Rotation Project

Compensatory Mutations in the Polymerase of *Mycobacterium Tuberculosis* increase Fitness above Wild Type Levels

Hilary Term 2022

Supervised by:
Dr. Philip Fowler
philip.fowler@ndm.ox.ac.uk
Nuffield Department of Medicine

Author:
Viktoria Brunner
viktoria.brunner@univ.ox.ac.uk

Acknowledgements

Thank you to Dr. Philip Fowler for inspiring the idea for this project and hosting me for the past three months in his group. Being also my primary supervisor during this time, he taught me a lot about AMR research and sustainable programming in designing packages. I am equally thankful for the lively exchange with the other group members, Alice, Charlotte and Matty, in the weekly group meetings as well as in our scarce but valued time in person on site at the JR. Lastly, I want to thank the DTC for supporting my studies and enabling me to conduct this project in the first place. I appreciate this opportunity and am doing my best to justify the confidence that was placed in us as DPhil students.

Abstract

Resistance mutations in *Mycobacterium tuberculosis (M. tuberculosis)* to various antibiotics often come with a fitness cost for the bacteria. Accordingly, resistance to the first-line drug rifampicin, arising predominantly through mutations within the β (*rpoB*) subunit of the RNA polymerase, leads to slower growth of *M. tuberculosis* compared to susceptible populations. This fitness cost can be alleviated by compensatory mutations (CMs) in other regions of the polymerase, most often in the β' (*rpoC*) subunit. These CMs are of particular interest clinically, since they have the potential to lock in resistance mutations, hence encouraging the spread of resistant strains worldwide. Here, we report the statistical inference of a large number of CMs in various polymerase subunits of *M. tuberculosis*. In contrast to previous studies in this area, we were able to conduct our investigations based on a much larger data set, employing the over 60.000 *M. tuberculosis* genomes collected and sequenced as part of the CRyPTIC project. This allowed us to use much more powerful statistical tests to investigate the association of putative CMs with resistance-conferring mutations. Overall, we confirmed 50 previously described and novel CMs by means of statistical association testing and propose hypotheses for how they work by mapping them onto the protein structure. In addition, we were able to confirm the positive effect of several of the CMs on the growth and hence presumably on the fitness of *M. tuberculosis*. Our results suggest that some CMs not only restore fitness, but also increase it above wild type levels, thereby potentially leading to increased virulence. This finding increases the urgency for accurate, cheap and widely accessible diagnostics for tuberculosis to not only improve patient outcomes but also to prevent the emergence of even more virulent strains.

Introduction

The rise of multidrug-resistant bacteria is one of the grand challenges we are facing as a global society. Almost a century after the fight against infectious diseases seemed to be near its end due to the discovery of potent antibiotics like penicillin, we are still far from reaching that goal [1, 2]. This is because short generation times, high mutation rates and genetic recombination, all trademarks of many pathogenic bacteria, make it possible for resistance to arise within extremely short time frames and spread rapidly throughout populations [3]. Inappropriate administration of antibiotics can potentially lock in new resistance mutations or contribute to the faster emergence and spread of multidrug-resistant strains through artificial selection [1]. With the prevalence of antibiotic resistance increasing and the discovery of potent new drugs slowing down, we are heading towards a crisis of our own making [3, 4].

M. tuberculosis is a bacterium prone to developing resistance to major antibiotics [5]. Although the first treatment for tuberculosis (TB) was identified in 1948 [6], it remains responsible for the death of about 1.5 million people per year. In addition, the percentage of people infected with *M. tuberculosis* strains resistant to the major first-line antibiotic drugs is rising [7–9]. Understanding how resistance emerges, how it is locked in and how it spreads is hence of high importance.

One of the four first-line antibiotics for treatment of TB is the drug rifampicin (RIF). The resistance to this compound arises through mutations in the rifampicin resistance determining region (RRDR) of the RNA polymerase (RNAP) [10, 11]. The RRDR is located within the RNAP gene *rpoB*, and codes for the binding site of RIF. In susceptible bacteria, amino acids within the RRDR allow RIF to form hydrogen bonds and van der Waals interactions with the polymerase (Figure 1). Bound RIF then

sterically obstructs the elongation of newly synthesized RNA, thereby stalling protein production in the bacteria [12]. Since *M. tuberculosis* does not exhibit horizontal gene transfer [13], resistance to this drug mostly arises through chromosomal mutation of amino acids within or close to the RRDR. This prevents rifampicin from binding [14], but at the same time changes this close to the active site of the RNAP introduce a fitness cost [15, 16]. To alleviate this cost, CMs emerge in other parts of the protein [17]. The existence of CMs gives a plausible explanation for the persistence of resistance mutations long after antibiotic treatment is stopped, when the fitness cost should lead to reversion to a susceptible phenotype.

Because of this potential to explain the persistence of resistance phenotypes and the epidemiological spread of resistance, CMs have been extensively investigated. Hundreds of potential candidates have been described in previous works [18–24], but often the statistical power was too low to confirm resistance association. And only for a small subset of putative CMs have the effects on growth and polymerase activity been confirmed experimentally [23, 25]. To further our understanding of resistance spread and persistence, it would be useful to close this gap and, if possible, to dissect the individual contributions of specific CMs to the fitness of *M. tuberculosis*.

In this project, I will identify new candidates based on a sequenced collection of more than 60.000 *M. tuberculosis* sample genomes processed as part of the CRyPTIC project [26], likely confirming previously described CMs along the way. The dataset size ensures the statistical analysis is about 10-100 times more powerful than previous studies in this field. The other big advantage of the approach is that there is growth data available for a large proportion of the samples [27]. This allowed me to correlate growth phenotypes with the respective genotype, i.e. the presence of resistance and compensatory mutations. As a consequence, I was able to quantify the amount of fitness increase through the presence of CMs in general, and even dissect the individual contributions of certain CMs to the higher fitness of resistant *M. tuberculosis*.

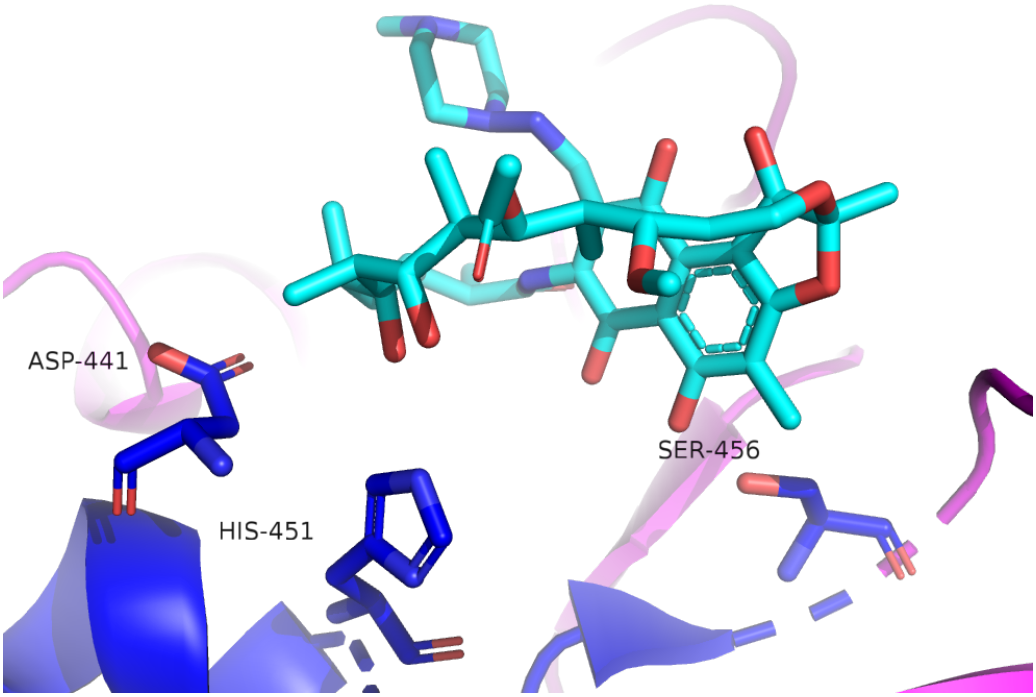

Figure 1: Close-up of the drug rifampicin (RIF) bound to the β subunit of the RNA polymerase (RNAP): The drug RIF is depicted in cyan. The β subunit of the RNAP is shown in magenta, with the RRDR highlighted in dark blue. The protruding amino acids Asp441, His451 and Ser456 are reported to form hydrogen bonds or Van der Waals interactions, respectively, with the drug [12]. Numbering of the amino acids on the β (*rpoB*) subunit is shifted by 6 amino acids in this representation, so the amino acids are in the following referred to as D435, H445 and S450. Due to the proximity of the RRDR to the RNAP active center, the binding of RIF causes a disruption of the RNA synthesis due to steric clash. (PDB: 5UHB)

Results

Rifampicin resistance mutations introduce a fitness burden in *M. tuberculosis*

In order to confirm that Rifampicin (RIF) introduces a fitness cost in *M. tuberculosis*, one needs to show that susceptible bacteria exhibit higher fitness than samples with resistance mutations. To measure the fitness of bacteria with and without resistance mutations, I used the average growth of *M. tuberculosis* samples in the positive control wells of the CRyPTIC projects' 96-well broth microdilution plates after two weeks incubation as a proxy for fitness [27]. Growth and fitness are thus often used interchangeably in the following results. Samples were assumed to be resistant if they contained mutations that were classified as RIF resistance mutations by the CRyPTIC project [26]. In theory I should exclude CMs to make sure that no additional mutations in those samples interfere with the fitness. Since I have not identified any CMs yet, I shall instead only allow synonymous mutations to co-occur with resistance mutations.

The growth distribution of these well-defined resistant samples is significantly different from that of pan-susceptible samples (Figure 2A), as confirmed by the low Mann-Whitney p-value (Table 1). Since the median growth of the resistant samples is visibly lower than the susceptible samples, I conclude that there is significantly better growth in non-resistant bacteria. By extrapolating from average growth to fitness, I can further state that fitness is reduced when resistance mutations are present.

Previous publications indicated that the magnitude of the fitness cost in *M. tuberculosis* is dependent on the type of resistance mutation. The resistance mutation associated with the lowest fitness cost was S450L in *rpoB* [16], the most prevalent RIF resistance mutation clinically. There were sufficient samples to obtain decent growth plots and p-values for the three most common resistance mutations in the dataset: S450L, H445Y and D435V (Figure 2B-D). I used the absolute difference between the respective growth distribution medians as a measure of fitness difference. In line with previous studies, the median growth of the *rpoB* S450L mutants was the highest out of the three resistance-conferring mutations.

mutation	median growth [%]	p-value	n
any resistance	17.57	2.95e-12	795
rpoB S450L	16.65	4.37e-03	196
rpoB H445Y	15.50	1.26e-02	42
rpoB D435V	16.40	3.92e-14	325

Table 1: Median growth of samples with different resistance mutations. Mann-Whitney p-value is calculated in reference to pan-susceptible sample growth and n indicates the sample size.

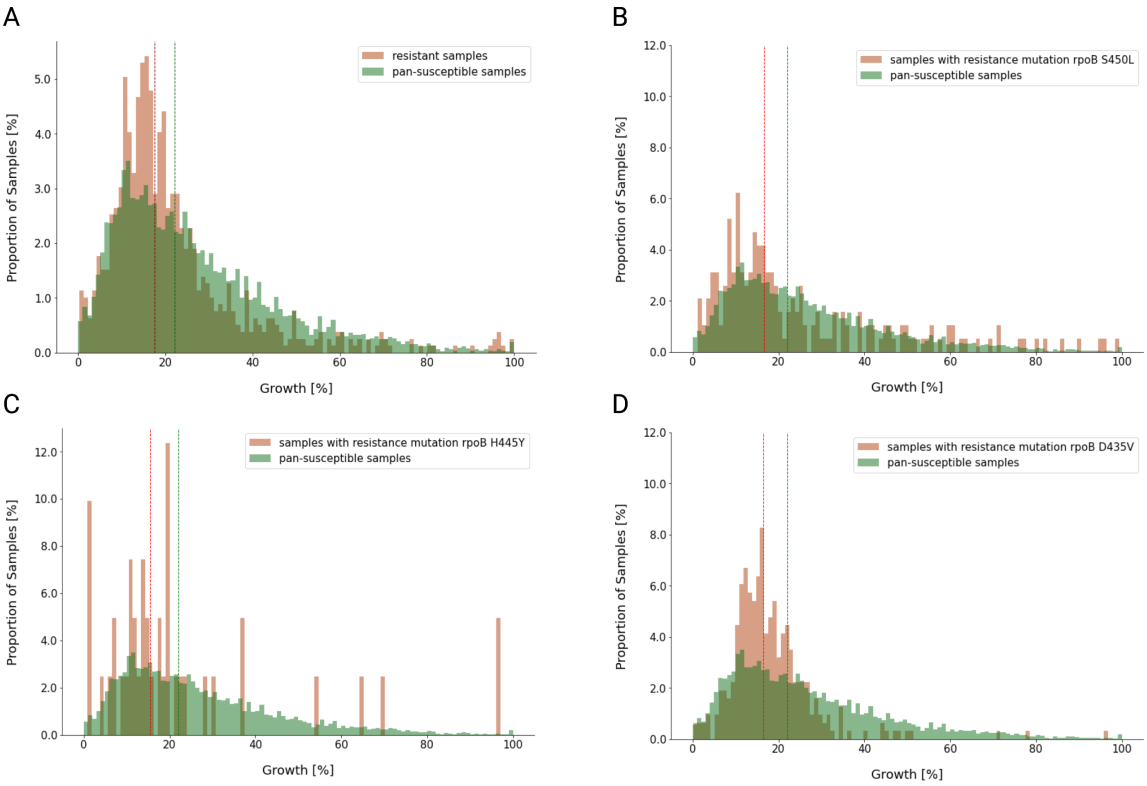

Figure 2: Growth distributions for pan-susceptible vs rifampicin (RIF) resistant samples in *M. tuberculosis* (A) Distribution of growth in percent of covered well area as measured in a previous project [27], plotted against the proportion of samples that display this amount of growth. Samples with RIF resistance mutations but no other potentially interfering mutations are plotted in red, samples that were classified as pan-susceptible are plotted in green. A dotted line indicates the respective median. This plot reflects the whole dataset and medians and the Mann-Whitney p-value are displayed in Table 1. **(B-D)** Plot layout as in A, but the red bar plot represents a subset of RIF resistant samples that exhibit only the resistance mutation indicated in the legend and no other potentially interfering mutations. The Mann-Whitney p-values and medians of the distributions are shown in Table 1.

Compensatory mutations can be identified through statistical association with resistance mutations

Since there is a significant fitness cost in samples that only contain non-synonymous mutations that are associated with resistance in the RNAP, we should observe CMs restoring fitness in a large proportion of resistant samples. The identification of CMs is not trivial and strongly depends on how they are defined. In many publications, due to the low number of available samples, CMs are simply assumed to be all mutations that co-occur with resistance mutations [20]. Higher numbers of samples allow CMs to be defined as mutations that exclusively occur with resistance mutations [18, 22, 24]. With a dataset of our size, where sequencing errors and wrong resistance classification of samples are likely to produce false positives and false negatives, it then becomes necessary to employ statistical association testing. I hence applied Fisher's exact test for a well-defined subset of pairs of resistance and co-occuring mutation, to determine if the latter is statistically associated with resistance or not.

I was able to recover up to 93 % of literature high-confidence CMs (Figure 3A), which is an indicator of overall good performance. After choosing cut-offs for mutation prevalence (Figure 3A, Supplementary Figure 1) and p-value (Figure 3B), I obtained a preliminary list of mutations that are significantly associated with RIF resistance.

To refine the list, I removed all synonymous and known lineage-defining mutations. The former are unlikely to have an effect on the phenotype, while the latter might interfere with our analysis. *M. tuberculosis* lineages exhibit distinct growth patterns, with some lineages growing much better than others. Lineage-defining mutations can hence be confounding factors, being mistaken for CMs if they occur mostly in lineages with high proliferation and resistance rates. The final list of putative CMs (Supplementary Table S2) can now be used for growth phenotype analysis.

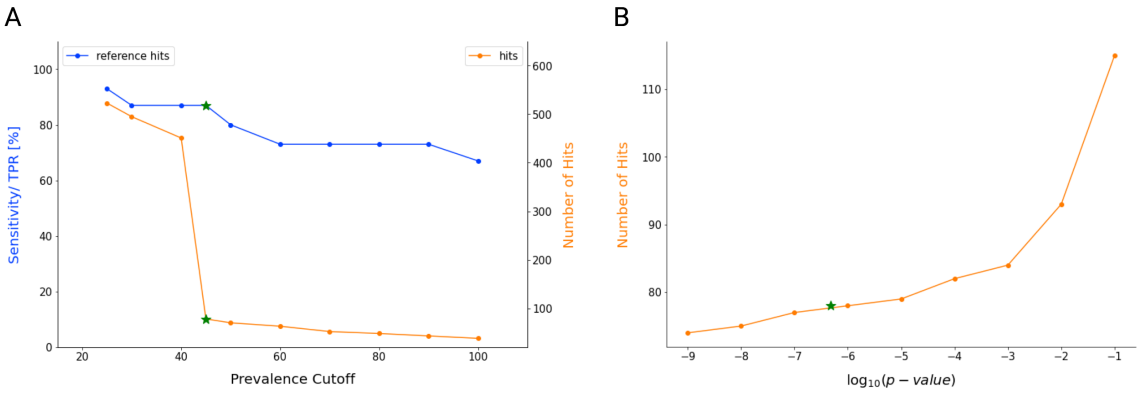

Figure 3: Sensitivity and number of significant hits (putative CMs) depending on prevalence cut-off decision and p-value (A) Influence of prevalence cut-off for individual mutations on sensitivity and number of hits. The left y-axis refers to the percentage of found reference hits from a compiled list, 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 chosen mutation prevalence cut-off. Green stars indicate the final cut-off choice, which is a consensus of considering both sensitivity and number of significant hits. **(B)** The graph shows the number of significant hits detected for the mutation prevalence cutoff of 45 and the $\log_{10}$ p-value shown on the x-axis. The green star indicates the p-value of choice, which is a p-value of 0.01 with Bonferroni correction for multiple testing.

Compensatory mutations enable resistant bacteria to recover and surpass wild-type growth levels

If I have correctly identified CMs, we should be able to see growth recover in the respective resistant samples. To test this, I analysed the growth distributions of pan-susceptible samples, as well as resistant samples that have at least one CM from our list (Supplementary Table S2), and resistant samples that show none.

We clearly see recovery of the resistant growth phenotype through the introduction of CMs (Figure 4A). The associated p-values indicate that growth, and hence we assume fitness, is significantly higher in resistant samples with CMs than in resistant samples without (Table 2). But the most surprising finding is probably that the average growth of *pan-susceptible* samples is significantly *lower* than the average growth of resistant samples with CMs (Figure 4A, Table 2). While it has been shown that CMs can restore growth and polymerase activity to wild-type levels, it has not yet been shown that CMs can increase growth to an extent that would allow resistant bacteria to out-compete the wild-type even in absence of antibiotics.

When looking at individual CMs and their effect on the growth phenotype, we want to be able to detect if increases in fitness are only a lineage effect. I attempted to filter out lineage-defining mutations, but it is possible other lineage-associated mutations remain in our analysis. In the overall dataset, lineage 4 makes up more than 50 % of samples (Figure 4B, left), while the most common resistance mutation, *rpoB* S450L, is found predominantly in Lineage 2 (Figure 4B, right). Since the majority of putative CMs are associated with this resistance mutation (Supplementary Table S2), we would expect a similar lineage distribution for samples containing these CMs.

I plotted growth distributions for a subset of samples with specific CMs against the pan-susceptible samples and samples with only the associated resistance mutation S450L. It is striking how much the magnitude of the effect on growth distributions varies, depending on the CM (Figure 4C-H, Supplementary Figure S2). For example, the amino acid at position 483 in the β' (*rpoC*) subunit exhibits two mutations that are both putative compensatory mutations. One mutates the amino acid Valine to Alanine and shows a strong fitness increase above wild type levels (Figure 4E). The other modification changes Valine to Glycine, and appears to restore fitness to the level of pan-susceptible samples, which correspond to the wild-type (Figure 4F). We observe similar effects for other CMs, with the strongest effect on growth resulting from the mutation of Isoleucine at position 491 on the β' subunit to either Valine or Threonine (Table 2).

As for the lineage distribution for CMs, the distribution is skewed towards Lineage 2 (Beijing lineage, blue areas in inlays of Figure 4). This is expected, since the Beijing lineage is known to accumulate resistance and is the most prevalent lineage in samples showing the resistance mutation S450L (Figure 4B, right). But if a specific mutation is found exclusively within one lineage, it can indicate that we are not dealing with a universal CM but rather a lineage-associated mutation. We find this strong correlation with one lineage for several of our putative CMs, most clearly for E1092D on *rpoC* (Figure 4C) and the *rpoB* promoter mutation c-61t (Supplementary Figure S2A). For now this observation does not rule out that these mutations are resistance-associated. Within their lineages, those mutations might still exert a compensatory effect, but this needs to be confirmed by growth data.

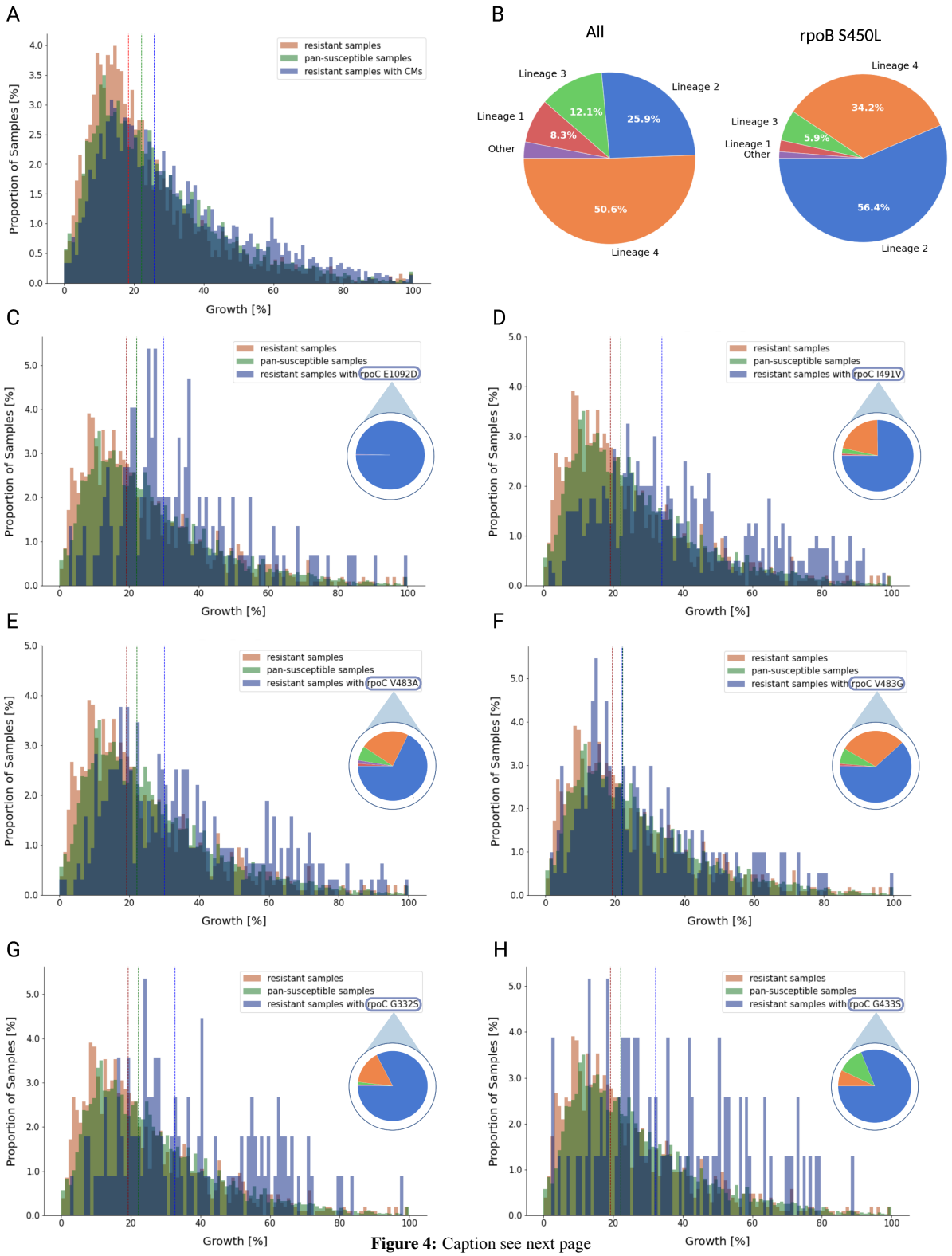

Figure 4: Caption see next page

Figure 4: Growth and lineage distributions of *M. tuberculosis* samples that are pan-susceptible, rifampicin (RIF) resistant and resistant with compensatory mutations (CMs) (A) Distribution of growth in percent as measured in the CRyPTIC project [27], plotted against the proportion of samples that display this amount of growth. Samples with RIF resistance mutations but no other potentially interfering mutations are plotted in red, samples that were classified as pan-susceptible are plotted in green. Samples that have RIF resistance mutations and at least one CM are shown in blue. This plot reflects the whole dataset and medians and Mann-Whitney p-values of the three distributions are shown in supplementary Table 2. **(B)** Lineage distribution for all samples (left) and for samples with the most common RIF resistance mutation S450L (right). **(C-H)** Plot structure as in A, but the blue bar plot represents a subset of samples that show the CM indicated in the legend, the resistance mutation S450L and no other putative CMs. The red bar plot only includes samples with the resistance mutations S450L and no CMs. The medians and Mann-Whitney p-values of the distributions are shown in Table 2. The pie plot below the legend indicates lineage distribution of samples that show the circled CM. The color code of the pie plot is equivalent to the one used in B.

mutation	median growth [%]	p-value to res	p-value to sus	n
susceptible (sus)	22.10			5283
resistant and no CMs (res)	18.37			2554
resistant and any CMs	25.73	3.753e-54	6.24e-22	2982
resistant and specific CM:				
rpoC E1092D	29.86	3.13e-13	2.78e-09	153
rpoC V483G	21.98	0.00637	0.458	204
rpoB c-61t	23.95	1.88e-08	0.000353	327
rpoC I491V	33.89	2.40e-34	9.27e-31	413
rpoC V483A	29.99	9.59e-19	1.80e-13	318
rpoC I491T	33.70	2.24e-17	6.51e-14	141
rpoC P1040R	22.87	8.01e-08	0.00101	274
rpoC G332S	32.69	3.89e-13	5.700e-10	117
rpoC G433S	32.20	4.57e-09	1.04e-06	89

Table 2: Median growth of samples with different compensatory mutations compared to susceptibles and samples with only resistance. Mann-Whitney p-value is calculated in reference to either growth of resistant samples without CMs or to pan-susceptible samples and n indicates the sample size.

The effect of compensatory mutations on fitness is interwoven with lineage association

Since the resistance phenotype is strongly correlated with lineage 2 (Figure 4B), we have to look at the differences in growth distributions between and within lineages to make sure the effect of CMs on growth is not merely an artifact of the imbalanced dataset. It is known that the Beijing lineage accumulates resistance mutations and spreads at a faster pace than other lineages [9]. If this is true, it could be highly problematic for the analysis, since I have identified CMs through correlation with resistance mutations and confirmed them through their growth phenotype as a proxy for sample fitness. Mutations that are associated with the Beijing lineage could hence show both of these attributes without necessarily being a CM.

mutation	median growth [%]	p-value to res	p-value to sus	n
susceptible (sus)	26.08			1331
resistant and no CMs (res)	22.66			506
resistant and specific CM:				
rpoC E1092D	29.86	3.33e-06	0.000152	153
rpoC I491V	37.17	2.93e-25	5.25e-28	360
rpoC V483A	32.63	3.18e-11	1.32e-09	248
rpoC I491T	34.43	2.03e-10	1.44e-09	139
rpoC G332S	34.19	1.55e-08	2.08e-07	109
rpoC G433S	33.09	5.77e-06	5.36e-05	82

Table 3: Median growth of samples from lineage 2 with different compensatory mutations, compared to growth of susceptible and resistant samples without CMs. Mann-Whitney p-value is calculated in reference to either growth of resistant samples without CMs or to pan-susceptible samples. All samples are from lineage 2 and n indicates the respective sample size.

To quantify the effect of lineage affiliation on growth and thereby on the fitness phenotype of samples, I plotted growth of the pan-susceptible samples for different lineages. Most importantly, we are interested in the growth of lineages 2 and 4, which dominate the overall dataset (Figure 4B). It is apparent that lineage 2 grows significantly better than lineage 4 in the *in vitro* 96-well experiments (Figure 5A), but it is out-performed by lineage 3 (Figure 5B). Except for the growth difference between lineages 1 and 2, the differences between lineages are all significant (Supplementary Table S3). The better growth of the resistance-prone Beijing lineage is unfortunate, since it makes lineage affiliation a confounding factor in the endeavour to confirm CMs and their positive effect on fitness through their growth phenotype.

In order to exclude that the fitness advantage seen in samples with CMs is purely caused by lineage association, I plotted the growth of samples with different genotypes and Beijing lineage background. If the better growth of resistant samples with CMs than wild-type *M. tuberculosis* is not solely based on lineage affiliation, we should see significant differences in growth distributions within this lineage as well.

In Beijing lineage background, six of our putative CMs (Figure 5, Supplementary Figure S3) are shown to push growth in resistant samples significantly higher than in susceptible samples (Table 3). We can hence confirm a significant fitness advantage, that is independent of the lineage, for all samples with any of these putative CMs. The most surprising finding might be that the mutation E1062D on the β' subunit, that almost exclusively occurs in lineage 2 and is hence very likely lineage-associated, still enhances growth within the lineage significantly over wild-type levels (Figure 5C, Table 3). This mutation thus appears to be both lineage-associated and a CM.

The fact that even within lineage 2 the average growth of resistant samples with CMs is exceeding wild-type growth levels confirms that this effect is most likely not purely lineage associated. We still decided to exclude mutations from our list that overwhelmingly occurred in one lineage. Only for hits where we have explicit evidence that they increase growth even within their lineage (e.g. E1092D), we can exclude that their effect on fitness of the *M. tuberculosis* samples is a result of lineage affiliation. This last filtering step yielded a final, high-confidence list of 50 putative CMs, of which 13 are novel candidates (Table 4).

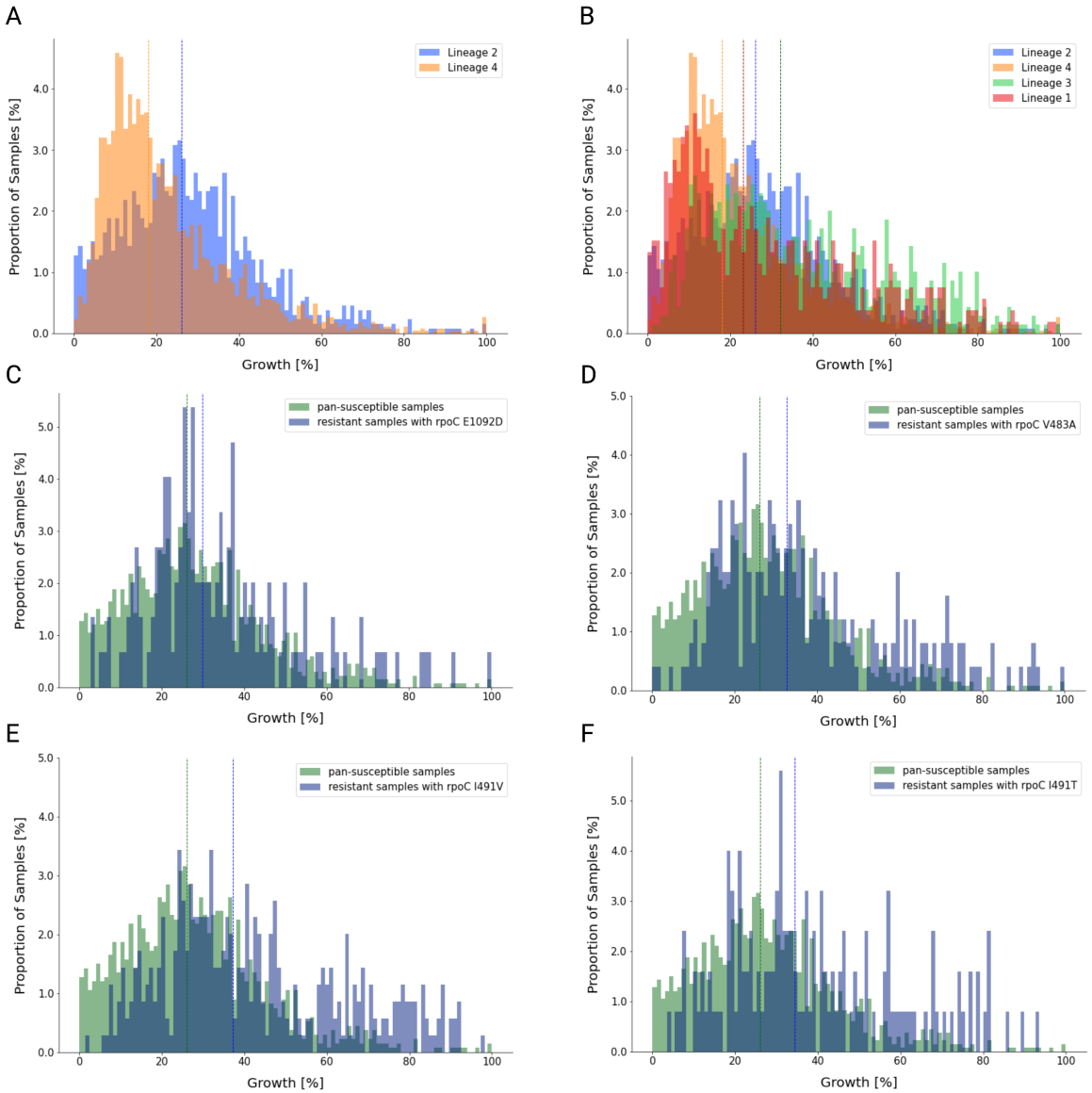

Figure 5: Growth distributions of *M. tuberculosis* samples for and within different lineages (A) Distribution of growth in percent as measured in the CRyPTIC project [27], plotted against the proportion of samples that display this amount of growth. Samples from lineage 2 are plotted in blue, samples from lineage 4 in orange. **(B)** Plot layout as in A, but lineage 3 (green) and 1 (red) are included. **(C-F)** Distribution of growth in percent as measured in the CRyPTIC project [27], plotted against the proportion of samples from lineage 2 that display this amount of growth. Lineage 2 samples were classified as pan-susceptible (green) and RIF resistant with the CM indicated in the legend and no other putative CMs (blue).

resistance mutation	putative CM	literature evidence	exp. evidence
rpoB S450L	*rpoC* D485Y	✓	x
rpoB S450L	*rpoB* I480V	✓	x
rpoB S450L	*rpoC* H525Q	✓	x
rpoB S450L	*rpoA* V183G	✓	x
rpoB S450L	*rpoC* E1092D	x	✓
rpoB L430P	*rpoC* E1092D	x	x
rpoB V170F	*rpoC* E1092D	x	x
rpoB S450L	*rpoB* A286V	✓	x
rpoB S450L	*rpoB* V496A	x	x
rpoB S450L	*rpoC* T812I	✓	x
rpoB S450L	*rpoC* G519D	✓	x
rpoB S450W	*rpoC* G519D	✓	x
rpoB S450L	*rpoC* W484G	✓	x
rpoB V170F	*rpoC* W484G	✓	x
rpoB S450L	*rpoC* P1040S	x	x
rpoB S450L	*rpoB* L731P	✓	x
rpoB S450L	*rpoC* L449V	✓	x
rpoB S450L	*rpoC* V1039A	x	x
rpoB S450L	*rpoB* I488V	x	x
rpoB S450L	*rpoC* P1040R	✓	✓
rpoB S450L	*rpoC* V431M	✓	x
rpoB S450L	*rpoC* A521D	✓	x
rpoB S450L	*rpoA*D190G	✓	x
rpoB S450L	*rpoC* L507V	✓	x
rpoB S450L	*rpoC* G332S	x	✓
rpoB S450L	*rpoB* K891E	x	x
rpoB S450L	*rpoC* G332R	✓	x
rpoB S450L	*rpoC* V1252L	✓	x
rpoB S450L	*rpoA* -40 indel	x	x
rpoB S450L	*rpoC* V517L	✓	x
rpoB S450L	*rpoA* A180V	x	x
rpoB S450L	*rpoC* V1252M	✓	x
rpoB S450L	*rpoB* P45S	✓	x
rpoB S450L	*rpoC* K445R	x	x
rpoB I491F	*rpoC* E1033A	x	x
rpoB S450L	*rpoC* F452C	✓	x
rpoB S450L	*rpoC* N416S	✓	x
rpoB S450L	*rpoB* Q409R	x	x
rpoB S450W	*rpoB* Q409R	x	x
rpoB S450L	*rpoC* L527V	✓	x
rpoB S450L	*rpoB* R827C	x	x
rpoB S450L	*rpoC* P1040A	✓	x
rpoB S450L	*rpoB* P45L	✓	x
rpoB S450W	*rpoB* P45L	✓	x
rpoB S450L	*rpoC* G433S	✓	✓
rpoB S450L	*rpoC* I491T	✓	✓
rpoB S450L	*rpoC* I491V	✓	✓
rpoB S450L	*rpoC* L516P	✓	x
rpoB S450L	*rpoC* P434R	✓	x
*rpoB*S450L	*rpoA* T187A	✓	x
rpoB S450L	*rpoC* D485N	✓	x
rpoB S450L	*rpoC* V483G	✓	✓
rpoB Q432P	*rpoC* V483G	✓	x
rpoB S450L	*rpoC* F452L	✓	x
rpoB S450L	*rpoC* V483A	✓	✓
rpoB S450L	*rpoC* N698K	✓	x
rpoB S450L	*rpoC* N698S	✓	x

Table 4: Final list of putative CMs and respective associated resistance mutation. Checkmarks indicate CMs that have been previously mentioned in reference literature and have experimental evidence from our available growth data.

Most compensatory mutations are found close to contact regions of the RNA polymerase subunits

I mapped all putative CMs (Table 4) onto the available protein structure in complex with RIF. The hits cluster in four different regions of the RNAP (Figure 6A): the interface between subunits β, β' and α (Figure 6B); close to the RRDR within the β subunit (Figure 6C); around the secondary channel on the β' subunit (Figure 6D); and at the DNA entry channel (Figure 6E-F). Overall, almost no CMs were buried below the protein surface, where they would not serve any obvious function. This gives further confidence in our results.

The overwhelming majority of CMs, some of them showing a significant growth phenotype (Figure 4D: I491V and 4E: V483A), are found close to the contact region of three RNAP subunits (Figure 6B). We also find many hits close to the RRDR, where RIF binds to the RNAP of susceptible bacteria (Figure 6C). These CMs are in the β subunit of the RNAP, which has long been rejected as a possible location for CMs [23].

The RNAP secondary channel (Figure 6D) and the DNA entry channel (Figure 6E-F) are other locations for CMs that are seldom mentioned in the literature. Overall, the CM areas largely conform with those previously reported, but we also managed to identify new clusters. There appear to be different strategies for enhancing the activity of the RNAP, but the CMs introducing the highest growth benefit for *M. tuberculosis*, I491V/T, G332S and G433S (Table 2) are all part of the interface between the β and the β' subunit (Figure 6B, Supplementary Figure S4).

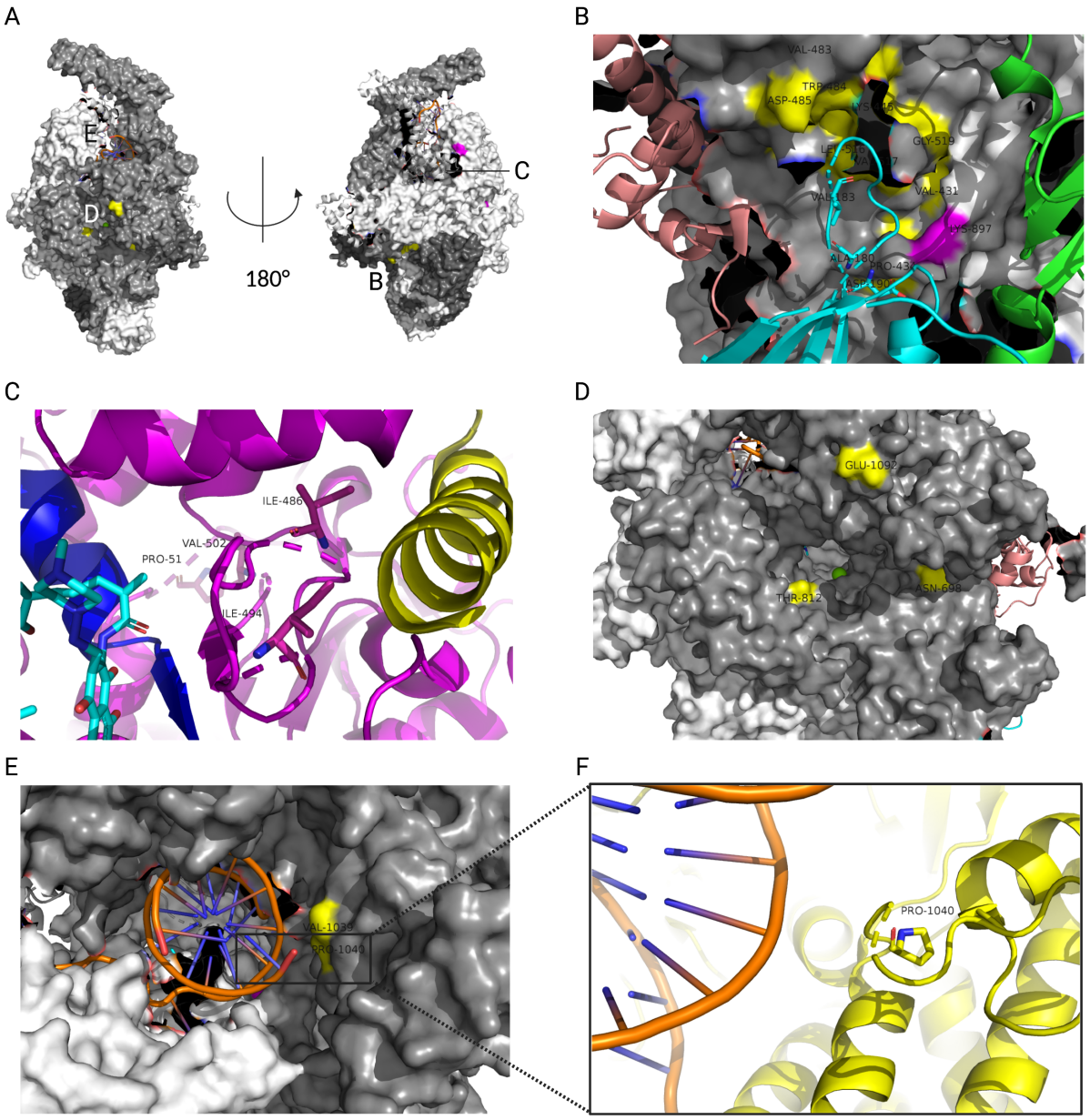

Figure 6: Locations of mapped compensatory mutations (CMs) on the RNA polymerase (RNAP) complex (A) Overview of clustering regions for CMs. Letters indicate where the close-ups shown in the rest of the figure are located. **(B)** The interaction region of subunits α (blue), β (dark grey) and β' (light grey). CMs can be found in all mentioned subunits and are highlighted by label and either in color (β: magenta and β': yellow subunits) or through stick representation (α subunit). **(C)** CMs close to the RRDR in the β subunit. Subunits are indicated by color, β in magenta, β' in yellow. RIF, here bound to the RRDR (dark blue), is shown in light blue. CMs are labelled and highlighted by stick representation. **(D)** Labelled CMs close to the RNAP secondary channel in the β' subunit (dark grey) are shown in yellow. The location of the active center can be deduced through the active site magnesium ion, indicated in green. **(E)** CMs close to the DNA entry channel. CMs on the β' subunit (dark grey) are labeled and indicated in yellow. The DNA helix is shown in blue and orange. **(F)** Close-up of the putative CM Proline (stick representation) to Arginine close to the DNA backbone.

Discussion

Integration of large sequencing datasets with carefully curated growth data gives enormous insight into genotype - phenotype correlations

Through statistical association testing, growth phenotype analysis and lineage affiliation studies, I obtained a final list of 50 high-confidence putative CMs (Table 4). Of these hits, 37 have been described in previous studies as being putative CMs [18–25], but mostly with insufficient statistical power to claim high confidence. For 9 of the overall hits I was in addition able to confirm the significantly increased fitness phenotype using the growth data (Figure 4, Supplementary Figure S2), showing that this effect is not solely based on lineage association (Figure 5, Supplementary Figure S3).

The high sensitivity of our approach for identifying CMs (Figure 3A) is thereby attributable to the uniquely large size and diverse nature of the sequencing dataset, giving our analysis much greater statistical power. But the inclusion of a range of laboratories in the sequencing endeavour also introduces a bias to the dataset. This, in addition to the fact that some lineages accumulate resistance, and are thereby over-represented among the resistant samples, makes the data slightly imbalanced. One way to correct for this is my attempt to dissect the influence of the respective sample lineage on resistance association. The next step would be the construction of a phylogenetic tree, annotating the positions of the high-confidence hits to confirm their status as CMs. CMs will be found at the tips of the phylogenetic tree, since they only arise upon RIF exposure in an evolutionary short time-frame.

The availability of growth data for a large number of samples is the other big advantage of our project. Having phenotypic data available for the different genotypes is vital for our validation of CMs. Since the growth data needed to be collected for a huge number of *M. tuberculosis* samples, the analysis was conducted in high-throughput fashion and *in vitro*. Bacteria were incubated in 96-well plates, which is not something we will see in a real infection and the data is hence not as reliable as carefully curated *in vivo* experiments. In addition, we made the assumption of growth being the equivalent of fitness, which is a strong simplification. We hence cannot necessarily translate our results to the epidemiological reality of *M. tuberculosis* spread in human populations. Polymerase activity assays could potentially close this gap, by testing laboratory-derived *M. tuberculosis* samples with an engineered combination of resistance mutations and CMs.

Lastly, we did not account for possible associations of mutations outside of the RNAP with RIF resistance. There have e.g. been experiments showing large-scale changes in gene expression following CM emergence in *rpoC* [28]. Nevertheless, the insights gained by combining the sequencing data with our available growth data already allowed us to make powerful conclusions concerning the phenotypic changes following emergence of resistance and CMs. The latter arise in a similar fashion in various other bacteria, such as *Salmonella enterica* and *Escherichia coli* [29, 30]. Our approach could hence be applied in a similar fashion to other organisms and drugs where resistance mutations are known and introduce a fitness cost.

Compensatory mutations can push fitness of *M. tuberculosis* samples over wild-type levels

I showed that the putative CMs can restore growth of *M. tuberculosis* samples at least to the level of susceptible sample growth (Figure 4). The mag-

nitude of this effect varied depending on the location and type of amino acid that were changed by the CM. The large differences observed, especially for different mutations at the same amino acid position (Figures 4E-F), indicate that the change of the amino acids' physio-chemical properties is of importance for the effect of each CM on growth. And in most cases, the presence of CMs pushed growth of resistant *M. tuberculosis* even higher than that of susceptible bacteria (Figure 4A). Since growth is our proxy for fitness, this demonstrates that CMs can restore fitness of resistant *M. tuberculosis* to levels higher than the wild-type.

Because the growth data was acquired *in vitro*, we have to be careful about any conclusions we draw from the previous observations. But assuming that these results can be reproduced *in vivo*, the increased fitness of rifampicin-resistant *M. tuberculosis* might explain the fast spread of multi-drug resistant strains. Evidence for this could be the Beijing lineage, which shows accumulation of resistance mutations as well as CMs. Due to its reported higher virulence, it currently out-competes other lineages [31].

If RIF resistance can ultimately lead to an evolutionary advantage through CM acquisition, the use of this antibiotic has to be more tightly controlled to prevent multi-resistant lineages from dominating global TB infections. Most importantly, resistance and lineage screening before drug administration should be self-evident.

Lineage association of compensatory mutations could explain the faster spread of certain *M. tuberculosis* lineages

The CM E1092D in the RNAP β' subunit is an interesting case, being classified both as a compensatory and a lineage-associated mutation in our analysis. It has been discovered in previous CM investigation studies, but was excluded due to evidence of strong association with the Beijing lineage [32]. Since we still see a positive effect of E1092D presence on growth of resistant samples within the Beijing lineage, we speculate that lineage effects might not be mutually exclusive with CMs. Excluding lineage-defining and associated mutations from analysis might therefore discard valuable information.

If a mutation that shows strong association with a lineage also exerts a compensatory effect, this could be yet another explanation for the faster spread of bacteria from the respective lineage. For instance, the large association of the high-confidence CM E1092D with the multi-drug resistant Beijing lineage might be one reason for its extensive dominance in eastern Asia. However, this must be investigated further before conclusions can be drawn. E1092D would need to be tested for resistance association within the Beijing lineage and ideally, one would wish to find a mechanistic explanation for the compensatory effect. Unfortunately, this CM is far from the subunit interfaces or other previously described CM locations (Figure 6D). While being reasonably close to the secondary channel of the RNAP, this is not enough to warrant any definite conclusions about the mode of action of this CM. The mechanistic function of E1092D remains unknown and we cannot confirm the effect it has on the Beijing lineage.

Compensatory mutations at the RNA polymerase secondary channel and entry tunnel can modulate fitness of *M. tuberculosis*

It has previously been suggested that CMs cluster in the RNAP structure, such as at the interfaces between the subunits or close to the RRDR. Mutations close to the RRDR have been suspected to alter the conformation of the active center, yielding an RNAP more efficient than before [23]. CMs in the interfacial region might alter binding of the sub-

units, without affecting the active center [24].

Fewer CMs are found at the DNA entry tunnel and the RNAP secondary channel and have not been described in detail before. The secondary channel serves as a direct connection from the outside of the protein to the active center and CMs at this location could hence influence diffusion in and out of the RNAP. It has been proposed that molecules entering through this channel could regulate RNAP activity [33]. This could play a role in both resistance to RIF, which may enter the RNAP through the secondary channel, as well as in fitness regeneration.

For the interaction of CMs with the DNA helix, we can come up with a hypothesis explaining the mechanistic effect of a CM. The mutation P1040R (Supplementary Figure S2C) changes the Proline to an Arginine and hence a bulky, neutral sidechain to a positively charged, elongated one. The Arginine could then possibly flip out of its averted position to interact with the negatively charged DNA backbone (Figure 6F).

The fact that the CMs group in certain regions of the protein suggests that each cluster may share the same mechanism of action. To investigate this, it would be useful to do statistical association testing for CMs from the same cluster. The dataset would easily allow an extension of the analysis to achieve this and many further objectives. As such, this project demonstrates the power of large genomic datasets, especially in combination with phenotypic data.

Materials and Methods

Figures and tables in this report can be reproduced using the jupyter Notebook "Recreate_figures.ipynb" available in the following GitHub repository (https://github.com/fowler-lab/tb-rnap-compensation.git)

Dataset sources

The dataset consists of 66.824 partially sequenced samples of *M. tuberculosis*, collected as a joint effort in multiple projects. The information on mutations present in different samples is available in the CRyPTIC data tables, assembled as part of the CRyPTIC project *M. tuberculosis* [26]. datasets also contain meta data, such as predicted resistances to various antibiotics, lineage association and the amount of growth in 96-well plates. All meta data can be traced back through a unique sample ID. Data tables are available in the linked GitHub repository in the subfolder "tb_rnap_compensation/tables/".

The 96-well plate growth data was obtained as part of a previous project, using the Automated Mycobacterial Growth Detection Algorithm (AMyGDA) [27]. It scans pictures of all 96-well plates incubated by the CRyPTIC project and quantifies the amount of bacterial growth in the respective wells. For a considerable number of samples from the CRyPTIC project there is hence high-throughput growth data available.

Statistical association testing of RNA Polymerase mutations with resistance mutations

The high-confidence hit list of CMs was obtained by performing pairwise statistical association testing of mutations in the RNA Polymerase genes that co-occur with RIF resistance mutations. The statistical test used was Fisher's exact test, which is one of the most common approach for testing categorical independence of two variables. Fisher's exact test is based on the hyper-geometric distribution and uses the binomial coefficients to calculate exact test statistics, hence it has a long running time. To reduce this, I used a package with a faster implementation of Fisher's test. To further lower the running time, I minimised the number of pair-wise tests that are calculated.

Optimising test parameters for single mutation prevalence and p-value

Several parameters need to be optimised to carry out the test statistics computations in reasonable time. WIth over 60.ooo samples, calculating Fisher's p-value for every single combination of resistance mutation and other mutation is computationally too expensive to carry out. I hence decided to set a cut-off for the prevalence of individual mutations within the dataset (Figure 3A) and optimised the significance cut-off for the p-value (Figure 3B).

To evaluate the performance of the tests with given cut-offs, I compiled a list of known, high-confidence putative CMs from reference studies (available as "Ref_CMs.xlsx" on the github). These are putative CMs that have been confirmed either experimentally or by thorough statistical investigation. The reference serves as a means to calculate the true positive rate (TPR), i.e. the percentage of previously described high-confidence CMs found with the current cut-offs. As a second measure of performance, I employed the number of detected hits in general, since I want to keep this reasonably low. This makes the approach more conservative and gives higher confidence in the resulting list of putative CMs.

Since there is a sharp drop in the number of hits re-

covered at a prevalence cut-off of 45 (Figure 3A), I decided to use this number for our downstream analysis. The drop indicates that there are many low prevalence mutations showing up as hits, but they are of minor interest. Low prevalence hits hint at an inflation of statistical significance due to low sample size (n) and even if they are a true positive, the respective CM does not appear to play an overall major role in restoring fitness.

After setting the single mutation prevalence cut-off to 45, I examined the effect of the significance threshold on the number of hits. For good practice, I chose a very conservative cut-off with p = 0.01 and Bonferroni correction for multiple testing. This ensures that the number of false positives and the number of hits above this threshold is reasonably low (Figure 3B). The resulting preliminary hit list with putative CMs and their respective associated resistance mutation is shown in Supplementary Table S1.

Optimising the threshold for mutation co-occurrence prevalence

Since we are most interested in highly relevant CMs, I applied another cut-off to our preliminary hit list. I tested the sensitivity and number of hits for different cut-offs of co-occurrence prevalence, which is the prevalence of the specific combination of resistance and putative compensatory mutation in our overall dataset. If this value is very low, the correlation is either a false positive or the resistance associated mutation does not appear to be an evolutionary highly relevant CM. I decided on a cut-off of 40, since this preserved 93 % of the previously found reference CMs. With a prevalence cut-off higher than 40, sensitivity decreased significantly (Supplementary Figure S1, blue graph). Filtering by the co-occurrence prevalence cut-off yielded a hit list of putative CMs that is amenable to growth analysis (Supplementary Table S2).

Plotting of growth data

The final list of putative, high-confidence CMs can now be used to classify samples as resistant with, and resistant without CMs (Supplementary Table S2). The corresponding unique sample IDs were then employed to extract the average sample growth from the 96-well plate measurements of the AMyGDA project. The growth of susceptible samples was obtained in a similar way and the three growth distributions were analysed as to their characteristic values. The characteristics were the median growth, as an indicator of overall fitness of the sample genotype, and the difference between the distributions. To quantify the difference between growth of samples that are susceptible, resistant or resistant with CMs, the pairwise Mann-Whitney p-value was calculated. If we obtain a significant p-value, the medians of the two tested categories differ significantly, i.e. the two tested sample categories are considered to originate from different distributions.

Mapping of putative compensatory mutations on RNA Polymerase structure

The crystal structure of the *M. tuberculosis* RNA polymerase in complex with Rifampicin was obtained from the protein data bank (5UHB). The open-source molecular visualisation software PyMOL was used to display the protein structure and annotate different subunits. High-confidence hits were mapped onto the structure to elucidate potential compensatory mechanisms. The numbering of the amino acids on the β (rpoB) subunit is shifted by 6 amino acids in the pdb file, so 6 must be added to the rpoB amino acids referred to in the text to obtain their label in the corresponding figure.

References

[1] Hiroshi Nikaido. Multidrug Resistance in Bacteria. *Annual Review of Biochemistry*, 78(1):119–146, 2009.

[2] Howard S. Gold and Robert C. Moellering. Antimicrobial-Drug Resistance. *New England Journal of Medicine*, 335(19):1445–1453, November 1996.

[3] Peter A. Smith and Floyd E. Romesberg. Combating bacteria and drug resistance by inhibiting mechanisms of persistence and adaptation. *Nature Chemical Biology*, 3(9):549–556, September 2007.

[4] Evelina Tacconelli, Elena Carrara, Alessia Savoldi, Stephan Harbarth, and et al. Discovery, research, and development of new antibiotics: the WHO priority list of antibiotic-resistant bacteria and tuberculosis. *The Lancet Infectious Diseases*, 18(3):318–327, March 2018.

[5] Pedro Eduardo Almeida Da Silva and Juan Carlos Palomino. Molecular basis and mechanisms of drug resistance in Mycobacterium tuberculosis: classical and new drugs. *Journal of Antimicrobial Chemotherapy*, 66(7):1417–1430, July 2011.

[6] E. Cambau and M. Drancourt. Steps towards the discovery of Mycobacterium tuberculosis by Robert Koch, 1882. *Clinical Microbiology and Infection*, 20(3):196–201, March 2014.

[7] Geneva: World Health Organization. Global tuberculosis report, 2021. Licence: CC BY-NC-SA 3.0 IGO.

[8] E. D. Chan and M. D. Iseman. Multidrug-resistant and extensively drug-resistant tuberculosis: a review. *Current Opinion in Infectious Diseases*, 21(6):587–595, 2008.

[9] Matthias Merker, Camille Blin, Stefano Mona, Nicolas Duforet-Frebourg, and et al. Evolutionary history and global spread of the Mycobacterium tuberculosis Beijing lineage. *Nature Genetics*, 47(3):242–249, March 2015. Number: 3 Publisher: Nature Publishing Group.

[10] Paul A. Aristoff, George A. Garcia, Paul D. Kirchhoff, and H. D. Hollis Showalter. Rifamycins – Obstacles and opportunities. *Tuberculosis*, 90(2):94–118, March 2010.

[11] David M. Rothstein. Rifamycins, Alone and in Combination. *Cold Spring Harbor Perspectives in Medicine*, 6(7):2157–1422, July 2016.

[12] Wei Lin, Soma Mandal, David Degen, Yu Liu, and et al. Structural Basis of Mycobacterium tuberculosis Transcription and Transcription Inhibition. *Molecular Cell*, 66(2):169–179, April 2017.

[13] Sebastian M. Gygli, Sonia Borrell, Andrej Trauner, and Sebastien Gagneux. Antimicrobial resistance in Mycobacterium tuberculosis: mechanistic and evolutionary perspectives. *FEMS Microbiology Reviews*, 41(3):354–373, May 2017.

[14] Lilly K. W. Yuen, David Leslie, and Peter J. Coloe. Bacteriological and Molecular Analysis of Rifampin-Resistant Mycobacterium tuberculosis Strains Isolated in Australia. *Journal of Clinical Microbiology*, 37(12):3844–3850, December 1999.

[15] O. J. Billington, T. D. McHugh, and S. H. Gillespie. Physiological Cost of Rifampin Resistance Induced In Vitro in Mycobacterium tuberculosis. *Antimicrobial Agents and Chemotherapy*, 43(8):1866–1869, August 1999.

[16] Sebastien Gagneux, Clara Davis Long, Peter M. Small, Tran Van, and et al. The Competitive Cost of Antibiotic Resistance in Mycobacterium tuberculosis. *Science*, 312(5782):1944–1946, June 2006.

[17] Amel Kevin Alame Emane, Xujun Guo, Howard E. Takiff, and Shengyuan Liu. Drug resistance, fitness and compensatory mutations in Mycobacterium tuberculosis. *Tuberculosis*, 129:102091, July 2021.

[18] Iñaki Comas, Sonia Borrell, Andreas Roetzer, Graham Rose, and et al. Whole-genome sequencing of rifampicin-resistant Mycobacterium tuberculosis strains identifies compensatory mutations in RNA polymerase genes. *Nature Genetics*, 44(1):106–110, January 2012.

[19] M. de Vos, B. Müller, S. Borrell, P. A. Black, and et al. Putative Compensatory Mutations in the rpoC Gene of Rifampin-Resistant Mycobacterium tuberculosis Are Associated with Ongoing Transmission. *Antimicrobial Agents and Chemotherapy*, 57(2):827–832, February 2013.

[20] Qin-jing Li, Wei-wei Jiao, Qing-qin Yin, Fang Xu, and 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(5):2807–2812, 2016.

[21] Nicola Casali, Vladyslav Nikolayevskyy, Yanina Balabanova, Olga Ignatyeva, and et al. Microevolution of extensively drug-resistant tuberculosis in Russia. *Genome Research*, 22(4):735–745, April 2012.

[22] Asho Ali, Zahra Hasan, Ruth McNerney, Kim Mallard, and et al. Whole Genome Sequencing Based Characterization of Extensively Drug-Resistant Mycobacterium tuberculosis Isolates from Pakistan. *PLOS ONE*, 10(2):e0117771, February 2015.

[23] Pengjiao Ma, Tao Luo, Liang Ge, Zonghai Chen, and et al. Compensatory effects of M. tuberculosis rpoB mutations outside the rifampicin resistance-determining region. *Emerging Microbes & Infections*, 10(1):743–752, January 2021.

[24] Ana Vargas, Angela Rios, Louis Grandjean, Daniela Kirwan, and et al. Determination of potentially novel compensatory mutations in rpoC associated with rifampin resistance and rpoB mutations in Mycobacterium tuberculosis clinical isolates from Peru. *International Journal of Mycobacteriology*, 9(2):121–137, 2020.

[25] Taeksun Song, Yumi Park, Isdore Chola Shamputa, Sunghwa Seo, and et al. Fitness costs of rifampicin resistance in Mycobacterium tuberculosis are amplified under conditions of nutrient starvation and compensated by mutation in the β subunit of RNA polymerase. *Molecular Microbiology*, 91(6):1106–1119, 2014.

[26] CRyPTIC Consortium and the 100,000 Genomes Project. Prediction of Susceptibility to First-Line Tuberculosis Drugs by DNA Sequencing. *The New England Journal of Medicine*, 379(15):1403–1415, October 2018.

[27] Philip W. Fowler, Ana Luíza Gibertoni Cruz, Sarah J. Hoosdally, Lisa Jarrett, Emanuele Borroni, Matteo Chiacchiaretta, Priti Rathod, Sarah Lehmann, Nikolay Molodtsov, Timothy M. Walker, Esther Robinson, Harald Hoffmann, Timothy E. A. Peto, Daniela Maria Cirillo, Grace E. Smith, and Derrick W.YR 2018 Crook. Automated detection of bacterial growth on 96-well plates for high-throughput drug susceptibility testing of Mycobacterium tuberculosis. *Microbiology*, 164(12):1522–1530, 2018. Publisher: Microbiology Society,.

[28] Zhihong Xu, Aiping Zhou, Jiawei Wu, Aiwu Zhou, and et al. Transcriptional Approach for Decoding the Mechanism of rpoC Compensatory Mutations for the Fitness Cost in Rifampicin-Resistant Mycobacterium tuberculosis. *Frontiers in Microbiology*, 9, 2018.

[29] Mary G Reynolds. Compensatory Evolution in Rifampin-Resistant Escherichia coli. *Genetics*, 156(4):1471–1481, December 2000.

[30] Gerrit Brandis and Diarmaid Hughes. Genetic characterization of compensatory evolution in strains carrying rpoB Ser531Leu, the rifampicin resistance mutation most frequently found in clinical isolates. *Journal of Antimicrobial Chemotherapy*, 68(11):2493–2497, November 2013.

[31] Lai Lai San, Khin Saw Aye, Nan Aye Thida Oo, Mu Mu Shwe, and et al. Insight into multidrug-resistant Beijing genotype Mycobacterium tuberculosis isolates in Myanmar. *International Journal of Infectious Diseases*, 76:109–119, November 2018.

[32] Matthias Merker, Thomas A. Kohl, Andreas Roetzer, Leona Truebe, and et al. Whole Genome Sequencing Reveals Complex Evolution Patterns of Multidrug-Resistant Mycobacterium tuberculosis Beijing Strains in Patients. *PLOS ONE*, 8(12):e82551, December 2013.

[33] Bryce E. Nickels and Ann Hochschild. Regulation of RNA Polymerase through the Secondary Channel. *Cell*, 118(3):281–284, August 2004.

1 Supplementary Materials

resistance mutation	putative CM
rpoB S450L	*rpoC* N826T
rpoB S450L	*rpoC* D485Y
rpoB S450L	*rpoB* A692T
rpoB S450L	*rpoB* I480V
rpoB S450L	*rpoC* H525Q
rpoB S450L	*rpoA* V183G
rpoB S450L	*rpoC* E1092D
rpoB L430P	*rpoC* E1092D
rpoB V170F	*rpoC* E1092D
rpoB S450L	*rpoB* A286V
rpoB S450L	*rpoB* V496A
rpoB S450L	*rpoC* T812I
rpoB S450L	*rpoC* G519D
rpoB S450W	*rpoC* G519D
rpoB D435Y	*rpoB* c-61t
rpoB D435G	*rpoB* I1106T
rpoB L452P	*rpoB* I1106T
rpoB S450L	*rpoC* W484G
rpoB V170F	*rpoC* W484G
rpoB S450L	*rpoC* P1040S
rpoB S450L	*rpoB* Q975H
rpoB S450L	*rpoC* P481T
rpoB D435V	*rpoC* P481T
rpoB S450L	*rpoC* L547V
rpoB S450L	*rpoB* A405P
rpoB S450L	*rpoC* R741S
rpoB H445D	*rpoC* R741S
rpoB S450L	*rpoB* L731P
rpoB S450L	*rpoC* L449V
rpoB S450L	*rpoC* V1039A
rpoB S450L	*rpoC* F452S
rpoB S450L	*rpoB* I488V
rpoB S450L	*rpoC* P1040R
rpoB S450L	*rpoB* E761D
rpoB S450L	*rpoC* V431M
rpoB S450L	*rpoC* A521D
rpoB S450W	*sigA* D146E
rpoB S450L	*rpoA* D190G
rpoB S450L	*rpoC* L507V

resistance mutation	putative CM
rpoB S450L	*rpoC* G332S
rpoB S450L	*rpoB* K891E
rpoB S450L	*rpoC* G332R
rpoB S450L	*rpoC* V1252L
rpoB S450L	*rpoB* V695L
rpoB D435Y	*rpoB* V695L
rpoB S450L	*rpoA* -40 indel
rpoB S441L	*rpoC* G1198S
rpoB S450L	*rpoC* V517L
rpoB S450L	*rpoA* A180V
rpoB S450L	*rpoC* V1252M
rpoB S450L	*rpoB* P45S
rpoB S450L	*sigA* 247 indel
rpoB S450L	*rpoC* K445R
rpoB I491F	*rpoC* E1033A
rpoB S450L	*rpoC* F452C
rpoB S450L	*rpoB* V403M
rpoB S450L	*rpoC* N416S
rpoB S450L	*rpoC* K1152Q
rpoB S450L	*rpoB* Q409R
rpoB S450W	*rpoB* Q409R
rpoB S450L	*rpoC* L527V
rpoB S450L	*rpoB* R827C
rpoB S450L	*rpoC* P1040A
rpoB S450L	*rpoB* P45L
rpoB S450W	*rpoB* P45L
rpoB S450L	*rpoB* R827L
rpoB S450L	*rpoC* G433S
rpoB S450L	*rpoC* I491T
rpoB S450L	*rpoC* I491V
rpoB S450L	*rpoC* L516P
rpoB S450L	*rpoC* P434R
rpoB S450L	*rpoA* T187A
rpoB S450L	*rpoC* D485N
rpoB S450L	*rpoC* V483G
rpoB Q432P	*rpoC* V483G
rpoB S450L	*rpoC* F452L
rpoB S450L	*rpoC* V483A
rpoB S450L	*rpoC* N698K
rpoB S450L	*rpoC* N698S

Table S1: Preliminary hit list of Fisher's exact test for association of resistance with co-occuring mutations. The cut-off for single mutation prevalence in the dataset is 45 and the p-value p = 0.01 with Bonferroni correction for multiple testing. Synonymous and lineage-defining mutations were removed.

resistance mutation	putative CM	prevalence
rpoB S450L	*rpoC* N826T	64
rpoB S450L	*rpoC* D485Y	194
rpoB S450L	*rpoB* A692T	79
rpoB S450L	*rpoB* I480V	78
rpoB S450L	*rpoC* H525Q	46
rpoB S450L	*rpoA* V183G	77
rpoB S450L	*rpoC* E1092D	1989
rpoB S450L	*rpoB* A286V	56
rpoB S450L	*rpoB* V496A	60
rpoB S450L	*rpoC* T812I	51
rpoB S450L	*rpoC* G519D	51
rpoB S450L	*rpoB* c-61t	737
rpoB S450L	*rpoC* W484G	80
rpoB S450L	*rpoC* P1040S	118
rpoB S450L	*rpoB* Q975H	65
rpoB S450L	*rpoC* P481T	69
rpoB S450L	*rpoC* L547V	80
rpoB S450L	*rpoB* A405P	48
rpoB S450L	*rpoB* L731P	226
rpoB S450L	*rpoC* L449V	50
rpoB S450L	*rpoC* V1039A	68
rpoB S450L	*rpoC* F452S	345
rpoB S450L	*rpoB* I488V	59
rpoB S450L	*rpoC* P1040R	396
rpoB S450L	*rpoB* E761D	304
rpoB S450L	*rpoC* V431M	62
rpoB S450L	*rpoC* A521D	65
rpoB S450L	*sigA* D146E	50
rpoB S450L	*rpoA* D190G	46
rpoB S450L	*rpoC* L507V	64
rpoB S450L	*rpoC* G332S	179
rpoB S450L	*rpoB* K891E	74
rpoB S450L	*rpoC* G332R	113
rpoB S450L	*rpoC* V1252L	175
rpoB S450L	*rpoB* V695L	81
rpoB S450L	*rpoA* -40 indel	64

resistance mutation	putative CM	prevalence
rpoB S450L	*rpoC* V517L	184
rpoB S450L	*rpoA* A180V	48
rpoB S450L	*rpoC* V1252M	59
rpoB S450L	*rpoB* P45S	65
rpoB S450L	*sigA* 247 indel	42
rpoB S450L	*rpoC* K445R	98
rpoB S450L	*rpoC* F452C	48
rpoB S450L	*rpoC* N416S	72
rpoB S450L	*rpoC* K1152Q	51
rpoB S450L	*rpoB* Q409R	69
rpoB S450L	*rpoC* L527V	113
rpoB S450L	*rpoB* R827C	123
rpoB S450L	*rpoC* P1040A	110
rpoB S450L	*rpoB* P45L	43
rpoB S450L	*rpoB* R827L	50
rpoB S450L	*rpoC* G433S	141
rpoB S450L	*rpoC* I491T	457
rpoB S450L	*rpoC* I491V	665
rpoB S450L	*rpoC* L516P	144
rpoB S450L	*rpoC* P434R	44
rpoB S450L	*rpoA* T187A	171
rpoB S450L	*rpoC* D485N	166
rpoB S450L	*rpoC* V483G	1206
rpoB S450L	*rpoC* F452L	96
rpoB S450L	*rpoC* V483A	586
rpoB S450L	*rpoC* N698K	48
rpoB S450L	*rpoC* N698S	205
rpoB L430P	*rpoC* E1092D	81
rpoB S450W	*rpoB* c-61t	50
rpoB D435Y	*rpoB* c-61t	165
rpoB D435G	*rpoB* I1106T	103
rpoB L452P	*rpoC* E1092D	48
rpoB L452P	*rpoB* I1106T	103
rpoB D435V	*rpoC* E1092D	60
rpoB D435V	*rpoB* c-61t	104
rpoB I491F	*rpoC* E1033A	46

Table S2: Hit list of Fisher's exact test for association of resistance with co-occuring mutations after applying a filter for combined prevalence of resistance and CM. The cut-off for combined mutation prevalence in the dataset is 40.

Lineage	median growth [%]	p-value to 1	p-value to 2	p-value to 3	p-value to 4
1	23.090	-	0.0717	8.68e-14	1.21e-05
2	26.080	0.0717	-	8.36e-16	7.98e-33
3	32.095	8.68e-14	8.36e-16	-	3.31e-64
4	18.065	1.22e-05	7.99e-33	3.31e-64	-

Table S3: Median growth of pan-susceptible samples from different *M. tuberculosis* lineages. Mann-Whitney p-values are calculated in reference to all other lineage growth distributions.

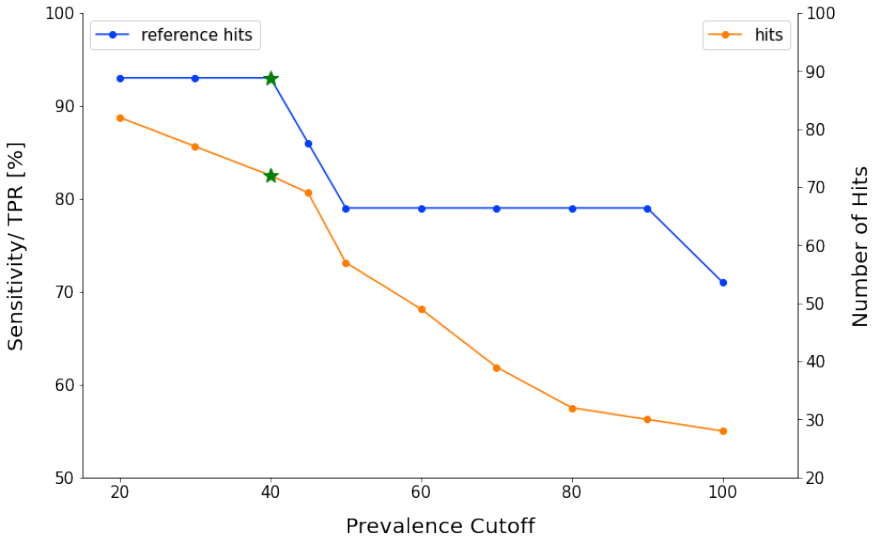

Figure S1: Sensitivity and number of significant hits (putative CMs) depending on the cut-off for prevalence of the co-occurring resistance mutations and CMs. Influence of prevalence cut-off for co-occurrence on sensitivity and number of hits. On the left y-axis, the percentage of found reference hits, also termed sensitivity or true positive rate (TPR), is plotted for different choices of prevalence cutoff. The dark blue graph represents the percentage of retained reference hits that appeared on the preliminary hit list. The right y-axis refers to the orange graph, which plots the number of mutations that were classified as significantly resistance associated under the chosen cut-off. Green stars indicate the final cut-off choice and is a consensus of considering the previously mentioned variables: sensitivity and number of detected hits in general.

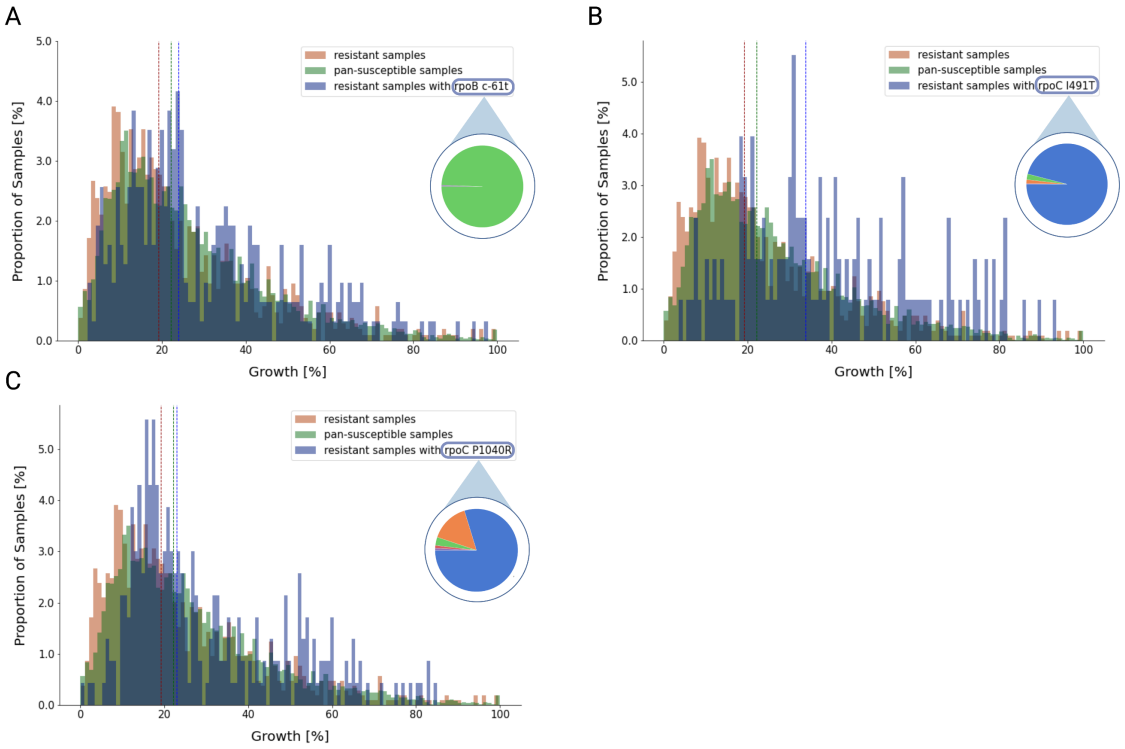

Figure S2: Growth and lineage distributions of *M. tuberculosis* samples that are pan-susceptible, rifampicin (RIF) resistant and resistant with compensatory mutations (CMs) (A-C) Distribution of growth in percent as measured in the CRyPTIC project, [27] when plotted against the proportion of samples that display this amount of growth. Samples with RIF resistance mutations but no other potentially interfering mutations are plotted in red, samples that were classified as pan-susceptible are plotted in green. Samples that have a RIF resistance mutations and a specific associated CM (indicated in the legend) are shown in blue. Medians and Mann-Whitney p-values of the different distributions are shown in Supplementary Table S3.

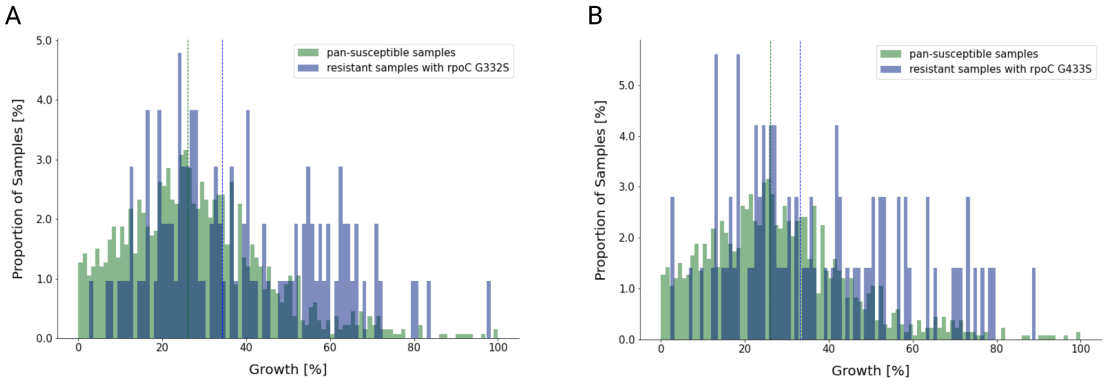

Figure S3: Growth distributions of *M. tuberculosis* samples within different lineages (A-B) Distribution of growth in percent as measured in the CRyPTIC project [27], plotted against the proportion of samples that display this amount of growth. Samples that were classified as pan-susceptible are plotted in green and the subset of samples with RIF resistance mutations and the CM indicated in the legend and no other putative CMs is shown in blue. Medians and Mann-Whitney p-values of the different distributions are shown in Supplementary Table S5.

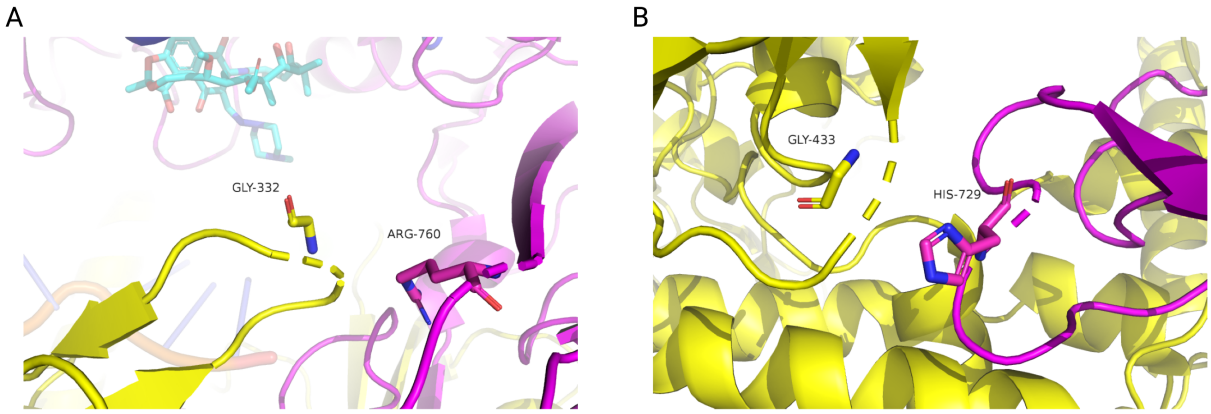

Figure S4: Location of two high-confidence compensatory mutations (CMs) on the RNA Polymerase (RNAP) (A) The CM G332S is located on the β' subunit, in a contact region to the β subunit (magenta). The change from Glycine (stick representation) to Serine (negatively charged side chain) might enable an interaction with the close-by Arginine (positively charged side chain) on the β subunit. The bound drug RIF (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.