	11.469461
rpoB D545E	*rpoB* L430L	18	5	11.469420
rpoB D545E	*rpoC* A1184A	18	5	11.469420
rpoB S441A	*rpoC* P41P	60	5	11.422447
rpoB D545E	*rpoC* R166R	19	5	11.368175
rpoB D545E	*rpoC* V162V	19	5	11.368175
rpoB D545E	*sigA* S305S	19	5	11.368175

rpoB D545E	*rpoC* P326P	19	5	11.368175
rpoB D545E	*rpoB* G489G	19	5	11.368175
rpoB L452P	*rpoB* D265G	6	7	11.301538
rpoB S441A	*rpoC* R506R	64	5	11.288958
rpoB S450L	*rpoB* R827H	10	28	11.279459
rpoB D545E	*sigA* I368I	20	5	11.271478
rpoB D545E	*rpoB* V555V	20	5	11.271478
rpoB D545E	*rpoC* L1180L	20	5	11.271478
rpoB D545E	*rpoC* V582V	3	4	11.229166
rpoB D545E	*rpoB* Q6Q	3	4	11.229166
rpoB D545E	*rpoB* A2A	3	4	11.229166
rpoB D545E	*rpoC* A553A	3	4	11.229166
rpoB S450L	*rpoZ* P85S	9	27	11.203015
rpoB D545E	*rpoC* a-32g	21	5	11.178936
rpoB D545E	*rpoB* R307R	21	5	11.178936
rpoB D545E	*rpoB* A728A	21	5	11.178936
rpoB D545E	*rpoB* V466V	21	5	11.178936
rpoB D545E	*rpoC* G652G	22	5	11.090208
rpoB D545E	*rpoC* G433G	22	5	11.090208
rpoB D545E	*rpoB* D703Q	22	5	11.090208
rpoB S428G	*sigA* G436S	3	3	11.052970
rpoB S441A	*rpoC* R387R	72	5	11.044421
rpoB S450L	*rpoB* I588V	1	18	11.011531
rpoB D545E	*rpoB* P810P	23	5	11.004990
rpoB D545E	*rpoC* A172A	23	5	11.004990
rpoB D545E	*sigA* R381R	23	5	11.004990
rpoB D545E	*rpoC* V1039V	24	5	10.923016
rpoB D545E	*rpoB* A498A	24	5	10.923016
rpoB S441A	*rpoB* E484E	16	4	10.919622
rpoB S441A	*rpoC* D119N	0	3	10.906796
rpoB S441A	*rpoB* V1091T	0	3	10.906796
rpoB S441A	*rpoC* N352T	0	3	10.906796
rpoB S441A	*rpoB* I717F	0	3	10.906796
rpoB S441A	*rpoB* S450S	0	3	10.906796
rpoB D545E	*rpoB* A629A	25	5	10.844047
rpoB D545E	*rpoC* H748R	25	5	10.844047
rpoB D545E	*rpoC* V775V	25	5	10.844047
rpoB D545E	*rpoC* S138A	25	5	10.844047
rpoB S450L	*rpoB* Y564H	0	16	10.838637
rpoB S450L	*rpoB* V534A	0	16	10.838637
rpoB S441A	*sigA* T365T	17	4	10.827916

rpoB D545E	*rpoB* L87L	26	5	10.767871
rpoB D545E	*rpoC* P827P	26	5	10.767871
rpoB D545E	*sigA* L411L	26	5	10.767871
rpoB D545E	*rpoC* V245Q	26	5	10.767871
rpoB S450L	*rpoC* I885V	2	19	10.742630
rpoB S450L	*rpoC* E750G	2	19	10.742630
rpoB S450L	*rpoC* K715T	2	19	10.742630
rpoB D545E	*rpoC* D462D	27	5	10.694298
rpoB S450N	*rpoB* G453G	137	4	10.665926
rpoB V170F	*rpoC* A861T	0	4	10.665926
rpoB D545E	*rpoB* G203G	28	5	10.623154
rpoB S450N	*rpoC* L399L	141	4	10.616813
rpoB Q432P	*rpoC* V483G	1230	13	10.614340
rpoB S450N	*rpoC* P390P	142	4	10.604748
rpoB S450N	*rpoB* S458S	143	4	10.592767
rpoB D545E	*rpoB* T585T	29	5	10.554286
rpoB S450N	*rpoB* L443L	148	4	10.534069
rpoB S441A	*sigA* E323E	92	5	10.532195
rpoB V170F	*rpoC* W484G	83	8	10.518823
rpoB S441A	*rpoC* D485D	93	5	10.509513
rpoB S441A	*rpoB* P1109P	21	4	10.503147
rpoB S450N	*rpoB* Q432Q	152	4	10.488499
rpoB D545E	*sigA* T395T	30	5	10.487552
rpoB H445L	*rpoC* Y61Y	169	12	10.483109
rpoB D545E	*rpoC* L566L	6	4	10.451643
rpoB D545E	*rpoC* P573P	6	4	10.451643
rpoB S450L	*sigA* 247$_i$*ndel*	33	42	10.434208
rpoB S441A	*rpoB* V466V	22	4	10.430661
rpoB D545E	*rpoB* V50V	31	5	10.422823
rpoB D545E	*rpoB* P112P	31	5	10.422823
rpoB D545E	*rpoC* R350R	32	5	10.359984
rpoB D545E	*rpoC* G1178G	32	5	10.359984
rpoB D545E	*rpoB* R627R	32	5	10.359984
rpoB D545E	*rpoB* G858G	32	5	10.359984
rpoB S450L	*rpoB* V970A	1	17	10.357031
rpoB S450L	*rpoB* A405P	44	48	10.340885
rpoB S441A	*rpoB* A760E	1	3	10.304802
rpoB S441A	*rpoB* L626R	1	3	10.304802
rpoB S441A	*rpoB* H723L	1	3	10.304802
rpoB S441A	*rpoB* M1022Q	1	3	10.304802
rpoB S441A	*rpoB* E761E	1	3	10.304802

rpoB S441A	*rpoB* P479P	1	3	10.304802
rpoB S441A	*rpoB* L1079L	1	3	10.304802
rpoB D545E	*rpoC* V738V	33	5	10.298927
rpoB S441A	*rpoC* R113R	105	5	10.254440
rpoB D545E	*sigA* G312G	34	5	10.239554
rpoB D545E	*rpoB* R671R	34	5	10.239554
rpoB D545E	*rpoC* G79G	34	5	10.239554
rpoB D545E	*rpoB* V1031V	34	5	10.239554
rpoB D545E	*rpoC* S377S	34	5	10.239554
rpoB D545E	*sigA* K292K	34	5	10.239554
rpoB D545E	*rpoC* V731V	34	5	10.239554
rpoB D545E	*rpoB* A544A	34	5	10.239554
rpoB D545E	*rpoB* A267A	34	5	10.239554
rpoB S441A	*rpoC* L507L	107	5	10.214712
rpoB H445R	*rpoB* A538V	0	4	10.199149
rpoB D545E	*rpoB* L735L	35	5	10.181775
rpoB S450L	*rpoB* D574E	0	15	10.161033
rpoB S450L	*rpoB* I873F	0	15	10.161033
rpoB D545E	*rpoC* R726R	36	5	10.125506
rpoB D545E	*rpoC* V429V	36	5	10.125506
rpoB D545E	*rpoB* H317N	36	5	10.125506
rpoB D545E	*rpoZ* A52A	36	5	10.125506
rpoB D545E	*rpoB* A586A	36	5	10.125506
rpoB S441A	*rpoC* G419G	114	5	10.081156
rpoB D545E	*rpoC* L585L	8	4	10.079676
rpoB D545E	*rpoB* L314L	37	5	10.070671
rpoB D545E	*rpoB* P856P	37	5	10.070671
rpoB S450N	*sigA* A360A	16	3	10.066681
rpoB S441A	*rpoB* L1083L	28	4	10.049869
rpoB S441A	*sigA* K401K	117	5	10.026356
rpoB D545E	*sigA* L387L	38	5	10.017198
rpoB S441A	*rpoB* D1088D	118	5	10.008395
rpoB S441A	*rpoB* G1103G	119	5	9.990581
rpoB D545E	*rpoB* G345G	40	5	9.914083
rpoB D545E	*rpoC* V110V	40	5	9.914083
rpoB S441A	*rpoB* E616D	2	3	9.906927
rpoB S441A	*rpoB* S615P	2	3	9.906927
rpoB S441A	*rpoB* T1078A	2	3	9.906927
rpoB S441A	*rpoB* V1017I	2	3	9.906927
rpoB S441A	*rpoB* E811E	2	3	9.906927
rpoB S441A	*rpoB* G602G	124	5	9.903659

rpoB S441A	*rpoB* R395R	127	5	9.853139
rpoB D545E	*rpoC* R325R	42	5	9.815685
rpoB D545E	*rpoC* L148L	42	5	9.815685
rpoB D545E	*rpoB* G687G	42	5	9.815685
rpoB D545E	*rpoB* G820G	42	5	9.815685
rpoB D545E	*rpoZ* L53L	42	5	9.815685
rpoB D545E	*rpoC* t-30a	42	5	9.815685
rpoB S441A	*rpoB* A603A	130	5	9.803777
rpoB D545E	*rpoC* Q22Q	10	4	9.774266
rpoB D545E	*rpoB* A776A	43	5	9.768125
rpoB D545E	*rpoC* g-43c	43	5	9.768125
rpoB H445Y	*rpoB* E460G	8	7	9.766466
rpoB S441A	*rpoC* K29K	133	5	9.755520
rpoB D545E	*rpoC* T1211T	44	5	9.721594
rpoB D545E	*rpoB* S88S	44	5	9.721594
rpoB D545E	*rpoC* A316A	44	5	9.721594
rpoB S450L	*rpoB* R552C	1	16	9.703860
rpoB S441A	*rpoC* G79G	35	4	9.690999
rpoB D545E	*rpoB* R598R	45	5	9.676049
rpoB D545E	*rpoC* L640L	45	5	9.676049
rpoB H445Y	*rpoB* E563A	0	5	9.668340
rpoB Q432P	*rpoB* H835R	35	5	9.667341
rpoB S441A	*rpoC* L39L	141	5	9.631882
rpoB D545E	*sigA* G314G	46	5	9.631450
rpoB S450W	*rpoB* Q409R	74	9	9.620475
rpoB S450L	*rpoB* V695L	118	81	9.615081
rpoB S441A	*sigA* G391G	3	3	9.605963
rpoB S441A	*rpoC* S138E	3	3	9.605963
rpoB S441A	*rpoB* Y1015D	3	3	9.605963
rpoB S441A	*sigA* I405M	3	3	9.605963
rpoB S441A	*rpoB* A586S	3	3	9.605963
rpoB S441A	*rpoC* V429V	37	4	9.600788
rpoB D545E	*rpoB* V865V	47	5	9.587757
rpoB D545E	*rpoC* G257G	47	5	9.587757
rpoB D545E	*rpoB* L235L	47	5	9.587757
rpoB D545E	*rpoB* L449L	47	5	9.587757
rpoB D545E	*rpoC* T150T	47	5	9.587757
rpoB S441A	*rpoC* H465H	146	5	9.558064
rpoB S441A	*rpoB* P1107P	38	4	9.557387
rpoB D545E	*rpoB* G504G	48	5	9.544935
rpoB D545E	*rpoB* V863V	48	5	9.544935

rpoB D545E	*rpoB* S641S	48	5	9.544935
rpoB D545E	*rpoC* A551A	12	4	9.515047
rpoB S441A	*rpoB* R1038R	39	4	9.515047
rpoB S441A	*sigA* L320L	149	5	9.514959
rpoB D435Y	*rpoB* H674D	1	5	9.512298
rpoB D545E	*rpoC* T83T	49	5	9.502950
rpoB D545E	*rpoC* A695S	49	5	9.502950
rpoB D545E	*rpoB* R389R	49	5	9.502950
rpoB D545E	*rpoC* G806G	49	5	9.502950
rpoB S450L	*rpoC* G519R	0	14	9.483454
rpoB S450L	*rpoC* E488Q	0	14	9.483454
rpoB D545E	*rpoC* M663I	50	5	9.461769
rpoB S450N	*rpoC* S401S	27	3	9.444534
rpoB S441A	*rpoC* V110V	41	4	9.433355
rpoB D545E	*rpoC* A694A	51	5	9.421364
rpoB D545E	*rpoZ* A58A	51	5	9.421364
rpoB D545E	*rpoC* G1042G	51	5	9.421364
rpoB D545E	*sigA* R384R	51	5	9.421364
rpoB D545E	*sigA* E271E	51	5	9.421364
rpoB D545E	*rpoC* N283S	51	5	9.421364
rpoB D545E	*rpoB* Y85D	13	4	9.398751
rpoB D545E	*rpoC* L487L	52	5	9.381705
rpoB D545E	*rpoB* A817A	52	5	9.381705
rpoB S441A	*rpoC* E59E	161	5	9.350683
rpoB D545E	*rpoB* T400T	53	5	9.342766
rpoB D545E	*rpoC* A753A	53	5	9.342766
rpoB D545E	*rpoB* R476R	53	5	9.342766
rpoB D545E	*rpoC* G388G	53	5	9.342766
rpoB K446T	*rpoC* 2213$_i$*ndel*	0	2	9.321067
rpoB 1278$_i$*ndel*	*rpoC* V68A	0	2	9.321067
rpoB L430P	*rpoB* V170A	10	6	9.318905
rpoB S441A	*sigA* A358A	44	4	9.317622
rpoB S441A	*rpoC* P502P	44	4	9.317622
rpoB S441A	*rpoB* A776A	44	4	9.317622
rpoB D545E	*rpoC* A99A	54	5	9.304520
rpoB D545E	*rpoB* R459R	55	5	9.266944
rpoB D545E	*sigA* S267A	55	5	9.266944
rpoB D545E	*rpoC* G1058G	55	5	9.266944
rpoB Q432K	*rpoB* T482I	0	3	9.260463
rpoB S441A	*rpoC* G408G	46	4	9.244555
rpoB S450N	*sigA* T395T	32	3	9.237176

rpoB D545E	*sigA* R418R	56	5	9.230015
rpoB D545E	*rpoC* A724A	57	5	9.193710
rpoB D545E	*rpoB* A204A	57	5	9.193710
rpoB D545E	*sigA* L399L	57	5	9.193710
rpoB D545E	*rpoB* L855L	57	5	9.193710
rpoB D545E	*rpoB* L99L	57	5	9.193710
rpoB D545E	*rpoC* A1144A	58	5	9.158010
rpoB D545E	*rpoB* R864R	58	5	9.158010
rpoB D545E	*rpoB* Y866Y	58	5	9.158010
rpoB D545E	*rpoB* S102S	58	5	9.158010
rpoB D545E	*rpoC* E1092E	58	5	9.158010
rpoB S441A	*rpoB* G1041G	179	5	9.125639
rpoB D545E	*rpoC* P122P	59	5	9.122895
rpoB D545E	*rpoB* I844I	59	5	9.122895
rpoB D545E	*rpoB* E852D	59	5	9.122895
rpoB D545E	*rpoC* A301A	59	5	9.122895
rpoB S441A	*rpoC* T83T	50	4	9.107075
rpoB D545E	*rpoB* D1170D	60	5	9.088345
rpoB D545E	*rpoB* A1166A	60	5	9.088345
rpoB D545E	*rpoC* S428S	60	5	9.088345
rpoB D545E	*rpoC* R295R	60	5	9.088345
rpoB D545E	*rpoC* S305S	60	5	9.088345
rpoB D545E	*rpoB* F847F	60	5	9.088345
rpoB D545E	*rpoB* E837E	60	5	9.088345
rpoB D545E	*rpoB* L1171L	61	5	9.054344
rpoB D545E	*rpoB* A1172A	61	5	9.054344
rpoB D545E	*rpoC* g-37t	61	5	9.054344
rpoB D545E	*rpoB* K870K	61	5	9.054344
rpoB D545E	*rpoB* R467R	61	5	9.054344
rpoB D545E	*rpoB* S1167S	61	5	9.054344
rpoB S450L	*rpoC* P434A	1	15	9.052173
rpoB S450L	*rpoA* T181A	1	15	9.052173
rpoB H445N	*rpoB* H674R	9	5	9.033043
rpoB D545E	*sigA* L379L	62	5	9.020873
rpoB S450F	*rpoB* A1075A	37256	144	9.004668
rpoB D545E	*rpoC* L46L	17	4	8.999102
rpoB D545E	*rpoC* D279E	63	5	8.987917
rpoB D545E	*rpoC* R474R	63	5	8.987917
rpoB D545E	*rpoB* A868A	63	5	8.987917
rpoB D545E	*rpoC* 3797$_i$*ndel*	1	3	8.961775
rpoB D545E	*rpoB* R65R	1	3	8.961775

rpoB D545E	*rpoC* G604A	1	3	8.961775
rpoB I491F	*rpoB* L490L	9	5	8.956569
rpoB D545E	*rpoC* c-38t	64	5	8.955461
rpoB S441A	*rpoC* A99A	55	4	8.949297
rpoB D545E	*rpoC* G1094G	65	5	8.923489
rpoB D435Y	*rpoC* D279G	8	6	8.895459
rpoB D545E	*rpoC* R356R	67	5	8.860942
rpoB D545E	*rpoB* R202R	67	5	8.860942
rpoB S450L	*rpoB* L554P	2	16	8.845199
rpoB D545E	*rpoC* V1158V	68	5	8.830340
rpoB D545E	*rpoB* V826V	68	5	8.830340
rpoB S450N	*sigA* A358A	45	3	8.815205
rpoB S450L	*rpoC* D747G	0	13	8.805901
rpoB S450L	*rpoB* T400S	0	13	8.805901
rpoB S450L	*rpoC* A1213E	0	13	8.805901
rpoB S441A	*rpoC* A336A	60	4	8.804720
rpoB D545E	*sigA* E397E	69	5	8.800170
rpoB D545E	*rpoC* t-19c	69	5	8.800170
rpoB D545E	*rpoC* T808T	69	5	8.800170
rpoB D545E	*sigA* L266M	70	5	8.770419
rpoB D545E	*rpoC* G728G	70	5	8.770419
rpoB D545E	*rpoB* P454P	70	5	8.770419
rpoB D545E	*rpoC* t-18a	70	5	8.770419
rpoB D545E	*rpoC* G13G	20	4	8.750424
rpoB D545E	*rpoC* V517I	20	4	8.750424
rpoB S450L	*rpoC* E757G	3	17	8.747229
rpoB D545E	*rpoC* R1183R	71	5	8.741077
rpoB D545E	*rpoB* G1136G	71	5	8.741077
rpoB D545E	*rpoC* S224N	71	5	8.741077
rpoB D545E	*rpoB* N1159N	71	5	8.741077
rpoB D545E	*rpoB* L1160L	71	5	8.741077
rpoB D545E	*rpoC* R214R	71	5	8.741077
rpoB D545E	*rpoC* R206R	72	5	8.712131
rpoB D545E	*rpoB* S1134S	72	5	8.712131
rpoB H445N	*rpoB* V170A	11	5	8.696675
rpoB S441A	*rpoC* P495P	8	3	8.689836
rpoB S441A	*rpoB* C1067V	8	3	8.689836
rpoB D545E	*rpoB* A1154A	73	5	8.683572
rpoB D545E	*rpoZ* Y56Y	73	5	8.683572
rpoB D545E	*sigA* L318L	73	5	8.683572
rpoB D435Y	*rpoB* L378R	9	6	8.676857

rpoB D545E	*rpoB* R105R	74	5	8.655389
rpoB V170F	*rpoC* E1092D	3971	30	8.644423
rpoB D545E	*rpoB* D108E	75	5	8.627573
rpoB D545E	*rpoC* R478R	75	5	8.627573
rpoB D545E	*rpoB* L1132L	75	5	8.627573
rpoB D545E	*rpoB* G1144G	76	5	8.600114
rpoB D545E	*rpoC* V1075V	76	5	8.600114
rpoB D545E	*rpoB* E1140E	76	5	8.600114
rpoB D545E	*rpoC* G1072S	76	5	8.600114
rpoB D545E	*rpoB* G164G	77	5	8.573003
rpoB D545E	*rpoB* E66D	2	3	8.564037
rpoB D545E	*rpoC* S654N	78	5	8.546232
rpoB D545E	*rpoB* E639E	78	5	8.546232
rpoB S450N	*rpoB* Y572F	0	2	8.542916
rpoB S450N	*rpoC* H653V	0	2	8.542916
rpoB S450N	*rpoB* E320S	0	2	8.542916
rpoB S450N	*rpoB* V517R	0	2	8.542916
rpoB D545E	*rpoB* R304R	79	5	8.519793
rpoB D545E	*rpoC* A156M	79	5	8.519793
rpoB D545E	*rpoC* V378V	79	5	8.519793
rpoB D545E	*sigA* P274P	80	5	8.493676
rpoB D545E	*rpoB* V534V	80	5	8.493676
rpoB D545E	*rpoB* E737E	80	5	8.493676
rpoB D545E	*rpoC* I208L	80	5	8.493676
rpoB D545E	*sigA* L257L	80	5	8.493676
rpoB D545E	*sigA* E297E	81	5	8.467876
rpoB D545E	*rpoB* V523H	81	5	8.467876
rpoB D545E	*rpoC* H1119H	81	5	8.467876
rpoB D545E	*rpoC* R459R	81	5	8.467876
rpoB D435Y	*rpoB* V695L	185	14	8.456613
rpoB S441A	*rpoC* V328V	74	4	8.454405
rpoB D545E	*rpoC* R194R	82	5	8.442383
rpoB D545E	*rpoC* P568P	82	5	8.442383
rpoB S450N	*rpoC* P122P	61	3	8.433471
rpoB D545E	*rpoB* E1121E	83	5	8.417191
rpoB D545E	*rpoC* A1201A	83	5	8.417191
rpoB D545E	*rpoB* A686A	83	5	8.417191
rpoB D545E	*rpoB* G516G	83	5	8.417191
rpoB D545E	*rpoB* T323T	83	5	8.417191
rpoB S441A	*rpoC* R478R	76	4	8.409691
rpoB D545E	*rpoC* P593D	25	4	8.402159

rpoB S450L	*rpoA* T187P	1	14	8.402156
rpoB S450N	*rpoC* P100P	63	3	8.392756
rpoB D545E	*rpoC* E1074E	84	5	8.392292
rpoB D545E	*rpoC* R1060R	84	5	8.392292
rpoB S450N	*sigA* L379L	64	3	8.372866
rpoB D545E	*rpoC* M672L	85	5	8.367681
rpoB D545E	*rpoB* C681S	85	5	8.367681
rpoB H445Y	*rpoB* P280L	2	5	8.354560
rpoB S441A	*rpoB* L490L	11	3	8.346415
rpoB S441A	*rpoB* V740V	11	3	8.346415
rpoB D545E	*rpoB* A617A	86	5	8.343351
rpoB D545E	*rpoC* R815R	86	5	8.343351
rpoB D545E	*rpoC* R641R	86	5	8.343351
rpoB D545E	*rpoB* P200P	86	5	8.343351
rpoB D545E	*rpoC* A710A	87	5	8.319295
rpoB D545E	*rpoB* T829T	87	5	8.319295
rpoB D545E	*rpoB* V351V	87	5	8.319295
rpoB D545E	*rpoB* D752D	87	5	8.319295
rpoB D545E	*rpoC* R1038R	87	5	8.319295
rpoB S450L	*rpoA* -40$_i$*ndel*	89	64	8.312436
rpoB D545E	*rpoC* S805S	88	5	8.295507
rpoB D545E	*rpoB* G442G	88	5	8.295507
rpoB D545E	*rpoC* S1222S	88	5	8.295507
rpoB D545E	*rpoC* G109G	88	5	8.295507
rpoB D435V	*sigA* R290H	0	5	8.281553
rpoB D545E	*rpoB* V243T	89	5	8.271982
rpoB D545E	*rpoC* V644A	89	5	8.271982
rpoB H445Y	*rpoC* L873L	118	13	8.252319
rpoB S441A	*rpoC* P31P	12	3	8.249570
rpoB S441A	*rpoC* R56R	12	3	8.249570
rpoB D545E	*rpoC* G680G	90	5	8.248713
rpoB D545E	*rpoB* R816R	90	5	8.248713
rpoB D545E	*rpoB* V695V	90	5	8.248713
rpoB D545E	*rpoC* G822G	90	5	8.248713
rpoB D545E	*rpoC* G361G	90	5	8.248713
rpoB D545E	*rpoB* R547R	91	5	8.225696
rpoB D545E	*rpoC* R386R	92	5	8.202925
rpoB D545E	*rpoC* V729V	92	5	8.202925
rpoB D545E	*sigA* E323E	92	5	8.202925
rpoB D545E	*rpoB* P104P	92	5	8.202925
rpoB D545E	*rpoB* R662H	92	5	8.202925

rpoB D545E	*rpoC* L264L	92	5	8.202925
rpoB D545E	*rpoC* R1231R	93	5	8.180394
rpoB D545E	*rpoC* L650L	93	5	8.180394
rpoB D545E	*sigA* R330R	93	5	8.180394
rpoB D545E	*rpoC* A569A	93	5	8.180394
rpoB S441A	*rpoB* T829T	88	4	8.163305
rpoB S441A	*rpoB* L612L	13	3	8.159459
rpoB S441A	*rpoC* L120L	13	3	8.159459
rpoB S441A	*sigA* A316L	13	3	8.159459
rpoB D545E	*rpoC* T814T	94	5	8.158099
rpoB D545E	*rpoB* V643V	94	5	8.158099
rpoB D545E	*rpoC* V1053V	95	5	8.136035
rpoB D545E	*rpoB* R734R	95	5	8.136035
rpoB D545E	*rpoC* A300A	95	5	8.136035
rpoB D545E	*rpoB* G315G	95	5	8.136035
rpoB S450L	*rpoC* N416T	0	12	8.128372
rpoB S450L	*rpoB* P45T	0	12	8.128372
rpoB S450L	*rpoC* V1147A	0	12	8.128372
rpoB S450L	*rpoC* D735E	0	12	8.128372
rpoB S450L	*rpoC* W484L	0	12	8.128372
rpoB S450L	*rpoC* V1124G	0	12	8.128372
rpoB S428G	*rpoC* T853T	45	3	8.116688
rpoB D545E	*rpoC* L789L	96	5	8.114198
rpoB D545E	*rpoC* E1096E	97	5	8.092582
rpoB D545E	*rpoC* V1076V	97	5	8.092582
rpoB D545E	*rpoB* R371R	97	5	8.092582
rpoB D545E	*rpoC* G809G	97	5	8.092582
rpoB D545E	*sigA* A417A	97	5	8.092582
rpoB D545E	*rpoC* G819G	98	5	8.071183
rpoB D545E	*rpoB* E337E	98	5	8.071183
rpoB S450N	*rpoB* D571E	1	2	8.065804
rpoB S450N	*rpoC* S919S	1	2	8.065804
rpoB S441A	*rpoB* I588I	94	4	8.052161
rpoB D545E	*rpoB* G139G	99	5	8.049998
rpoB D545E	*rpoC* S570S	99	5	8.049998
rpoB D435V	*rpoB* N673S	3	6	8.039275
rpoB S441A	*sigA* E402E	95	4	8.034314
rpoB D545E	*rpoB* P321P	100	5	8.029021
rpoB S441A	*rpoC* L330L	15	3	7.996088
rpoB V170F	*rpoB* I488L	0	3	7.994774
rpoB D545E	*rpoC* R1041R	102	5	7.987678

rpoB D545E	*rpoB* V262A	102	5	7.987678
rpoB D545E	*rpoB* L342L	102	5	7.987678
rpoB D545E	*sigA* T348T	102	5	7.987678
rpoB S450N	*rpoC* P481P	88	3	7.968845
rpoB D545E	*rpoB* S568S	103	5	7.967304
rpoB D545E	*rpoC* D192D	103	5	7.967304
rpoB D545E	*rpoB* L775L	103	5	7.967304
rpoB D545E	*rpoB* P700P	103	5	7.967304
rpoB S450N	*rpoB* D752D	89	3	7.954452
rpoB D545E	*rpoB* R415R	104	5	7.947123
rpoB D545E	*rpoC* I1080L	105	5	7.927133
rpoB D545E	*rpoB* V613V	105	5	7.927133
rpoB D545E	*rpoC* I1132I	105	5	7.927133
rpoB D545E	*rpoC* R113R	105	5	7.927133
rpoB S441A	*sigA* A360A	16	3	7.921520
rpoB S450N	*rpoC* G361G	92	3	7.912205
rpoB H445C	*rpoB* E561E	562	9	7.911335
rpoB D545E	*rpoB* V333V	106	5	7.907328
rpoB D545E	*rpoB* L600L	106	5	7.907328
rpoB D545E	*rpoC* V1165V	107	5	7.887706
rpoB D545E	*rpoB* G306G	107	5	7.887706
rpoB D545E	*rpoC* P1068P	107	5	7.887706
rpoB D545E	*rpoC* V1135V	107	5	7.887706
rpoB S450L	*rpoC* P481T	104	69	7.876354
rpoB S450N	*rpoB* I588I	95	3	7.871285
rpoB D545E	*rpoC* R703R	108	5	7.868264
rpoB D545E	*rpoC* S403S	108	5	7.868264
rpoB D545E	*rpoB* G623G	109	5	7.848998
rpoB D545E	*rpoB* T374T	109	5	7.848998
rpoB D545E	*rpoB* G804G	109	5	7.848998
rpoB D545E	*rpoC* T1230T	110	5	7.829906
rpoB D545E	*rpoC* Q1145Q	110	5	7.829906
rpoB D545E	*rpoC* R144R	110	5	7.829906
rpoB D545E	*rpoB* S648S	110	5	7.829906
rpoB S441A	*rpoC* L527L	108	4	7.817696
rpoB D545E	*rpoB* P589P	111	5	7.810983
rpoB D545E	*rpoB* Y308Y	111	5	7.810983
rpoB D545E	*rpoC* D1101D	111	5	7.810983
rpoB D545E	*rpoC* A1188A	112	5	7.792228
rpoB D545E	*rpoB* T130T	112	5	7.792228
rpoB S450L	*rpoC* R741S	30	34	7.789671

rpoB S441A	*rpoC* P54P	1341	7	7.789182
rpoB S441A	*rpoC* L118L	110	4	7.786661
rpoB S441A	*rpoC* R857R	18	3	7.784123
rpoB S450W	*rpoB* E592D	3	4	7.782050
rpoB D545E	*rpoB* V610V	113	5	7.773637
rpoB D545E	*rpoB* P483P	113	5	7.773637
rpoB S450N	*rpoB* Q346H	2	2	7.764783
rpoB S450N	*rpoB* A617R	2	2	7.764783
rpoB S450N	*sigA* T348T	104	3	7.755720
rpoB D545E	*rpoC* L125L	114	5	7.755207
rpoB S441A	*sigA* R352R	113	4	7.741132
rpoB D545E	*rpoC* R89R	115	5	7.736936
rpoB D545E	*sigA* L423L	115	5	7.736936
rpoB D545E	*rpoB* R614R	115	5	7.736936
rpoB D545E	*rpoB* L591L	115	5	7.736936
rpoB D545E	*rpoB* L206L	115	5	7.736936
rpoB H445Y	*rpoB* R557H	0	4	7.733532
rpoB H445Y	*sigA* S442R	0	4	7.733532
rpoB D545E	*rpoC* A511A	38	4	7.731616
rpoB S441A	*rpoC* T38T	19	3	7.720519
rpoB D545E	*rpoC* I777I	116	5	7.718822
rpoB D545E	*rpoB* G759G	116	5	7.718822
rpoB D545E	*rpoC* L1176L	117	5	7.700860
rpoB D545E	*rpoC* G1198G	117	5	7.700860
rpoB S441L	*rpoC* A701V	4	3	7.693082
rpoB D545E	*rpoC* E163E	118	5	7.683050
rpoB D545E	*sigA* A255A	118	5	7.683050
rpoB D545E	*rpoB* R167R	118	5	7.683050
rpoB S441A	*rpoB* V790V	117	4	7.682241
rpoB R448Q	*sigA* A86E	1	2	7.667873
rpoB D545E	*rpoB* S431S	119	5	7.665387
rpoB D545E	*rpoC* F1175F	119	5	7.665387
rpoB D545E	*sigA* R301R	119	5	7.665387
rpoB S450F	*rpoB* N75N	0	3	7.660113
rpoB D545E	*rpoB* G693G	120	5	7.647871
rpoB D545E	*rpoB* S519T	120	5	7.647871
rpoB H445R	*rpoC* D714N	0	3	7.645795
rpoB S450L	*rpoC* E1140D	6	18	7.641946
rpoB D545E	*rpoC* G115G	121	5	7.630498
rpoB D545E	*rpoB* R253M	121	5	7.630498
rpoB D545E	*rpoB* A125A	121	5	7.630498

rpoB D545E	*rpoB* D543D	121	5	7.630498
rpoB D545E	*rpoB* V174V	121	5	7.630498
rpoB D545E	*rpoC* A212A	121	5	7.630498
rpoB S450N	*rpoB* P483P	115	3	7.627129
rpoB H445Q	*rpoB* F424L	9	3	7.618394
rpoB D545E	*rpoB* G546G	122	5	7.613267
rpoB D545E	*rpoB* V562V	122	5	7.613267
rpoB S450N	*rpoB* L591L	117	3	7.605055
rpoB S441A	*rpoB* G489G	21	3	7.601960
rpoB S450N	*rpoB* G759G	118	3	7.594156
rpoB D545E	*rpoB* P682P	124	5	7.579219
rpoB D545E	*rpoC* A230A	124	5	7.579219
rpoB D545E	*rpoB* G602G	124	5	7.579219
rpoB D545E	*rpoB* T410T	124	5	7.579219
rpoB D545E	*rpoC* A174A	124	5	7.579219
rpoB D545E	*rpoB* T288T	125	5	7.562399
rpoB D545E	*rpoC* G254G	125	5	7.562399
rpoB S450N	*sigA* R301R	121	3	7.561998

Table S1: List resulting from Fisher's exact test for association of resistance with co-occuring mutations. The first column indicates the resistance mutation that the putative compensatory mutation (CM) in the second column is associated to. 'Only CM' indicates how often the CM occurs on its own, without the corresponding resistance mutation, and 'both' indicates how often we see the two mutations occur together. The negative logarithm of the p-value is displayed in the last column. Entries with an association p-value below the Bonferroni correction (-log10(0.01/360000) = 7.56) are not displayed, but can be accessed through the github repository.[168]

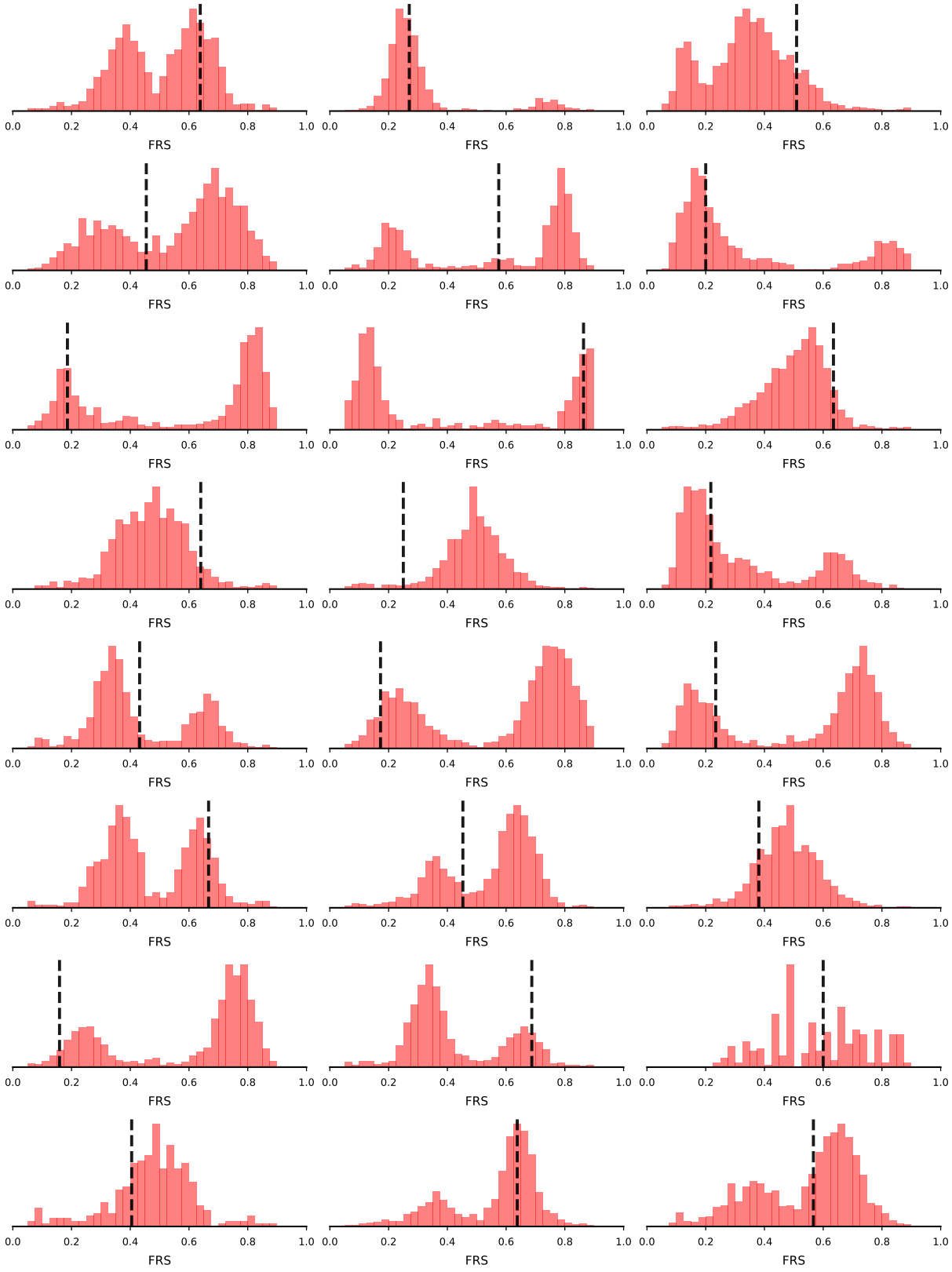


Figure S1: FRS distribution of mutations and the respective RAV in heterogeneous samples with multiple (sub)lineages. The number of (heterogeneous) mutations at each FRS are shown in histogram representation. A black dashed line indicates the FRS of the RAV in the sample. This plot is showing samples 1-24.

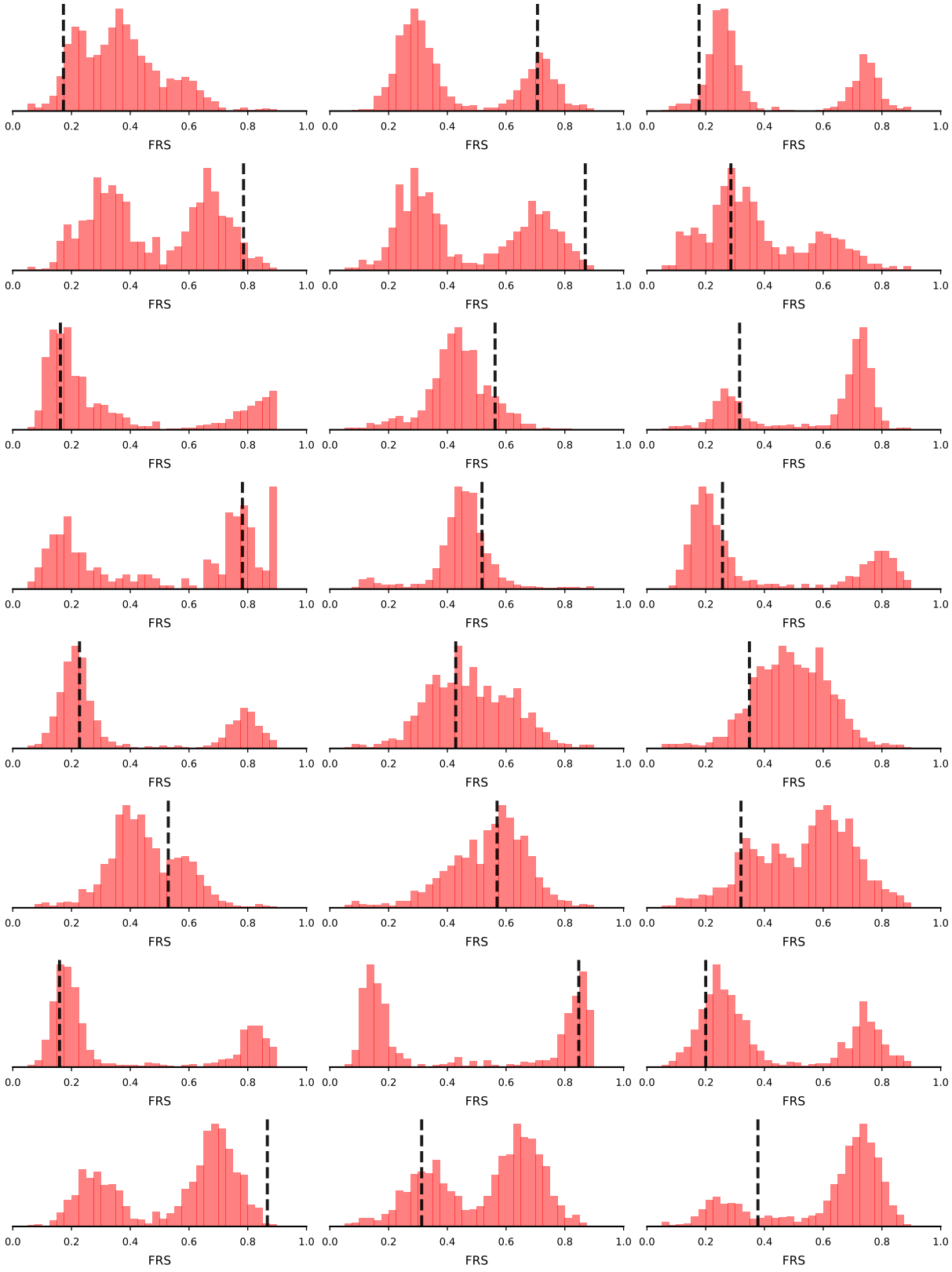


Figure S2: FRS distribution of mutations and the respective RAV in heterogeneous samples with multiple (sub)lineages. The number of (heterogeneous) mutations at each FRS are shown in histogram representation. A black dashed line indicates the FRS of the RAV in the sample. This plot is showing samples 25-48.

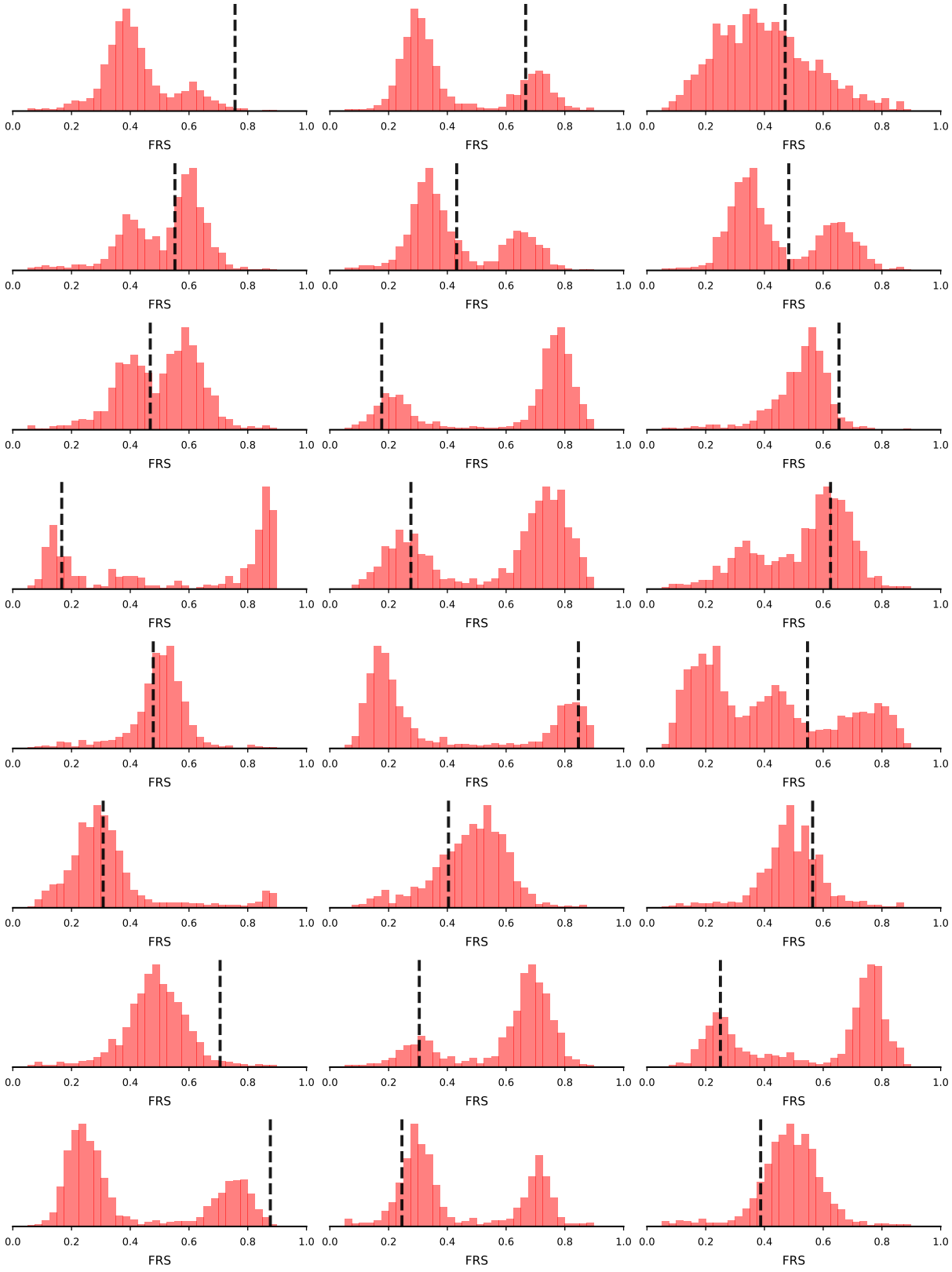


Figure S3: FRS distribution of mutations and the respective RAV in heterogeneous samples with multiple (sub)lineages. The number of (heterogeneous) mutations at each FRS are shown in histogram representation. A black dashed line indicates the FRS of the RAV in the sample. This plot is showing samples 49-72.

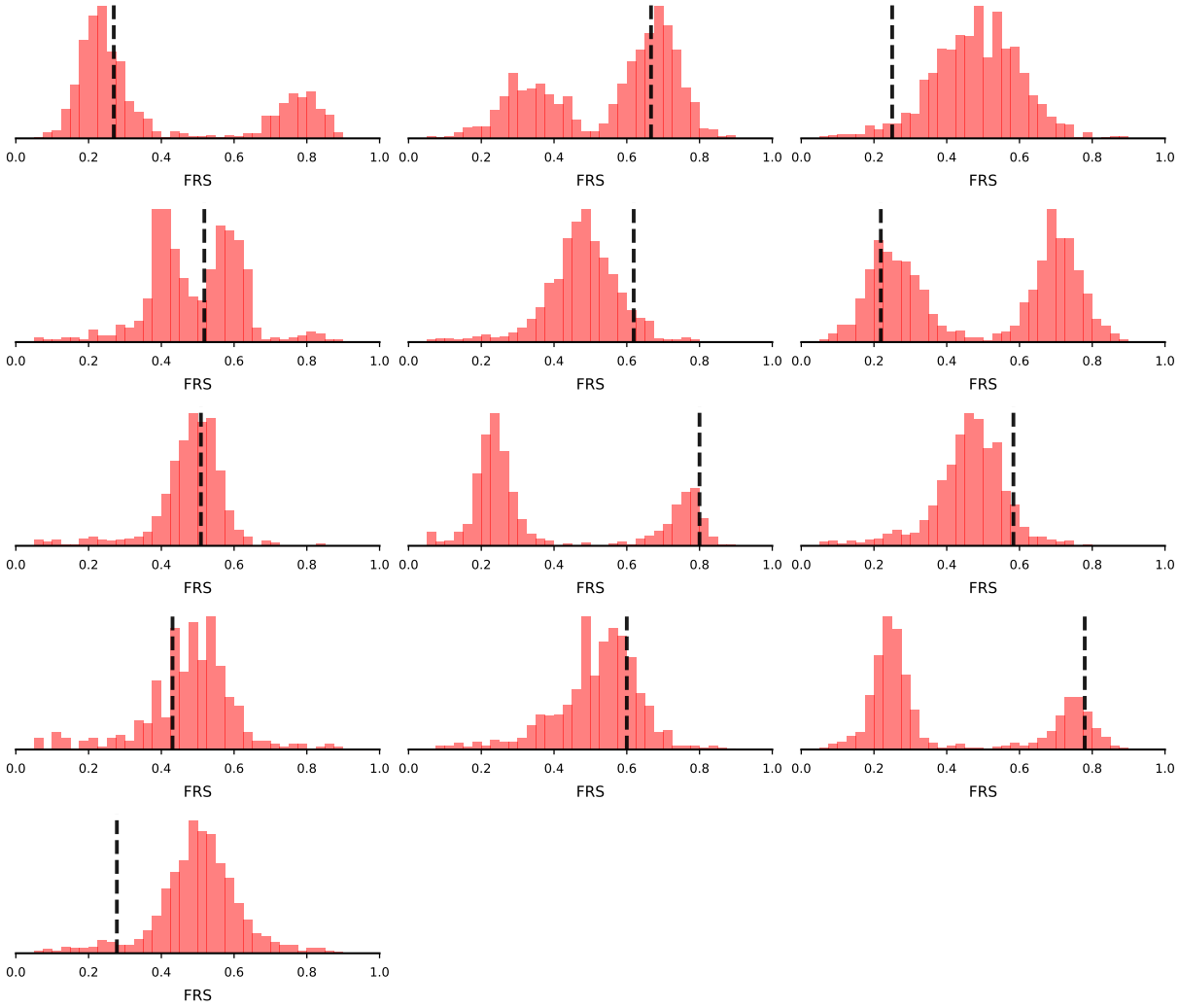


Figure S4: FRS distribution of mutations and the respective RAV in heterogeneous samples with multiple (sub)lineages. The number of (heterogeneous) mutations at each FRS are shown in histogram representation. A black dashed line indicates the FRS of the RAV in the sample. This plot is showing samples 72-85.