HOMOPLASY IN BACTERIAL EVOLUTION

A Dissertation

by

YI-PIN LAI

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Thomas R. Ioerger
Committee Members,	James J. Cai
	Jyh-Charn (Steve) Liu
	Sing-Hoi Sze
Head of Department,	Scott D. Schaefer

May 2020

Major Subject: Computer Science

ABSTRACT

The appearance of homoplasy occurs when mutations are not derived from a common ancestor but arise independently in multiple branches of a phylogenetic tree. For bacteria, it suggests that genetic recombination events occur or positive selection exists during evolution, affecting the accuracy of phylogeny estimation. Without considering recombination, the reconstruction of phylogenetic trees based on an alignment of bacterial strains could be misleading. Hence, to better understand their true evolutionary histories among a bacterial population, it is essential to identify recombination breakpoints before estimating their phylogeny.

We developed an average compatibility ratio method with a permutation test, ptACR, to detect recombination breakpoints in a multiple sequence alignment without requiring a tree. We use a sliding window to evaluate the local compatibility of adjacent polymorphic sites to locate potential breakpoints and then assess the statistical significance of candidate breakpoints by applying a permutation test. We evaluate the performance of ptACR on both simulated and empirical datasets. The simulation results show that it has similar sensitivity but higher specificity and better F1 score compared to existing methods. Also, ptACR detects recombination events in a collection of clinical isolates of *Mycobacterium avium* and *Staphylococcus aureus*, and identifies boundaries of regions with statistical significance, where the adjacent regions exhibit distinct phylogenies.

For clonal species, since recombination is less likely to occur, the occurrence of homoplasy is a strong indicator of positive selection, such as antibiotic resistance. To identify mutations conferring resistance, genome-wide association studies are commonly applied to identify statistically significant associations between genotypes (polymorphisms) and phenotypes of interests (antibiotic resistance) across the entire genome. However, homoplasy is not well accounted for by most bacterial genome-wide association analyses, producing false positives or false negatives. Also, existing association methods usually use an individual site or group polymorphisms within a gene as genotypes without considering the frequency of evolutionary convergence and the mutation rate in different regions.

To better exploit homoplasy, we developed a two-phase evolutionary cluster-based convergence test (ECC) to identify regions harboring mutations under selection pressure associated with antibiotic resistance. In the first-phase step, we apply a Poisson distribution to detect regions exhibiting more changes (distinct mutational events) than expected by optimizing the grouping of SNPs within windows. Next, we test associations between the clustered regions and drug resistance using a hypergeometric distribution based on the concept of convergence test in the second phase. We model the distribution of changes occurring in the resistant or sensitive branches for each clustered region and compare it to the background. We evaluate the ECC method on empirical datasets of clinical isolates of *Mycobacterium tuberculosis* with seven phenotypes from drug susceptibility tests. Our two-phase evolutionary cluster-based convergence method is able to identify known resistant-associated sites within genes or intergenic regions corresponding to seven anti-tuberculous drugs. It also identifies two novel clustered regions in Rv2571 and Rv1830, potentially linked to isoniazid resistance. It improves the potential over existing methods for association tests to find more novel resistant-associated mutations, which will ultimately help in developing new antibiotic treatments.

In sum, we present two models for identifying genomic regions affected by recombination (ptACR) and clustered regions associated with antibiotic resistance driven by selection pressure (ECC) in bacterial genomes.

DEDICATION

To my mother, father, partner, family and friends.

ACKNOWLEDGMENTS

I would first like to thank my advisor, Dr. Thomas R. Ioerger, for his continual guidance, invaluable insights and endless support throughout my studies. His hardworking attitude, extensive knowledge in bioinformatics, and impressive research work in the fields of infectious diseases and bacterial genomics have inspired me to become a better scientist and keep sharpening my skills. I am so honored to have many opportunities to work with him. I sincerely thank my committee members, Dr. James J. Cai, Dr. Jyh-Charn (Steve) Liu, and Dr. Sing-Hoi Sze, for their insightful advice and great support.

I am also thankful for my labmates, classmates, and friends, Michael A. DeJesus, Eric Nelson, Ivan Fuentes, Siddharth Subramaniyam, Esha Dutta, Sanjeevani Choudhery, Katrina Wu, Donny Chung, Szu-Ting Kuo, Yu-Ya Liang, En-Tzu Lee, Hsin-Yi Li, Shen-Yu Hu, Sarah Yeh, Jason Lin, Jasmine Cheng, Jay Chou, Shu-Hao Yeh, Sophie Hsu, Kathy Pai, and all the members in TSA badminton team, for working together and making graduate school fun. Additionally, I am so grateful for my best friend, Ching-Hua Wang, for her unwavering support through thick and thin.

Lastly, I would like to thank my family and in-laws for their lasting support. Particularly, I am grateful for my mother for developing my courage, my father for setting the bar high, my uncle for motivating me to study science and engineering, and finally my partner, Hsin-Hung Huang, for always being there for me.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supervised and supported by a dissertation committee consisting of Professors Thomas R. Ioerger, Jyh-Charn (Steve) Liu, and Sing-Hoi Sze of the Department of Computer Science and Engineering, and Professor James J. Cai of the Department of Veterinary Integrative Biosciences.

All bioinformatics analyses and interpretation were carried out by the student and her advisor.

Funding Sources

Graduate study was supported by a graduate research assistantship in the Department of Computer Science and Engineering at Texas A&M University. Funding for this research was provided in part by an NIH CETR grant (NIAID U19 AI109755) from the National Institutes of Health.

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION *

1.1 Motivation

In a phylogeny, the appearance of homoplasy occurs when mutations/polymorphisms are not from a common ancestor but arise independently in multiple branches. Homoplasy occurs due to evolution with recombination and recurrent mutations driven by selection pressures [2]. Estimating a phylogeny accurately helps to intepret the evolutionary history of bacterial species. Bacteria are prokaryotes which have a single set of chromosomes, i.e., haploid. The evolution of bacterial species is influenced by the extent of clonality varying between vertical inheritances and horizontal transfers. During evolution, some bacteria tend to reproduce clonally by replicating DNA through cell division with a few random point mutations. Conversely, some become divergent by exchanging DNA through recombination [3, 4]. Growing evidence has shown that several bacteria exhibit homoplasy in their genomes, including *Mycobacterium avium* [5], *Mycobacterium intracellulare* [6], *Neisseria meningitidis* [7, 8], *Salmonella enterica* [9], *Staphylococcus aureus* [10, 11, 12], *Streptococcus pneumoniae* [13] and *Streptococcus pyogenes* [14]. For strains exhibiting recombinant genomes, the inferred phylogenetic tree may be misleading since some polymorphisms are incongruent with a single tree [15]. Hence, it is essential to identify recombination breakpoints to obtain local regions of distinct phylogenies. We will describe an approach (ptACR) based on incompatibility and a permutation test for finding boundaries of recombination regions. It is more efficient than other computational approaches. This will help studies of bacterial species where recombination is prevalent.

For some pathogens, their evolution processes are believed to be highly clonal across time, meaning that most genetic materials descend vertically through cell division. However, they har-

*Part of the data reported in this chapter is reprinted with permission from "A statistical method to identify recombination in bacterial genomes based on SNP incompatibility" by Y.-P. Lai and T. R. Ioerger, 2018. *BMC Bioinformatics*, 19, 450, Copyright [2018] by BioMed Central. DOI:10.1186/s12859-018-2456-z.
Part of the data reported in this chapter is reprinted with permission from "A compatibility approach to identify recombination breakpoints in bacterial and viral genomes" by Y.-P. Lai and T. R. Ioerger, 2017. *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pp. 11-20, Copyright [2017] by Association for Computing Machinery. DOI:10.1145/3107411.3107432.

bor some mutations occurring in more than one branch in the tree, i.e. homoplasy. Homoplasy occurs when mutations do not evolve randomly during DNA replication, suggesting positive selection pressure. For example, *Mycobacterium tuberculosis* is thought to be highly clonal in general, but it has acquired homoplasic mutations driven by the emergence of antibiotic resistance [16]. The occurrence of homoplasy is a strong indicator of selection pressures in clonal species, yet it is not exploited in current genome-wide association studies (GWAS). GWAS is developed to statistically find genotypes associated with phenotypes of interest in whole genomes. Humans are diploid eukaryotes while bacteria are haploid prokaryotes. Commonly used methods in human GWAS cannot be applied directly to bacterial association mappings without considering confounders of population stratification, linkage disequilibrium and homoplasy [17, 18]. In addition, the genotypes used in an association test are usually an individual polymorphic site or a grouping of sites within a single gene. However, the known resistant-associated variants vary in groupings of sites (clusters) under different phenotypes. Furthermore, co-resistance may exist, resulting in ambiguous associations. Studies have shown that isoniazid-resistant strains have a higher propensity to have resistant mutations to rifampicin in *M. tuberculosis*, i.e., multidrug-resistant strains [19]. Therefore, in a dataset exhibiting co-resistance, the identified polymorphisms associated with a particular drug may be confounded by another drug, resulting in ambiguous associations. We show that optimizing the grouping of SNPs can enhance the statistical significance. However, this must be done efficiently, to avoid complexity of testing too many windows. Hence, we develop a two-phase evolutionary cluster-based convergence (ECC) approach to test associations between genotypes as clustered regions against phenotypes of interest. The clustering gives a benefit to homoplasic sites because they are often in clusters and hence get tested for significance. Our approach considers the effects of homoplasy and population stratification using a Poisson distribution and a hypergeometric model along with a reconstructed phylogenetic tree. We evaluate our method in empirical datasets of *M. tuberculosis*. It is not only able to identify known resistant-associated loci but identify novel loci potentially linked to antibiotic resistance. It helps to increase the power of bacterial association tests to determine novel causal variants responsible for drug resistance.

In sum, we develop algorithms to characterize homoplasy in bacteria from two aspects: the detection of recombination breakpoints in recombinant genomes and the identification of polymorphisms associated with antibiotic resistance in clonal genomes considering homoplasy.

2. RECOMBINATION IN BACTERIAL GENOMES *

2.1 Background

Recombination is an important force of evolution in prokaryotes that results in mosaic genomes through exchanging genetic materials between strains [20]. In bacterial populations, when some strains acquire genetic changes from other strains, it can produce the appearance of homoplasy (where the same change at a site appears to have occurred multiple times independently, in separate branches). In a multiple sequence alignment, the polymorphic sites may have different phylogenetic relationships compared with other sites, i.e., phylogenetic incongruence [2, 15]. Studies have explored the effect of recombination in phylogeny estimation and indicated that the impact depends on the extent of recombinant events and the relatedness of taxa [20, 21, 22]. The true evolutionary history of a set of taxa may not be reflected if recombination events occurred during evolution yet are ignored. Growing evidence indicates that recombination has occurred in the evolution of many pathogenic bacterial species, including *Mycobacterium avium* [5], *Mycobacterium intracellulare* [6], *Neisseria meningitidis* [7, 8], *Salmonella enterica* [9], *Staphylococcus aureus* [10, 11, 12], *Streptococcus pneumoniae* [13] and *Streptococcus pyogenes* [14]. Hence, it is essential to identify recombination regions among bacterial isolates before inferring a phylogeny, to better understand their evolutionary histories.

Over the last four decades, many methods have been proposed to detect the presence of recombination in bacterial genomes, applying concepts of maximum likelihood, phylogenetic incongruence, substitution patterns, distance-based approach, or character compatibility [23, 24, 25, 26, 27, 28]. Commonly used methods to identify recombination breakpoints include ClonalFrameML [26], RDP [27] and GARD [28]. All are phylogenetic-based programs. ClonalFrameML uti-

*Reprinted with permission from "A statistical method to identify recombination in bacterial genomes based on SNP incompatibility" by Y.-P. Lai and T. R. Ioerger, 2018. *BMC Bioinformatics*, 19, 450, Copyright [2018] by BioMed Central. DOI:10.1186/s12859-018-2456-z.
Part of the data reported in this chapter is reprinted with permission from "A compatibility approach to identify recombination breakpoints in bacterial and viral genomes" by Y.-P. Lai and T. R. Ioerger, 2017. *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pp. 11-20, Copyright [2017] by Association for Computing Machinery. DOI:10.1145/3107411.3107432.

lizes a maximum-likelihood tree to reconstruct ancestral states of internal nodes. It then applies a hidden Markov model (ClonalFrame) to infer the recombination parameters and recombination locations of each branch of the tree using an Expectation-Maximization (EM) algorithm [26]. RDP characterizes homoplasy signals using pairwise scanning of the alignment, with the integration of several non-parametric recombination detection methods [27]. GARD applies Akaike's Information Criterion with a genetic algorithm to search the recombinant locations heuristically [28]. Compatibility-based methods are considered to be more efficient than phylogenetic-based methods to identify recombination, since they do not require the reconstruction of phylogenetic trees [23]. The Reticulate program uses compatibility matrices to calculate neighbor similarity score (NSS) and clusters compatible sites by randomly shuffling the matrices [24]. Bruen et al. define the pairwise homoplasy index (PHI) in terms of pairwise incompatibility score of each site and its downstream sites in entire alignment globally, and then they obtain the Monte Carlo p-value by permuting the entire alignment, or by computing the cumulative probability under a normal distribution generated from expected mean and variance of the PHI statistic [25]. Both programs are compatibility-based methods and able to detect recombination and report informative sites, but they do not report breakpoints.

We introduce an average compatibility ratio (ACR) method to identify the potential recombination breakpoints in a bacterial genome by analyzing the pattern of SNPs among a collection of isolates [29]. The ACR method detects the presence or absence of recombination by calculating an overall compatibility score among pairs of sites. Next, ACR will scan the entire alignment with a sliding window of fixed size to identify regions where the local compatibility among pairs of sites in the region decreases and reaches a local minimum. However, the local minima that are below a fixed threshold may include false positives. To reduce false positives, we apply a permutation test on the positions of local minima to assess the statistical significance of potential breakpoints in the genome. We also extend the ACR method to test the compatibility of multi-state characters by applying an efficient algorithm based on Buneman's theorem [30]. The performance of ptACR is evaluated on simulated datasets with varying mutation rates and rate heterogeneity among sites.

The sequences are simulated by evolving along distinct trees with changes in topology, where a group of taxa have been moved from one branch to another randomly. The simulation results show that the integration of the permutation test has lower false positive rate than basic ACR method. Yet both methods have a similar level of sensitivity for the detection of recombination breakpoints. We use ptACR [31] to identify genomic regions of recombination in clinical isolates of *Mycobacterium tuberculosis*, *Mycobacterium avium* and *Staphylococcus aureus*.

2.2 Methods

2.2.1 Characters and Compatibility

For a multiple DNA sequence alignment, a character is defined as a set of states (nucleotides) for all taxa at a given site. The definitions of pairwise compatibility for binary characters and multi-state characters are given as follows [32].

Definition 1. Pairwise compatibility for binary characters: Two sites of binary characters are compatible if and only if there exists a tree for which each site can be explained by one change.

Definition 2. Pairwise compatibility for multi-state characters: Two sites of multi-state characters are compatible if and only if there exists a tree for which each site can be explained by the number of change that equals to the number of distinct states minus one (the minimum number of changes required for a site with n nucleotides is n-1).

For a pair of binary characters at two sites, the four gamete test is a quick way in polynomial time to determine their compatibility [33]. It converts the state of taxa at each site to 0 and 1, and concatenates the states at two sites for a given taxon as one of the following combinations: {00, 01, 10, 11}. If at most three combinations exist, then the two sites are compatible. For a set of binary characters in an alignment, there exists a perfect phylogeny if all characters are jointly compatible. To determine the compatibility of a pair of multi-state characters (two sites at a time), the problem can be reduced to triangulating colored graphs problem [34] and then solved in polynomial time [30]. Two characters are first converted to a partition intersection graph by the following steps. For each character, the taxa of the same state are denoted as a vertex. An edge between two vertices

is added if the vertices contain the same taxon/taxa to form the partition intersection graph. Next, if their derived partition intersection graph is acyclic, then they are determined to be compatible [30]. The method to determine the compatibility of two characters is illustrated in Algorithm 1.

Algorithm 1 Pairwise compatibility of two multi-state characters

Input: Characters χ_p and χ_q at the site p and site q

Output: *True* if they are jointly compatible and *False* if they are incompatible;

function CHARCOMPAT(χ_p, χ_q)

Collect the sets of taxon/taxa of the same state (nucleotide), where the number of unique states are denoted as r_1 and r_2:

$\chi'_p \leftarrow \{x_i\}, i = 1, ..., r_1$

$\chi'_q \leftarrow \{y_j\}, j = 1, ..., r_2$

Initialize an undirected graph *G* by the adjacency list

Add sets in χ'_p and χ'_q as nodes to *G*

Add an edge between node u and node v by $G(u, v)$ to update the graph G:

for all x_i in χ'_p **do**

for all y_j in χ'_q **do**

if $x_i \cap y_j \neq \emptyset$ **then**

$G \leftarrow G(x_i, y_j)$

end if

end for

end for

Check for cycles in *G* by depth first search (DFS)

return *True* if there is no cycle in *G*, *False* otherwise

end function

2.2.2 Recombination Algorithm Using Compatibility

Given a multiple sequence alignment of n taxa and m informative sites (i.e., with more than one nucleotide among the taxa), at each informative site i, ACR calculates a pairwise compatibility score between all pairs of informative sites within a sliding window of size $2w$ centered on the i^{th} SNP (from i-w to i+w). The pairwise compatibility score is 1 if two characters χ_p and χ_q are compatible; otherwise, the score is 0 (Equation 2.1). Next, it averages the scores of all pairs of sites within the region to obtain the average compatibility ratio, σ_{i_w}, for the region (Equation 2.2).

$$CompatPW_{pq} = \begin{cases} 1, & \text{if characters } \chi_p \text{ and } \chi_q \text{ are compatible} \\ 0, & \text{otherwise} \end{cases} \tag{2.1}$$

$$\sigma_{i_w} = \frac{1}{\binom{2w+1}{2}} \sum_{p=i-w}^{i+w-1} \sum_{q=p+1}^{i+w} CompatPW_{pq} \tag{2.2}$$

The lower the value of the average compatibility ratio (σ_{i_w}), the less jointly compatible the sites in a window are. Hence, a site of local minimum means that sites in the region are least compatible locally, suggesting phylogenetic incongruence between the upstream and downstream regions. Sites with local minima of average compatibility ratio are regarded as potential breakpoints. An example of applying ACR on a recombined alignment of 5200 sites using the window size of 200 is demonstrated in Figure 2.1.

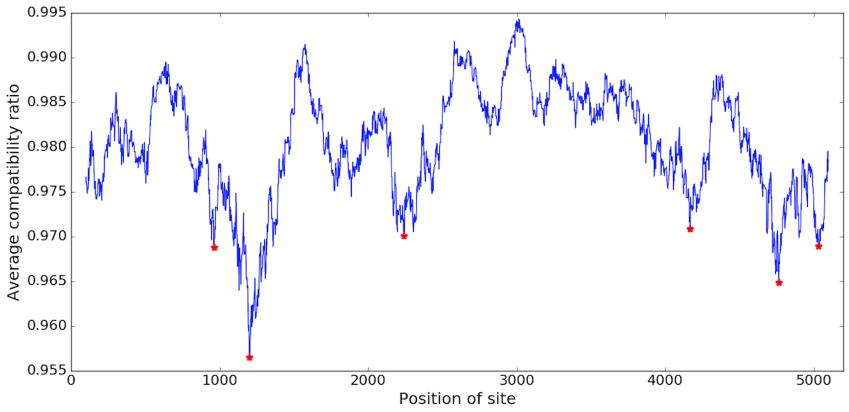


Figure 2.1: Example of applying ACR on an alignment of several recombined regions using the window size of 200. Among 5200 sites, six sites are identified as the potential breakpoints and labeled in red.

2.2.3 Permutation Test for Statistical Significance of Candidate Breakpoints

To assess the statistical significances of potential breakpoints, we apply a permutation test. The test statistic, s_{i_w}, for a potential breakpoint at the site i is defined as the summation of all compatibility scores of pairs composed of a site from the upstream region $[i-w, i-1]$ with the other site from the downstream region $[i+1, i+w]$ (Equation 2.3).

$$s_{i_w} = \sum_{p=i-w}^{i-1} \sum_{q=i+1}^{i+w} CompatPW_{pq} \tag{2.3}$$

This statistic is compared to a null distribution generated by permuting the sites in the window. The null hypothesis is that the level of compatibility between the sites in the window is independent of the sequential order of the sites, i.e., whether sites are compared from upstream or downstream of site i does not matter. The alternative hypothesis is that the order of the sites in the local sequences is crucial and does not happen by chance. So the sites within the region are randomly shuffled multiple times (default: 10,000) to produce the sampling distribution of values s_{i_w} obtained under the null hypothesis. Let the distribution of values from random permutations on sites in the window be denoted by D_s. The significance of observed value s_{i_w} is determined by computing the proportion

of times that the permuted statistics in D_s are less than or equal to the observed value to get the empirical p-value (Equation 2.4).

$$p = P(x \leq s_{i_w} \text{ for } x \in D_s) \tag{2.4}$$

If the p-value is lower than a given threshold (default: 0.05), then it rejects the null hypothesis of no recombination, hence ptACR will report the site as a probable/significant breakpoint. To correct the p-value threshold due to multiple comparisons, we use the Bonferroni correction and set the adjusted p-value cutoff to 0.05/n, where n is the number of local minima identified by ACR, to limit the false discovery rate to at most 5%. An example of a statistic determined as significant in the histogram of a null distribution is illustrated in Figure 2.2. To make the permutation test more efficient, we convert all characters in nucleotides of the alignment to patterns in numbers and make character patterns as a unique set. Then we record pairwise compatibility information among all pairwise patterns in the set in a hash table. Hence, the compatibility information of any two shuffled sites can be looked up in the hash table in constant time.

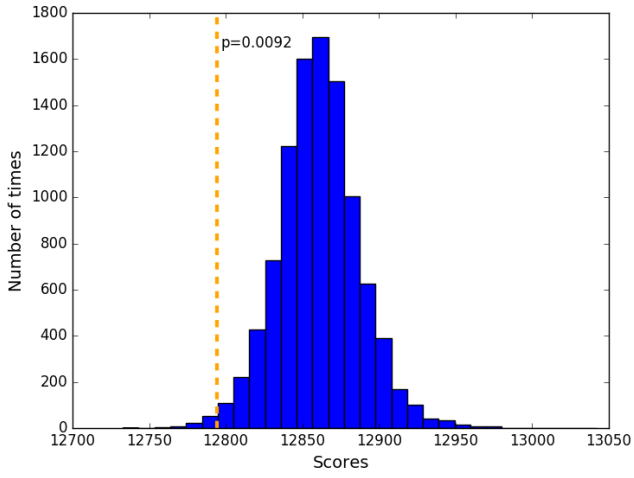


Figure 2.2: Example of the assessment of statistical significance for a compatibility score in the histogram of a null distribution (N=10k). Observed compatibility score at the site i was 12800, among pairs selected upstream and downstream sites. Distribution shows scores from randomly selected pairs in window of $[i - w, i + w]$. The p-value in this case is 0.0092 (at the tail).

2.2.4 Estimation of Phylogenies and Homoplasy

Given a sorted list of candidate breakpoints, local phylogenetic trees of each region between two adjacent breakpoints is constructed by the maximum parsimony method using the function of *dnapars* in PHYLIP 3.66 [35]. To estimate the level of homoplasy for each region, the homoplasy ratio and excess changes is calculated by applying the Sankoff Algorithm [36] on each local tree. The *homoplasy ratio*, which is also called the ratio of changes per site, is defined as the summation of actual state changes (Sankoff score) divided by the summation of minimum number of changes (number of nucleotides at each site minus one). The number of *excess changes* for a site is defined as the difference between the number of actual changes and the minimum number of changes. For a given region, the homoplasy ratio of 1.0 means all sites are congruent (homoplasy-free); a homoplasy ratio > 1.0 means some sites are homoplasic, requiring excess changes in the maximum-parsimony tree.

2.3 Performance on Simulated Datasets

To evaluate the performance of ptACR, we generated simulated sequence data with known recombinations by random branch swaps. Our goal was to evaluate the sensitivity and specificity of detecting known breakpoints, and how this depends on mutation rate and differences in topology. To simulate sequences with predetermined recombination events, a bifurcating tree with 10 taxa is generated by GenPhyloData [37] under a birth-death process with a birth rate of 0.2 and a death rate of 0.1. Next, 300 alternative trees with recombination between a random pair of donor and acceptor branches based on the original tree are obtained using HGT-Gen [38]. Then, Seq-Gen 1.3.4 [39] is applied to generate aligned sequences of 1000 sites evolved along each tree. Parameters for substitution rate and heterogeneity are varied in the experiment, as described below. The sequences are simulated under the Hasegawa-Kishino-Yano model (HKY85) [40] with nucleotide frequencies A:0.2, G:0.3, C:0.3, T:0.2 and 2-to-1 ratio of transitions to transversions. Lastly, we concatenate sequences for the original tree, one of the modified trees, and the original tree again to obtain a simulated alignment with 3000 total sites that has recombination breakpoints around coordinates 1000 and 2000 and a distinct phylogeny in the middle.

The true positive rate (*sensitivity*), false positive rate (1-*specificity*), and F1 score for the ptACR method are defined as follows. For an alignment with a predetermined recombination region, the inferred breakpoint that is located within 50 bp of an actual breakpoint (ground truth) is counted as true positive (TP), and one that is identified by our method but not within this range is denoted as false positive (FP). Failure to detect a known breakpoint at any site within 50 bp is counted as false negative (FN). The true and false positive rates are defined by dividing by the total number of true breakpoints, and the total number of negative sites outside the breakpoint windows, respectively, $\frac{TP}{TP+FN}$ and $\frac{FP}{FP+TN}$. The precision is defined as the number of accurately inferred breakpoints to the number of identified breakpoints, $\frac{TP}{TP+FP}$. The F1 score, which is the harmonic mean of sensitivity and precision, is $\frac{TP}{2TP+FP+FN}$; higher F1 is better. For each scenario, we average the statistics over all the replicates.

2.3.1 Effect of Evolutionary Branch Swapping Distance

Because recombination events among deeper branches should involve strains with more differences and make incompatibility easier to detect, we expect that sensitivity and specificity will be a function of the magnitude of the changes in the simulated trees. To quantify this, we defined an metric called evolutionary branch swapping distance (EBSD) to divide the alternative trees into 3 groups: small, medium, and large evolutionary changes. While there are several generalized methods for comparing topologies of arbitrary labeled trees (sharing the same taxa) [41, 42, 43], assuming that the change between two trees involves only a single branch swap (as generated by HGT-Gen, simulating recombination), we developed a quantitative measure that reflects the magnitude of evolutionary distance involved in the change. First, we identify the group of taxa that changes position in the tree. Call this group A, and let B be the complement in the tree (rest of the taxa). We define the evolutionary branch swapping distance between the two trees (T1 and T2) as the average absolute value of the difference in distances between each pair of taxa i in A and j in B in trees T1 and T2 (Equation 2.5).

$$EBSD(T1, T2) = \frac{1}{|A| * |B|} \sum_{i \in A} \sum_{j \in B} |dist_{T1}(i, j) - dist_{T2}(i, j)| \quad (2.5)$$

The distances (sum of branch lengths on connecting path) between pairs of taxa that are both in A or both in B should be unaffected by the branch swap; only pairs of strains between the two groups will exhibit changes in relative position and hence changes in distance. If a strain or group of taxa recombines with a nearby branch, the average change of distances will be low; however, if they recombine with a more remote branch of the tree, representing exchange of genetic material with a more divergent ancestor strain, then the relationships among the strains will be larger. The distribution of EBS distances between the original tree and the 300 alternative trees ranged from 0.77 to 9.22 (Figure 2.3). The alternative trees are categorized into three groups according to the tree distance with the original one, including small (0.77-2.99), medium (3.02-4.80) and large distance (4.80-9.22) groups. There are about 100 trees in each category.

The true positive rate, false positive rate and F1 score of replicates in the three groups are shown in Figure 2.4. Importantly, there is a great reduction in false positives (2.4b) without much loss of true positives (2.4a) for ptACR on ACR. In general, a replicate in the large evolutionary branch swapping distance group has sequences simulated from a more distinct alternative topology compared to the original tree, which makes the sites in the middle of the alignment tend to exhibit more homoplasy. Thus, the boundaries of the recombination event are easier to detect. In contrast, replicates in the small distance group have closer relatedness of taxa since the alternative tree is less different to the original tree. As evolutionary branch swapping distance decreases, both sensitivity and specificity are reduced.

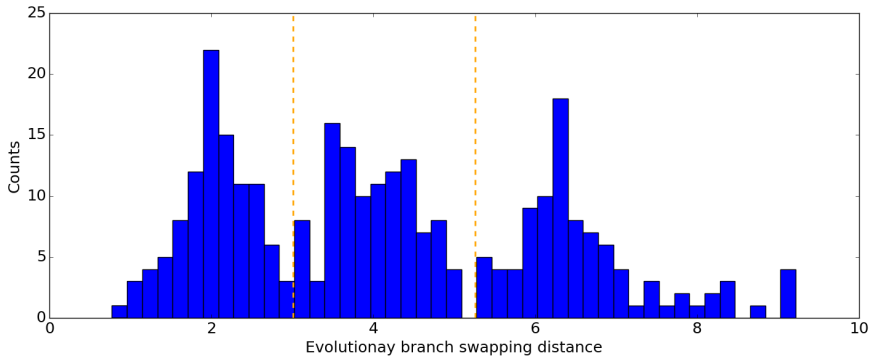
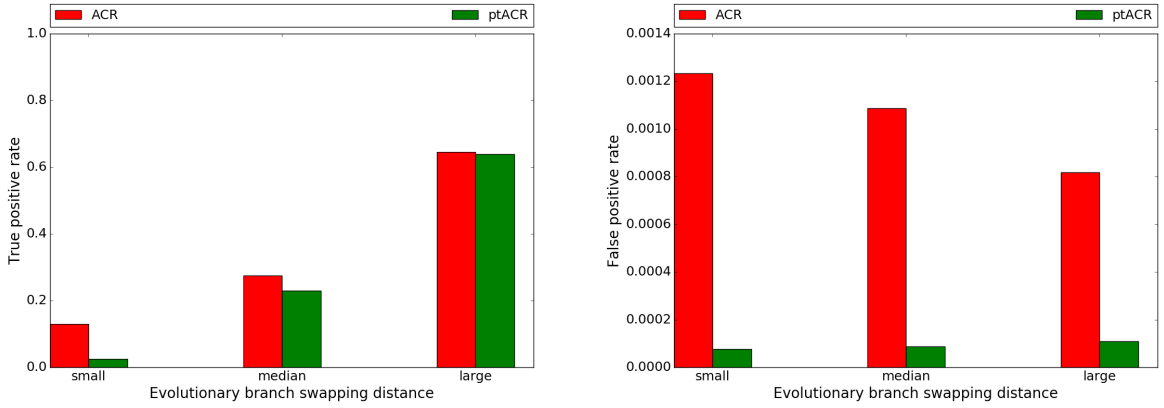
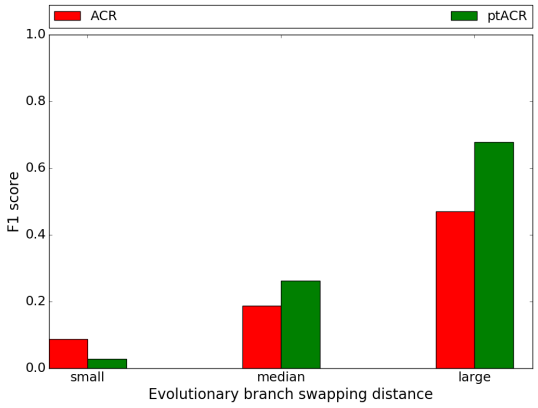


Figure 2.3: Histogram of evolutionary branch swapping distance between the original tree and 300 alternative trees generated using HGT-Gen.



(a) (b)



(c)

Figure 2.4: True positive rate (a), false positive rate (b) and F1 score (c) of 3 scenarios of increasing evolutionary branch swapping distance (no heterogeneity).

2.3.2 Effect of Substitution Rate and Heterogeneity

Sequences were simulated in four scenarios by setting the substitution rate parameter of Seq-Gen to 0.01, 0.02, 0.04 and 0.08. The default substitution rate heterogeneity parameter in Seq-Gen was used ($\alpha = \infty$, which means no heterogeneity). The proportion of nucleotides in each scenario is shown in Figure 2.5. With low substitution rate, there are 62% monomorphic sites. As substitution rate increases, the fraction of informative sites increases. The true positive rate, false positive rate and F1 score of the four scenarios are plotted in Figure 2.6. With low substitution rate, the true positive rate is high, the false positive rate is low and the F1 score is high. The ptACR approach performs better than the ACR in terms of lower false positive rate and higher F1 score.

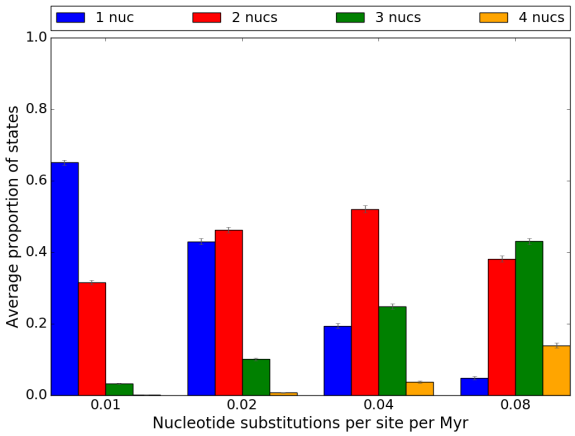
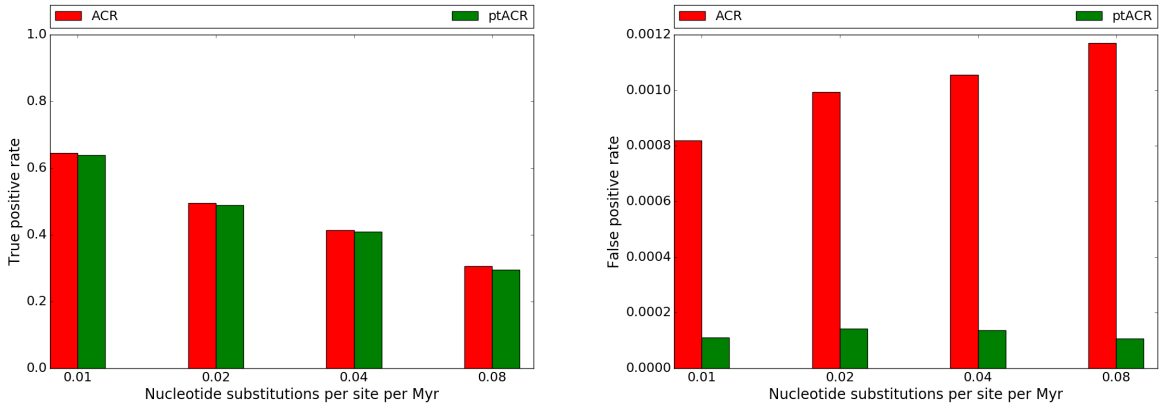
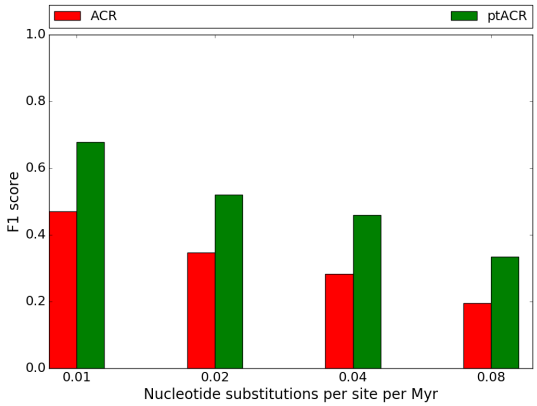


Figure 2.5: Proportion of nucleotides in 4 scenarios of increasing substitution rate.



(a)

(b)



(c)

Figure 2.6: True positive rate (a), false positive rate (b) and F1 score (c) of 4 scenarios of increasing substitution rate (large evolutionary branch swapping distance group).

To examine how substitution rate heterogeneity affects ptACR performance, we varied the heterogeneity α (shape parameter of the gamma distribution) in Seq-Gen, which influences the variability of substitution rates among individual sites. Sequences are simulated in four scenarios of heterogeneity parameter α ranging from 0.2, 0.8, 1.6 to ∞ (with the fixed substitution rate of 0.01). The scenario where α is equal to ∞ represents sequences simulated with a uniform rate at all sites. The proportion of nucleotides in alignments in each scenario is listed in Figure 2.7. With low heterogeneity (α=∞), there are 37% polymorphic sites and 12% of there are multi-state characters. As heterogeneity increases, the fraction of informative sites decreases. The true

positive rate, false positive rate and F1 score of four scenarios are plotted in Figure 2.8. The red bars stand for the results from the previous ACR method while the green bars show the results of incorporating the permutation test (ptACR). With low heterogeneity, the true positive rate is high, the false positive rate is low and the F1 score is high. Only at the highest heterogeneity are the sensitivity and specificity reduced. Hence, ptACR accurately detects recombination breakpoints in the alignments, including multi-state characters, except in the most extreme divergent situations (where there is more background homoplasy) occurring stochastically even without recombination.

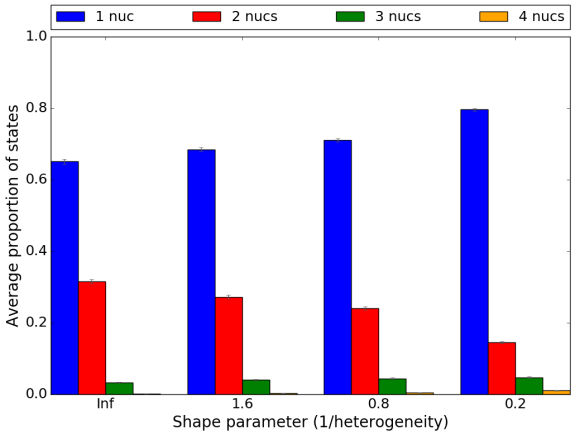
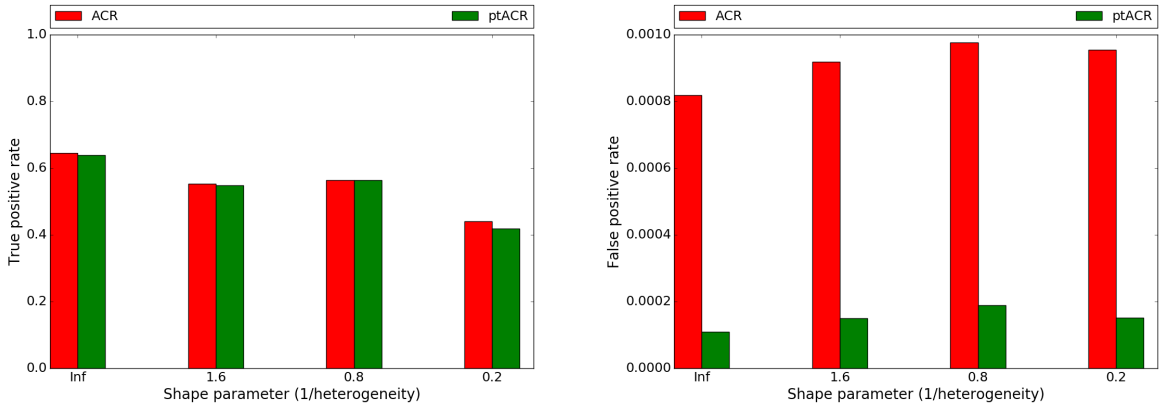
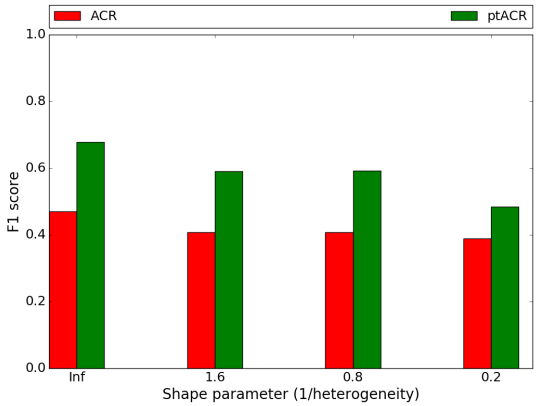


Figure 2.7: Proportion of nucleotides in 4 scenarios of increasing heterogeneity.



(a)

(b)



(c)

Figure 2.8: True positive rate (a), false positive rate (b) and F1 score (c) of 4 scenarios of increasing heterogeneity (fixed substitution rate and large evolutionary branch swapping distance group).

3. IDENTIFICATION OF RECOMBINATION IN COLLECTIONS OF PATHOGENS *

To evaluate our ptACR method, we use it to characterize homoplasy in three species: *Mycobacterium tuberculosis*, *Mycobacterium avium* and *Staphylococcus aureus*.

3.1 *Mycobacterium tuberculosis*

The bacterial species *M. tuberculosis* is thought to be highly clonal and have shown basically no recombination events in previous studies [44, 45]. It is used as a negative control.

The dataset is composed of 50 worldwide clinical isolates [46]. We aligned them to the reference genome H37Rv (accession NC_000962.2) of size 4.4M bp. There are 10565 SNP sites in the alignment and the number of changes per site is 1.006 (10633/10565). The global phylogenetic tree is reconstructed from 10565 informative sites and shown in Figure 3.1. The tree was produced using SplitsTree [47] where an acyclic graph suggests that the tree is monophyletic. The overall compatibility ratio is 0.999, reflecting the clonal nature of *M. tuberculosis* strains worldwide. Hence, we should expect to find no recombination. The plot of average compatibility ratio of three window sizes is shown in Figure 3.2. Since the average compatibility ratio of the entire alignment is over 99.5%, our approach will report no combination breakpoints. In addition, RDP4 reported that no evidence of recombination event was found in the alignment.

---

*Reprinted with permission from "A statistical method to identify recombination in bacterial genomes based on SNP incompatibility" by Y.-P. Lai and T. R. Ioerger, 2018. *BMC Bioinformatics*, 19, 450, Copyright [2018] by BioMed Central. DOI:10.1186/s12859-018-2456-z.
Part of the data reported in this chapter is reprinted with permission from "A compatibility approach to identify recombination breakpoints in bacterial and viral genomes" by Y.-P. Lai and T. R. Ioerger, 2017. *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pp. 11-20, Copyright [2017] by Association for Computing Machinery. DOI:10.1145/3107411.3107432.

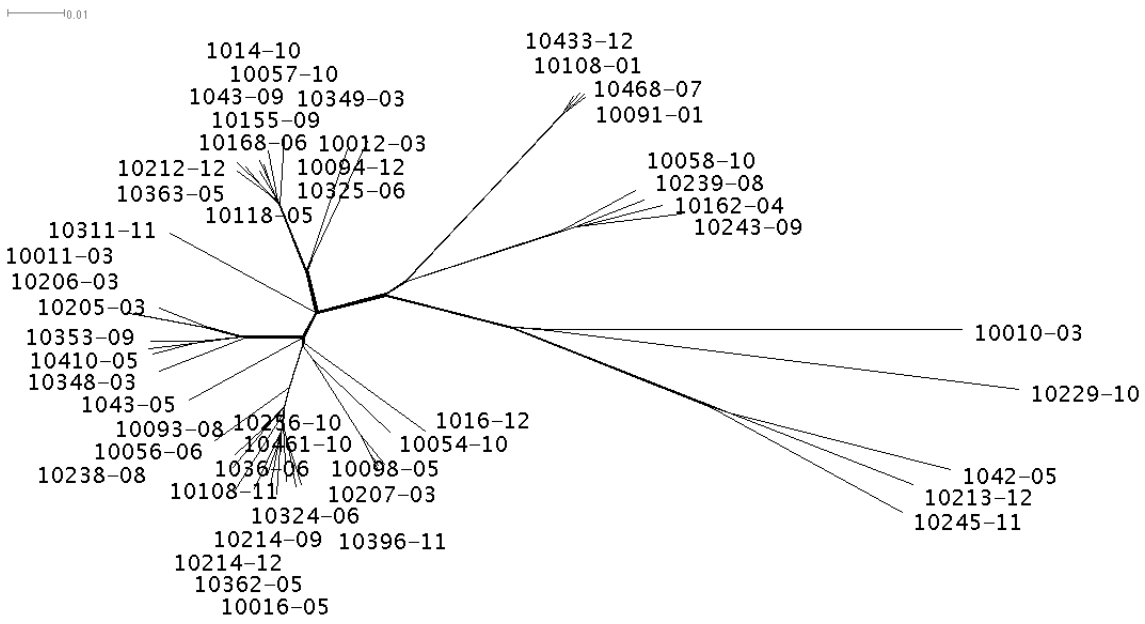


Figure 3.1: Global phylogenetic tree of 50 isolates for *M. tuberculosis*.

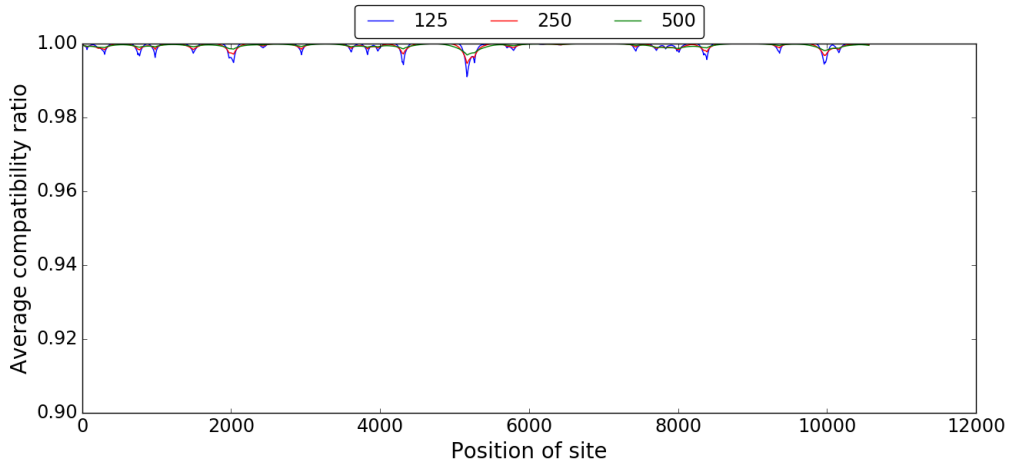


Figure 3.2: Average compatibility ratio for each site using window sizes of 125, 250 and 500 for *M. tuberculosis*.

3.2 *Mycobacterium avium*

The second dataset we evaluated consists of a set of 18 clinical isolates of *Mycobacterium avium (M. avium)* from our collaborators at St. Olav's Hospital in Trondheim, Norway [48]. The isolates were collected from sputum samples of the patients diagnosed with *M. avium* infections between 2007 and 2009. The isolates were sequenced by an Illumina sequencer (HiSeq 4000) to obtain paired-end reads of a length of 150 bp, and then the reads were assembled by an in-house method [49]. The contigs were aligned to the reference genome avium104 (accession NC_008595.1) together with two other reference strains of TH135 (AP012555.1) and H87 (CP018363.1).

The isolates are highly diverse. In the alignment of length 5.5 Mb, there are 70722 polymorphic sites, and 510 sites (0.72%) have more than two nucleotides (multi-state). The overall compatibility ratio over the whole genome is 78.65%, and the average homoplasy ratio is 1.6799. The global phylogenetic tree is reconstructed from 70722 informative sites and shown in Figure 3.3. The tree is produced using SplitsTree [47]. The cluster of edges (circles in the graph) in the middle indicates that sites exist that are not congruent with a perfect monophyletic tree, suggesting recombination or non-clonality. The ptACR algorithm is applied to scan the alignment using a window size of 250 SNPs. Figure 3.4 shows that it identifies 71 local minima as the potential recombination boundaries (labeled in red). Next, 70 breakpoints (labeled in green) are identified as statistically significant with permutation test where the threshold of the corrected p-value is 0.0007 (0.05/71).

To validate the level of phylogenetic congruence of 71 regions from the global tree to the regional tree, the plot of the homoplasy ratio for each region based on the global tree and a regional tree is shown in Figure 3.5. The homoplasy ratio for each region decreases from the global tree to each regional tree. Further analysis of the consecutive regions from the 34^{th} to 36^{th} segments shows that the excess changes are reduced in each region using the corresponding local tree. Statistics are listed in Table 3.1. The phylogenetic trees of the consecutive regions are shown in Figure 3.6. Seven isolates that do not share a common branch point across the three regions are labeled in rectangles of the same color. For example, MAV07 and MAV09 are clustered with avium104 in the

34^{th}region, but they are clustered with H87 in the 35^{th} region, indicating a probable recombination event. An interesting example related to antibiotic resistance is that, in the 34^{th} region, there is a gene named *MAV_3128* (Lysyl-tRNA synthetase LysS), which has been shown to be sensitive to antibiotics and prone to mutation in the *M. avium* subspecies *hominissuis* [50].

Lastly, the plot of the most closely related reference strain for each isolate in each region is shown in Figure 3.7. Changes of the most closely related reference strain across the regions for all isolates suggeste mosaic structures in the population. Five isolates, MAV21, MAV38, MAV18, MAV32 and MAV23, are not only divergent but considerably mosaic, with similarities alternating among avium104, H87 and TH135.

The analysis of recombination from ClonalFrameML is shown in Figure 3.8 where dark blue horizontal bars indicate recombination events for each branch and white vertical bars represent substitutions. Strains MAV23, MAV32, MAV18, MAV38 and MAV21 have several recombination events across the genomes. The locations of recombinations in strains MAV18 and MAV38 are close to each other. The ClonalFrameML identifies 601 recombinant regions in 15 internal branches and 332 recombinant regions in 7 strains. The sizes of regions range from 5 to 6510 SNPs and 341 regions are smaller than 200 SNPs. It shows that the ClonalFrameML identifies more small recombinant regions and more breakpoints than ptACR.

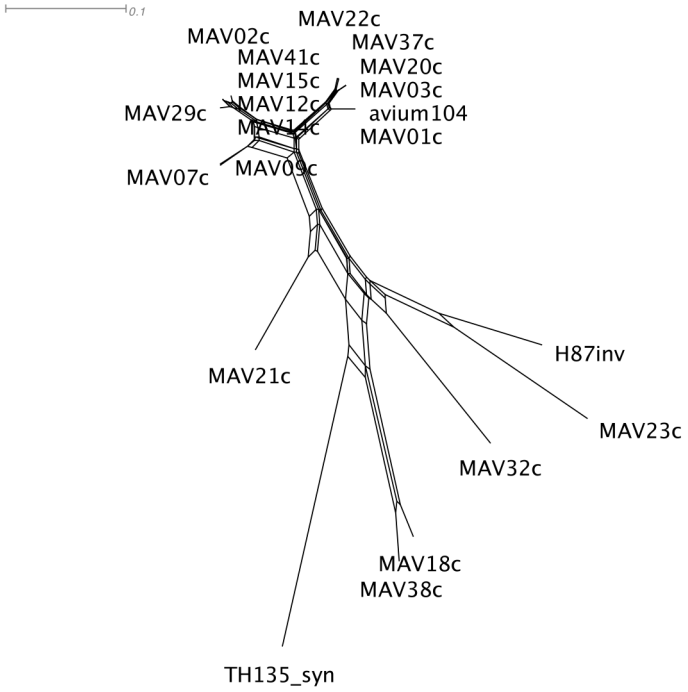


Figure 3.3: Global phylogenetic tree of 18 isolates for *M. avium*. The cluster of edges in the middle indicates that sites exist that are not congruent with a perfect monophyletic tree.

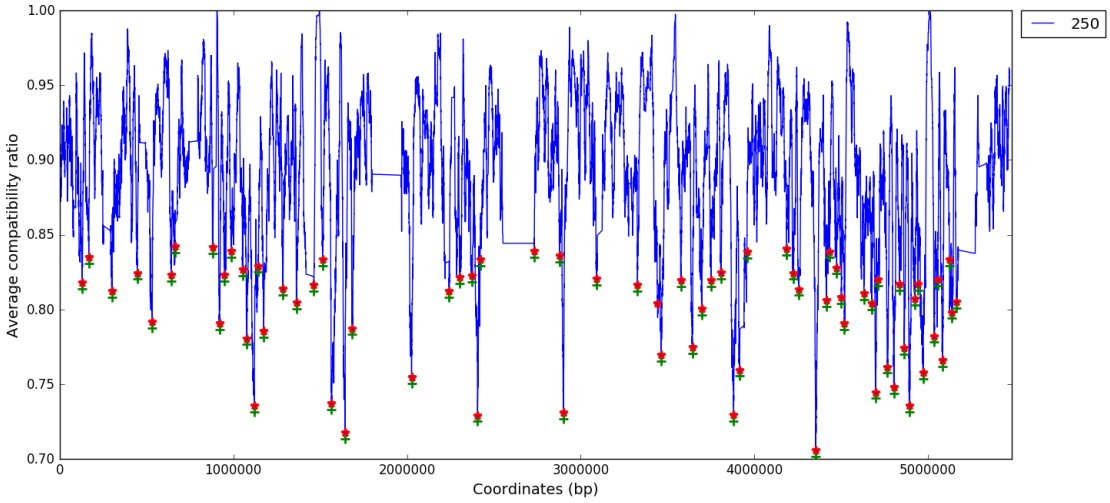


Figure 3.4: Identified breakpoints using window sizes of 250 bp for *M. avium*.

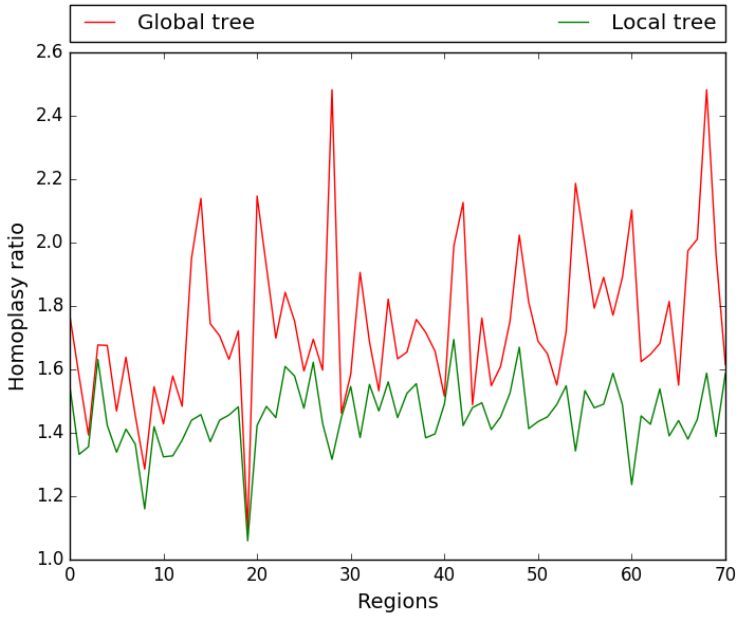


Figure 3.5: Homoplasy ratio based on global and regional trees for each region of *M. avium*.

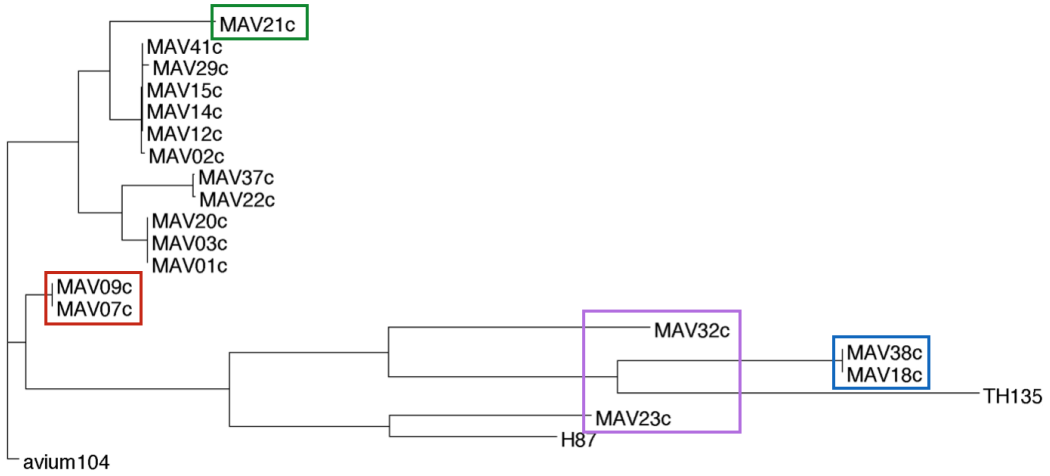
Table 3.1: Information for regions of *M. avium*.

Region	Size (kb) [a]	SNPs [b]	Genes [c]	Compat [d]	EC_G [e]	EC_L [f]	Ratio [g]
34^{th}	237.16	2964	*MAV_3053-3224*	84.98%	1597	1407	11.90%
35^{th}	134.98	1895	*MAV_3225-3319*	85.20%	1577	1076	31.77%
36^{th}	114.24	1588	*MAV_3320-3429*	87.19%	1014	717	29.29%

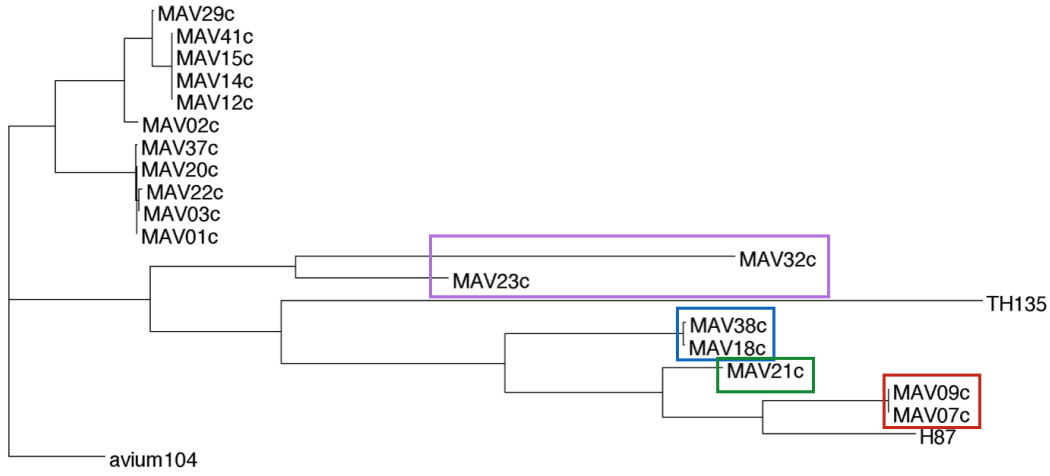
[a] region size; [b] number of informative sites; [c] genes in the region;

[d] regional compatibility ratio; [e] the excess changes based on the global tree;

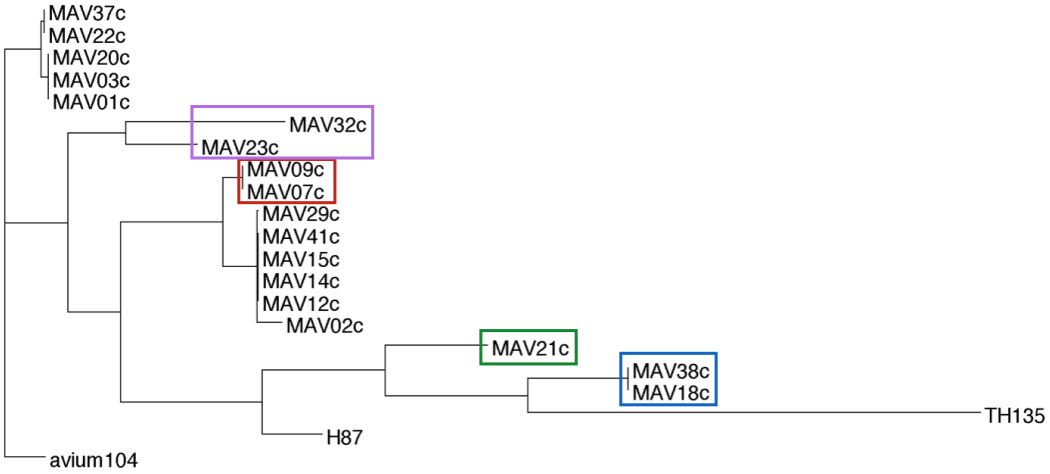
[f] the excess changes based on the local tree; [g] the reduction ratio of excess changes, $1\text{-}\frac{EC_{local}}{EC_{global}}$.



(a)



(b)



(c)

Figure 3.6: Phylogenetic trees in the 34^{th}-36^{th} regions (a-c) of *M. avium*.

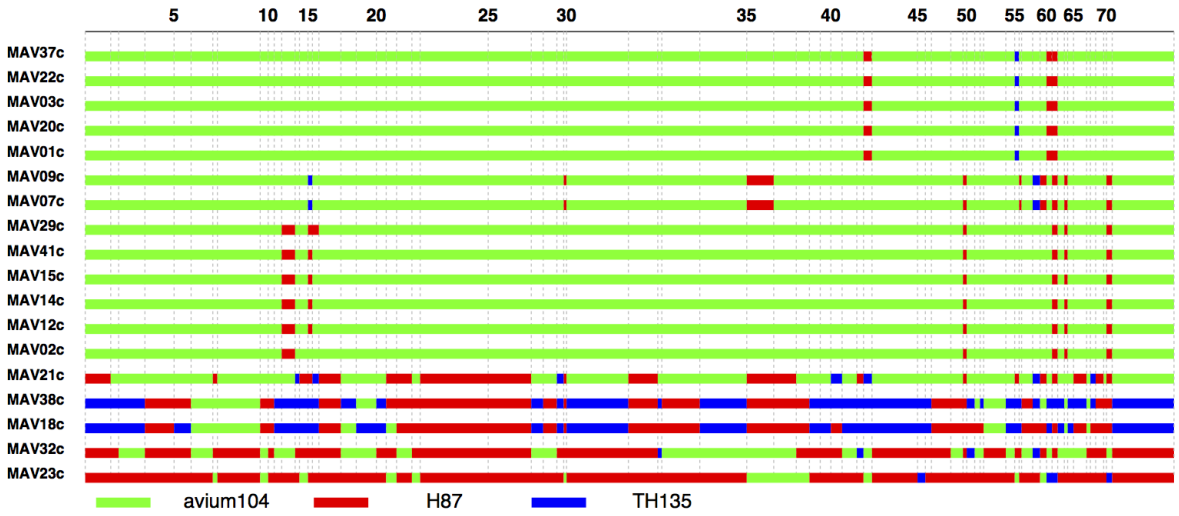


Figure 3.7: Mosaic patterns plotted from the most closely related reference strains across 71 regions for 18 *M. avium* strains.

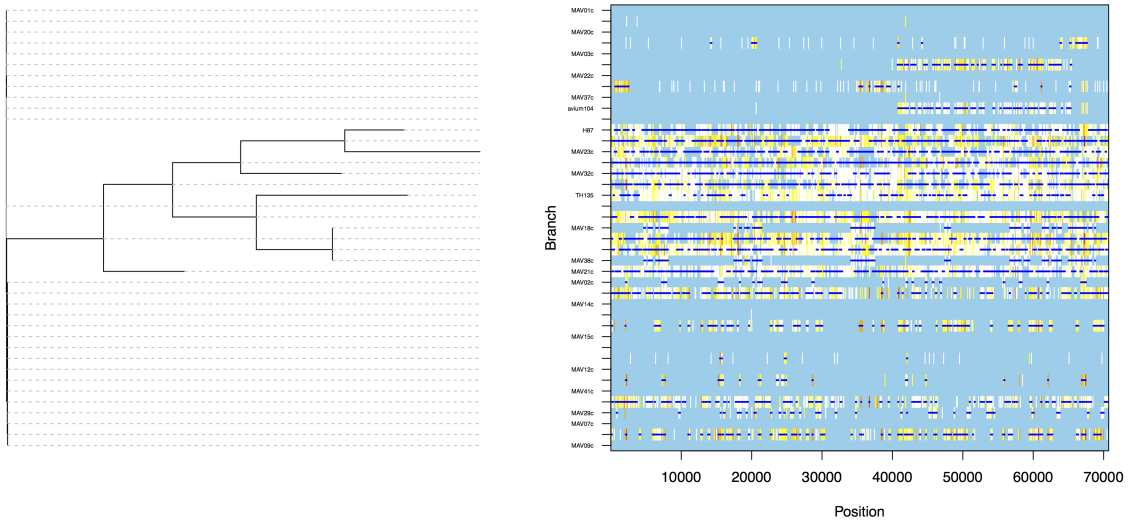


Figure 3.8: ClonalFrameML analysis in *M. avium*. Recombination events are marked in dark blue horizontal bars.

Mycobacterium avium complex is a group of pathogenic mycobacteria, including *M. avium*, *M. intracellulare* and *M. chimaera*. It is characterized as non-tuberculous mycobacteria (NTM). Clinical isolates of *M. avium* exhibit high genetic diversity [51]. The recombination that we see in *M. avium* contrasts with *Mycobacterium tuberculosis*, for which it has been shown that isolates worldwide fit into a well-defined tree (lineage structure) without the evidence of recombination, likely due to the lack of functional recombination pathways [52, 53] or conjugation [54]. In general, *M. tuberculosis* is believed to be highly clonal during evolution [55]. However, recombination has been observed in other mycobacterial species such as *M. canetti* [56, 57] and *M. smegmatis* [58]. Recombination in some mycobacterial strains mediates the exchange of genetic materials and drives rapid genetic evolution. Recombination in *M. avium* has been reported [5], but the recombinant regions we detect with ptACR are much larger than individual genes. In this study, we reveal that frequent recombination events are observed in *M. avium*. The identification of breakpoints contributes to obtaining regional phylogenies that are different from the global tree, explaining homoplasy in the clinical isolates.

3.3 *Staphylococcus aureus*

Staphylococcus aureus is a human pathogen that causes lung and skin infections. Studies have revealed that *S. aureus* contains many types of mobile genetic elements that drive recombination hotspots, including plasmids, bacteriophages, pathogenicity genomic islands and islets, transposons, insertion sequences and staphylococcal cassette chromosomes (SCC) [11, 12].

We applied ptACR to analyze a collection of 30 clinical isolates of *S. aureus* [11] aligned with 5 reference strains, including ST8:USA300 (NC_010079.1), SACOL (CP000046.1), EMRSA-15 (HE681097.1), N315 (BA000018.3) and ATCC 25923 (NZ_CP009361.1). Recombination has previously been observed for the species [11, 12]. The alignment of *Staphylococcus aureus* contains 2.87 Mb nucleotides where 113,936 sites are informative (polymorphic) and 3,625 sites (3.18%) have over two nucleotides. The overall compatibility ratio over the genome is 88.34% and the homoplasy ratio is 1.4484, suggesting recombination occurs among the population. The global phylogenetic tree is shown in Figure 3.9. This figure is produced using SplitsTree [47]. Figure

3.10 illustrates that 86 local minima (labeled in red) are identified by ACR as potential breakpoints using a window size of 250 informative sites, and then 65 breakpoints (labeled in green) are identified as statistically significant by ptACR with permutation test where the threshold of the corrected p-value is 0.000581 (0.05/86). Hence, 66 regions are obtained. Any two adjacent regional phylogenetic trees constructed by their corresponding local alignments have distinct tree topologies, reflecting the identified boundaries are confident, since changes in phylogenetic relationships occur between each pair of adjacent regions.

The plot of the homoplasy ratio for each region based on the global tree and a regional tree is shown in Figure 3.11. For each region, both homoplasy ratio and excess changes decrease from the global tree to the regional tree, showing that the regions identified by ptACR have different topologies from the global tree, and each local tree is able to accommodate more sites within the corresponding region. Figure 3.12 shows local phylogenetic trees for three consecutive regions, starting from the 37^{th} segment, as an example for further analysis. The recombined groups of isolates are labeled in rectangles of the same color. According to the tree topologies, the 37^{th} region shows that the strain ERR410042 receives a copy from an ancestor of two strains, ERR410056 and ERR410060. Yet in the 38^{th} region the strain ERR410042 receives a copy from an ancestor of three strains, ERR410044, ERR410046 and N315, while a parent of ERR410056 and ERR410060 receives a copy from an ancestor of ERR410038, ERR410039 and EMRSA-15. In the 39^{th} region the strain ERR410042 receives the copies from parents of the strain ERR410058 instead. The information of region size, number of informative sites (SNPs), genes, overall compatibility ratio (Compat), the excess changes based on global tree (EC_{global}) and local tree (EC_{local}), and the reduction ratio of excess changes (Ratio) for the three regions is listed in Table 3.2. The number of excess changes decreases from the global tree to the local tree, showing that the local trees significantly reduce the apparent homoplasy based on the global tree.

To visualize the relationships among strains, a plot of the most closely related reference strain for each strain in each region is shown in Figure 3.13. Strains ST8:USA300, EMRSA-15, ATCC 25923 and N315 were used as references, spanning several different lineages/strain types world-

wide. For each strain, the most closely related reference strain is defined as the one that has the least differences in a region. Figure 3.13 shows that for several strains, the most closely related reference strain changes across the genome (i.e., pattern is mosaic), indicating that they are likely recombined (especially ERR410042). This is consistent with previous studies that found extensive recombination in this collection of *S. aureus* isolates [11, 12]. In the collection we studied, the 28^{th} region contains *mecA* (*USA300HOU_0956*) gene that is located on SCC and most commonly known as encoding methicillin resistance in *S. aureus* [59, 60]. Also, the *scpA* gene, which is on a plasmid-associated island and contributes to staphylococcal virulence [61], is in the 37^{th} region.

The analysis of recombination from ClonalFrameML is shown in Figure 3.14 where dark blue horizontal bars indicate recombination events for each branch and white vertical bars represent substitutions. It shows that lots of recombination events are detected in several internal branches and three strains, ERR410035, ERR410042 and ERR410058. Each of three strains receives a copy from different ancestors in consecutive regions identified by ptACR. The ClonalFrameML identifies 1264 recombinant segments in 18 internal branches and 307 recombinant segments in 10 strains. The sizes of segments range from 2 to 20052 SNPs and 519 segments are smaller than 200 SNPs. In sum, the ClonalFrameML identifies more breakpoints than ptACR.

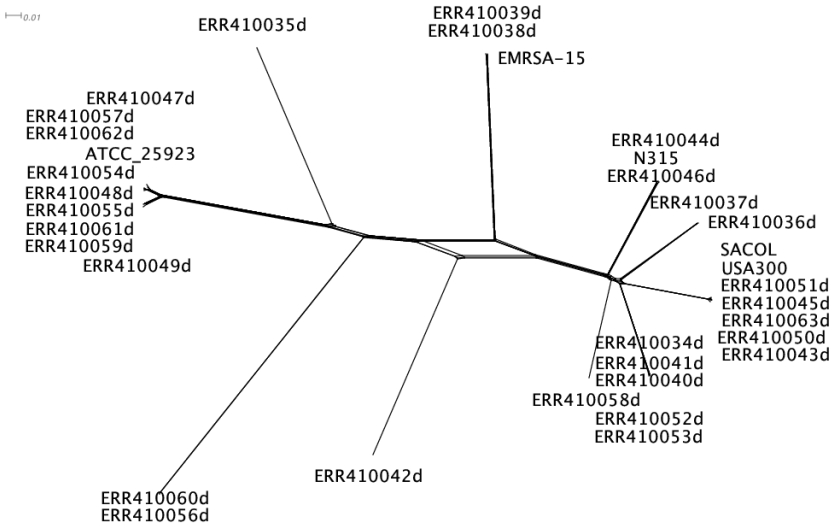


Figure 3.9: Global phylogenetic tree of 35 strains for *S. aureus*. The cluster of edges in the middle indicates that sites exist that are not congruent with a perfect monophyletic tree.

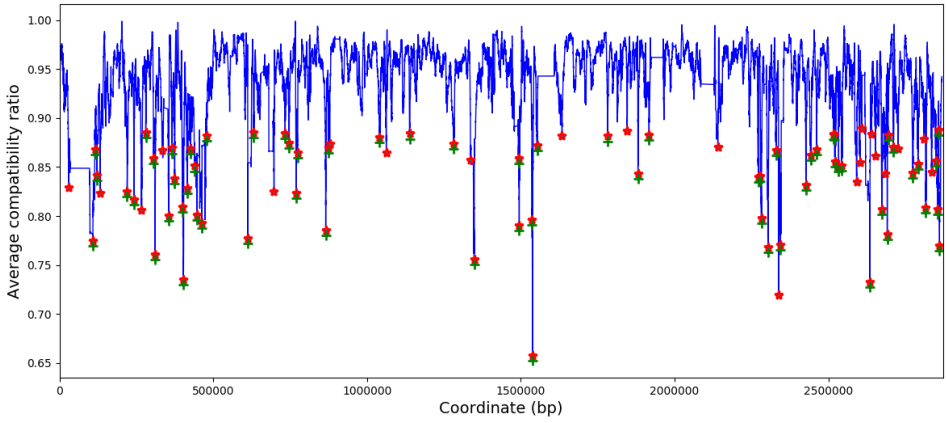


Figure 3.10: Identified breakpoints using window sizes of 250 informative sites for *S. aureus*.

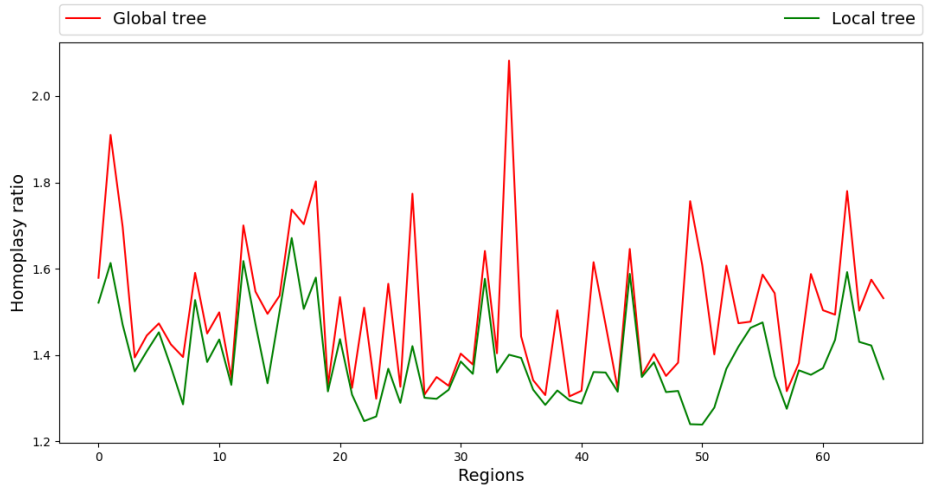


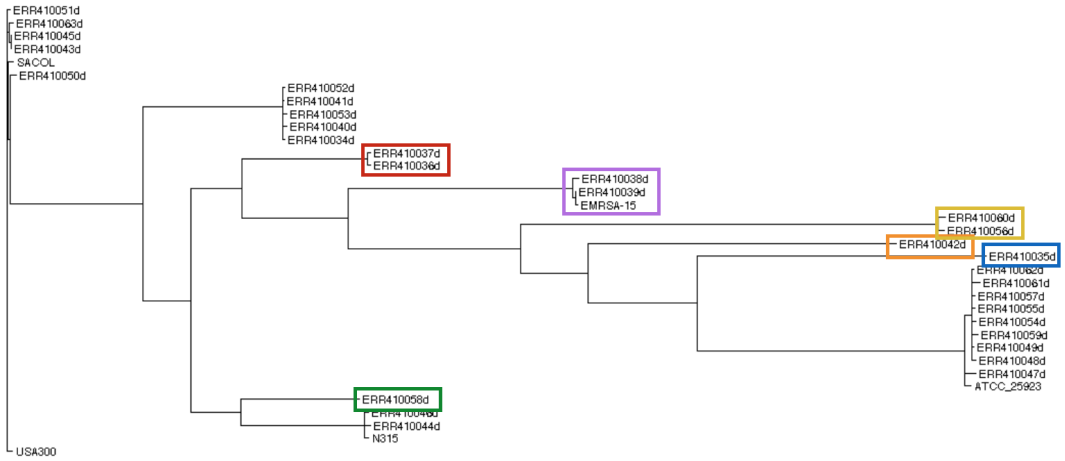
Figure 3.11: Homoplasy ratio based on global and regional trees for each region of *S. aureus*.

Table 3.2: Information for regions of *S. aureus*.

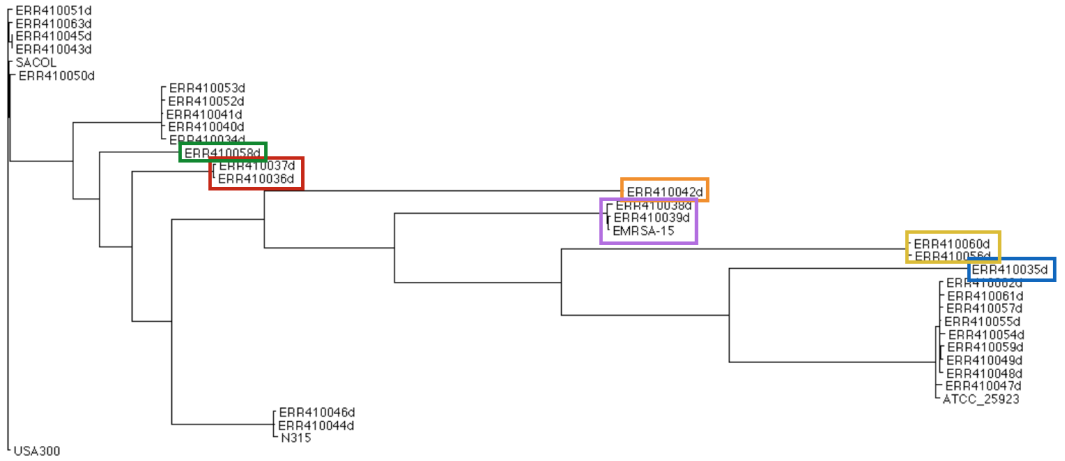
Region	Size (kb) [a]	SNPs [b]	Genes [c]	Compat [d]	EC_{global} [e]	EC_{local} [f]	Ratio [g]
37^{th}	228.41	5526	*USA300_1420-1668*	94.59%	1993	1808	9.28%
38^{th}	97.74	4777	*USA300_1669-1747*	93.63%	1512	1400	7.41%
39^{th}	36.17	1745	*USA300_1747-1778*	89.93%	914	577	36.87%

[a] region size; [b] number of informative sites; [c] genes in the region;

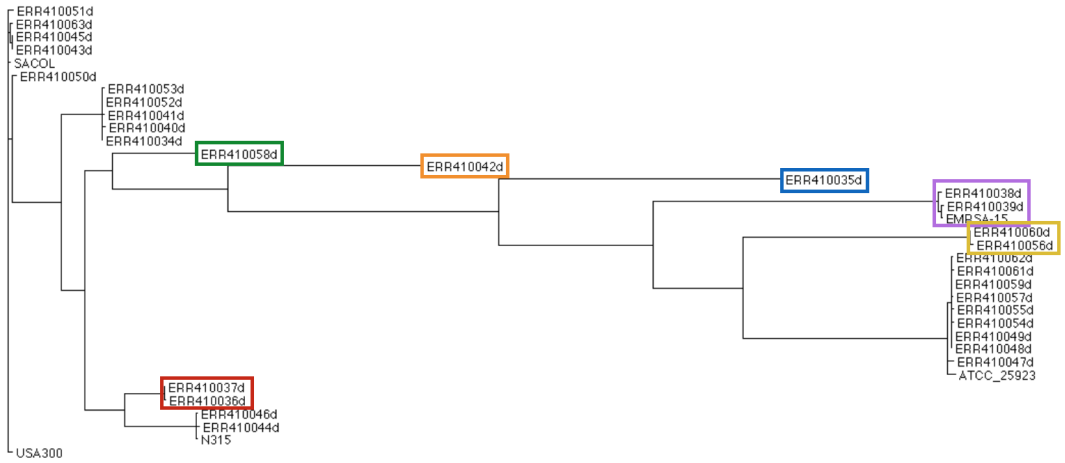
[d] regional compatibility ratio; [e] the excess changes based on the global tree; [f] the excess changes based on the local tree; [g] the reduction ratio of excess changes, 1-$\frac{EC_{local}}{EC_{global}}$.



(a)



(b)



(c)

Figure 3.12: Phylogenetic trees in the 37^{th}-39^{th} regions (a-c) of *S. aureus*.

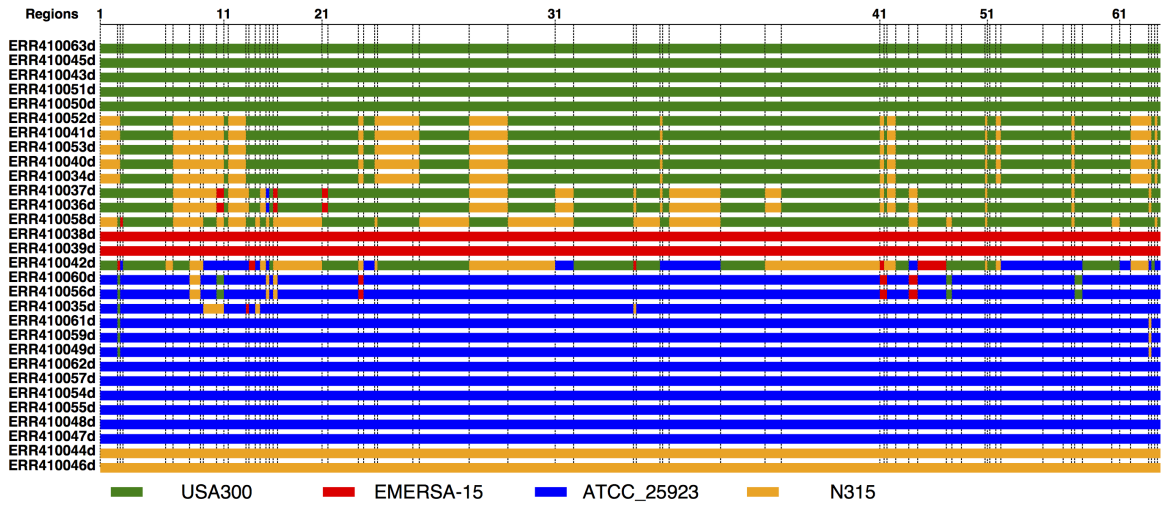


Figure 3.13: Mosaic patterns plotted from the most closely related reference strains across 66 regions for 30 *S. aureus* strains.

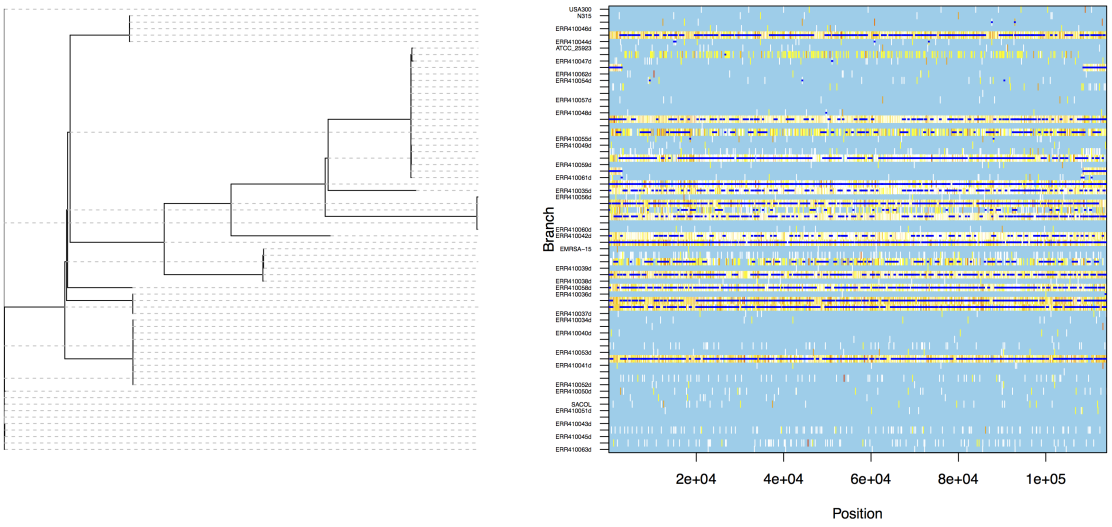


Figure 3.14: ClonalFrameML analysis in *S. aureus*. Recombination events are marked in dark blue horizontal bars.

4. HOMOPLASY IN DRUG-RESISTANT POLYMORPHISMS IN PATHOGENS

4.1 Background

To infer the causality between genotypes and phenotypes in genomes of bacterial pathogens, methods for genome-wide association studies have been developed to statistically find the genetic variants (mutants) associated with the phenotypic traits, including antibiotic resistance, host specificity and virulence [17, 18, 62]. Bacteria accumulate heritable genetic variants during evolution. Since bacteria are haploid and their reproduction is asexual, the occurrence of homoplasy is an important signal in genome evolution for bacterial species. The genetic mechanisms of homoplasy include horizontal gene transfer (usually involving transformation, transduction and conjugation), recombination (through conjugation) and recurrent mutation [17]. Some bacteria tend to exchange DNA frequently through recombination and therefore their genomes are more diversified. In contrast, some bacteria generally replicate DNA vertically so they remain highly clonal. Their homoplasic signals in genomes are mainly from recurrent mutations driven by selection pressures [18]. Hence, for clonal bacteria, homoplasy plays a role in understanding antibiotic resistance through the statistical associations between polymorphic sites and resistant phenotypes. It indicates positive selections yet it is not well accounted by most methods.

4.1.1 Bacterial Genome-Wide Association Studies

Genome-wide association studies identify statistically significant associations between genotypes and phenotypes among the entire genomes without prior assumptions on causal associations [63]. The genotypes are genetic variants among samples, such as gene expressions from microarray, single nucleotide polymorphisms (SNPs), insertions or deletions (indels) from next-generation sequencing (NGS). The phenotypes are traits of interests from binary (e.g., resistant versus sensitive to a drug) to different levels of quantitative values (e.g., growth rates, minimal inhibitory concentrations). The first GWAS was proposed and applied in human genomes in 2005 [64]. Human genomes are eukaryotic with diploid chromosomes. Through meiosis, parental cells pass on

genetic materials to descendants by chromosomal crossover or recombination to achieve linkage equilibrium, i.e., no correlation between genetic sites. Typical human GWAS categorizes samples at each polymorphic site into a two-by-two contingency table according to the genotypes of major and minor allele frequencies and phenotypes of cases and controls. It then commonly applies statistical tests such as the chi-squared test, Fisher's exact test or hypergeometric test to calculate the test statistics. By comparing with expectations, the statistical significance of the association could be assessed. Other regression-based methods apply linear models to regress genotypes (covariates) against phenotypes to estimate the significance of correlations [65]. Main confounding factors in human GWAS are population stratification and linkage disequilibrium (LD) [66, 67]. Stratification in a population represents that some subpopulations exist and individuals in the subgroups are relatively closer to each other than others. Linkage disequilibrium occurs when some regions of the genome are descended together, forming LD blocks with correlated alleles. Current methods to reduce the impact of confounders are genomic control (λ_{GC}) [68], principal component analysis (PCA) [69], LD score regression [67], and linear mixed model [70, 71]. Well-known and frequently-used programs include PLINK [65], EMMA [70] and GEMMA [71].

Recently GWAS has begun to be applied to bacterial genomes to dissect the genetic variants associated with traits of antibiotic resistance, virulence and bacterial-host interaction [17, 18, 62]. Yet approaches in eukaryotic studies cannot be applied directly to bacteria due to the differences of genome compositions. Humans are diploid eukaryotes while bacteria are haploid prokaryotes. The reproduction of bacteria is asexual and the clonality of genomes is shaped by replicating DNA vertically and exchanging DNA horizontally. During evolution, some bacterial genomes tend to be more divergent through recombination, while some bacteria remain clonal through cell division [3, 4]. For clonal bacteria, the extent of linkage disequilibrium is larger, the impact of population structure is stronger and the recombination is less likely to occur. Hence, if a homoplasic polymorphism exists, it shows that a recurrent mutation evolves along different tree branches, indicating the selection pressure. Ignoring confounders like population structure or homoplasy in bacterial GWAS may produce false positives or false negatives.

A conventional linear model tests the effect size β between two random variables, assuming the null hypothesis H_0: $\beta = 0$ and the alternative hypothesis H_1: $\beta \neq 0$. Given n individuals, regressing phenotypes against genotypes can be modeled as

$$\boldsymbol{y} = \alpha + \boldsymbol{x}\beta \tag{4.1}$$

where $\boldsymbol{y}$ is an n-vector of phenotypic traits, $\boldsymbol{x}$ is an n-vector of genotypes at a given locus, β is the effect size and α is the intercept. The top principal components of genotypes capture genetic distances between individuals, representing the ancestry. To reduce the impact of population stratification in bacterial GWAS, regression-based approaches apply the PCA as covariates or fixed effects in linear regression test. It is usually modeled as

$$\boldsymbol{y} = \boldsymbol{W}\alpha + \boldsymbol{x}\beta \tag{4.2}$$

where $\boldsymbol{W} = (w_1, \ldots, w_k)$ is an n x k matrix of top k principal components as covariates and α is a k-vector of coefficients of corresponding covariates. In addition, to account for population structure, a genetic relatedness (kinship) matrix is applied to the linear mixed models (LMMs) as a random effect. Let genotypes $\boldsymbol{X}$ be an n x p matrix of n samples and p genetic loci, the kinship matrix $\boldsymbol{K}$ can be estimated as

$$\boldsymbol{K} = \boldsymbol{X}\boldsymbol{X}^T. \tag{4.3}$$

$\boldsymbol{K}$ is an n x n matrix that captures genetic covariances between individuals and is also named as a genetic relatedness matrix. Then the LMM can be described as

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{x}\beta + \boldsymbol{u} + \boldsymbol{\epsilon}, \\ \boldsymbol{u} &\sim \mathrm{MVN}_n(0, \sigma_a^2 \boldsymbol{K}), \\ \boldsymbol{\epsilon} &\sim \mathrm{MVN}_n(0, \sigma_e^2 \boldsymbol{I}_n), \end{aligned} \tag{4.4}$$

where $\boldsymbol{y}$ is an n x 1 vector of phenotypes, $\boldsymbol{x}$ is a matrix of genotypes, β represents the effect size

of genotypes, $\boldsymbol{u}$ presents the random effect modeled by a multivariate normal distribution (MVN) with the genetic variance (σ_a^2) and the genetic relatedness matrix ($\boldsymbol{K}$), $\boldsymbol{\epsilon}$ represents a vector of environmental errors with the variance (σ_e^2), and $\boldsymbol{I}_n$ is an n x n identity matrix. The significance of coefficients can be determined by the Wald test or likelihood ratio test [71]. For example, an R package, bugwas, not only utilizes LMM but also considers lineage-effect associations by decomposing the kinship to principal components [72].

4.1.2 Phylogenetic Convergence Tests

For clonal bacterial species, a single phylogeny exists, which can be used to account for homoplasy. Thus, phylogeny-based approaches have also been developed, including phyC [16], phyOverlap [73] and treeWAS [74]. The phylogenetic convergence test (phyC) obtains the internal nodes where the mutations occur for all polymorphic sites, and then it determines the drug susceptibility of all internal nodes by maximum parsimony. For each site, it utilizes a permutation test to assess the significance by calculating the empirical p-value from background signals of all polymorphic sites [16]. For example, we assign both a phenotype and a genotype at a site to 15 strains, assuming they evolve along the tree shown in Figure 4.1. The phenotype for each branch is determined from the maximum parsimony approach. The allele substitutions occur in 6 strains. We apply Sankoff's algorithm on the genotype to the tree and then obtain 3 branches (changes) where the substitution/mutation occurs. Two occur in sensitive branches and one in a resistant branch (2S, 1R). Subsequently, we test the significance of the association between the genotype and the phenotype by computing how likely this observation occur by chance compared to the background. The concept of phyOverlap is similar to the phyC. It identifies the tree branches where the changes occur, and calculates how many strains underneath the branches have the phenotypic traits to determine the overlapping score. The significance of the score is estimated from the permutation of redistributing mutations across the tree [73]. The treeWAS tool tests three statistics of genotypic variants correlated with phenotypic traits from leaves (terminal score) to branches (simultaneous score) to the entire tree (subsequent score) [74]. These three scores rely on the permutation test to estimate the statistical significance. The loci of associations that do not occur by chance from

three tests are pooled as the candidates. The above methods are usually applied to genotypes of individual sites or sites grouped by a whole gene without considering interactions between genotypes (epistasis). They also do not consider correlations among phenotypes, i.e., co-resistance of drugs.

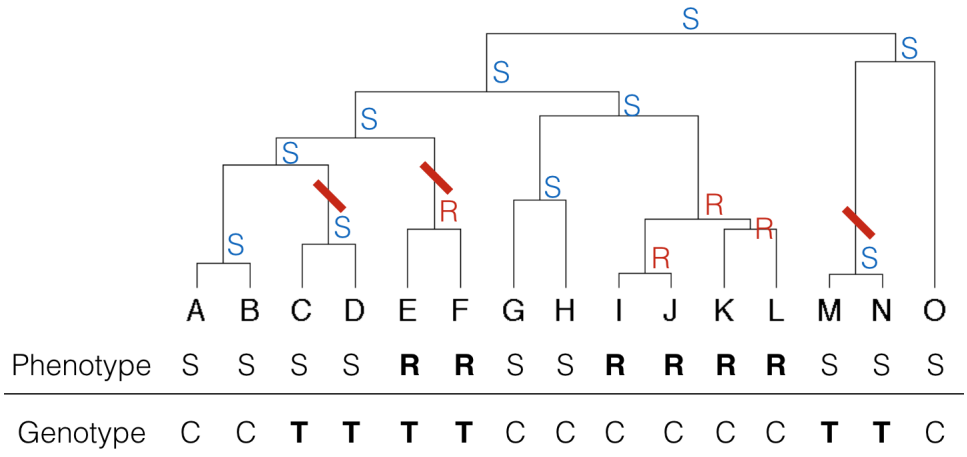


Figure 4.1: Tree of 15 strains with a pair of a binary phenotype (R/S) and a genotype (C/T) at a site. The R/S labeled in each branch is determined by the maximum parsimony approach. A red bar in the branch presents where allele substitution occurs in the tree estimated by applying the Sankoff's algorithm. In this example, we obtain three branches where a change occur from nucleotide C to T. One branch is resistant-associated and two are sensitive-associated.

4.1.3 Association Mapping in *Mycobacterium tuberculosis*

Mycobacterium tuberculosis is a causative pathogen of tuberculosis that primarily infects human lung. The *M. tuberculosis* genome is about 4.4M base pairs and believed to be highly clonal with low mutation rate in previous studies [75, 76]. There is also no obvious evidence of recombination or horizontal gene transfer in the *M. tuberculosis* genome. Worldwide *M. tuberculosis* complex in human is classed to four major lineages by spoligotype families: lineage 1 (East African-Indian (EAI)), lineage 2 (Beijing), lineage 3 (Central Asian (CAS)), and lineage 4 that includes Latin American-Mediterranean (LAM), Haarlem, T clade, X clade and H clade [77].

To treat tuberculosis infection, current anti-tuberculous drugs include 5 first-line drugs and several second-line drugs. The five first-line drugs are isoniazid (INH), rifampicin (RIF), streptomycin (STR), ethambutol (EMB) and pyrazinamide (PZA). Other second-line drugs include fluoroquinolones (ofloxacin (OFX), moxifloxacin (MOX) and ciprofloxacin (CPX)), ethionamide (ETH), cycloserine (CS), amikacin (AMK), kanamycin (KAN), capreomycin (CAP) and para-aminosalicylic acid (PAS). If the strain is resistant to both INH and RIF, it is defined to be multidrug-resistant (MDR). If it is further resistant to any second-line antibiotics, then it is defined to be extensively drug-resistant (XDR). Mechanisms of resistance to several antibiotics in *M. tuberculosis* have been discovered and conferred by some SNPs and indels [78]. The well-known annotated loci associated with anti-tuberculous drugs are listed in Table 4.1.

Since *M. tuberculosis* genome is clonal, association mappings in *M. tuberculosis* could be impacted by population stratification, linkage disequilibrium and selection pressure. Positive selection such as drug resistance is a driving force for evolution with causal mutations [16]. Sites that are incongruent with the tree are homoplasic and their impact needs to be accounted for in association mappings. Currently treating polymorphic sites (SNPs) individually as a genotypic input does not exploit proximity/clustering of SNPs, as often observed in an active site. Yet applying genotypes at the gene level (sites grouped by a gene) may not maximize homoplasy signals. Also, the correlation of multidrug resistance is not well-considered in current GWAS methods. That is, a strain that is resistant to one of the first-line anti-tuberculous drugs has a higher propensity to be resistant to others, i.e., a multidrug-resistant strain [19]. For example, studies have shown that strains resistant to INH have a higher propensity to be resistant to RIF. Thus the identified polymorphisms associated with a given drug may be conferred by another drug, resulting in false positives. Hence, we aim to develop a phylogeny-based probabilistic model to identify novel polymorphisms within an optimal window that maximizes homoplasy impact associated with drug resistances driven by positive selections. Our approach is expected to account for the population stratification and the co-resistance between drugs to find the sets of loci conferring corresponding drug resistance with higher sensitivity and specificity.

Table 4.1: Most frequent resistance mutations observed for several anti-tuberculous drugs.

Antibiotics	Mutations
INH	*katG*: S315T, S315R; *inhA* promoter: t-8c, c-15t, g-17t; *inhA*: S94A, I194T, I21T
RIF	*rpoB*: RDRR; *rpoC*
EMB	*embB*: M306V, M306I, G406S, G406A; intergenic region between *embC* and *embA*
STR	*rrs*: A514C; *rpsL*: K43R, K88T; *gidB*: nonsynonymous mutations
PZA	*pncA*: nonsynonymous mutations and indels
KAN	*rrs*: A1401G; mutations in the upstream of *eis* (UTR)
Fluoroquinolones	*gryA*: A90, D94
ETH	*ethA*, *inhA* promoter
PAS	*folC*; *thyA*;

4.2 Methods

To identify loci associated with drug resistance, a linear mixed model and a phylogenetic-based convergence test (phyC) are evaluated by three types of genotypes, including an individual polymorphic site (site-based), a single gene (gene-based) and a k-mer (k-mer-based). A k-mer means a sequence of k neighboring sites pooled as a pseudo site. For pooling sites as a gene or a pseudo site of k-mer, a strain that has at least one mutation in the sites within the boundaries will be marked as having a mutant.

We implement a convergence test using the concept in phyC as follows. Given a set of strains, we align them with a reference genome to obtain a multiple sequence alignment. We apply the maximum parsimony or maximum likelihood method on the alignment of polymorphisms to reconstruct the phylogenetic tree. We then apply Sankoff's algorithm [36] to determine both the genotypic and phenotypic states for each internal nodes in the tree. We traverse the tree to obtain branches where changes occur and record corresponding phenotypic states for each polymorphic site. The numbers of phenotypic states of branches are used in the hypergeometric model to calculate the probability of each site.

4.3 Results

4.3.1 Evaluation of Three Existing Methods Using Simulated Datasets

To better understand the performance of the linear model with principal components, linear mixed model and convergence test, we create a simulated dataset with homoplasy patterns. We first generate a bifurcating tree of 15 taxa (Figure 4.2) using GenPhyloData [37]. We then simulate 10,000 sites based on the tree using a HKY85 model [40] with unequal nucleotide substitution frequencies and unequal frequencies of transitions and transversions (2:1) by Seq-gen[39]. We obtain 5616 polymorphic sites. Six taxa are labeled as cases (R) while nine are labeled as controls (S). The linear model with PCA and linear mixed model are implemented by python scripts and the program GEMMA [71]. The plot of accumulated variances of PCA and the scatter plot of the top two components of 15 taxa are shown in Figure 4.3 (a)-(b), respectively. We apply the top two principal components which account for 60% of variances as the covariants in the linear regression model to reduce the effect of population structure. The heatmap of genetic relatedness matrix is shown in Figure 4.4. It is calculated by the simulated genotypes and used as random effects in the LMM.

Six genotypic patterns are shown in Table 4.2. Taxa of mutants are all resistant in the genotype 1, making a perfect correlation (100%). The rest of the genotypes have four mutants in different taxa, but all are labeled as resistant. Thus the frequency distribution of phenotypes and genotypes in the contingency table are the same for the genotypes 2 to 6, which will result in the same p value by the linear regression, the Fisher's exact test, and the chi-squared test. However, the homoplasy levels are different among them. The genotype 2 has no homoplasy. The mutations in the genotypes 3 and 4 occur in two branches so the number of excess changes is 1. The mutations in the genotypes 5 and 6 occur in three branches, indicating a higher level of homoplasy. Results tested by the linear model with principal components (LM_PCA), the LMM and the convergence test (phyC) are shown in Table 4.3. The linear model with PCA distinguishes the genotype 2 from the genotypes 3 to 6. It estimates a lower significance for the genotype 2 since taxa with

mutations are more clustered in the genotype 2 than other genotypes. However, the genotypes 2, 3 and 4 have higher significances estimated from the LMM while the genotypes 5 and 6 have higher homoplasic extents. In the convergence test, patterns of higher homoplasic extents have lower p values, suggesting a higher level of significance. In sum, the genotypes 2 to 5 have the same proportion of mutants associated with resistance, but the genotypes 3 to 6 should be more significant because they are homoplasic.

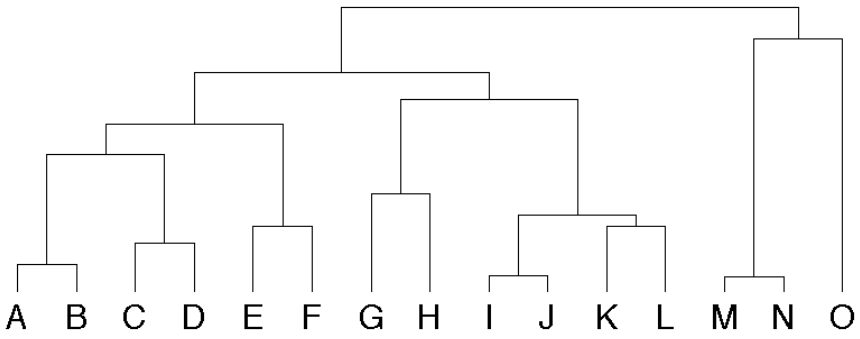
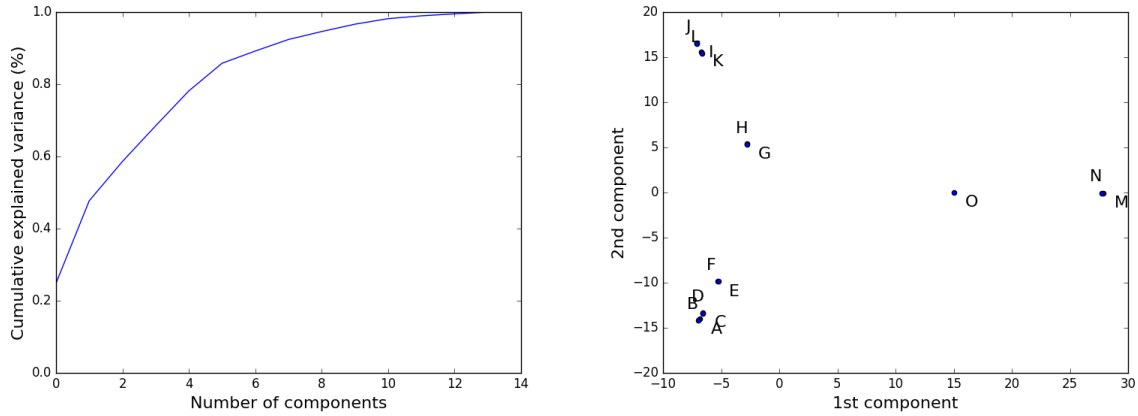


Figure 4.2: Tree of 15 taxa generated based on a birth-death process of rate 3:1 for evaluation.



(a) (b)

Figure 4.3: Plot of accumulated variances (a) and the scatter plot of the top two components (b) for 15 taxa.

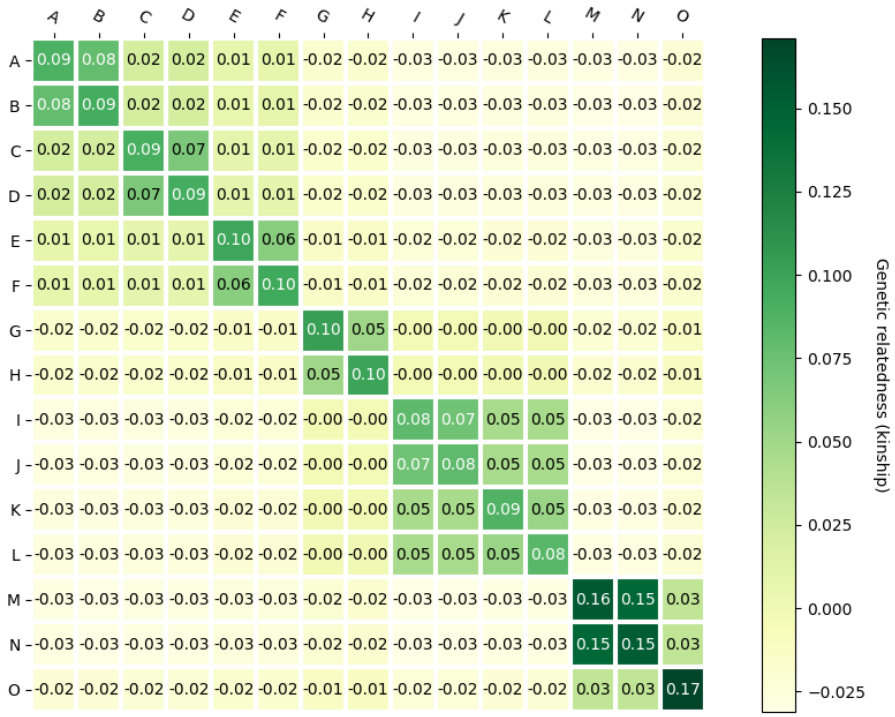


Figure 4.4: Heatmap of the genetic relatedness matrix (kinship).

Table 4.2: Phenotypes and genotypes of 15 taxa.

Taxa	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	Excess changes
Phenotype	S	S	S	S	**R**	**R**	S	S	**R**	**R**	**R**	**R**	S	S	S	
Genotype 1	C	C	C	C	T	T	C	C	T	T	T	T	C	C	C	1
Genotype 2	C	C	C	C	C	C	C	C	T	T	T	T	C	C	C	0
Genotype 3	C	C	C	C	T	T	C	C	T	T	C	C	C	C	C	1
Genotype 4	C	C	C	C	T	T	C	C	C	C	T	T	C	C	C	1
Genotype 5	C	C	C	C	T	T	C	C	T	C	T	C	C	C	C	2
Genotype 6	C	C	C	C	T	C	C	C	T	T	T	C	C	C	C	2

Table 4.3: Results estimated from LM_PCA, LMM and phyC.

Genotype	LM_PCA_es [a]	LM_PCA_p [b]	LMM_es [c]	LMM_p [d]	N [e]	R [f]	S [g]	phyC_p [h]
1	1.0	0.0	1.0	0.0	2	2	0	0.085
2	0.924	0.084	0.841	0.011	1	1	0	0.284
3	0.655	0.003	0.416	0.036	2	2	0	0.085
4	0.667	0.002	0.450	0.018	2	2	0	0.085
5	0.660	0.003	0.137	0.247	3	3	0	0.025
6	0.586	0.041	0.241	0.172	3	3	0	0.025

[a] effect size estimated from the linear model with PCA;

[b] p value estimated from the Wald test in the linear model with PCA;

[c] effect size estimated from the linear mixed model; [d] p value estimated from the Wald test in the linear mixed model; [e] number of branches in which changes occur;

[f] number of branches in which changes occur and are labeled as resistant;

[g] number of branches in which changes occur and are labeled as sensitive;

[h] p value estimated from the convergence test

4.3.2 Identifications of Antibiotic Resistant Polymorphisms in *Mycobacterium tuberculosis*

The empirical dataset contains 660 *M. tuberculosis* strains from Lima, Peru. They are aligned to the reference genome H37Rv. There are 19933 polymorphisms, excluding gaps and ambiguous sites. The phylogenetic tree labeling with lineages is shown in Figure 4.5 where most strains are categorized to lineage 2 or lineage 4.

The clinical phenotypic dataset consists of drug susceptibility test (DST) of 7 antibiotics. The distributions of drug susceptibility tests of INH, RIF, EMB, STR, PZA, KAN and CPX are listed in Figure 4.6. The DST data of KAN and CPX are only available for a subset of strains.

The heatmap of correlation coefficients between pairs of anti-tuberculous drugs is shown in Figure 4.7. The correlation coefficient between INH and RIF is about 0.87, indicating high co-resistance of antibiotic susceptibilities. Also, many coefficients of pairwise first-line drugs are larger than 0.5, suggesting that they have medium to high levels of correlation with most of each other.

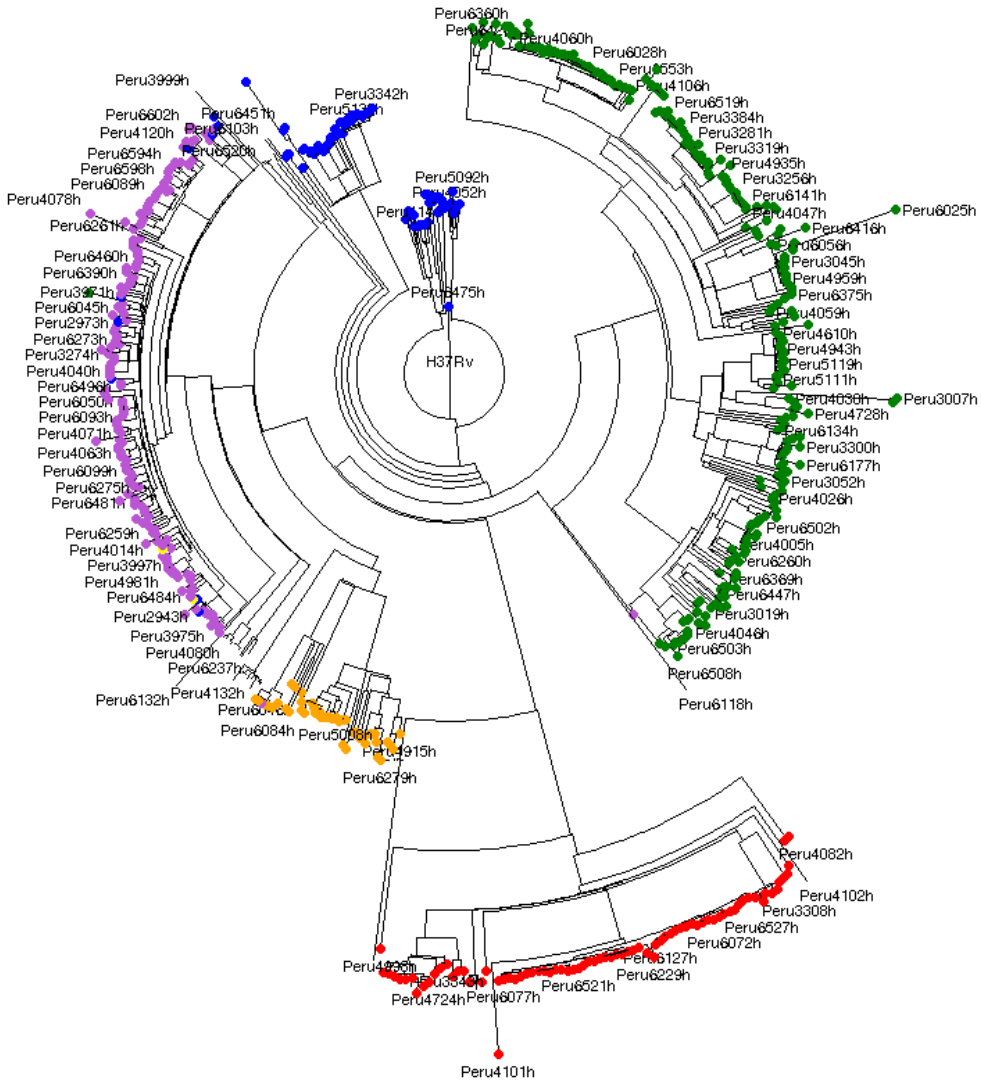


Figure 4.5: Phylogenetic tree and the distribution of lineages of 660 clinical isolates from Peru. The number of isolates and labeling color for each lineage is as follows: Red: Beijing (78); green: LAM (255); purple: Haarlem (167); blue: T-clade (82); orange: X-clade (42); yellow: H-clade (2); none: unrecognized (34).

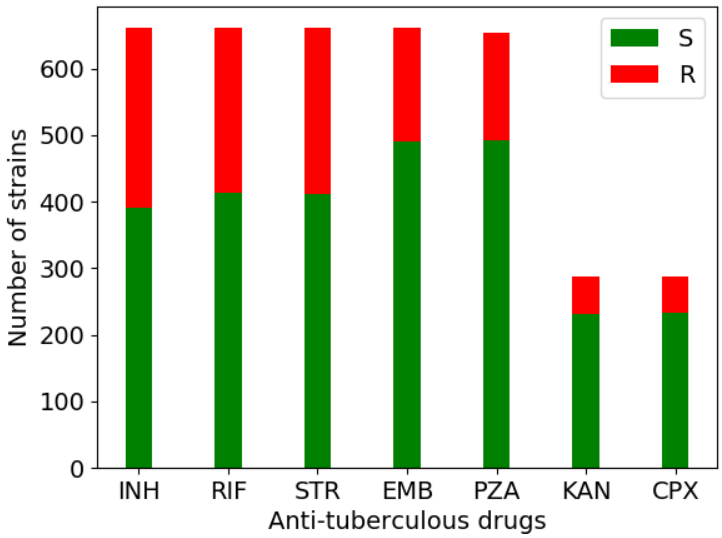


Figure 4.6: Distribution of drug susceptibility in the Peru dataset of 660 strains. KAN and CPX are available for only a subset of 286 strains.

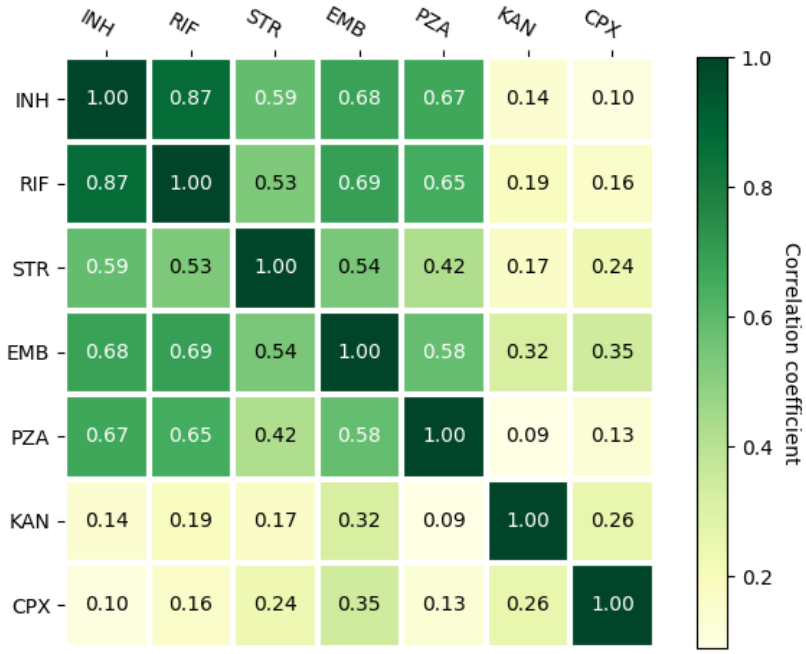


Figure 4.7: Heatmap plot of pairwise correlations between drugs. Each cell represents the correlation between a pair of drug susceptibilities. Darker green presents stronger co-resistance between drugs for strains. The correlation between INH and RIF is 0.87, suggesting that many strains are resistant to INH and RIF or sensitive to both of the drugs.

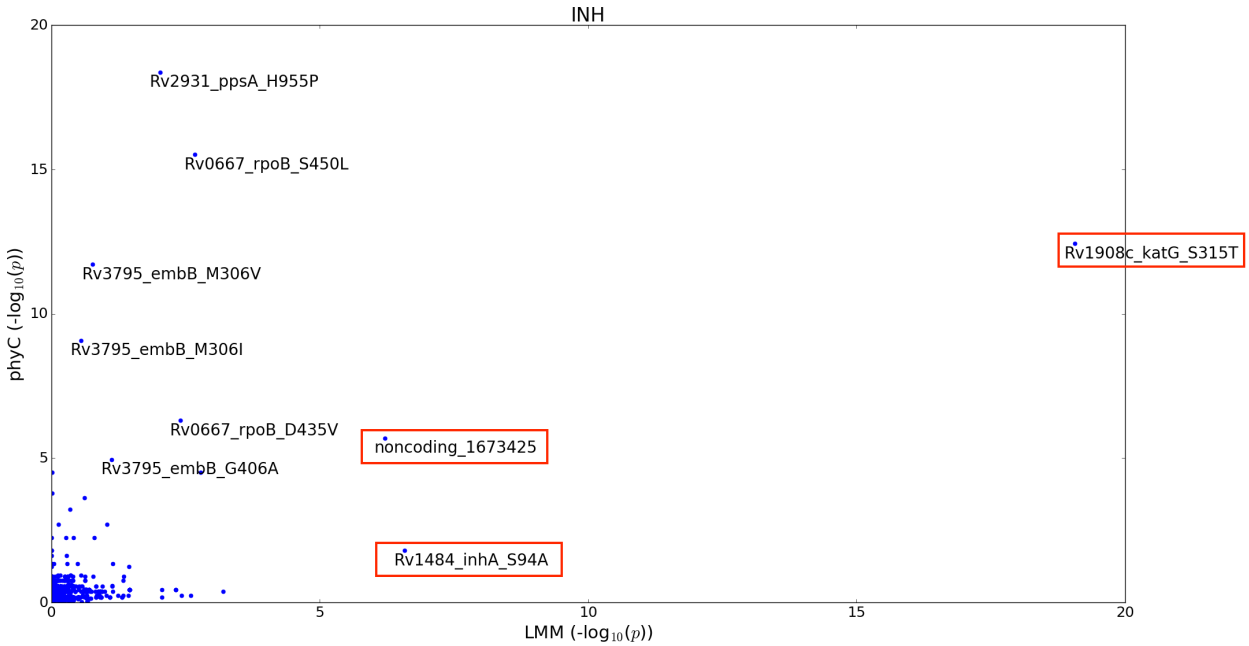
Three types of genotypes, site-based, gene-based and k-mer-based, and three anti-tuberculous drugs are tested in association mappings using LMM and the convergence test (Figure 4.8, Figure 4.9 and Figure 4.10). The scatter plots of negative logarithm of *p* value for each site, gene or pseudo site associated with isoniazid are shown in Figure 4.8 (a-c), respectively. The x-axis is the negative logarithm of the *p* value estimated from the LMM and the y-axis represents the negative logarithm of the *p* value estimated from the convergence test. In these plots, ideally, we want known resistant-associated mutations to be significant based on both tests (LMM in x-axis and phyC in y-axis), and hence to appear as far toward upper-right as possible. Annotated loci conferring INH resistance are *katG* mutations at codon 315 (S315T), the promoter region of the gene *inhA* (c-8t, c-15t, g-17t), and *inhA* mutations at codons 21, 94 and 194. In the site-based association test (Figure 4.8 (a)), both methods report *katG* mutations at codon 315 associated with INH, yet they also report other polymorphisms that are involved in other drug resistances, suggesting false positives. In Figure 4.8 (b), noncoding_1699 is the intergenic region between *Rv1482c* and *fabG1*, noncoding_4191 is the intergenic region between *embC* and *embA*, noncoding_2693 is in the *Rv2418-Rv2419c* region and noncoding_2948 is at coordinate near 2965856. Gene *katG* and intergenic region between *Rv1482c* and *fabG1* (noncoding_1699) are identified with strong associations from both methods in the gene-level test. Figure 4.8 (c) shows the results of the k-mer-based test where mutations at codon 315, noncoding_1673423 and noncoding_1673425 (loci in the *Rv1482c-fabG1* intergenic region) are identified by both methods.

Figure 4.9 (a-c) are the scatter plots of negative logarithm of *p* value for each site, gene or pseudo site associated with rifampicin estimated from convergence test (y-axis) and LMM (x-axis), respectively. Well-known mutants involving in RIF resistance are loci within *rpoB* RDRR region (region determining rifampicin resistance, codons 435-450). In Figure 4.9 (a), both methods identify S450L, D435V and H445D from the site-based test. The noncoding_4243217 is in the intergenic region between *embC* and *embA*. Pooling sites within a gene enhances the overlapping level of allele counts and resistant counts yet decreases the homoplasic extent for *rpoB* in Figure 4.9 (b). But when we group adjacent sites together (k-mers = 3), more loci in the *rpoB* RDRR region

(D435Y, D435V, H445D, R448Q) show stronger associations (rank near top) in Figure 4.9 (c). The mutations at the loci only occur in a few isolates with low level of homoplasy, but they are known to contribute RIF-resistance in RDRR.

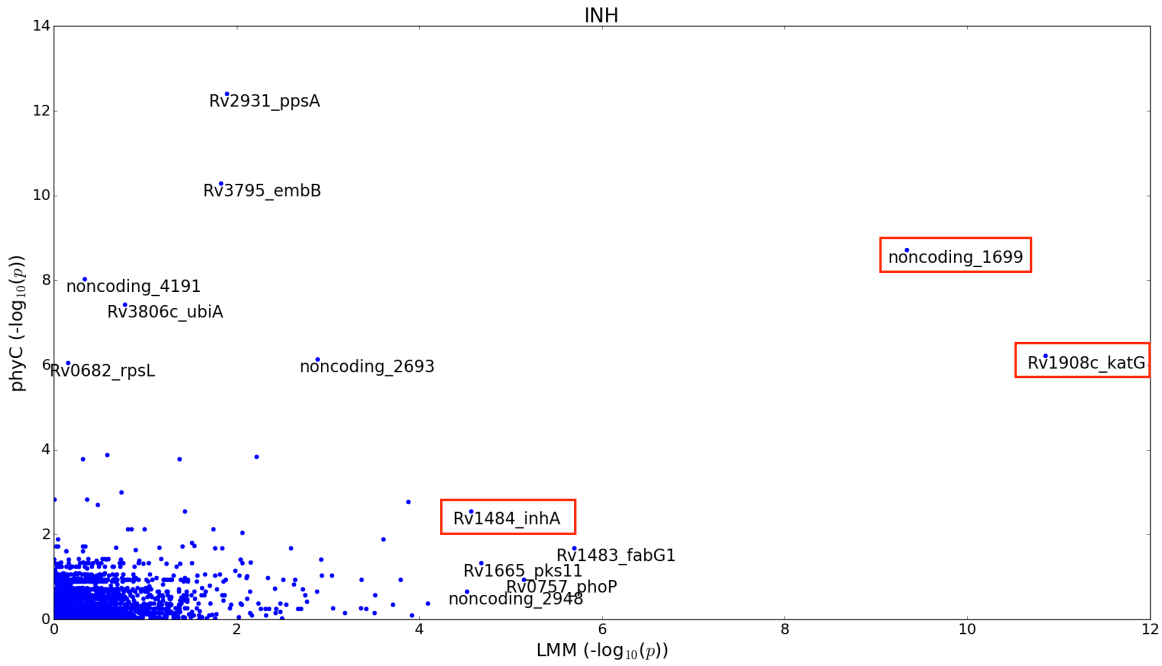
The scatter plots of negative logarithm of p value for each site, gene or pseudo site associated with ethambutol estimated from convergence test (y-axis) and LMM (x-axis) are shown in Figure 4.10 (a-c), respectively. Mutations in the *embB* operon (codons 306 and 406) and *embC-embA* intergenic region are known to be related to EMB resistance. In the site-based association test, convergence test identifies strong associations between *embB* mutations at codon 316 and EMB resistance while LMM is unable to do that (Figure 4.10 (a)). The gene-based test in Figure 4.10 (b) indicates gene *embB* has the highest rank in both tests, yet the *embC-embA* intergenic region (noncoding_4191) is significant in the convergence test but not in the LMM. Results of the k-mer-based association test in Figure 4.10 (c) demonstrate that the convergence test performs better than the LMM in terms of identifying causal mutations associated with EMB resistance.

The results in Figure 4.8-Figure 4.10 show that the k-mer-based association test performs better than the site-based or gene-based test, suggesting the feasibility of grouping part of SNPs within a gene optimally to identify loci associated with drug resistance.

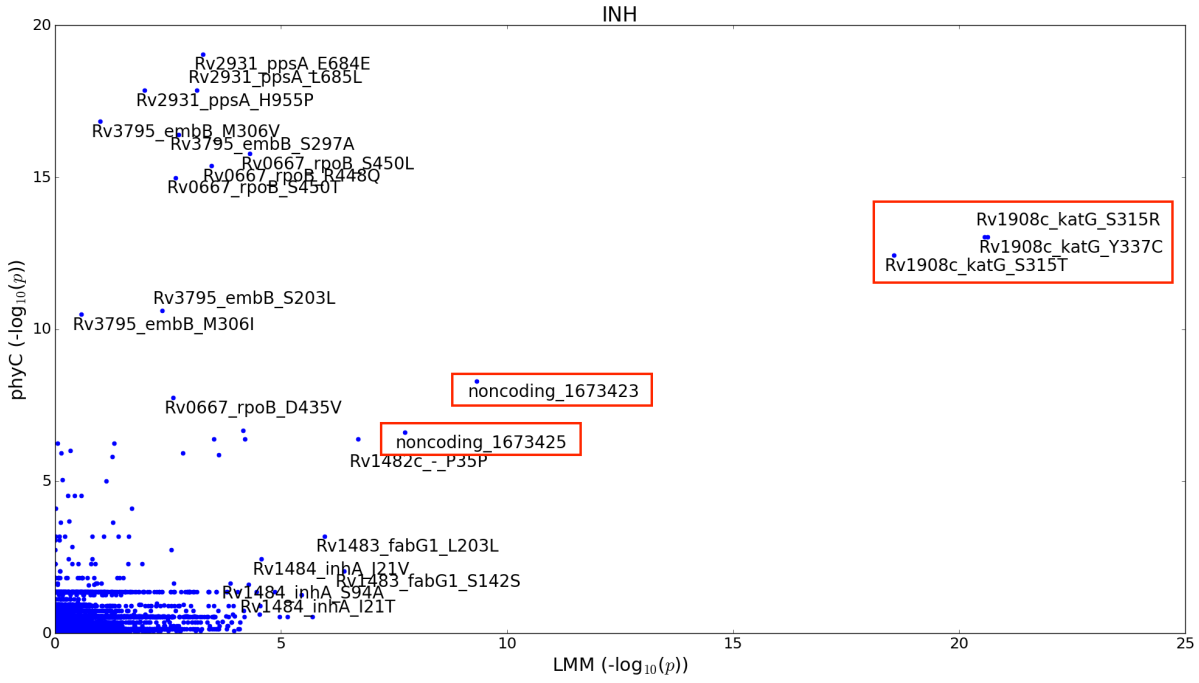


(a)

Figure 4.8: Scatter plots of association mapping between INH and (a) single site, (b) individual gene and (c) pseudo site of 3-mer in *M. tuberculosis* using LMM and phyC. The x-axis and y-axis represent the negative logarithm of p values from two association tests, respectively. Genotypic traits that are relatively associated with the phenotype are labeled with the gene annotations or coordinates for intergenic regions.

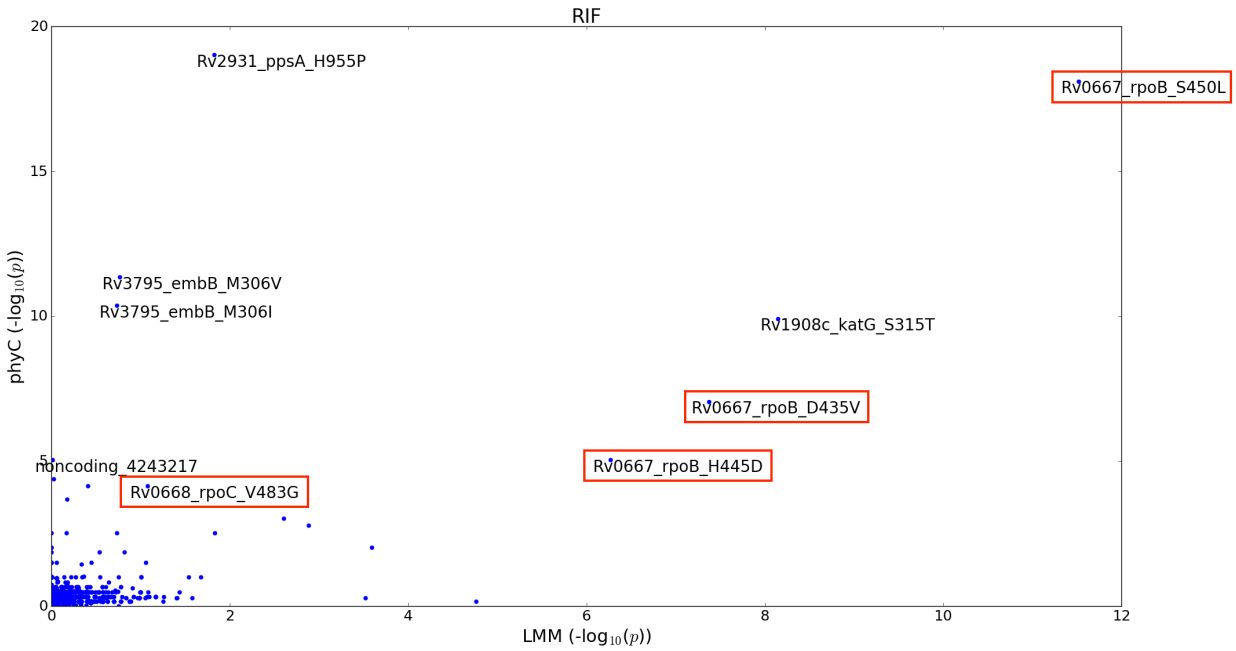


(b)



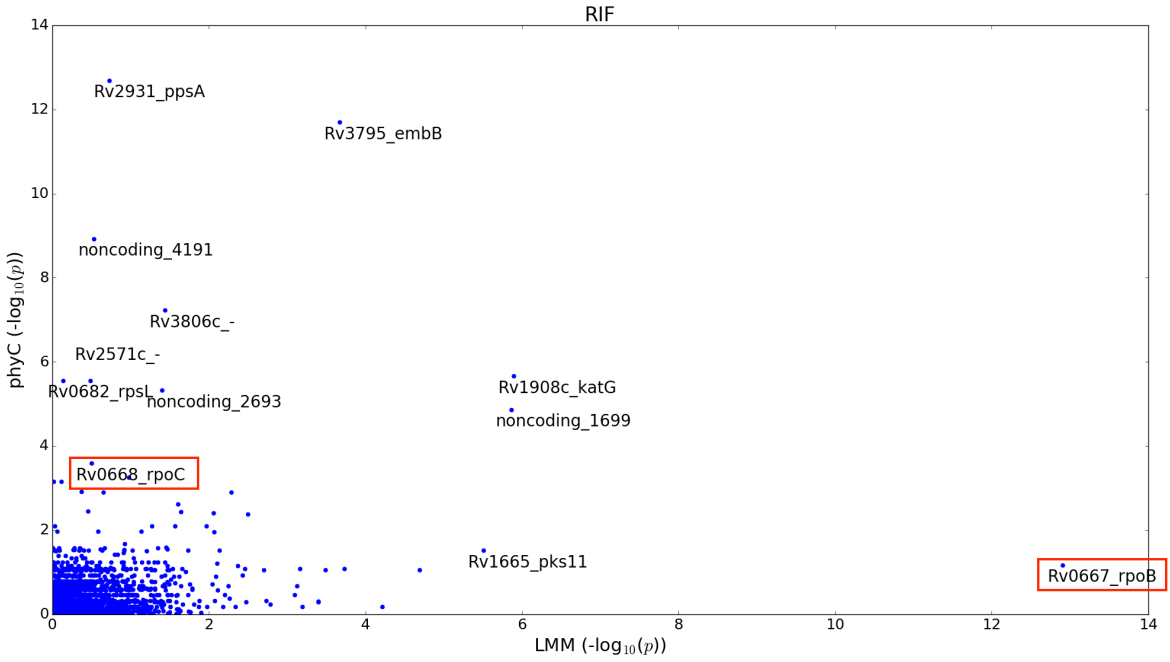
(c)

Figure 4.8: Continued.

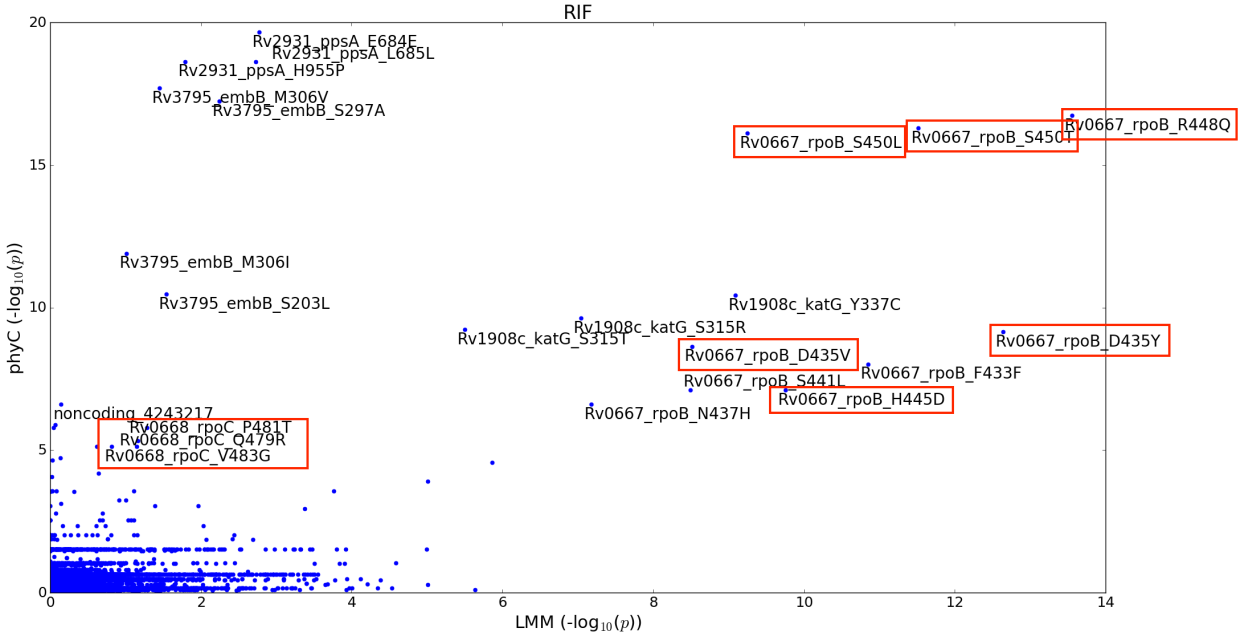


(a)

Figure 4.9: Scatter plots of association mapping between RIF and (a) single site, (b) individual gene and (c) pseudo site of 3-mer in *M. tuberculosis* using LMM and phyC. The x-axis and y-axis represent the negative logarithm of p values from two association tests, respectively. Genotypic traits that are relatively associated with the phenotype are labeled with the gene annotations or coordinates for intergenic regions.

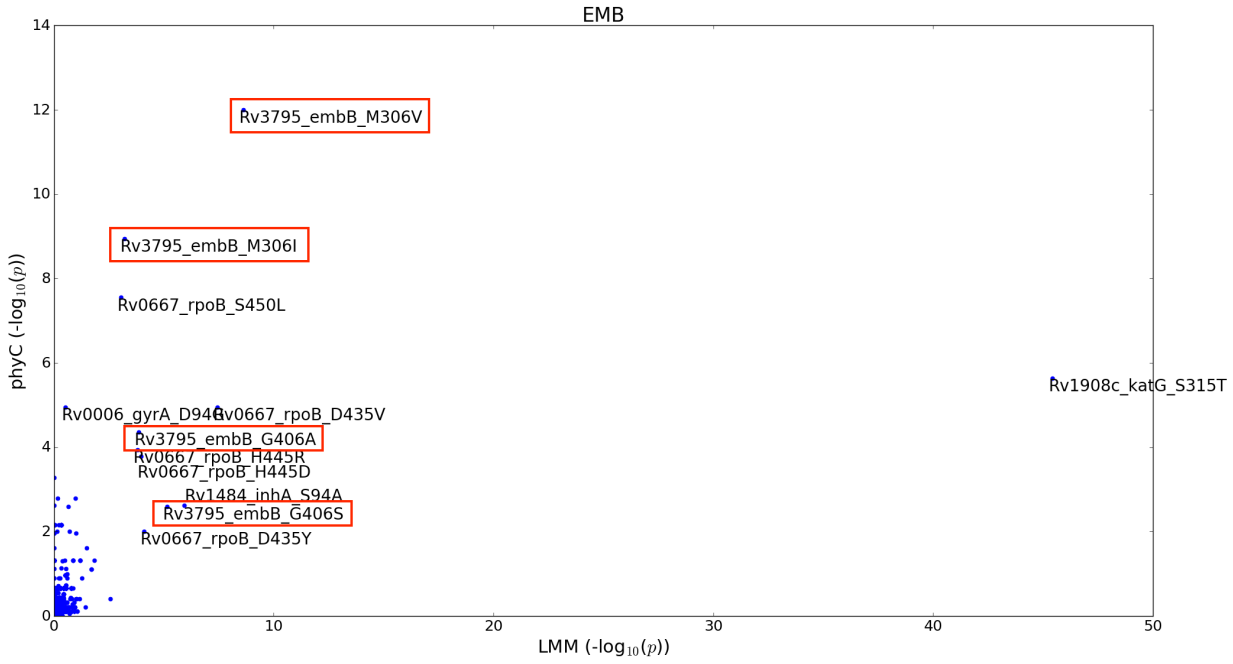


(b)



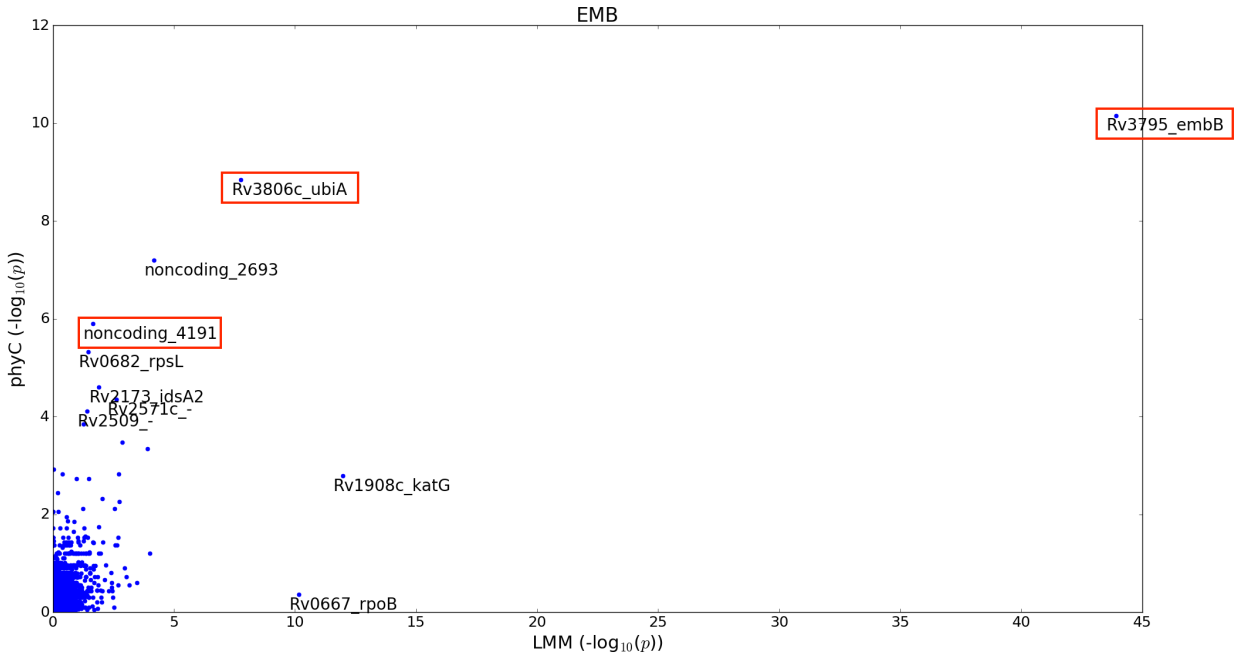
(c)

Figure 4.9: Continued.

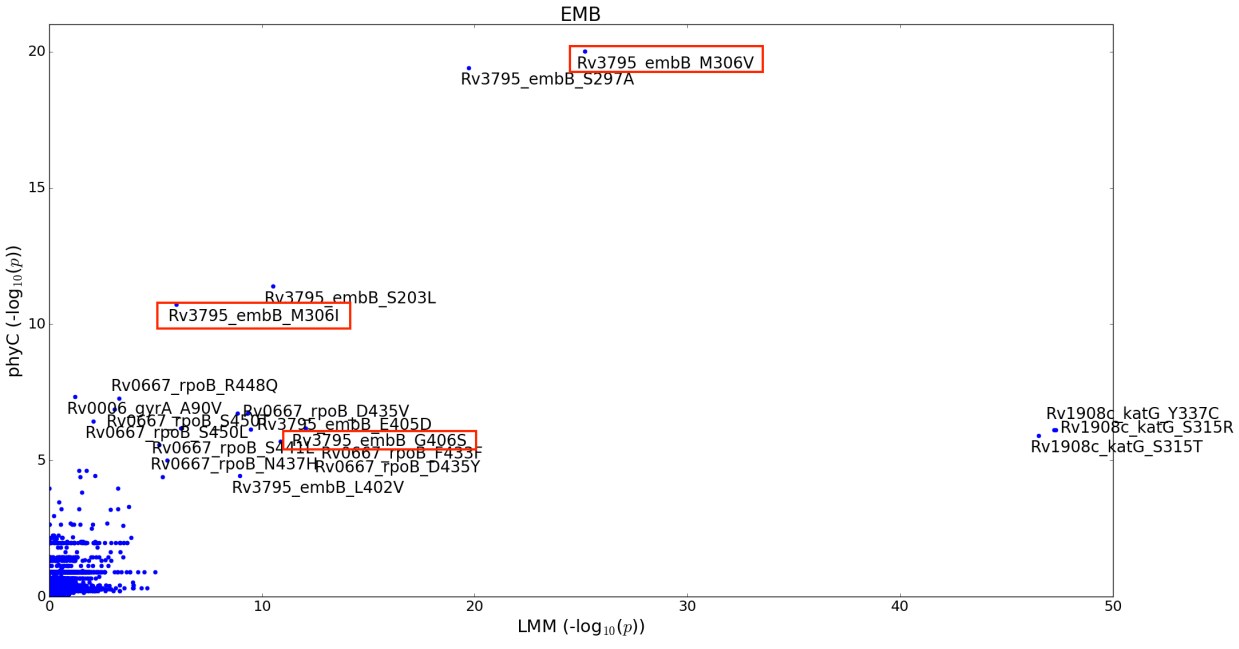


(a)

Figure 4.10: Scatter plots of association mapping between EMB and (a) single site, (b) individual gene and (c) pseudo site of 3-mer in *M. tuberculosis* using LMM and phyC. The x-axis and y-axis represent the negative logarithm of p values from two association tests, respectively. Genotypic traits that are relatively associated with the phenotype are labeled with the gene annotations or coordinates for intergenic regions.



(b)



(c)

Figure 4.10: Continued.

4.4 Optimized Grouping of SNPs for Genome-wide Convergence Test

To identify drug-resistant loci across an entire genome, a genotypic trait used in an association test is usually an individual allelic SNP or a gene-level grouping of SNPs. However, previous results show the k-mer-based approach performs better than either site-based or gene-based methods. The k in the k-mer is a fixed number of k adjacent sites as a window, yet sites that confer drug resistance vary from individual locus to a group of sites. Thus, we optimize the grouping of SNPs to maximize significance in some cases where a homoplasic site by itself is enough and in other cases where several adjacent sites with mutations are related to antibiotic resistance. That is, we extend the window size of a fixed k to every possible k-mer where k ranges from 1 to the size of a gene to test the optimized grouping of k SNPs against a certain phenotype. The pseudocode of the algorithm is in Algorithm 2 and the time complexity is $O(s^2m)$, where s is the total number of SNPs within a gene or an intergenic region, and m is the number of all genes and intergenic regions.

Algorithm 2 Optimized Grouping of SNPs within a Gene or an Intergenic Region

Input: An alignment (Aln) of n SNPs (χ_j) in m genes and intergenic regions (G_i)
function OPTIMIZEDGROUPING(G_i in Aln)
Aln has m regions (G_i), where i ranges from 1 to m;
G_i has s SNPs (χ_j), where j ranges from 1 to s;
for all $i \leftarrow 1$ to m **do** ▷ for each gene, $gene_i$
for all $k \leftarrow 1$ to s **do** ▷ s is the number of SNPs in each gene
for all $l \leftarrow k$ to s **do** ▷ all combinations of groupings
χ'_l = changes of χ_l ▷ number of branches where changes occur in the tree
$g_{ikl} \leftarrow \sum_{p=k}^{l} \chi'_p$ ▷ a sub-region between SNP_k and SNP_l in the $gene_i$
Conduct an association test on the sub-region (g_{ikl}) with a particular phenotype
end for
end for
Identify the sub-region of strongest association (g'_{ikl}) within the $gene_i$ with a particular phenotype
end for
return sub-regions of optimized grouping of SNPs for all genes or intergenic regions (g'_{ikl} for each G_i)
end function

4.4.1 Associations between Groupings of SNPs within *rpoB* and RIF Resistance

We evaluate the optimized grouping method using the phenotypes of rifampicin susceptibility. The associations between all possible groupings of SNPs within the *rpoB* gene and the rifampicin resistance are shown in Figure 4.11. Each cell in the heatmap represents a negative logarithm p value of association of a sub-region g_{ijk} for the window from SNP_j to SNP_k within a region G_i. Forty-four SNPs are found in the *rpoB* gene and eleven SNPs are found in the RDRR region (codons 435-450). The most significant association (presented in the darkest green) occurs in the grouping of eight SNPs from N437H to S450L locating within the RDRR region. The association of the entire gene against RIF resistance is presented in the most bottom-right cell with light green, suggesting little association. The association of the mutation at codon S450L by itself against RIF resistance is presented in the 27^{th} diagonal cell with relatively darker green, yet it is not the darkest one. The p values of the grouping of SNPs from N437H to S450L, mutation at the codon S450L, and the grouping of all SNPs within the *rpoB* are 4.84×10^{-19}, 7.41×10^{-17}, and 0.10, respectively. In other words, mutation at codon S450L by itself has a strong association with RIF resistance, but it is strengthened by grouping SNPs within the RDRR region for *rpoB*.

4.4.2 Associations between Groupings of SNPs and Other Anti-tuberculous Drugs

For INH resistance, the best grouping of SNPs within *katG* that maximizes the association is from codons G120S to Y337C with 17 SNPs in total, where forty-four SNPs occur in *katG* spanning 2223 bp. The subregion has 63 changes where 53 occur in the INH-resistant branches. This is consistent with reports that other nonsynonymous mutations besides S315T in *katG* can confer resistance [79]. The p values from the association test for three genotypes within the *katG* of the best grouping of SNPs, S315T by itself, and the grouping of SNPs within the whole gene are 2.27×10^{-15}, 3.69×10^{-13} and 2.85×10^{-6}, respectively. The size of the upstream of the *inhA* promoter is 134 bp and 3 positions harbor mutations with 33 changes, where 28 are in the INH-resistant branches. The strongest association happens when grouping all three mutations together as a sub-region where the p value equals to 7.12×10^{-9}.

For EMB resistance, we expect to see higher association occurring in *embB* at codons 306 and 406. The size of the *embB* gene is 3297 bp and 40 mutations are obtained during evolution. The strongest association happens in the sub-region of 16 consecutive mutations between codons S297A and V493M resulting in 72 changes where 56 are resistant to EMB. The sub-region includes known causal loci of M306V/I and G406S/A. The p values for the best grouping (S297A-V493M), local grouping of M306V/I, and local grouping of M306V/I are 1.20×10^{-23}, 2.25×10^{-19}, and 8.20×10^{-7}, respectively.

4.4.3 Summary

We perform association tests on groupings of SNPs against anti-tuberculous drugs and the results provide evidence that optimizing the grouping of allelic sites enhances the detection of drug-resistant loci. However, three limitations arise. Since we optimize over many windows per gene, the time complexity is quadratic for looping over all possible consecutive groupings. Also, the empirical p values will be reduced by the multiple testing correction to adjust the *p* values since so many significance tests are being performed in parallel within each gene. Lastly, the grouping of adjacent polymorphic sites in a gene does not consider the span of the gene. To overcome these challenges, the method will be extended in the next chapter.

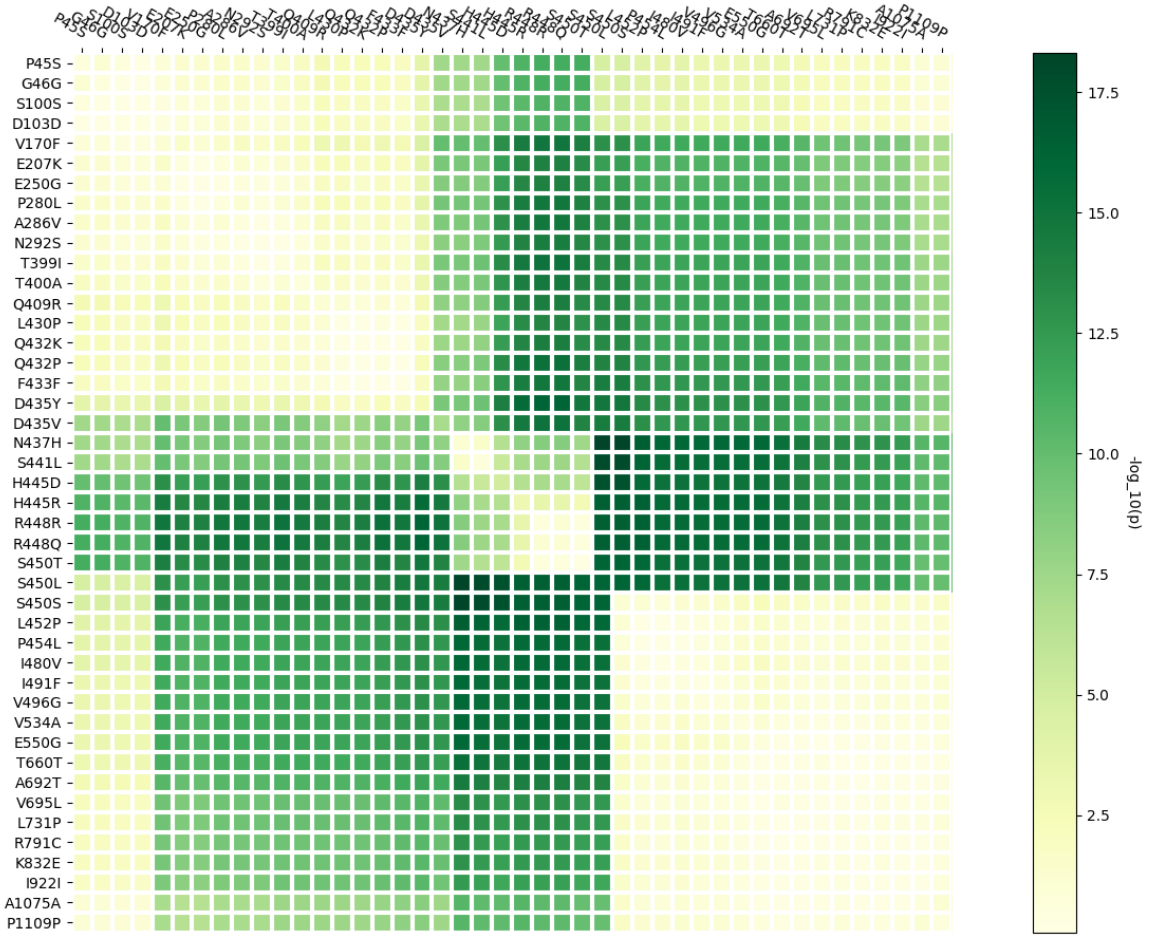


Figure 4.11: Heatmap plot of associations between the genotypes of all possible groupings of SNPs within the *rpoB* gene and the phenotype of rifampicin suscetibility. A square cell represents the negative logarithm of p value from the association test of the grouping of SNPs between two codons. A cell in diagonal presents the association between phenotype and genotype of an individual site while the most bottom-right cell presents the genotype of grouping of all SNPs within the gene. The darker the green, the higher the association. The most significant association occurs in the region of grouping SNPs between codons N437H and S450L.

5. IDENTIFICATION OF DRUG-RESISTANT POLYMORPHISMS USING EVOLUTIONARY CONVERGENCE CLUSTERING

5.1 Introduction

In this chapter, we combine statistical tests of association of SNPs with drug resistance phenotypes with clustering of SNPs, as a secondary indicator of positive selection. As a bacterial genome evolves, it accumulates genetic variants across time. According to the neutral theory of evolution, the evolutionary changes are assumed to be randomly occurred and spread out the genome. A neutral change is a synonymous or silent nucleotide substitution (mutation) which encodes the same amino acid and thus is presumed to have little or no effects on fitness. The other type of change is a nonsynonymous mutation that encodes different amino acid, which is more likely to be affected by positive selection, including virulence, adaptation to immune response and drug resistance.

To identify regions harboring mutations under selection pressure, Wagner developed a method called variation clusters by assuming that genomes neutrally evolve with mutations spontaneously [80]. The null hypothesis is that mutations/nucleotide substitutions occurring within a given span of sites in the DNA sequence follow a Poisson distribution. If nucleotide substitutions observed in a region are more abundant than expectations, then the mutations are unlikely to occur by chance within the region. Higher deviations from the expectation in the Poisson model indicate more adjacent mutations are clustered closely within a region, suggesting positive selection exists.

In a bacterial population, the evolutionary history can be estimated from a phylogenetic tree reconstructed based on the whole-genome polymorphisms. Some polymorphisms are incongruent with the tree topology and are defined as homoplasic changes. Homoplasy occurs when genetic changes are descended in different branches instead of from a common ancestor during evolution. Homoplasic polymorphisms often reflect positive selection. Grandjean et. al. detect homoplasic polymorphisms across the entire genome using a convergence method based on the disruption of a phylogenetic tree. Many well-known drug-resistant loci are observed as being homoplasic among

the dataset of a high proportion of multidrug-resistant *Mycobacterial tuberculosis* strains [81], suggesting they are under selection pressure. The convergence test, phyC, looks for evolutionary changes in a phylogenetic tree that are related to antibiotic resistance [16]. However, phyC as a statistical test does not get the advantage of homoplasy since it applies genotypic traits as either an individual polymorphism or a gene-level pooling of SNPs by a burden test. It does not consider that in some cases a locus has enough homoplasic changes by itself while in other cases some adjacent sites have fewer homoplasic changes yet are related to drug resistance. It also does not consider the size of a gene, thus loses the ability to distinguish the cases where the same number of changes occurs within a large span or a small span of genes. Moreover, not all SNPs within a gene are related to the phenotype, and pooling them with the real associated SNPs together may decrease the true positive rate. Thus, optimizing the grouping of SNPs to maximize the significances of associations is essential.

A challenge arises, that is, how to identify optimized groupings of sites for association testings. Our previous optimization using window sizes of all possible k-mer within a gene (Ch. 4) has issues of multiple testing correction and computational complexity. All possible combinations of groupings in an alignment of *n* polymorphisms and *m* genes of size *p* at most take quadratic time ($O(mp^2)$). Thus, in this chapter, we develop a two-phase evolutionary cluster-based convergence test (ECC). We firstly identify regions of clustered SNPs across the whole genome using a Poisson distribution. Secondly, we test the associations of clustered regions as genotypes against antibiotic susceptibilities as phenotypes by applying a hypergeometric model. The association of a region with drug resistance is determined by the probability of observed changes relative to resistant than sensitive branches under the null distribution of the population modeled by a hypergeometric distribution. For correcting the multiple testing, we estimate false discovery rate (FDR) to adjust *p* values from both tests using the Benjamini-Hochberg procedure. We apply the cluster-based convergence method in three empirical datasets and evaluate its performance with previous site-based, gene-based and k-mer-based methods.

5.2 Methods

Our two-phase evolutionary cluster-based convergence algorithm is shown in Algorithm 3.

5.2.1 Phase 1: Clustered Region Identification

Given a phylogeny reconstructed from polymorphic sites in a multiple sequence alignment (excluding ambiguous sites, repetitive regions and known drug-resistant loci), we apply Sankoff's algorithm to determine ancestral states of internal nodes/branches in the phylogenetic tree. For each locus, we obtain the number of branches where changes occur. If mutations occur in more than one branch in the tree, then the locus has over one change and is defined as homoplasic. The probability of adjacent changes occurring within a span of a genome can be modeled as a Pearson type III distribution [80]. Given *m* mutations occurring in a genome of *n* nucleotides, the mutation rate λ is estimated by the total number of mutations evolved over the entire genome, which equals to $\frac{m}{n}$. The probability density of a span of length *x* containing *k* consecutive mutations equals to

$$P(x) = \frac{\lambda}{\Gamma(k-1)}(\lambda x)^{k-2}e^{-\lambda x} \tag{5.1}$$

, where $\lambda = \frac{m}{n}$ and $\Gamma(k) = (k-1)!$.

The probability of *k* changes occurring within regions smaller than the size of d_k can be estimated from the accumulated Pearson type III distribution as

$$P(d_k) = \frac{\lambda}{\Gamma(k-1)}\int_0^{d_k}(\lambda x)^{k-2}e^{-\lambda x}dx \tag{5.2}$$

It is also equivalent to be modeled as a Poisson distribution by estimating the accumulated probabilities where more abundant changes occur within a given span d_k.

$$P(d_k) = 1 - \sum_{i=0}^{k-2}\frac{(\lambda d_k)^i}{i!}e^{-\lambda d_k} \tag{5.3}$$

The k-cluster method is applied to the alignment of polymorphisms where *k* is the total number of changes within a region by grouping adjacent loci up to a given number of nucleotides. The

p values are adjusted by the FDR method (Benjamini-Hochberg procedure). We sort all regions by the adjusted p values into a list and then apply a greedy algorithm to examine each ordered region to obtain non-overlapping clustered regions. For each candidate region sorted in order, if the candidate region does not overlapped with previously examined regions, then the region is marked as examined. In contrast, the overlapping region is discarded from the list. Lastly, we obtain optimized clustered regions of changes from the examined regions where their adjusted *p* values are less than 0.05 as the default cutoff. Thus, a gene or an intergenic region might have none, one or more than one non-overlapping clustered sub-regions.

5.2.2 Phase 2: Association Test Based on the Evolutionary Convergence

We test associations of genotypes as clustered regions against phenotypes using a evolutionary convergence method similar to the phyC [16]. The permutation test in phyC to test the convergence of genotypes against phenotypes can be modeled as a hypergeometric distribution. The probability density of a clustered region whether the changes are related to a particular phenotype like drug resistance equals to

$$P(X = r) = \frac{\binom{N_R}{r}\binom{N-N_R}{k-r}}{\binom{N}{k}} \tag{5.4}$$

, where *k* is the number of changes in the region, *r* is the number of changes labeled as resistant in the region, *N* is the total number of changes, and *N_R* is the total number of changes labeled as resistant.

To determine whether a cluster of homoplasic changes resistant to a particular drug is overrepresented in the population, the *p* value is calculated as

$$P(X \geq r) = 1 - \sum_{i=0}^{r-1} \frac{\binom{N_R}{i}\binom{N-N_R}{k-i}}{\binom{N}{k}}, r \geq 1 \tag{5.5}$$

Algorithm 3 Two-phase evolutionary cluster-based convergence test (ECC)

Input: An alignment (Aln) of n polymorphisms (χ_j), a phylogenetic tree, phenotypes (DST)

Phase 1 – ClusteredRegions

Aln has n polymorphic sites
Given a maximum window size k
Calculate mutation rate λ
for all $i \leftarrow 1$ to n **do** ▷ for each locus
  for all $j \leftarrow$ i to $i + k$ **do** ▷ all possible groupings up to k adjacent loci
    χ'_j = changes of χ_j
    $g_{ij} \leftarrow \sum_{p=i}^{j} \chi'_p$
    Calculate the accumulated probability for the region (g_{ij}) by a Poisson distribution
  end for
end for
Apply FDR correction to adjust *p* values
Obtain non-overlapping clustered regions by a greedy algorithm (g'_{ij})
Identify significant clustered regions G_{ij} where $p_{adjusted}$ < 0.05 (the set of g'_{ij})

Phase 2 – ConvergenceTest

for all g'_{ij} in G_{ij} **do**
  Conduct an association test on g'_{ij} against DST using a hypergeometric model
end for
Adjust *p* values by the FDR correction
Identify regions of strong associations with a particular drug based on the adjusted *p* values (cutoff = 0.05)
Obtain clustered regions associated with drug resistance

5.3 Results

5.3.1 Genetic Variants, Lineages Distribution and Anti-tuberculous Drugs

We test our proposed two-stage model on an empirical dataset of *M. tuberculosis* from Lima, Peru, where pre-dominantly multi-drug resistant (MDR) strains were sampled. A MDR strain is defined as being resistant to both isoniazid and rifampicin. It consists of 660 strains with 7 drug susceptibility tests and minimum inhibitory concentrations. For genotypes, there are 21,501 polymorphic sites in the alignment, excluding ambiguous sites and repetitive regions in *PPE* and *PGRS* genes. We locate branches where mutations occur in the tree and obtain 23,847 changes (some sites have multiple changes). The phylogenetic tree labeling with lineages is shown in Figure 4.5 where most strains are categorized to lineage 2 (Beijing) or lineage 4 (LAM, Haarlem, X-clade, T-clade and H-clade). For phenotypes, the proportions of drug-resistant strains are shown in Figure 5.1. The proportion ranges from 18.2% (CPX) to 40.8% (INH). The MDR-TB accounts for 35.9% (237 / 660) in the population.

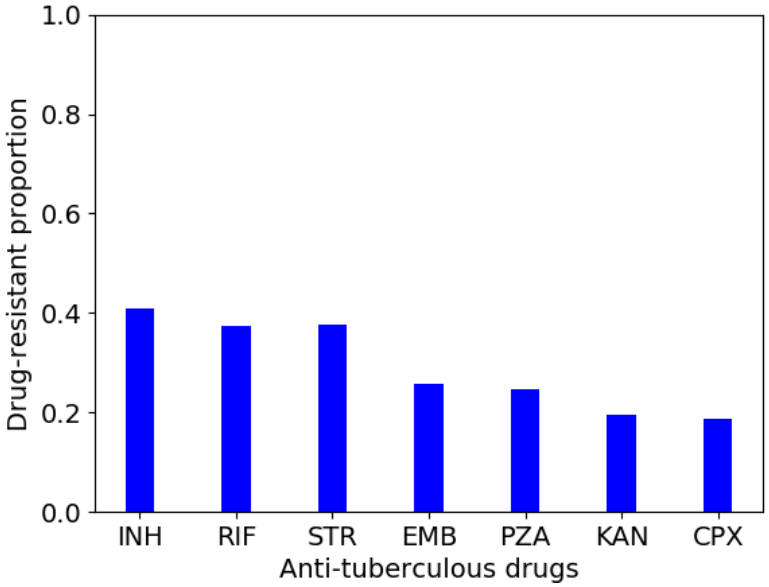


Figure 5.1: Proportion of drug-resistant strains for 7 drugs. The proportion ranges from 18.2% (CPX) to 40.8% (INH).

5.3.2 Identification of Optimized Clusters of SNPs

In the multiple alignment of 660 clinical isolates, regions that harbor variants occurring in the same strains in adjacent sites repeatedly are viewed as artifacts of sequencing and thus are also masked out, including Rv0095c, Rv1148c, Rv1945, Rv2048c (*pks12*), Rv2081c, Rv2543 (*lppA*), Rv2791c, Rv2931 (*ppsA*), and intergenic regions around coordinates of 104782, 332704, 838186, 1277749, 2030239, and 2338512.

The overall mutation rate is 0.54 % (23,847 changes / 4,411,532 bp). By applying the Poisson model, we obtain 643 clustered regions where their adjusted p values are less than 0.05 by the Benjamini-Hochberg correction. The distribution of clustered regions across the entire genome is shown in Figure 5.2 and the top clustered regions are listed in Table 5.1. The smallest size of a region is one site harboring from 3 up to 53 changes. The largest region spans 1213 bp with 17 changes in 15 sites. The well-known drug-resistant loci are identified as homoplasic with the specific clustered regions, including *gyrA*, *embB*, intergenic region of *embC-embA*, *rpoB* RDRR, *rpoC*, *katG*, *inhA* promoter region (*Rv1482c-fabG1*), *rpsL*, *gidB*, *pncA* and the upstream of *eis*. The gene *lldD2* is also identified as homoplasic, though it does not have any known relation to drug resistance [82].

5.3.3 Convergence Test for Clustered Regions for Individual Drugs

Not all clustered polymorphic regions are associated with drug resistance (e.g. *lldD2*). Therefore the second phase is to test regions for the significance of association with resistant to individual drugs. The convergence test is conducted on the 643 clustered regions for seven drugs individually and Manhattan plots of genetic associations with anti-tuberculous resistance of INH, RIF, STR, EMB, PZA, KAN and CPX are shown in Figure 5.3-5.18, respectively. In each figure, the y-axis presents the negative logarithm of the adjusted p value from our clustered association test. And the x-axis stands for the coordinates in base pairs over the whole genome. A blue point represents a clustered region that groups sites of more changes than expected from a Poisson distribution in terms of the size of the region regarding the mutation rate in a population. Points ranked higher in

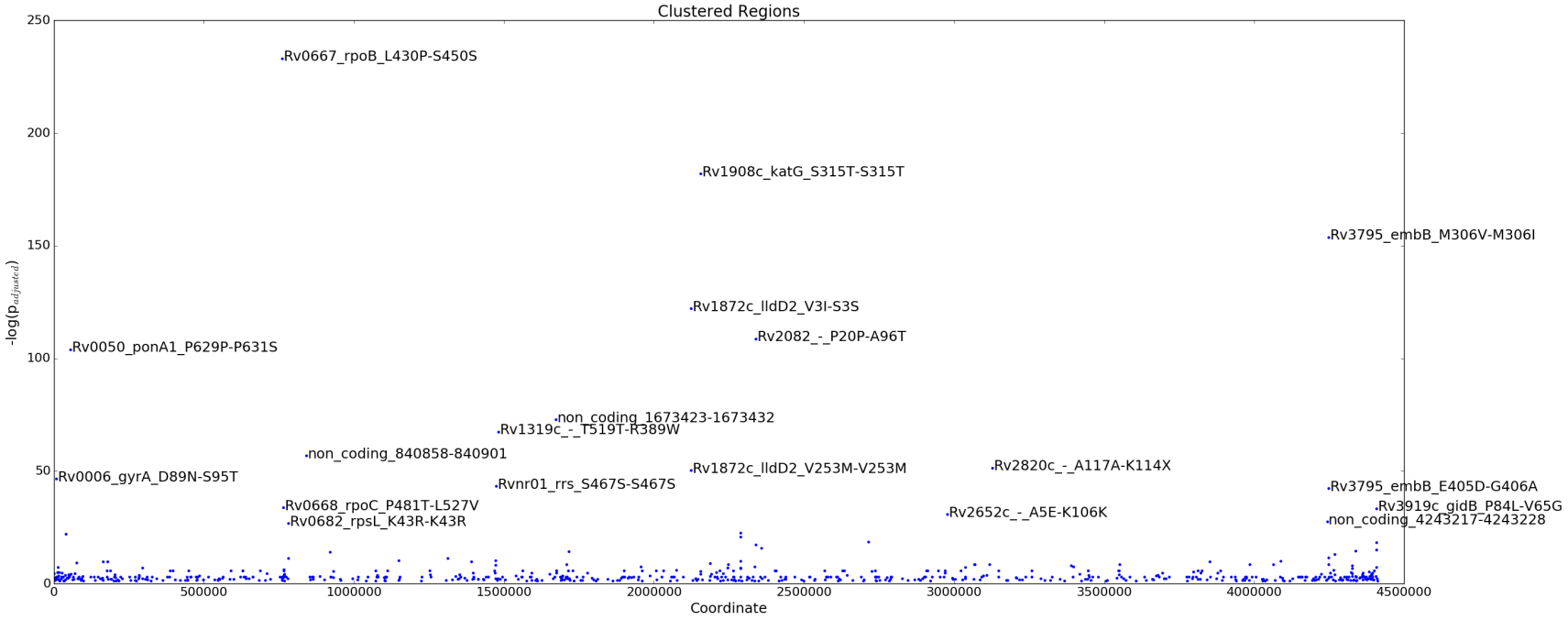


Figure 5.2: Manhattan plot showing non-overlapping clustered regions across the genome. Clustered regions of adjusted p values less than 5×10^{-19} are listed in Table 5.1.

Table 5.1: Top 25 non-overlapping clustered regions of 660 *M. tuberculosis* strains from Peru.

Coordinates	Gene / Intergenic	Codons	Changes	Size (bp)	Adjusted p value
761095-761156	Rv0667_*rpoB*	C:[L430P]-T:[S450S]	114	62	7.34×10^{-234}
2155168-2155168	Rv1908c_*katG*	G:[S315T]-G:[S315T]	53	1	9.45×10^{-183}
4247429-4247431	Rv3795_*embB*	G:[M306V]-A:[M306I]	52	3	1.56×10^{-154}
2123145-2123182	Rv1872c_*lldD2*	T:[V3I]-C:[S3S]	62	38	7.19×10^{-123}
2338768-2338994	Rv2082_-	T:[P20P]-A:[A96T]	81	227	1.78×10^{-109}
55549-55553	Rv0050_*ponA1*	T:[P629P]-T:[P631S]	40	5	1.02×10^{-104}
1673423-1673432	*inhA* promoter	1673423-1673432	33	10	1.58×10^{-73}
1480945-1481337	Rv1319c_-	G:[T519T]-A:[R389W]	66	393	3.50×10^{-68}
840858-840901	non_coding	840858-840901	35	44	1.36×10^{-57}
3127922-3127931	Rv2820c_-	G:[A117A]-A:[K114X]	25	10	3.95×10^{-52}
2122395-2122395	Rv1872c_*lldD2*	T:[V253M]-T:[V253M]	18	1	5.23×10^{-51}
7566-7585	Rv0006_*gyrA*	A:[D89N]-C:[S95T]	26	20	2.44×10^{-47}
2637541-2637541	non_coding	2637541-2637541	16	1	4.15×10^{-44}
1473246-1473246	Rvnr01_*rrs*	G:[S467S]-G:[S467S]	16	1	4.15×10^{-44}
4247728-4247730	Rv3795_*embB*	C:[E405D]-C:[G406A]	18	3	5.54×10^{-43}
764810-764948	Rv0668_*rpoC*	A:[P481T]-G:[L527V]	31	139	1.49×10^{-34}
4407952-4408009	Rv3919c_*gidB*	A:[P84L]-C:[V65G]	25	58	4.15×10^{-34}
2976541-2976592	Rv2652c_-	T:[A5E]-C:[K106K]	23	52	1.98×10^{-31}
4243217-4243228	*embC-embA*	4243217-4243228	16	12	3.63×10^{-28}
781687-781687	Rv0682_*rpsL*	G:[K43R]-G:[K43R]	11	1	1.86×10^{-27}
2289069-2289216	Rv2043c_*pncA*	C:[F58C]-C:[V9G]	24	148	2.70×10^{-23}
39022-39030	non_coding	39022-39030	13	9	8.62×10^{-23}
2288805-2288934	Rv2043c_*pncA*	A:[A146E]-C:[Y103X]	22	130	1.40×10^{-21}
2715340-2715346	*eis* promoter	2715340-2715346	11	7	3.29×10^{-19}
4408054-4408156	Rv3919c_*gidB*	C:[L50R]-C:[L16R]	19	103	4.96×10^{-19}

the figure indicate that changes within the clustered region occur more frequently in the resistant branches relative to sensitive branches compared to the background (more associated with drug resistance) or vice versa.

5.3.3.1 Isoniazid

Isoniazid is a prodrug for the treatment of tuberculosis infection by interfering the cell wall synthesis through blocking *inhA*. The *katG* gene encodes the catalase-peroxidase enzyme, which is responsible for the activation of isoniazid. The enoyl-acyl carrier protein reductase enzyme is encoded by the *inhA* gene and required for the biosynthesis of mycolic acid, an essential component of the cell wall. Mutations at the *katG* gene, *inhA* gene and its promoter region have been linked with resistance, so mycobacterial strains harboring these mutations may develop resistance to INH. The causal variants for isoniazid resistance are mutations in *katG* at the codon Ser315Thr

(S315T) and substitutions in positions of -8, -15 and -17 in the *inhA* promoter region which increase expression of the target [83, 84, 85, 86].

Our analysis of genetic associations with INH resistance across the genome is shown in Figure 5.3, where the top identified regions are listed in Table 5.2. Two sub-regions within *katG* are identified to be statistically significant clustered from the test (see Figure 5.4). The first one is the mutation at the codon S315T by itself. It has 53 homoplasic changes where 49 occur in the branches labeled as isoniazid resistance. The other one is the region grouping 15 SNPs from codons P57R to A202A that spans 437 bp (marked in Figure 5.4). It has 16 changes and 13 occur in the INH-resistant branches. Though S315T is the highest frequency, other amino acid substitutions in *katG* have been shown to inhibit binding, decrease activation and increase resistance [79], and this has been observed in INH-resistant clinical isolates [87]. For the *inhA* promoter region, 3 mutations exist and the best clustering results in the grouping of all 3 loci (g-17t, c-15t, c-8t) spanning 10 bp. There are 33 changes in the intergenic region and 28 are related to resistance (Figure 5.5).

For lineage-specific mutations, our method does not identify them as associated with the phenotype. Codon R463L at *katG* has been shown no association against isoniazid resistance. It is a lineage-specific marker associated with Beijing strains [88]. In the dataset, 82 strains have mutations at the site. Yet, there are only two changes in the tree and one is labeled as resistant. Thus, our method reports low or no association between codon R463L and INH resistance.

By comparison, in the previous association test on INH resistance, the site-based method reports one site S315T, the gene-based method reports the gene *katG*, and the k-mer-based method reports codons S315R, Y337C and S315T as the starting positions of a 3-mer (see Figure 4.8 in Chapter 4). Identified associations between INH resistance and genotypes related to the *inhA* promoter region by site-based, gene-based, and k-mer-based (k=3) methods are one site g-17t (coordinate 1673423), the whole intergenic region, and loci from g-17t (coordinate 1673423) and c-15t (coordinate 1673425) for 3 mers, respectively. Our cluster-based method identifies all known mutations that confer resistance to INH, but the previous methods do not. Therefore, the cluster-based method identifies regions associated with resistant to INH more comprehensively with known loci

either by itself or as a region than other methods. It is also able to disregard mutations in lineage-specific codons to decrease the false positives in the association test.

Several clustered regions ranked higher in Table 5.2 are known resistant loci related to other drugs, for example, *rpoB* associated with rifampicin resistance and *embB* associated with ethambutol resistance, indicating co-resistance in the population. That is, resistance to one drug co-exists with another drug resistance and makes association ambiguous. An indication of the accuracy of the ECC method is that there are very few false positives on this list. Almost everything is known to be associated with resistance to isoniazid or another anti-tuberculous drug.

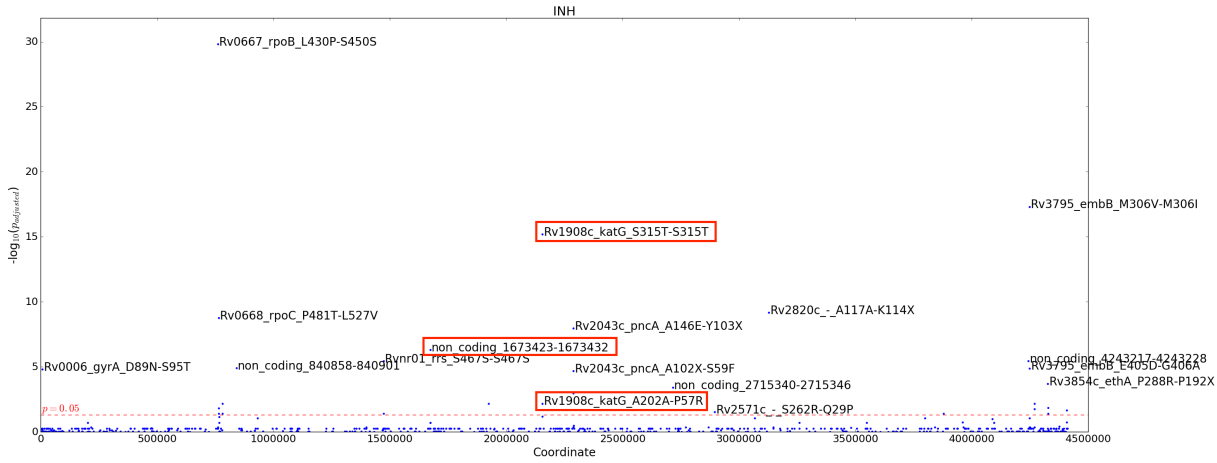


Figure 5.3: Genetic associations between clustered regions and INH resistance for 660 strains from Peru. Top resistance-associated regions are labeled in texts and listed in Table 5.2.

Table 5.2: Top regions most associated with INH resistance ($p_{assoc} < 0.05$).

Coordinates	Gene / Intergenic	Codons	R [a]	S [b]	Adjusted p value	Known drug effect [c]
761095-761156	Rv0667_$rpoB$	C:[L430P]-T:[S450S]	102	12	1.49×10^{-30}	RIF
4247429-4247431	Rv3795_$embB$	G:[M306V]-A:[M306I]	50	2	4.82×10^{-18}	EMB
3248308-3248308	Rv2931_$ppsA$	C:[H955P]-C:[H955P]	45	1	3.12×10^{-17}	NA
2155168-2155168	**Rv1908c_*katG***	**G:[S315T]-G:[S315T]**	**49**	**4**	6.46×10^{-16}	**INH**
3127922-3127931	Rv2820c_$-$	G:[A117A]-A:[K114X]	25	0	6.79×10^{-10}	NA
764810-764948	Rv0668_$rpoC$	A:[P481T]-G:[L527V]	29	2	1.78×10^{-9}	RIF
2288805-2288934	Rv2043c_$pncA$	A:[A146E]-C:[Y103X]	22	0	1.10×10^{-8}	PZA
2289069-2289216	Rv2043c_$pncA$	C:[F58C]-C:[V9G]	23	1	5.39×10^{-8}	PZA
1673423-1673432	***inhA* promoter**	**1673423-1673432**	**28**	**5**	4.92×10^{-7}	**INH**
1473246-1473246	Rvnr01_rrs	G:[S467S]-G:[S467S]	16	0	3.57×10^{-6}	KAN
4243217-4243228	*embC-embA*	4243217-4243228	16	0	3.57×10^{-6}	EMB
840858-840901	non_coding	840858-840901	0	35	1.20×10^{-5}	NA
4247728-4247730	Rv3795_$embB$	C:[E405D]-C:[G406A]	17	1	1.28×10^{-5}	EMB
7566-7585	Rv0006_$gyrA$	A:[D89N]-C:[S95T]	22	4	1.58×10^{-5}	CPX
2288937-2289066	Rv2043c_$pncA$	A:[A102X]-A:[S59F]	14	0	2.09×10^{-5}	PZA
2288696-2288778	Rv2043c_$pncA$	A:[L182X]-C:[V155X]	10	0	1.05×10^{-3}	PZA
781687-781687	Rv0682_$rpsL$	G:[K43R]-G:[K43R]	10	1	6.55×10^{-3}	STR
4269292-4269317	Rv3806c_$ubiA$	A:[A181V]-C:[S173A]	10	1	6.55×10^{-3}	EMB
2155506-2155942	**Rv1908c_*katG***	**T:[A202A]-C:[P57R]**	**13**	**3**	6.55×10^{-3}	**INH**
1923918-1924684	Rv1699_$pyrG$	T:[S30S]-C:[T286P]	13	3	6.55×10^{-3}	NA
4327289-4327520	Rv3854c_$ethA$	T:[L62X]-*	12	3	1.46×10^{-2}	ETH
765462-765669	Rv0668_$rpoC$	G:[N698S]-C:[H767P]	9	1	1.54×10^{-2}	RIF
4269090-4269148	Rv3806c_$ubiA$	T:[F248L]-G:[V229A]	7	0	1.73×10^{-2}	EMB
4407952-4408009	Rv3919c_$gidB$	A:[P84L]-C:[V65G]	17	8	2.20×10^{-2}	STR
2895175-2895875	**Rv2571c_-**	**C:[S262R]-G:[Q29P]**	**11**	**3**	2.85×10^{-2}	**NA**
781822-781822	Rv0682_$rpsL$	C:[K88T]-C:[K88T]	6	0	3.86×10^{-2}	STR
1472750-1472753	Rvnr01_rrs	A:[S302*]-G:[K303R]	6	0	3.86×10^{-2}	NA
3877947-3877969	Rv3457c_$rpoA$	T:[T187T]-A:[A180V]	6	0	3.86×10^{-2}	RIF
4327034-4327082	Rv3854c_$ethA$	G:[Y147S]-T:[C131Y]	6	0	3.86×10^{-2}	ETH
764660-764725	Rv0668_$rpoC$	A:[V431M]-G:[F452L]	6	0	3.86×10^{-2}	RIF

[a] Number of changes occurring in the resistant branches

[b] Number of changes occurring in the sensitive branches [c] INH: Isoniazid, RIF: Rifampicin, STR: Streptomycin, EMB: Ethambutol, PZA: Pyrazinamide, ETH: Ethionamide, KAN: Kanamycin, CPX: Ciprofloxacin, NA: Not Available.

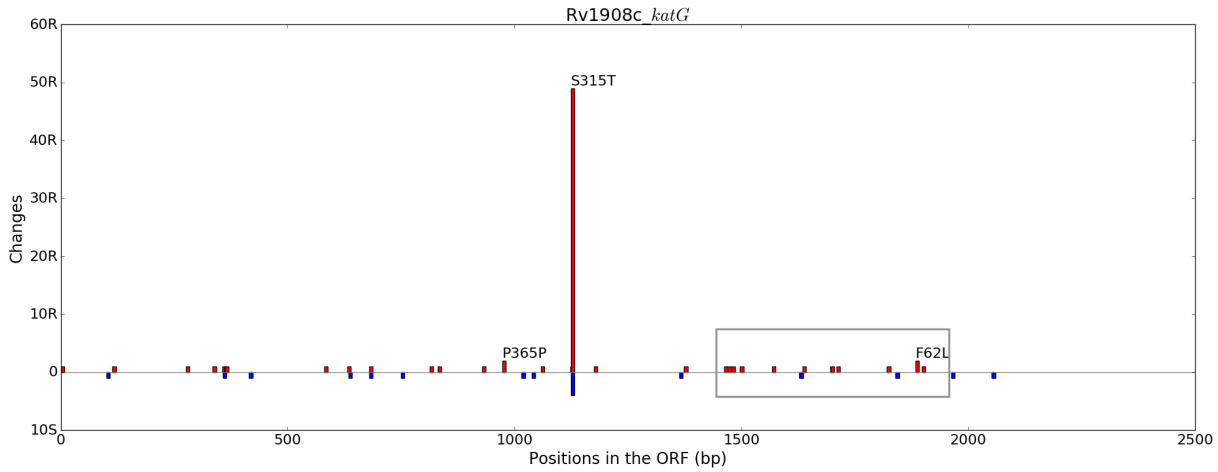


Figure 5.4: The distribution of changes occurring in branches associated with INH susceptibility (R/S) for each polymorphic site in the gene *katG*. The y-axis presents number of changes linked to resistance or sensitivity and the x-axis represents the position of a site in the ORF in bp. A codon exhibiting over one change (homoplasic site) in the resistant branch is labeled in text. The cluster (besides S315T) is boxed.

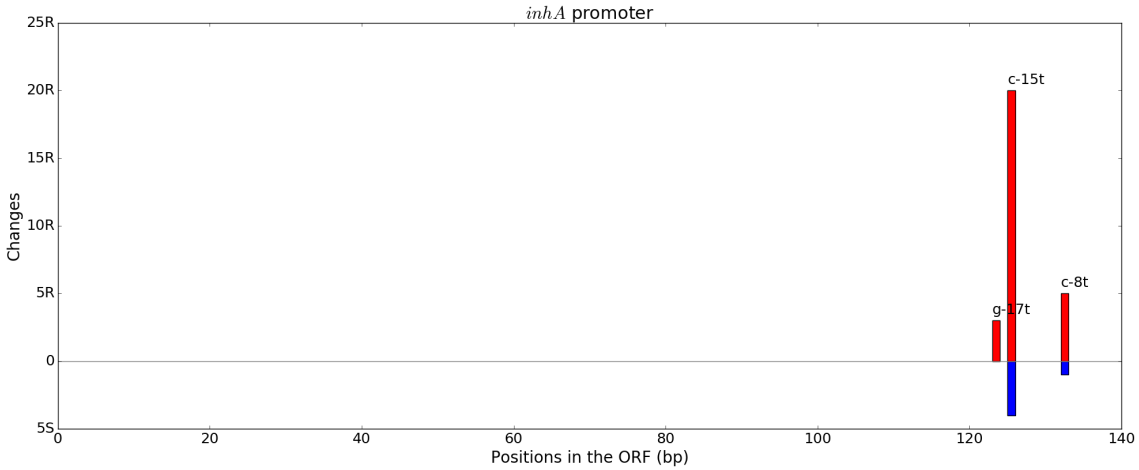


Figure 5.5: The distribution of changes occurring in branches associated with INH susceptibility (R/S) for each polymorphic site in the promoter region of *inhA*. The y-axis presents number of changes linked to resistance or sensitivity and the x-axis represents the position of a site in the ORF in bp. A codon exhibiting over one change (homoplasic site) in the resistant branch is labeled in text.

5.3.3.2 Rifampicin

Rifampicin blocks DNA-dependent RNA synthesis (transcription) in *M. tuberculosis* by binding to the RNA polymerase. Since the β-subunit of the RNA polymerase is encoded by the gene *rpoB*, the known mutations that mediate RIF resistance are mostly located within the RDRR region of *rpoB* (amino acids 435-450, region determining rifampicin resistance) [89, 90]. Additionally, several compensatory mutations are observed in *rpoC* and *rpoA* associated with RIF resistance[91].

We found that the region of the strongest association with rifampicin resistance is within the *rpoB* gene by grouping 15 mutations at loci starting from L430P to S450S that covers the RDRR region (Figure 5.7). The identified region has 114 changes spanning 62 bp and 100 changes are associated with RIF resistance. The distributions of changes occurring in branches associated with RIF susceptibility (R/S) for each polymorphic site in the genes *rpoC* and *rpoA* are shown in Figures 5.8 and 5.9, respectively. Mutations in four regions within the gene *rpoC* and one region within the gene *rpoA* which have the compensatory effects [91] are strongly associated with RIF resistance (FDR < 0.05). In *rpoC*, they are regions P481T-L527V with 29 resistant and 2 sensitive changes, N698S-H767P with 10 changes that are all resistant to RIF, E1033A-A1047P with 7 resistant and 1 sensitive changes, and V431M-F452L with 6 changes that are all related to RIF-resistant (marked in Figure 5.8). In *rpoA*, the clustered region contains 6 mutations between codons A180V and T187A spanning 23 bp, and all changes are resistant to RIF. Comas et. al. [91] also noted that compensatory mutations in *rpoA* tended to be clustered around amino acid 187, whereas they were distributed throughout the *rpoC*.

In the previous association tests (see Figure 4.9), codons D435V, H445D and S450L are identified as highly associated with RIF resistance in the site-based analysis. These are the most frequently observed RIF-resistant mutations clinically [92]. Grouping all sites within *rpoB* shows less association in the gene-based analysis since multiple changes at loci outside of the RDRR region occur in sensitive branches. Nine regions are identified from the k-mer-based association test (because we restricted to windows of size k that equals 3 adjacent SNPs) and all are located within the RDRR.

For *rpoC*, it is identified to be slightly associated with RIF resistance in 3 previous analyses. In the site-based analysis, only codon V483G is reported since there are 11 changes where 10 occur in the resistant branches. Though some changes at other loci in *rpoC* are related to resistance, they are not abundant enough to pass the test. That is, a SNP at an individual site does not have enough changes occurring in resistant branches. In the gene-based analysis, grouping all SNPs within the gene results in more changes (18) yet 5 changes occur in sensitive branches, showing a weak association (p-value = 2.54×10^{-3}). In the k-mer-based analysis, 3 groupings of consecutive sites are identified to be slightly associated with RIF resistance, all include the codon V483G.

For *rpoA*, none of the previous analyses reports they are associated with the RIF resistance. There are 12 SNPs in total occurring from codons L80V to E319K, where each individual site has exactly 1 change (see Figure 5.9). In the site-based analysis, 7 sites harbor changes occurring in the resistant branches and 5 changes are in the sensitive branches. Since sites within *rpoA* do not have many changes occurring in resistant branches by itself, no association is identified. Similarly, in the k-mer-based test where k equals to 3, the grouping of adjacent 3 SNPs still does not have enough changes related to resistance. Furthermore, grouping all SNPs within the gene *rpoA* increases the total number of changes to 12 yet 5 changes occur in the sensitive branches, suggesting little or no association from the gene-based analysis. In contrast, our cluster-based approach focuses significance testing on 6 SNPs between amino acids 180 to 187, all of which are associated with resistant branches (adjusted p value = 0.0246).

Thus, for cases like *rpoB*, where changes in an individual site are abundant enough to be strongly linked with resistance, our cluster-based method performs better than others in terms of identifying the best grouping that maximizes the association. For other cases like *rpoC* and *rpoA*, where changes in an individual site are not enough to be identified from the association test, locating the locally clustered region with more changes occurring in resistant branches helps to identify resistant-associated regions within genes.

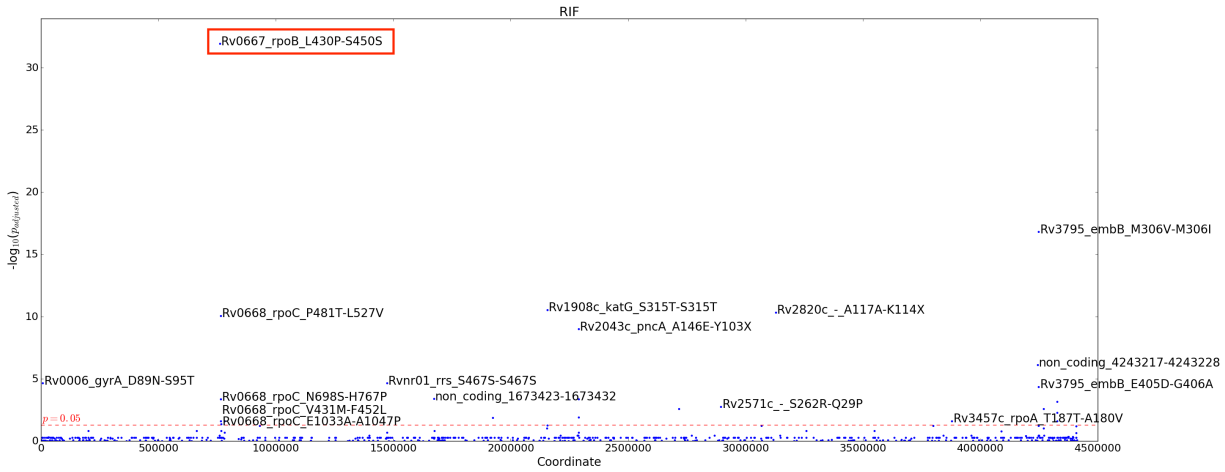


Figure 5.6: Genetic associations between clustered regions and rifampicin resistance for 660 strains from Peru. Top resistance-associated regions are labeled in texts.

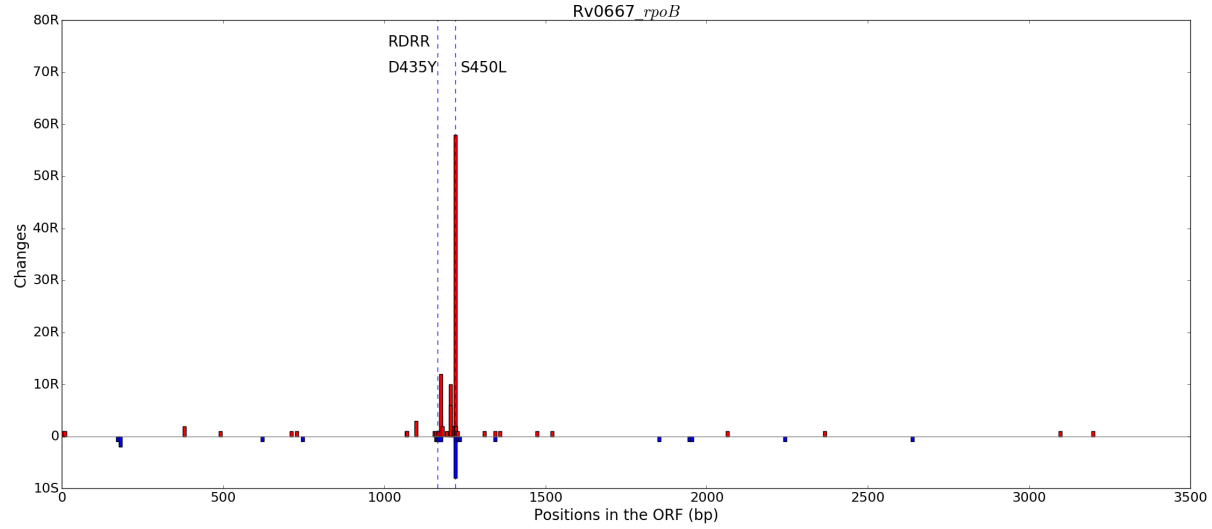


Figure 5.7: The distribution of changes occurring in branches associated with RIF susceptibility (R/S) for each polymorphic site in the gene *rpoB*. The y-axis presents number of changes linked to resistance or sensitivity and the x-axis represents the position of a site in the ORF in bp. The region between two blue vertical dashed lines is the RDRR region.

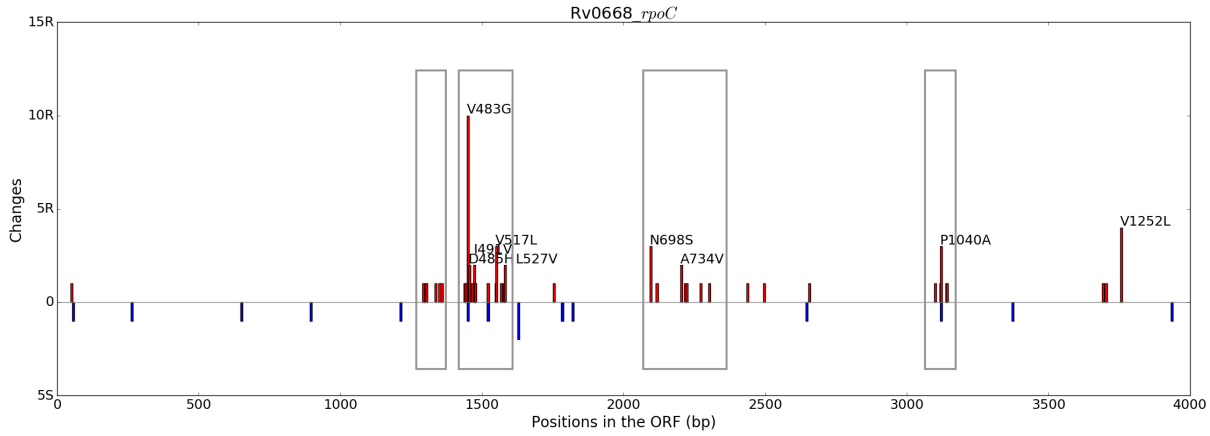


Figure 5.8: The distribution of changes occurring in branches associated with RIF susceptibility (R/S) for each polymorphic site in the gene *rpoC*. The y-axis presents number of changes linked to resistance or sensitivity and the x-axis represents the position of a site in the ORF in bp. A codon exhibiting over one change (homoplasic site) in the resistant branch is labeled in text. Clusters are boxed.

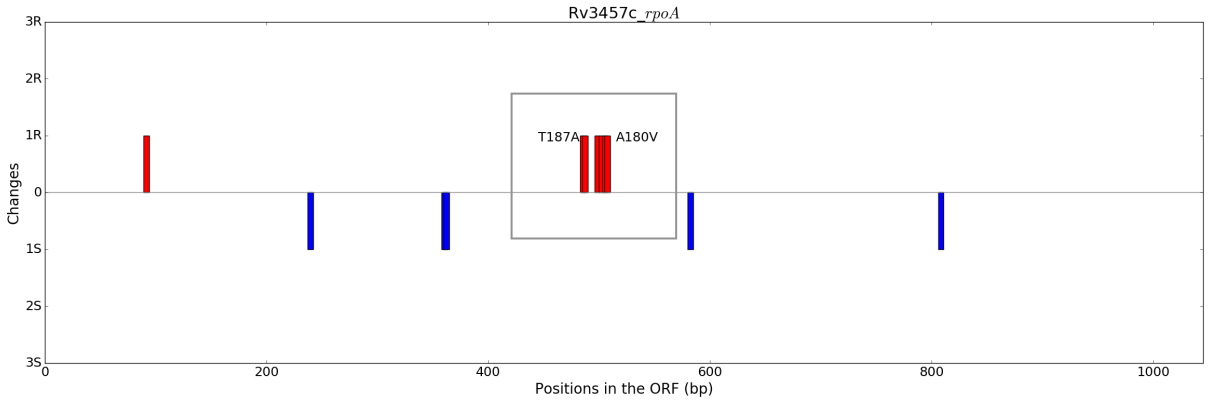


Figure 5.9: The distribution of changes occurring in branches associated with RIF susceptibility (R/S) for each polymorphic site in the gene *rpoA*. The y-axis presents number of changes linked to resistance or sensitivity and the x-axis represents the position of a site in the ORF in bp. The codons in the clustered region are labeled in text. The clustered region of amino acids 180-187 is boxed.

5.3.3.3 Ethambutol

Ethambutol is involved in the inhibition of cell wall synthesis by targeting the arabinogalactan biosynthesis. The *embB* gene (especially codons 306 and 406, which are the highest frequency) and the *embC-embA* intergenic region are responsible for mediating EMB resistance [93, 94]. Genes *embABC* form a cell-wall complex that is involved in transferring lipoarabinomannan (LAM) precursors to the outer membrane. Growing evidence shows that *ubiA* appears to be associated with EMB resistance, especially high-level EMB resistance [95]. Gene *ubiA* encodes decaprenyl-phosphate 5-phosphoribosyltransferase in the pathway for synthesizing LAM (a cell wall glycolipid in *M. tuberculosis*) [96].

In Figure 5.10, the locally clustered region of two codons M306V and M306I in *embB* gene shows the strongest association with EMB resistance. There are 52 changes in the region of 3 bp, and 43 changes are related to resistance. The other locally clustered region within *embB* containing codons G406S and G406A is also identified in the association test analysis. It has 18 changes that span 3 bp where 15 changes are associated with EMB-resistant (Figure 5.11). Seven polymorphisms exist within the *embC-embA* intergenic region spanning 39 bp and 18 changes are obtained from the tree. In the first-phase, a clustered region is obtained consisting of 6 SNPs with 16 changes spanning 12 bp. In the second-phase, the clustered region is identified to be associated with EMB resistance since 12 changes occur in the resistant branches (Figure 5.12). The Rv3806c (*ubiA*) gene is also found to be homoplasic and associated with resistance. Three clustered regions are obtained and two have overrepresented resistant changes, including 8 SNPs clustered from codons A181V to S173A with 11 changes (9R, 2S) spanning 26 bp, and 7 SNPs clustered from codons F248L to V229A with 7 changes (6R, 1S) across 59 bp (Figure 5.13).

In previous results (see Figure 4.10), codons M306V, M306I, G406A and G406S are identified to be associated with EMB resistance by the site-based test. Gene *embB* is identified to be strongly associated with resistance to EMB from the gene-based association test. Several 3-mer groupings show associations with EMB resistance from the k-mer-based test, including loci starting from M306V, S297A, S203L, M306I, G406S, E405D and L402V, ranked by association levels. None of

the mutations in gene Rv3806c (*ubiA*) is identified to be involved in EMB resistance by previous tests in site-based and k-mer-based analyses, but the gene-based association test (combining all changes in a gene together by a burden test) found it to be significant (p value = 1.44×10^{-9}). Mutations are found in 22 sites spanning from A35S to L284L with 26 changes in total. In the site-based analysis, each site has only 1 or 2 changes, resulting in little or no association from the test. In addition, grouping 3 adjacent SNPs together increases the significance of the association between sites within Rv3806c and EMB resistance but is still not significant enough. Thus, our method detected *ubiA*, intergenic region of *embC-embA*, whereas other methods did not.

Figure 5.10: Genetic associations between clustered regions and ethambutol resistance for 660 strains from Peru. Top resistance-associated regions are labeled in texts.

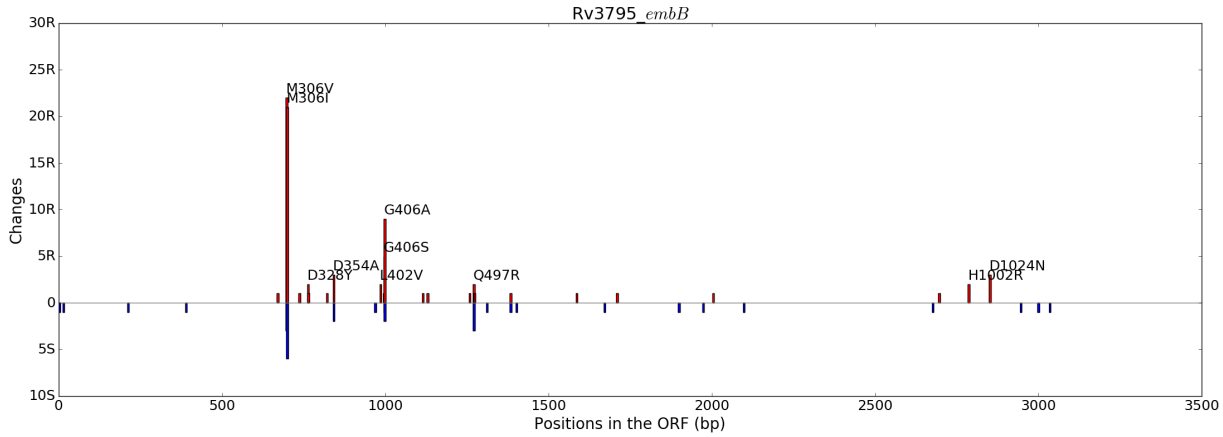


Figure 5.11: The distribution of changes occurring in branches associated with EMB susceptibility (R/S) for each polymorphic site in the gene *embB*. The y-axis presents number of changes linked to resistance or sensitivity and the x-axis represents the position of a site in the ORF in bp. A codon exhibiting over one change (homoplasic site) in the resistant branch is labeled in text.

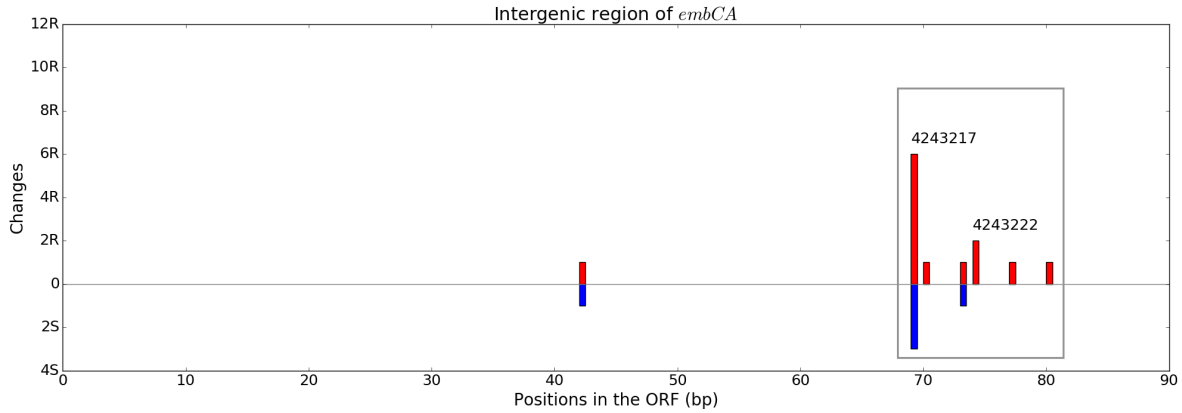


Figure 5.12: The distribution of changes occurring in branches associated with EMB susceptibility (R/S) for each polymorphic site in the intergenic region between *embC* and *embA*. The y-axis presents number of changes linked to resistance or sensitivity and the x-axis represents the position of a site in the ORF in bp. A codon exhibiting over one change (homoplasic site) in the resistant branch is labeled in text. The cluster is boxed.

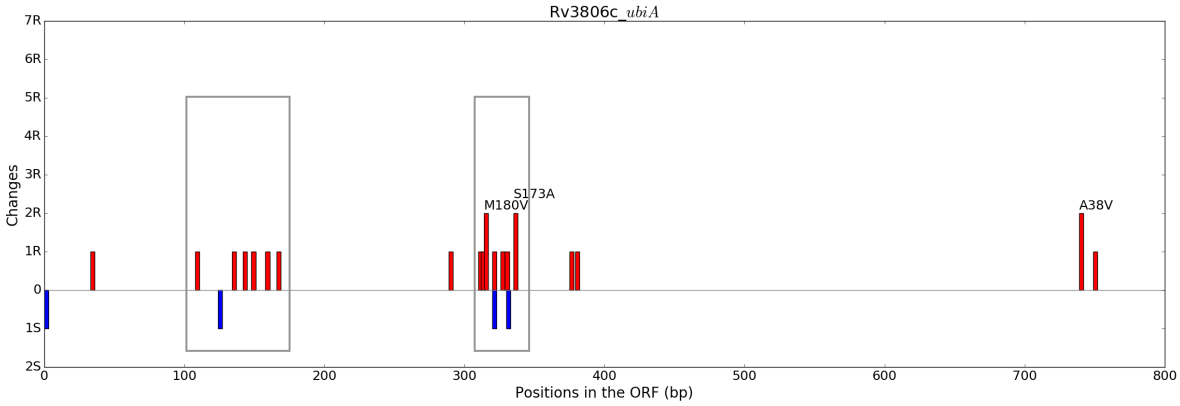


Figure 5.13: The distribution of changes occurring in branches associated with EMB susceptibility (R/S) for each polymorphic site in the gene *ubiA*. The y-axis presents number of changes linked to resistance or sensitivity and the x-axis represents the position of a site in the ORF in bp. A codon exhibiting over one change (homoplasic site) in the resistant branch is labeled in text. Clusters are boxed.

5.3.3.4 Streptomycin

Streptomycin binds to 16S ribosomal RNA (*rrs*) and ribosomal protein S12 (*rpsL*) to inhibit protein synthesis (translation). Resistance to streptomycin has shown to be conferred by the mutations in A514C of *rrs* (16S rRNA) and codons K43R and K88T in *rpsL* (S12 ribosomal protein) [97, 98]. Also, nonsynonymous mutations at *gidB* (a 16S rRNA methyltransferase) have been discovered for conferring streptomycin resistance [99]. The Methyl group is needed for optimal binding of streptomycin, so loss-of-function mutations in the methyltransferase mediate resistance to STR.

In Figure 5.14, two codons at K43R and K88T in *rpsL* and one clustered region within the *gidB* are identified from the cluster-based analysis. Two mutations occur in *rpsL* and apart with each other by 135 bp, so they are locally clustered by itself for each. Codon K43R has 11 changes and all are resistant to STR. Codon K88T has 6 changes and all occur in the STR-resistant branches. For *gidB*, 59 mutations are acquired across 528 bp. Four clustered regions are obtained yet only one is associated with STR resistance. The region ranges from codons P84L to V65G spanning 58 bp with 25 changes where 18 are in the resistant branches. Seven strains have mutations in

S172R of *rrs* evolving along 5 branches independently in the tree where 4 branches are labeled as resistant, resulting in a weak association.

In the previous results, since both codons K43R and K88T at *rpsL* harbor multiple changes mostly occurring in the resistant branches, they are identified to be strongly associated with STR resistance in all three analyses. For *gidB*, most individual sites harbor only 1 or 2 changes, so the site-based method fails to report the association between *gidB* and STR resistance. This is because *gidB* is a ribosome methyltransferase, and resistance is conferred by loss-of-function mutations, which can occur anywhere throughout the ORF. Similarly, in the k-mer-based analysis, any grouping of adjacent 3 SNPs within *gidB* still does not have enough changes occurring in resistant branches. Yet, grouping all SNPs within *gidB* turns out to be slightly associated with resistance to streptomycin since it obtains 32 changes where 21 occur in the resistant branches. Thus, our cluster-based method does a much better job identifying the significance of association of mutations in *gidB* with STR resistance.

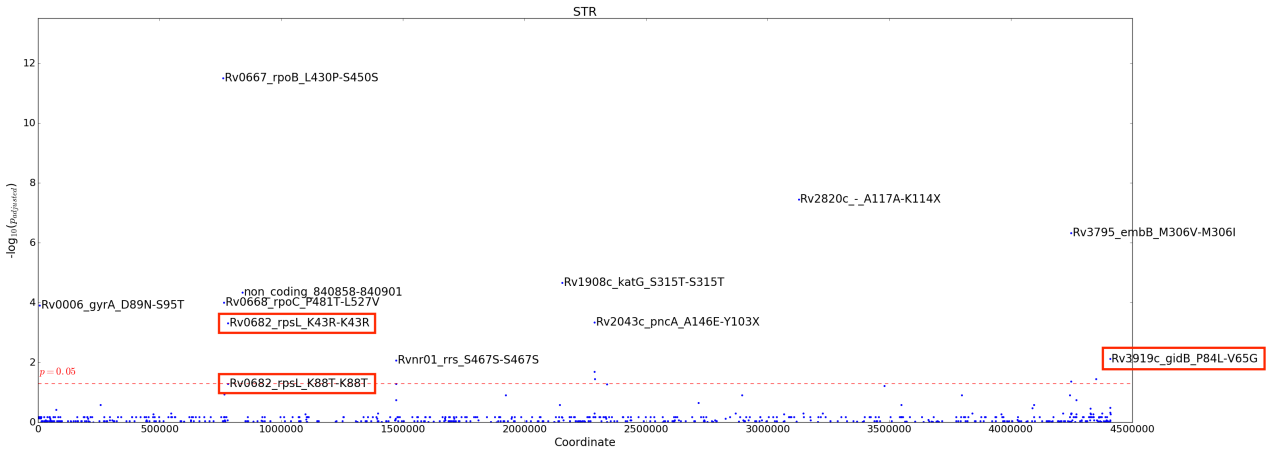


Figure 5.14: Genetic associations between clustered regions and streptomycin resistance for 660 strains from Peru. Top resistance-associated regions are labeled in texts.

5.3.3.5 Pyrazinamide

Pyrazinamide is one of the first-line anti-tuberculous drugs. It must be activated by the pyrazinamidase (PZase), an enzyme encoded by the *pncA* gene, to become pyrazinoic acid (POA), an active form of PZA. Mutations, especially loss-of-function, occurring in *pncA* have been observed primarily to confer PZA resistance [100, 101].

In the dataset of 660 strains from Peru, 166 strains have at least 1 mutation in *pncA*. Figure 5.16 shows the distribution of changes occurring in branches associated with PZA susceptibility (R/S) for each polymorphic site in the gene *pncA*. The y-axis presents number of changes linked to resistance or sensitivity and the x-axis represents the position of a site in the ORF in bp. We observed 55 mutations occurring in *pncA* across 536 bp, where one occurs in the stop codon (Q10*). Among 55 loci, 11 polymorphic sites are homoplasic, exhibiting distinct mutational events. For insertions and deletions (indels), 15 strains have indels (14 indels < 10 bp) and 13 of them are resistant to pyrazinamide. In our cluster-based analysis, 4 sub-regions in *pncA* are identified from the first phase. In the second phase, 3 sub-regions of *pncA* are strongly associated with PZA resistance, where their FDR-corrected p values are all less than 0.05, including codons F58C-V9G (18R, 5S), A102X-S59F (10R, 3S) and A146E-Y103X (19R, 3S).

In previous analyses, the site-based method did not identify any association between loci in *pncA* and PZA resistance since changes in loci are not abundant enough (less than 4) to pass the test. Several 3-mer groupings show weak associations with PZA resistance from the k-mer-based test, including loci starting from F58L, H57L, and D12G. In the gene-based analysis, *pncA* is identified to be associated with resistance to PZA. There are 51 changes in total and 34 occur in the resistant branches. Like *gidB*, loss-of-function mutations can occur anywhere throughout the ORF, since *pncA* is an activator of the prodrug PZA to pyrazinoic acid (POA) [100, 101].

Recently, it has been reported that PZA-resistant mutations have a low frequency of mutations in *panD* (pantothenate pathway) [102], and that the true target of pyrazinoic acid (activated species of PZA) is *coaBC* in coenzyme A biosynthesis pathway [103]. However, there are only two SNPs in *panD* and both are in the sensitive branches. In comparison, loss-of-function mutations in *pncA*

are more prevalent than *panD* mutations in PZA-resistant isolates.

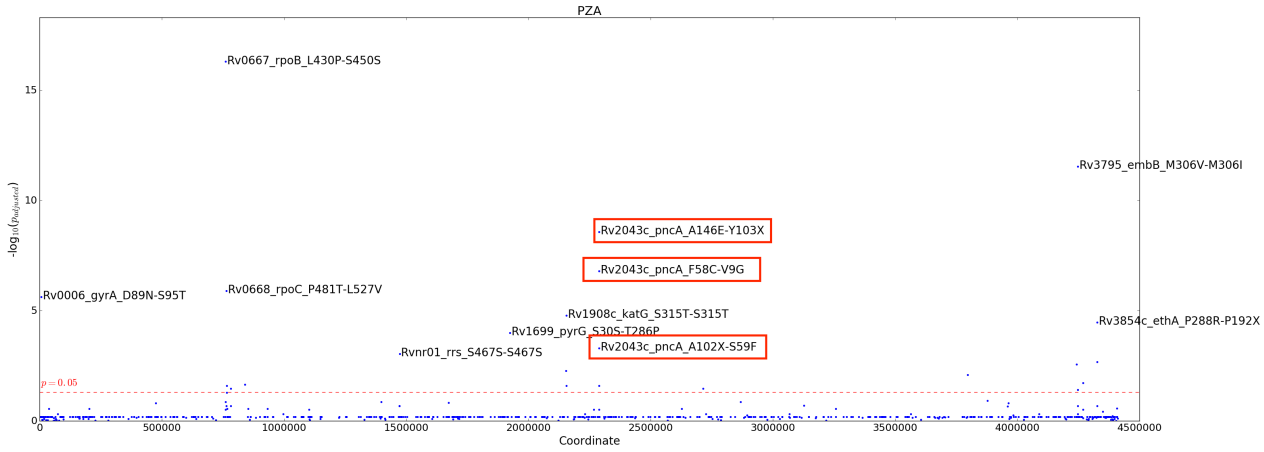


Figure 5.15: Genetic associations between clustered regions and pyrazinamide resistance for 660 strains from Peru. Top resistance-associated regions are labeled in texts.

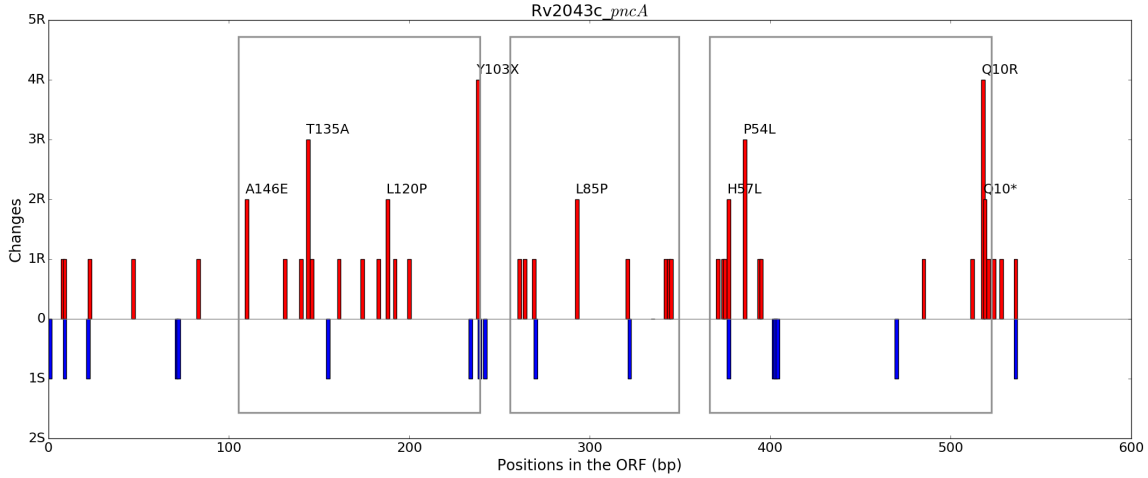


Figure 5.16: The distribution of changes occurring in branches associated with PZA susceptibility (R/S) for each polymorphic site in the gene *pncA*. The y-axis presents number of changes linked to resistance or sensitivity and the x-axis represents the position of a site in the ORF in bp. A codon exhibiting over one change (homoplasic site) in the resistant branch is labeled in text. Clusters are boxed.

5.3.3.6 Kanamycin

Kanamycin, one of the second-line injectable drugs, binds to ribosomes and results in protein synthesis inhibition. Mutations within the *eis* promoter region and *rrs* at codon S467S (nucleotide 1401, A1401G) are reported to be involved in the kanamycin resistance [104]. Gene *eis* is an efflux pump, and mutations in the promoter increase expression.

Six mutations evolve within the upstream of *eis* across 93 bp (2715340-2715432). One clustered region is identified where 4 mutations are grouped within 7 bp (2715340-2715346). The strongest association with the KAN resistance results in the clustered region of *eis* promoter with 11 changes consisting of 10 resistant branches and 1 sensitive branch. Four clustered regions within the gene *rrs* are obtained for the first phase analysis, and 1 region that contains a single site at the codon S467S by itself has a strong association with KAN resistance. It has 15 changes and 11 occur in the resistant branches.

By comparison, in the previous association test on kanamycin resistance, the site-based method only identified one site in the *eis* promoter (coordinate 2715346) to be associated with KAN resistance because it exhibits 5 changes that all occur in the resistant branches. Even though other 3 sites (2715342, 2715344 and 2715376) harbor changes in the resistant branches, they do not pass the test due to the exhibition of only 1 or 2 changes per site. In the gene-based analysis, the grouping of all SNPs within the *eis* promoter is strongly associated with resistance to KAN. In the k-mer-based analysis, 4 groupings of 3 consecutive sites are identified to be associated with KAN resistance, starting from codons 2715340 to 2715346.

Figure 5.17: Genetic associations between clustered regions and kanamycin resistance for 660 strains from Peru. Top resistance-associated regions are labeled in texts.

5.3.3.7 Ciprofloxacin

Ciprofloxacin is one of the fluoroquinolones categorized to the second-line drug. It is a DNA synthesis inhibitor that targets DNA gyrase enzyme encoded by *gyrA*. Resistance to CPX is known to be mediated by mutations occurring in the *gyrA* gene, primarily codons A90V and D94G [105].

In the Peru dataset, there are 14 mutations spreading over 2361 bp in *gyrA*. Only one clustered region is reported within the gene *gyrA*. That is, the grouping of codons from D89N to S95T of 5 mutations spanning 20 bp. This sub-region of *gyrA* is highly homoplasic with 23 changes and 20 are related to CPX resistance.

Mutation at codon S95T is a lineage-specific mutation and not related to CPX resistance [88]. However, it is located within the clustered region. The reason is probably that S95T is just 3 bp downstream of D94G, in the first phase clustering it together with its upstream 4 mutations (D89N, A90V, D94N, D94G) would be a region with the highest local mutation rate among others than expected.

In the previous analyses of three methods, the site-based method only identified codon D94G at *gyrA* to be strongly associated with CPX resistance since it exhibits 13 changes and 12 occur in the

resistant branches. Not all known mutations that confer resistance to CPX are identified because codon like A90 or D94N does not have enough changes by itself. In the k-mer-based analysis where k equals 3, the grouping of 3 codons A90V, D94N and D94G maximizes the association between *gyrA* and CPX resistance. We observed 19 changes and 18 are in the resistant branches. Grouping of all SNPs within *gyrA* also shows association with CPX in the gene-based analysis yet it is less significant than the best grouping (A90V-D94G) from the k-mer-based analysis which is less significant than the grouping of 5 codons (D89N-S95T) from our cluster-based analysis.

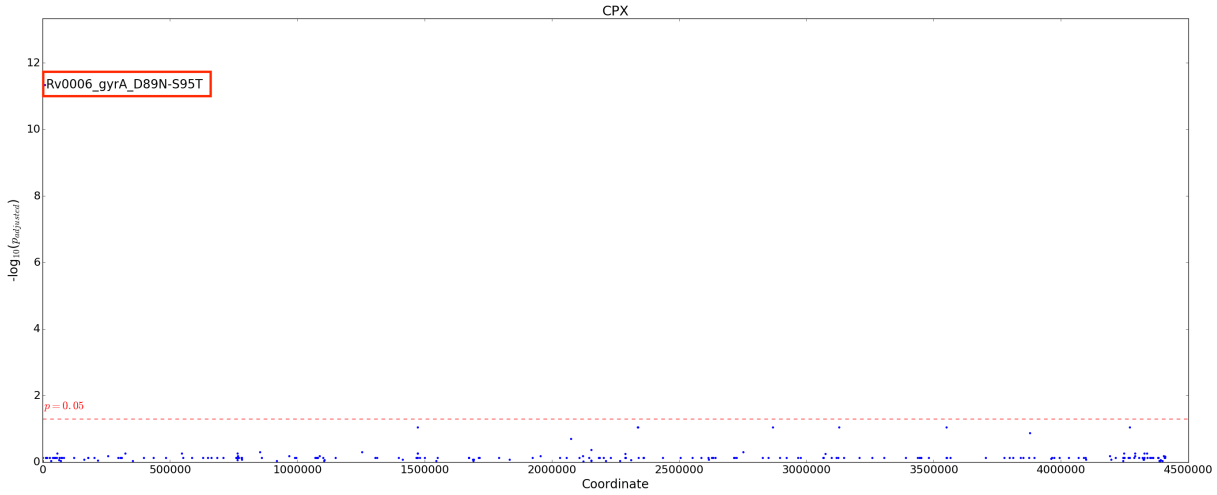


Figure 5.18: Genetic associations between clustered regions and ciprofloxacin resistance for 660 strains from Peru. Top resistance-associated regions are labeled in texts.

5.3.4 Novel Genetic Variant Associated with Anti-tuberculous Drugs: Rv2571c

Rv2571c was identified as a novel gene significantly linked with resistance to multiple anti-tuberculous drugs in the analysis of MDR clinical isolates from Peru. Rv2571c is a gene of unknown function categorized as a nonessential gene [106] of length 1068 bp (coordinates 2894893-2895960). It is a membrane protein consisting of 355 amino acids with 6 predicted transmembrane regions (alanine, valine and leucine-rich, see Figure 5.19) located adjacent to *aspS*, aspartyl-tRNA synthetase (see Figure 5.20). In the dataset of 660 strains from Peru, it is identified as a clustered region spanning 701 bp with 14 changes (distinct mutational events) distributed throughout the open reading frame (ORF) from Q29P to S262R (see Figure 5.21). Among fourteen changes, two are stop-codon (loss of function) mutations and one is synonymous (A57A).

Rv2571c shows up on the list of associations for several drugs but is not previously linked to resistance. The FDR-adjusted *p* values of associations with resistance are listed in Table 5.3. In the association test analysis, the changes in the region occur in more branches related to INH, RIF and EMB resistance of adjusted *p* values less than 0.03 (see Figures 5.3, 5.6 and 5.10). For comparison, *InhA* promoter region is shown, mostly associated with INH resistance, but also with other drugs due to co-resistance. *LldD2* is shown because it is a known homoplasic locus, but is clearly not associated with resistance to any of these anti-tuberculous drugs [82].

In the clustered region of Rv2571 from Q29P to S262R, Table 5.4 shows the distribution of phenotypes for strains harboring mutations distributed throughout the 14 codons in Rv2571c. For each codon, the number of HRES resistant strains and the number of sensitive strains are listed. An HRES resistant strain represents it is at least resistant to one of the anti-tuberculous drugs that include isoniazid (H), rifampicin (R), ethambutol (E) and streptomycin (S). The topological distribution of mutations in *katG*, *inhA* promoter region and Rv2571c are shown in Figure 5.22. Lineages are labeled in colors in the leaves of the tree. Strains from Peru are mostly categorized to lineage 2 (Beijing) and lineage 4 (LAM, Haarlem, T-clade, X-clade and H-clade). Strains resistant to drugs (INH, RIF, EMB, and STR) are labeled in red, strains harboring mutations in *katG* or *inhA* promoter region are labeled in green, and strains exhibiting mutations in sites within Rv2571c are

labeled in blue.

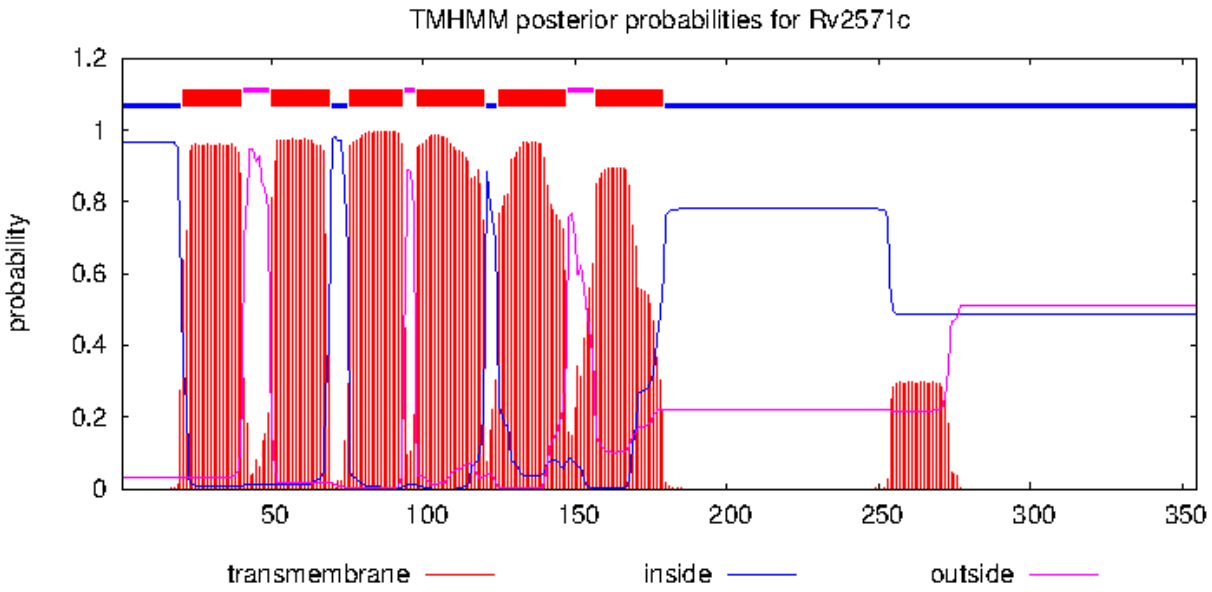


Figure 5.19: Prediction of transmembrane helices in proteins for Rv2571c from TMHMM [1]. Six transmembrane regions are predicted in Rv2571c across 355 amino acids.

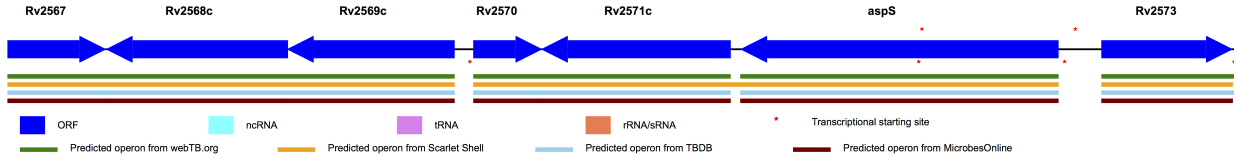


Figure 5.20: The genomic location of Rv2571c and its adjacent genes in the *M. tuberculosis* genome.

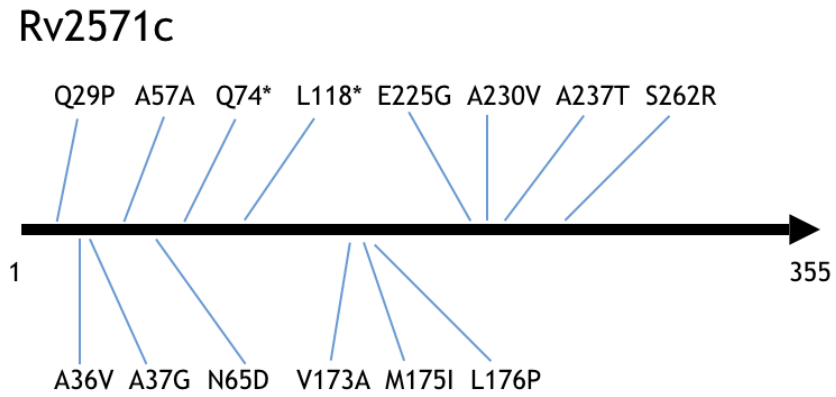


Figure 5.21: Relative locations of observed changes within the clustered region of Rv2571c in the dataset of 660 strains from Peru. Rv2571c has 355 amino acids.

Table 5.3: Associations with resistance and clustered regions of Rv2571c, *InhA* promoter and *LldD2* of *M. tuberculosis*. The adjusted p values are listed for pairs of SNP clusters and drugs along with the number of changes at resistant branches (R) and the number of changes at sensitive branches (S).

SNP cluster	INH	RIF	EMB	STR
Rv2571c (Q29P-S262R)	0.02851 (11R, 3S)	0.0017 (12R, 2S)	0.0039 (10R, 4S)	0.1221 (10R, 4S)
InhA promoter (-8...-17)	4.92×10^{-7} (28R, 5S)	0.0004 (23R, 10S)	0.0204 (16R, 17S)	0.7482 (13R, 20S)
LldD2 (V3I-S3S)	0.9431 (18R, 44S)	0.9599 (15R, 47S)	0.9973 (6R, 56S)	0.9970 (13R, 49S)

Table 5.4: Distribution of phenotypes for strains harboring mutations in Rv2571c. An HRES resistant strain represents it is at least resistant to one of the following anti-tuberculous drugs: isoniazid (H), rifampicin (R), ethambutol (E) and streptomycin (S).

Sites	Number of strains with any HRES resistance	Number of sensitive strains
Q29P	1	0
L36V	3	4
A37G	1	0
A57A (synonymous)	0	3
N65D	1	0
Q74* (stop codon)	1	0
L118* (stop codon)	0	1
V173A	1	0
M175I	4	0
L176P	1	0
E225G	1	0
A230V	3	0
A237T	1	0
S262R	1	0

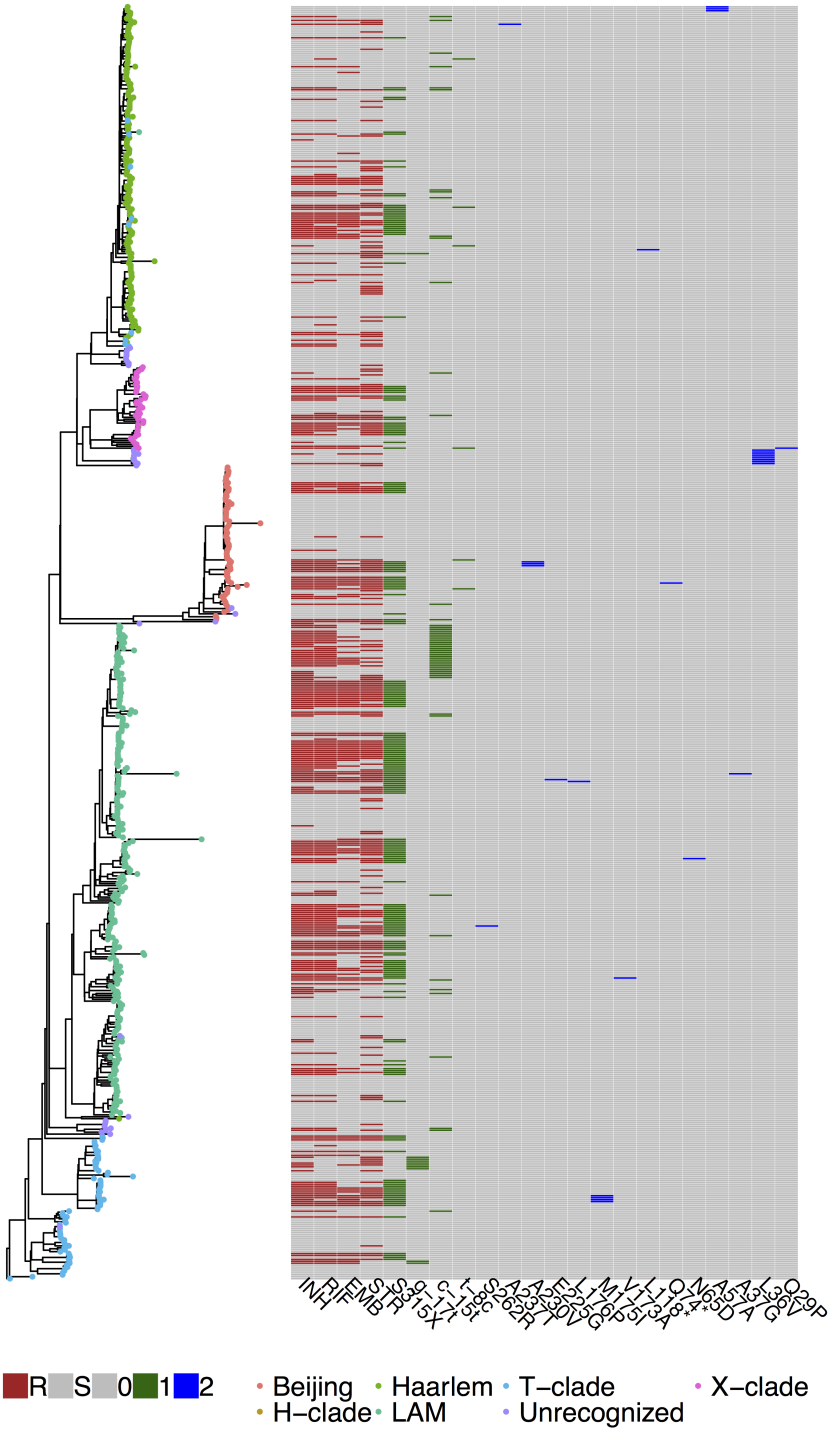


Figure 5.22: Distribution of lineages, phenotypes and mutations in Rv2571c in the phylogenetic tree. Lineages are labeled in colors in the leaves of the tree. Strains resistant to four drugs (INH, RIF, EMB, and STR) are labeled in red, strains that harbor mutations in *katG* or *inhA* promoter region are labeled in green, and strains that have mutations in locus within Rv2571c are labeled in blue.

We also expanded the study to a larger set of 1525 isolates from Lima, Peru by conducting the analysis of the consistency of drug susceptibility and polymorphisms with a focus on Rv2571c. The proportion of INH-resistant strains accounts for 70.9% in the population. Fifty-four strains exhibit at least one nonsynonymous mutation in Rv2571c and forty-seven of them are resistant to INH. Also, eleven strains have insertions in the Rv2571 and nine are resistant to INH.

Although we do not know the function of Rv2571c, this analysis suggests it is associated with resistance. We cannot confidently determine which drug is related to the mutations at Rv2571c. The association of Rv2571c with drug resistance has not been reported yet. However, Grandjean et. al. noticed that Rv2571c was homoplasic in another MDR dataset, though they did not have sensitive strains for comparison of association testings [81]. The distribution of sites throughout ORF, along with mutations at 2 stop codons and 1 indel of Rv2571c, suggests resistance is conferred through loss-of-function, similar to an activator.

For validation of Rv2571c, we used a worldwide dataset of 3651 clinical isolates of *M. tuberculosis* with susceptibility to anti-tuberculous drugs [46]. We aligned them to the reference genome H37Rv (accession NC_000962.2) of size 4.4M bp. There are 197,519 polymorphic sites in the alignment, excluding ambiguous sites, repetitive regions of *PPE* and *PGRS* genes. The global phylogenetic tree is reconstructed from informative sites using PAUP [107] (Figure 5.23). For five phenotypes of resistance to INH, RIF, EMB, STR and PZA, the proportion of drug-resistant strains ranges between 6.86% (PZA) to 20.45% (STR) (Figure 5.24). The number of strains that are resistant both to INH and RIF (MDR-TB) accounts for 10.5% (382/3651) in the population. Nonsynonymous nucleotide substitutions occur in forty-seven sites within Rv2571c. Fifty-seven changes from nonsynonymous mutations are found in the gene Rv2571c and 16 changes occur in the resistant branches in the phylogenetic tree by maximal parsimony. Because this is close to the background, there is little or no association with INH resistance. Therefore, our hypothesis is not confirmed in this dataset.

There are several possible reasons for failing to observe Rv2571c associated with drug resistance in the worldwide dataset. First of all, in the first dataset, clinical isolates are from Lima, Peru

locally and all are categorized to either lineage 2 (Beijing) or lineage 4 (LAM, Haarlem, T-clade, X-clade, H-clade). Yet clinical isolates in the second dataset are collected throughout the world and categorized to all major *M. tuberculosis* lineages (lineages 1-7 and bovis) [108], which are more divergent. In addition, the dataset of strains from Peru has a higher proportion of MDR strains (35.9%) than the dataset of worldwide strains (10.5%). Lastly, differences in the accuracy of drug susceptibility test could have an impact on the association test. The result of a drug susceptibility test for a clinical isolate is binary, resistant or sensitive. However, the DST is not robust, especially when the isolate does not grow well in culture, yielding poor reliability [109]. The DST results are less reproducible or reliable for EMB and PZA [110]. An isolate determined as resistant from the DST may be slightly or strongly resistant to the drug. To determine the drug susceptibility for strains more accurately, the minimal inhibitory concentration quantitatively tests the lowest concentration of antibiotic for killing a strain.

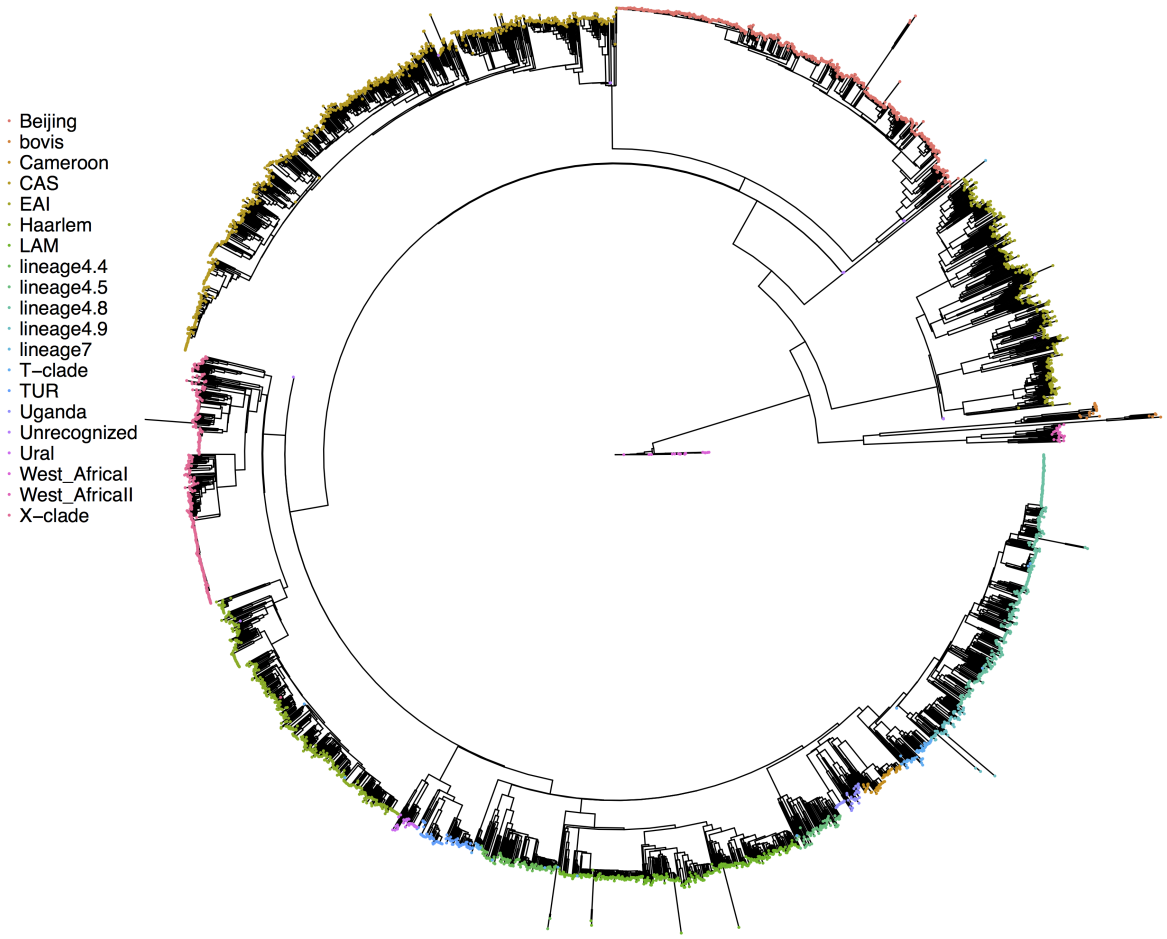


Figure 5.23: Phylogenetic tree and the distribution of lineages of the worldwide dataset of 3651 *M. tuberculosis* clinical isolates.

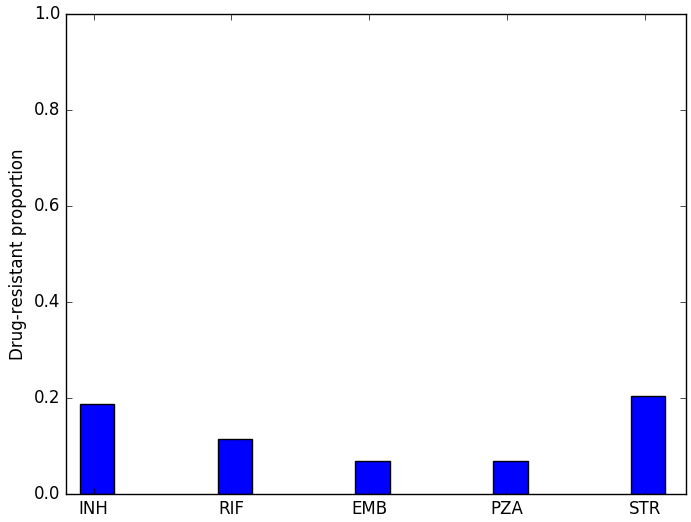


Figure 5.24: Proportion of drug-resistant strains for 5 drugs in the worldwide dataset of 3651 *M. tuberculosis* clinical isolates.

5.3.5 Novel Genetic Variant Associated with Anti-tuberculous Drugs: Rv1830

We applied the two-phase approach on a dataset consisting of 550 *M. tuberculosis* clinical isolates from China with DST data of 5 antibiotics (accession number PRJNA268900) [73]. We subsampled strains randomly to obtain a subset of 175 INH-resistant and 201 sensitive strains, where the proportion of INH-resistant strains equals 46.5% (175/376). We obtained 25,794 non-synonymous polymorphic sites, excluding ambiguous sites and repetitive regions. By applying Sankoff's algorithm on the reconstructed tree, we acquired 29,535 changes. The overall mutation rate is 0.67% (29,535 changes / 4,411,532 bp). In the first-phase analysis, 617 clustered regions are obtained with adjusted p values below 0.05. In the second-phase association test, we identified the known mutations in the *katG* gene and in the *inhA* promoter region associated with isoniazid resistance (see Figure 5.25). We further observed that polymorphic sites within a sub-region of Rv1830 exhibit more changes as a clustered region than are likely to occur by chance alone, and more changes occur in INH-resistant branches than expected.

Rv1830 is an essential gene [106] with unknown function, annotated as a putative helix-turn-helixt type transcriptional regulator. Its gene length is 678 bp (coordinates 2074841-2075518) comprising of 225 amino acids. Figure 5.26 presents the genomic location of Rv1830 and its adjacent genes. In the dataset of 376 strains from China, there are 18 nonsynonymous mutations spanning 493 bp within Rv1830. In the sub-region throughout the open reading frame from P68L to H128P, 16 changes (distinct mutational events) are observed spanning 181 bp (see Figure 5.27), where 13 changes occur in the INH-resistant branches. The distribution of INH susceptibility for strains harboring mutations in the clustered region of Rv1830 are listed in Table 5.5. In the clustered region of 15 codons, strains exhibiting mutations are resistant to INH in 12 codons. Hicks et. al. also reported that Rv1830 was near significant in the dataset using phyOverlap [73]. By using our cluster-based method, it is identified to be slightly associated with resistance to isoniazid.

For validation, we used two datasets of clinical isolates of *M. tuberculosis*. In the dataset of 660 strains from Peru, there are 12 nonsynonymous mutations from codons P20L to *226W spanning 619 bp. We obtain 13 changes and 7 occur in the INH-resistant branches, showing little or no association. In the dataset of 400 worldwide strains, the proportion of INH-resistant strains equals 39.8% (159/400). For Rv1830, 5 mutations are observed from codons L121I to D221A spanning 302 bp. We obtain 5 changes where 3 occur in the INH-resistant branches, implying a weak association between Rv1830 and INH resistance. Results from the above analyses using three empirical datasets suggest that the gene Rv1830 may be slightly associated with resistance to isoniazid. Further experiments need to be performed to understand the contribution of mutations in Rv1830 on INH susceptibility. Mutations in the upstream of Rv1830 were hypothesized to be putative markers of amoxicillin and clavulanate susceptibility [111].

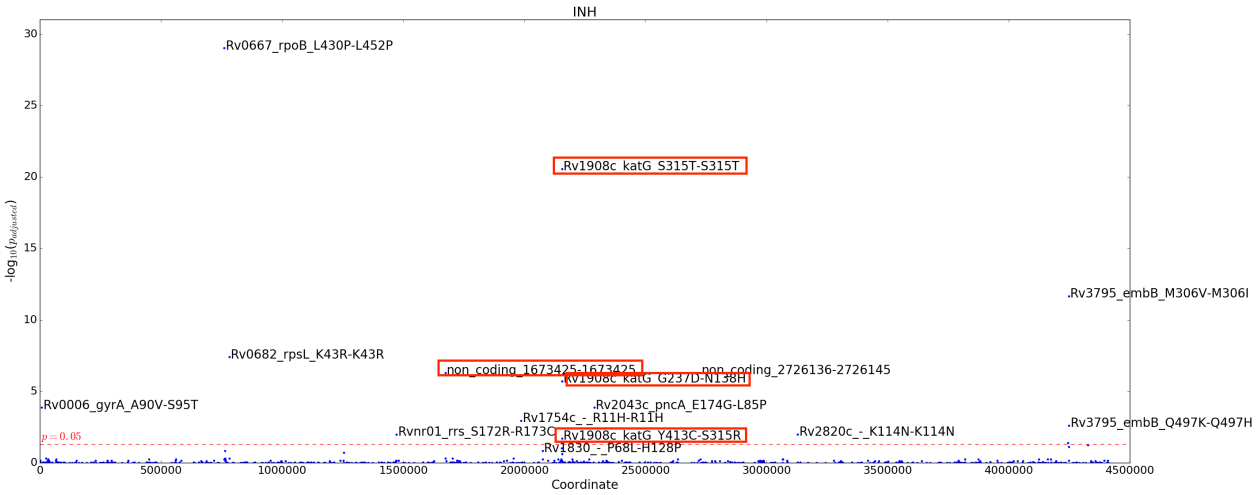


Figure 5.25: Genetic associations between clustered regions and isoniazid resistance for 376 strains from China.

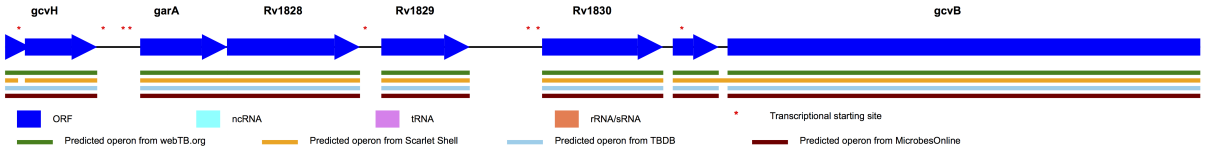


Figure 5.26: The relative location of Rv1830 and its adjacent genes in the *M. tuberculosis* genome.

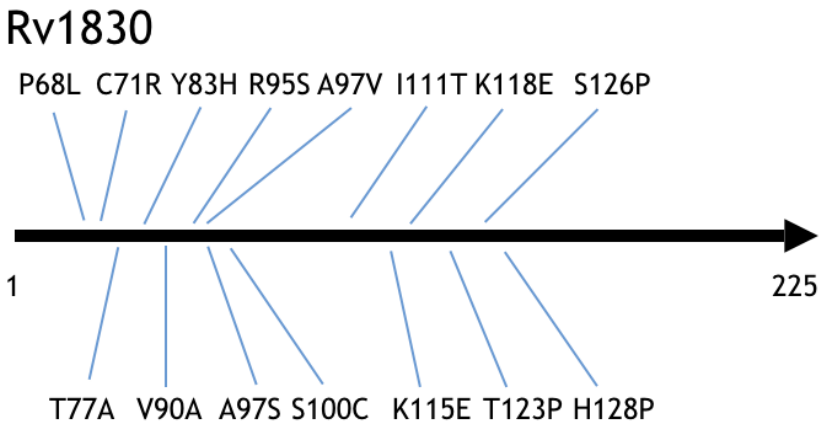


Figure 5.27: Relative locations of observed changes within the clustered region of Rv1830 in the dataset of 376 strains from China. Rv1830 has 225 amino acids.

Table 5.5: Distribution of phenotypes for strains harboring mutations in Rv1830. An INH-resistant strain represents that it is resistant to isoniazid.

Sites	Number of INH-resistant strains	Number of sensitive strains
P68L	1	0
C71R	1	0
T77A	1	0
Y83H	2	0
V90A	1	0
R95S	1	0
A97S	1	0
A97V	1	0
S100C	1	0
I111T	1	0
K115E	1	0
K118E	0	1
T123P	0	1
S126P	1	0
H128P	0	1

5.4 Discussion

We presented four types of methods for association tests:

1. site-based method (test polymorphisms at each individual site against phenotypes of interest)
2. gene-based method (pooling polymorphisms within a gene using a burden test)
3. k-mer-based method (grouping all windows of k adjacent polymorphisms)
4. two-phase cluster-based method (optimization of clustered polymorphisms using the ECC algorithm).

The known loci conferring drug resistance differ in sizes of loci within a gene or intergenic region. In one case, a single polymorphism within the gene is associated with drug resistance, for example, the S315T codon at *katG*. In another case, causal variants are involved in non-contiguous parts of the sequence, such as K43R and K88T codons at *rpsL*. In other cases, a group of locally adjacent polymorphisms (a SNP cluster) is associated with drug resistance, for instance, the RDRR region of *rpoB*, intergenic region between *embC-embA*, codons of M306V and M306I or G406S and G406A at *embB*, the upstream region of *eis*. In the association tests, regions like *eis* promoter (associated with KAN resistance) and *pncA* (associated with PZA resistance) are not detected by site-based or gene-based testings, yet they are detected in our evolutionary cluster-based convergence test (ECC). Since the size of the grouping of polymorphic sites is different for each case and homoplasic sites suggest the bacteria in the population are under positive selection, our two-phase optimization method (ECC) locates regions harboring more changes than others and then tests for associations subsequently. The proposed method of optimization of clustered polymorphisms as genotypic traits performs better than using other genotypes (individual sites or grouping by a gene) in terms of accuracy of detecting the known resistant-associated loci. In the association test against RIF, the cluster-based method successfully groups sites within the RDRR region of *rpoB* as a region and then reports that it is statistically significantly associated with RIF resistance. It also identifies mutations in *rpoA* (A180V-T187T) and *rpoC* (P481T-L527V, N698S-F831L, V431M-F452L and E1033A-A1047P) as clustered regions where most changes are resistant to RIF. Several compensatory mutations have already been described in *rpoA* and *rpoC* that involved in RNA polymerase subunits [91]. However, the other three methods fail to report associations between RIF resistance and *rpoA*. Previous methods identify a weak association between *rpoC* and RIF resistance.

There are at least two limitations of our method: epistasis and co-resistance of drugs. Epistasis is the interaction of two or more loci. If codons that confer resistance to a certain drug are spread out in several positions across a gene or several genes, our method might not identify them since it is only able to identify polymorphisms that are clustered within a consecutive region. Non-resistance-related mutations in between could decrease the significance. For example, there are 8

polymorphic sites between codons 306 and 406 in *embB*. We obtained 14 changes yet 3 of them occur in the sensitive branches.

The co-resistance of drugs means a strain that is resistant to one drug may have higher propensity to be resistant to another one. For example, INH and RIF resistance in *M. tuberculosis* strains, which is defined as the multi-drug resistant (MDR) strains. The dataset of strains from Peru contains up to 35.9% of MDR strains. Thus, ambiguity exists in the association test. Loci that confer INH resistance are also ranked top in the association test of RIF resistance and vice versa.

It might be possible to combine the two-phase method in the future to detect clustered regions associated with antibiotic resistance simultaneously. Since the current method detects the clusterings of polymorphisms first without considering DST, some regions are sub-optimal, for example, *gyrA*. Our method identifies a clustered region within *gyrA* from codons to D89N to S95T in the first phase. Thirty-three changes exist in *gyrA* spanning 2361 bp. The region is detected as being clustered since it harbors 23 changes spanning 5 bp. In the second phase, the region is also identified as associated with CPX resistance since 20 changes occur in the resistant branches. However, mutation at codon S95T at *gyrA* has been reported that it is a lineage-specific mutation and not linked to CPX resistance [112]. In the dataset from Peru, 278 strains have the mutation at codon S95T, but the mutation is inherited from a common ancestor for all these strains. That is, there is only one change in the tree and the change occurs in the CPX-sensitive branch. Thus, the clustered region is sub-optimal in the association test. In the future, it would be desirable to modify the algorithm to identify regions that maximize clustering and association simultaneously.

6. CONCLUSION *

The neutral theory of evolution assumes that bacteria acquire nucleotide substitutions spread throughout the genome spontaneously during evolution. In many bacteria, these are inherited clonally. Bacteria may also obtain genetic materials from the environment or exchange with other organisms via transduction, transmission, or conjugation. The evolutionary history of bacteria in a population can be inferred by the global phylogenetic tree estimated from the polymorphisms. If a point mutation or a segment of nucleotide substitutions is incongruent with the tree, it creates the appearance of homoplasy. Homoplasy occurs when a mutation by itself or a segment of consecutive mutations (recombination) does not descend from a common ancestor but appears on at least two branches independently. Homoplasy often indicates positive selection. For divergent species, identifying regions that involve recombination would provide a better understanding of evolutionary history. For clonal species like *Mycobacterium tuberculosis*, the homoplasic mutations may be associated with drug resistance.

To characterize homoplasy in bacterial genomes, in the first part of this work, we developed a polymorphism incompatibility method to identify any region where a recombination event occurs without requiring the reconstruction of a phylogenetic tree. We use a sliding window to scan for potential breakpoints throughout the genome by locating sites with lower compatibility scores, and then we assess the statistical significance of the breakpoints by a permutation test. Our method (ptACR) is able to practically determine the compatibility of sites of binary- and multi-state characters and detect the recombination boundaries of lower average compatibility ratio with the assessment of statistical significance as candidate breakpoints. The evaluation of our method on the simulated datasets of varying substitution rates and heterogeneity shows that ptACR is sensitive,

*Part of the data reported in this chapter is reprinted with permission from "A statistical method to identify recombination in bacterial genomes based on SNP incompatibility" by Y.-P. Lai and T. R. Ioerger, 2018. *BMC Bioinformatics*, 19, 450, Copyright [2018] by BioMed Central. DOI:10.1186/s12859-018-2456-z.
Part of the data reported in this chapter is reprinted with permission from "A compatibility approach to identify recombination breakpoints in bacterial and viral genomes" by Y.-P. Lai and T. R. Ioerger, 2017. *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pp. 11-20, Copyright [2017] by Association for Computing Machinery. DOI:10.1145/3107411.3107432.

yet has a relatively lower false positive rate than method without using a permutation test, supporting its ability to characterize mosaic genomes and identify the regions of distinct phylogenetic histories. With the detection of recombination events in clinical isolates of *Staphylococcus aureus* and *Mycobacterium avium*, it could provide a better understanding of evolutionary relationships among bacterial isolates that are not clonal.

In the second part of this work, we studied methods for exploiting homoplasy to identify sites/genes linked with drug resistance. We developed a two-phase evolutionary cluster-based convergence test (ECC) to estimate associations between genetic variants and drug-resistant phenotypes with accounting for mutation rates, homoplasy, and population stratification. We locate regions where changes cluster within a smaller span more strongly than expected from a Poisson distribution. Homoplasic sites tend to nucleate clustered regions, and can even be clustered on their own. Then we test the genotypes as clustered regions of changes against phenotypic traits of drug resistance using a hypergeometric test to identify potential causal variants. We evaluate the performance of ECC on three empirical datasets of clinical isolates of *Mycobacterium tuberculosis* and compare its results to those from the site-based, gene-based and k-mer-based methods. The results show that our cluster-based method is able to identify known drug-resistant loci more accurately compared to other methods. The clustering in phase one and the focus of association testing on the clustered regions in phase two give an advantage to homoplasic sites. It has the potential to identify novel polymorphisms that confer drug resistance.

REFERENCES

[1] E. L. Sonnhammer, G. Von Heijne, A. Krogh, *et al.*, "A hidden markov model for predicting transmembrane helices in protein sequences.," in *ISMB*, vol. 6, pp. 175–182, 1998.

[2] M. C. Brandley, D. L. Warren, A. D. Leaché, and J. A. McGuire, "Homoplasy and clade support," *Systematic Biology*, vol. 58, no. 2, pp. 184–198, 2009.

[3] J. M. Smith, N. H. Smith, M. O'Rourke, and B. G. Spratt, "How clonal are bacteria?," *Proceedings of the National Academy of Sciences*, vol. 90, no. 10, pp. 4384–4388, 1993.

[4] B. J. Shapiro, "How clonal are bacteria over time?," *Current Opinion in Microbiology*, vol. 31, pp. 116–123, 2016.

[5] E. Krzywinska, J. Krzywinski, and J. S. Schorey, "Naturally occurring horizontal gene transfer and homologous recombination in *Mycobacterium*," *Microbiology*, vol. 150, no. 6, pp. 1707–1712, 2004.

[6] B. Marklund, D. Speert, and R. Stokes, "Gene replacement through homologous recombination in *Mycobacterium intracellulare*," *Journal of Bacteriology*, vol. 177, no. 21, pp. 6100–6105, 1995.

[7] E. C. Holmes, R. Urwin, and M. Maiden, "The influence of recombination on the population structure and evolution of the human pathogen *Neisseria meningitidis*," *Molecular Biology and Evolution*, vol. 16, no. 6, pp. 741–749, 1999.

[8] Y. Kong, J. H. Ma, K. Warren, R. S. Tsang, D. E. Low, F. B. Jamieson, D. C. Alexander, and W. Hao, "Homologous recombination drives both sequence diversity and gene content variation in *Neisseria meningitidis*," *Genome Biology and Evolution*, vol. 5, no. 9, pp. 1611–1627, 2013.

[9] X. Didelot, R. Bowden, T. Street, T. Golubchik, C. Spencer, G. McVean, V. Sangal, M. F. Anjum, M. Achtman, D. Falush, *et al.*, "Recombination and population structure

in *Salmonella enterica*," *PLoS Genetics*, vol. 7, no. 7, p. e1002191, 2011.

[10] S. Takuno, T. Kado, R. P. Sugino, L. Nakhleh, and H. Innan, "Population genomics in bacteria: a case study of *Staphylococcus aureus*," *Molecular Biology and Evolution*, vol. 29, no. 2, pp. 797–809, 2011.

[11] R. G. Everitt, X. Didelot, E. M. Batty, R. R. Miller, K. Knox, B. C. Young, R. Bowden, A. Auton, A. Votintseva, H. Larner-Svensson, *et al.*, "Mobile elements drive recombination hotspots in the core genome of *Staphylococcus aureus*," *Nature Communications*, vol. 5, 2014.

[12] E. M. Driebe, J. W. Sahl, C. Roe, J. R. Bowers, J. M. Schupp, J. D. Gillece, E. Kelley, L. B. Price, T. R. Pearson, C. M. Hepp, *et al.*, "Using whole genome analysis to examine recombination across diverse sequence types of *Staphylococcus aureus*," *PLoS One*, vol. 10, no. 7, p. e0130955, 2015.

[13] C. Chaguza, J. E. Cornick, and D. B. Everett, "Mechanisms and impact of genetic recombination in the evolution of *Streptococcus pneumoniae*," *Computational and Structural Biotechnology Journal*, vol. 13, pp. 241–247, 2015.

[14] A. Kalia, B. G. Spratt, M. C. Enright, and D. E. Bessen, "Influence of recombination and niche separation on the population genetic structure of the pathogen *Streptococcus pyogenes*," *Infection and Immunity*, vol. 70, no. 4, pp. 1971–1983, 2002.

[15] J. Maynard Smith and N. H. Smith, "Detecting recombination from gene trees.," *Molecular Biology and Evolution*, vol. 15, no. 5, pp. 590–599, 1998.

[16] M. R. Farhat, B. J. Shapiro, K. J. Kieser, R. Sultana, K. R. Jacobson, T. C. Victor, R. M. Warren, E. M. Streicher, A. Calver, A. Sloutsky, *et al.*, "Genomic analysis identifies targets of convergent positive selection in drug-resistant *Mycobacterium tuberculosis*," *Nature Genetics*, vol. 45, no. 10, p. 1183, 2013.

[17] T. D. Read and R. C. Massey, "Characterizing the genetic basis of bacterial phenotypes using genome-wide association studies: a new direction for bacteriology," *Genome Medicine*, vol. 6, no. 11, p. 109, 2014.

[18] P. E. Chen and B. J. Shapiro, "The advent of genome-wide association studies for bacteria," *Current Opinion in Microbiology*, vol. 25, pp. 17–24, 2015.

[19] F. Coll, J. Phelan, G. A. Hill-Cawthorne, M. B. Nair, K. Mallard, S. Ali, A. M. Abdallah, S. Alghamdi, M. Alsomali, A. O. Ahmed, *et al.*, "Genome-wide analysis of multi-and extensively drug-resistant *Mycobacterium tuberculosis*," *Nature Genetics*, vol. 50, no. 2, p. 307, 2018.

[20] E. J. Feil and B. G. Spratt, "Recombination and the population structures of bacterial pathogens," *Annual Reviews in Microbiology*, vol. 55, no. 1, pp. 561–590, 2001.

[21] D. Posada and K. A. Crandall, "The effect of recombination on the accuracy of phylogeny estimation," *Journal of Molecular Evolution*, vol. 54, no. 3, pp. 396–402, 2002.

[22] X. Didelot and M. C. Maiden, "Impact of recombination on bacterial evolution," *Trends in Microbiology*, vol. 18, no. 7, pp. 315–322, 2010.

[23] D. Posada and K. A. Crandall, "Evaluation of methods for detecting recombination from DNA sequences: computer simulations," *Proceedings of the National Academy of Sciences*, vol. 98, no. 24, pp. 13757–13762, 2001.

[24] I. B. Jakobsen and S. Easteal, "A program for calculating and displaying compatibility matrices as an aid in determining reticulate evolution in molecular sequences," *Bioinformatics*, vol. 12, no. 4, pp. 291–295, 1996.

[25] T. C. Bruen, H. Philippe, and D. Bryant, "A simple and robust statistical test for detecting the presence of recombination," *Genetics*, vol. 172, no. 4, pp. 2665–2681, 2006.

[26] X. Didelot and D. J. Wilson, "ClonalFrameML: efficient inference of recombination in whole bacterial genomes," *PLoS Computational Biology*, vol. 11, no. 2, p. e1004041, 2015.

[27] B. L. Maidak, J. R. Cole, T. G. Lilburn, C. T. Parker Jr, P. R. Saxman, J. M. Stredwick, G. M. Garrity, B. Li, G. J. Olsen, S. Pramanik, *et al.*, "The RDP (ribosomal database project) continues," *Nucleic Acids Research*, vol. 28, no. 1, pp. 173–174, 2000.

[28] S. L. Kosakovsky Pond, D. Posada, M. B. Gravenor, C. H. Woelk, and S. D. Frost, "GARD: a genetic algorithm for recombination detection," *Bioinformatics*, vol. 22, no. 24, pp. 3096–3098, 2006.

[29] Y.-P. Lai and T. R. Ioerger, "A compatibility approach to identify recombination breakpoints in bacterial and viral genomes," in *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology,and Health Informatics*, ACM-BCB '17, (New York, NY, USA), pp. 11–20, ACM, 2017.

[30] T. J. Warnow, "Constructing phylogenetic trees efficiently using compatibility criteria," *New Zealand Journal of Botany*, vol. 31, no. 3, pp. 239–247, 1993.

[31] Y.-P. Lai and T. R. Ioerger, "A statistical method to identify recombination in bacterial genomes based on SNP incompatibility," *BMC Bioinformatics*, vol. 19, no. 1, p. 450, 2018.

[32] G. F. Estabrook, C. Johnson, and F. McMorris, "A mathematical foundation for the analysis of cladistic character compatibility," *Mathematical Biosciences*, vol. 29, no. 1-2, pp. 181–187, 1976.

[33] R. R. Hudson and N. L. Kaplan, "Statistical properties of the number of recombination events in the history of a sample of DNA sequences," *Genetics*, vol. 111, no. 1, pp. 147–164, 1985.

[34] P. Buneman, "A characterisation of rigid circuit graphs," *Discrete Mathematics*, vol. 9, no. 3, pp. 205–212, 1974.

[35] J. Felsenstein, "PHYLIP-phylogeny inference package (version 3.2)," *Cladistics*, vol. 5, no. 163, p. 6, 1989.

[36] D. Sankoff, "Simultaneous solution of the RNA folding, alignment and protosequence problems," *SIAM Journal on Applied Mathematics*, vol. 45, no. 5, pp. 810–825, 1985.

[37] J. Sjöstrand, L. Arvestad, J. Lagergren, and B. Sennblad, "GenPhyloData: realistic simulation of gene family evolution," *BMC Bioinformatics*, vol. 14, no. 1, p. 209, 2013.

[38] T. Horiike, D. Miyata, Y. Tateno, and R. Minai, "HGT-Gen: a tool for generating a phylogenetic tree with horizontal gene transfer," *Bioinformation*, vol. 7, no. 5, p. 211, 2011.

[39] A. Rambaut and N. C. Grass, "Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees," *Bioinformatics*, vol. 13, no. 3, pp. 235–238, 1997.

[40] M. Hasegawa, H. Kishino, and T.-a. Yano, "Dating of the human-ape splitting by a molecular clock of mitochondrial DNA," *Journal of Molecular Evolution*, vol. 22, no. 2, pp. 160–174, 1985.

[41] M. K. Kuhner and J. Felsenstein, "A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates.," *Molecular Biology and Evolution*, vol. 11, no. 3, pp. 459–468, 1994.

[42] D. F. Robinson and L. R. Foulds, "Comparison of phylogenetic trees," *Mathematical Biosciences*, vol. 53, no. 1-2, pp. 131–147, 1981.

[43] K. Zhang and D. Shasha, "Simple fast algorithms for the editing distance between trees and related problems," *SIAM Journal on Computing*, vol. 18, no. 6, pp. 1245–1262, 1989.

[44] M. M. Gutacker, J. C. Smoot, C. A. L. Migliaccio, S. M. Ricklefs, S. Hua, D. V. Cousins, E. A. Graviss, E. Shashkina, B. N. Kreiswirth, and J. M. Musser, "Genome-wide analysis of synonymous single nucleotide polymorphisms in *Mycobacterium tuberculosis* complex organisms: resolution of genetic relationships among closely related microbial strains," *Genetics*, vol. 162, no. 4, pp. 1533–1543, 2002.

[45] C. Arnold, "Molecular evolution of *Mycobacterium tuberculosis*," *Clinical Microbiology and Infection*, vol. 13, no. 2, pp. 120–128, 2007.

[46] T. M. Walker, C. L. Ip, R. H. Harrell, J. T. Evans, G. Kapatai, M. J. Dedicoat, D. W. Eyre, D. J. Wilson, P. M. Hawkey, D. W. Crook, *et al.*, "Whole-genome sequencing to delineate

Mycobacterium tuberculosis outbreaks: a retrospective observational study," *The Lancet Infectious Diseases*, vol. 13, no. 2, pp. 137–146, 2013.

[47] D. H. Huson and D. Bryant, "Application of phylogenetic networks in evolutionary studies," *Molecular Biology and Evolution*, vol. 23, no. 2, pp. 254–267, 2005.

[48] N. Kannan, Y.-P. Lai, M. Haug, M. Lilleness, S. S. Bakke, A. Marstad, H. Hov, T. Naustdal, J. E. Afset, T. R. Ioerger, *et al.*, "Genetic variation/evolution and differential host responses resulting from in-patient adaptation of *Mycobacterium avium*," *Infection and Immunity*, pp. IAI–00323, 2019.

[49] T. R. Ioerger, Y. Feng, K. Ganesula, X. Chen, K. M. Dobos, S. Fortune, W. R. Jacobs, V. Mizrahi, T. Parish, E. Rubin, *et al.*, "Variation among genome sequences of H37Rv strains of *Mycobacterium tuberculosis* from multiple laboratories," *Journal of Bacteriology*, vol. 192, no. 14, pp. 3645–3653, 2010.

[50] F. A. Khattak, A. Kumar, E. Kamal, R. Kunisch, and A. Lewin, "Illegitimate recombination: An efficient method for random mutagenesis in *Mycobacterium avium* subsp. *hominissuis*," *BMC Microbiology*, vol. 12, no. 1, p. 204, 2012.

[51] K.-i. Uchiya, S. Tomida, T. Nakagawa, S. Asahi, T. Nikai, and K. Ogawa, "Comparative genome analyses of *Mycobacterium avium* reveal genomic features of its subspecies and strains that cause progression of pulmonary disease," *Scientific Reports*, vol. 7, p. 39750, 2017.

[52] T. Dos Vultos, O. Mestre, T. Tonjum, and B. Gicquel, "DNA repair in *Mycobacterium tuberculosis* revisited," *FEMS Microbiology Reviews*, vol. 33, no. 3, pp. 471–487, 2009.

[53] T. Parish and N. G. Stoker, "The common aromatic amino acid biosynthesis pathway is essential in *Mycobacterium tuberculosis*," *Microbiology*, vol. 148, no. 10, pp. 3069–3077, 2002.

[54] V. Balasubramanian, M. S. Pavelka, S. S. Bardarov, J. Martin, T. R. Weisbrod, R. A. McAdam, B. R. Bloom, and W. R. Jacobs, "Allelic exchange in *Mycobacterium tubercu-*

losis with long linear recombination substrates.," *Journal of Bacteriology*, vol. 178, no. 1, pp. 273–279, 1996.

[55] S. Gagneux and P. M. Small, "Global phylogeography of *Mycobacterium tuberculosis* and implications for tuberculosis product development," *The Lancet Infectious Diseases*, vol. 7, no. 5, pp. 328–337, 2007.

[56] D. N. Muttucumaru and T. Parish, "The molecular biology of recombination in *Mycobacteria*: what do we know and how can we use it?," *Current Issues in Molecular Biology*, vol. 6, no. 2, pp. 145–158, 2004.

[57] P. Supply, M. Marceau, S. Mangenot, D. Roche, C. Rouanet, V. Khanna, L. Majlessi, A. Criscuolo, J. Tap, A. Pawlik, *et al.*, "Genomic analysis of smooth tubercle bacilli provides insights into ancestry and pathoadaptation of *Mycobacterium tuberculosis*," *Nature Genetics*, vol. 45, no. 2, pp. 172–179, 2013.

[58] T. A. Gray, J. A. Krywy, J. Harold, M. J. Palumbo, and K. M. Derbyshire, "Distributive conjugal transfer in *Mycobacteria* generates progeny with meiotic-like genome-wide mosaicism, allowing mapping of a mating identity locus," *PLoS Biology*, vol. 11, no. 7, p. e1001602, 2013.

[59] J. M. Musser and V. Kapur, "Clonal analysis of methicillin-resistant *Staphylococcus aureus* strains from intercontinental sources: association of the mec gene with divergent phylogenetic lineages implies dissemination by horizontal transfer and recombination.," *Journal of Clinical Microbiology*, vol. 30, no. 8, pp. 2058–2063, 1992.

[60] C. Wielders, A. Fluit, S. Brisse, J. Verhoef, and F. Schmitz, "mecA gene is widely disseminated in *Staphylococcus aureus* population," *Journal of Clinical Microbiology*, vol. 40, no. 11, pp. 3970–3975, 2002.

[61] S. Murray, B. Pascoe, G. Méric, L. Mageiros, K. Yahara, M. D. Hitchings, Y. Friedmann, T. S. Wilkinson, F. J. Gormley, D. Mack, *et al.*, "Recombination-mediated host adaptation

by avian *Staphylococcus aureus*," *Genome Biology and Evolution*, vol. 9, no. 4, pp. 830–842, 2017.

[62] R. A. Power, J. Parkhill, and T. de Oliveira, "Microbial genome-wide association studies: lessons from human GWAS," *Nature Reviews Genetics*, vol. 18, no. 1, p. 41, 2017.

[63] P. M. Visscher, M. A. Brown, M. I. McCarthy, and J. Yang, "Five years of GWAS discovery," *The American Journal of Human Genetics*, vol. 90, no. 1, pp. 7–24, 2012.

[64] R. J. Klein, C. Zeiss, E. Y. Chew, J.-Y. Tsai, R. S. Sackler, C. Haynes, A. K. Henning, J. P. SanGiovanni, S. M. Mane, S. T. Mayne, *et al.*, "Complement factor H polymorphism in age-related macular degeneration," *Science*, vol. 308, no. 5720, pp. 385–389, 2005.

[65] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. De Bakker, M. J. Daly, *et al.*, "PLINK: a tool set for whole-genome association and population-based linkage analyses," *The American Journal of Human Genetics*, vol. 81, no. 3, pp. 559–575, 2007.

[66] A. L. Price, N. A. Zaitlen, D. Reich, and N. Patterson, "New approaches to population stratification in genome-wide association studies," *Nature Reviews Genetics*, vol. 11, no. 7, p. 459, 2010.

[67] B. K. Bulik-Sullivan, P.-R. Loh, H. K. Finucane, S. Ripke, J. Yang, N. Patterson, M. J. Daly, A. L. Price, B. M. Neale, S. W. G. of the Psychiatric Genomics Consortium, *et al.*, "LD score regression distinguishes confounding from polygenicity in genome-wide association studies," *Nature Genetics*, vol. 47, no. 3, p. 291, 2015.

[68] M. L. Freedman, D. Reich, K. L. Penney, G. J. McDonald, A. A. Mignault, N. Patterson, S. B. Gabriel, E. J. Topol, J. W. Smoller, C. N. Pato, *et al.*, "Assessing the impact of population stratification on genetic association studies," *Nature Genetics*, vol. 36, no. 4, p. 388, 2004.

[69] A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich, "Principal components analysis corrects for stratification in genome-wide association studies," *Nature Genetics*, vol. 38, no. 8, p. 904, 2006.

[70] H. M. Kang, N. A. Zaitlen, C. M. Wade, A. Kirby, D. Heckerman, M. J. Daly, and E. Eskin, "Efficient control of population structure in model organism association mapping," *Genetics*, vol. 178, no. 3, pp. 1709–1723, 2008.

[71] X. Zhou and M. Stephens, "Genome-wide efficient mixed-model analysis for association studies," *Nature Genetics*, vol. 44, no. 7, p. 821, 2012.

[72] S. G. Earle, C.-H. Wu, J. Charlesworth, N. Stoesser, N. C. Gordon, T. M. Walker, C. C. Spencer, Z. Iqbal, D. A. Clifton, K. L. Hopkins, *et al.*, "Identifying lineage effects when controlling for population structure improves power in bacterial association studies," *Nature Microbiology*, vol. 1, no. 5, p. 16041, 2016.

[73] N. D. Hicks, J. Yang, X. Zhang, B. Zhao, Y. H. Grad, L. Liu, X. Ou, Z. Chang, H. Xia, Y. Zhou, *et al.*, "Clinically prevalent mutations in *Mycobacterium tuberculosis* alter propionate metabolism and mediate multidrug tolerance," *Nature Microbiology*, vol. 3, no. 9, p. 1032, 2018.

[74] C. Collins and X. Didelot, "A phylogenetic method to perform genome-wide association studies in microbes that accounts for population structure and recombination," *PLoS Computational Biology*, vol. 14, no. 2, p. e1005958, 2018.

[75] P. Supply, R. M. Warren, A.-L. Bañuls, S. Lesjean, G. D. Van Der Spuy, L.-A. Lewis, M. Tibayrenc, P. D. Van Helden, and C. Locht, "Linkage disequilibrium between minisatellite loci supports clonal evolution of Mycobacterium tuberculosis in a high tuberculosis incidence area," *Molecular Microbiology*, vol. 47, no. 2, pp. 529–538, 2003.

[76] T. Dos Vultos, O. Mestre, J. Rauzier, M. Golec, N. Rastogi, V. Rasolofo, T. Tonjum, C. Sola, I. Matic, and B. Gicquel, "Evolution and diversity of clonal bacteria: the paradigm of *Mycobacterium tuberculosis*," *PloS One*, vol. 3, no. 2, p. e1538, 2008.

[77] F. Coll, M. Preston, J. A. Guerra-Assunção, G. Hill-Cawthorn, D. Harris, J. Perdigão, M. Viveiros, I. Portugal, F. Drobniewski, S. Gagneux, *et al.*, "PolyTB: a genomic variation map for *Mycobacterium tuberculosis*," *Tuberculosis*, vol. 94, no. 3, pp. 346–354, 2014.

[78] T. M. Walker, T. A. Kohl, S. V. Omar, J. Hedge, C. D. O. Elias, P. Bradley, Z. Iqbal, S. Feuerriegel, K. E. Niehaus, D. J. Wilson, *et al.*, "Whole-genome sequencing for prediction of *Mycobacterium tuberculosis* drug susceptibility and resistance: a retrospective cohort study," *The Lancet Infectious Diseases*, vol. 15, no. 10, pp. 1193–1202, 2015.

[79] C. E. Cade, A. C. Dlouhy, K. F. Medzihradszky, S. P. Salas-Castillo, and R. A. Ghiladi, "Isoniazid-resistance conferring mutations in *Mycobacterium tuberculosis katG*: Catalase, peroxidase, and INH-NADH adduct formation activities," *Protein Science*, vol. 19, no. 3, pp. 458–474, 2010.

[80] A. Wagner, "Rapid detection of positive selection in genes and genomes through variation clusters," *Genetics*, vol. 176, no. 4, pp. 2451–2463, 2007.

[81] L. Grandjean, R. H. Gilman, T. Iwamoto, C. U. Köser, J. Coronel, M. Zimic, M. E. Török, D. Ayabina, M. Kendall, C. Fraser, *et al.*, "Convergent evolution and topologically disruptive polymorphisms among multidrug-resistant tuberculosis in Peru," *PloS One*, vol. 12, no. 12, p. e0189838, 2017.

[82] N. S. Osório, F. Rodrigues, S. Gagneux, J. Pedrosa, M. Pinto-Carbó, A. G. Castro, D. Young, I. Comas, and M. Saraiva, "Evidence for diversifying selection in a set of *Mycobacterium tuberculosis* genes in response to antibiotic- and nonantibiotic-related pressure," *Molecular Biology and Evolution*, vol. 30, no. 6, pp. 1326–1336, 2013.

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

[84] A. Banerjee, E. Dubnau, A. Quemard, V. Balasubramanian, K. S. Um, T. Wilson, D. Collins, G. de Lisle, and W. R. Jacobs, "*inhA*, a gene encoding a target for isoniazid and ethionamide

in *Mycobacterium tuberculosis*," *Science*, vol. 263, no. 5144, pp. 227–230, 1994.

[85] A. S. Pym, B. Saint-Joanis, and S. T. Cole, "Effect of *katG* mutations on the virulence of *Mycobacterium tuberculosis* and the implication for transmission in humans," *Infection and Immunity*, vol. 70, no. 9, pp. 4955–4960, 2002.

[86] I. Mokrousov, O. Narvskaya, T. Otten, E. Limeschenko, L. Steklova, and B. Vyshnevskiy, "High prevalence of *katG* Ser315Thr substitution among isoniazid-resistant *Mycobacterium tuberculosis* clinical isolates from northwestern russia, 1996 to 2001," *Antimicrobial Agents and Chemotherapy*, vol. 46, no. 5, pp. 1417–1424, 2002.

[87] F. Lanzas, P. C. Karakousis, J. C. Sacchettini, and T. R. Ioerger, "Multidrug-resistant tuberculosis in Panama is driven by clonal expansion of a multidrug-resistant *Mycobacterium tuberculosis* strain related to the KZN extensively drug-resistant *M. tuberculosis* strain from South Africa," *Journal of Clinical Microbiology*, vol. 51, no. 10, pp. 3277–3285, 2013.

[88] S. Sreevatsan, X. Pan, K. E. Stockbauer, N. D. Connell, B. N. Kreiswirth, T. S. Whittam, and J. M. Musser, "Restricted structural gene polymorphism in the *Mycobacterium tuberculosis* complex indicates evolutionarily recent global dissemination," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 94, no. 18, pp. 9869–9874, 1997.

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

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

[91] I. Comas, S. Borrell, A. Roetzer, G. Rose, B. Malla, M. Kato-Maeda, J. Galagan, S. Niemann, and S. Gagneux, "Whole-genome sequencing of rifampicin-resistant *Mycobacterium*

tuberculosis strains identifies compensatory mutations in RNA polymerase genes," *Nature Genetics*, vol. 44, no. 1, pp. 106–110, 2012.

[92] S. Ramaswamy and J. Musser, "Molecular genetic basis of antimicrobial agent resistance in *Mycobacterium tuberculosis*: 1998 update," *Tubercle and Lung Disease*, vol. 79, no. 1, pp. 3–29, 1998.

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

[94] S. Sreevatsan, K. E. Stockbauer, X. Pan, B. N. Kreiswirth, S. L. Moghazeh, W. R. Jacobs, A. Telenti, and J. M. Musser, "Ethambutol resistance in *Mycobacterium tuberculosis*: critical role of *embB* mutations.," *Antimicrobial Agents and Chemotherapy*, vol. 41, no. 8, pp. 1677–1681, 1997.

[95] H. Safi, S. Lingaraju, A. Amin, S. Kim, M. Jones, M. Holmes, M. McNeil, S. N. Peterson, D. Chatterjee, R. Fleischmann, and D. Alland, "Evolution of high-level ethambutol-resistant tuberculosis through interacting mutations in decaprenylphosphoryl-β-D-arabinose biosynthetic and utilization pathway genes," *Nature Genetics*, vol. 45, no. 10, pp. 1190–1197, 2013.

[96] L. He, X. Wang, P. Cui, J. Jin, J. Chen, W. Zhang, and Y. Zhang, "ubiA (Rv3806c) encoding dppr synthase involved in cell wall synthesis is associated with ethambutol resistance in *Mycobacterium tuberculosis*.," *Tuberculosis*, vol. 95, no. 2, pp. 149–154, 2015.

[97] M. Finken, P. Kirschner, A. Meier, A. Wrede, and E. C. Böttger, "Molecular basis of streptomycin resistance in *Mycobacterium tuberculosis*: alterations of the ribosomal protein S12 gene and point mutations within a functional 16S ribosomal RNA pseudoknot," *Molecular Microbiology*, vol. 9, no. 6, pp. 1239–1246, 1993.

[98] S. Okamoto, A. Tamaru, C. Nakajima, K. Nishimura, Y. Tanaka, S. Tokuyama, Y. Suzuki, and K. Ochi, "Loss of a conserved 7-methylguanosine modification in 16S rRNA confers

low-level streptomycin resistance in bacteria.," *Molecular Microbiology*, vol. 63, no. 4, pp. 1096–1106, 2007.

[99] S. Y. Wong, J. S. Lee, H. K. Kwak, L. E. Via, H. I. M. Boshoff, and C. E. Barry, "Mutations in gidB confer low-level streptomycin resistance in *Mycobacterium tuberculosis*," *Antimicrobial Agents and Chemotherapy*, vol. 55, no. 6, pp. 2515–2522, 2011.

[100] A. Scorpio and Y. Zhang, "Mutations in *pncA*, a gene encoding pyrazinamidase/nicotinamidase, cause resistance to the antituberculous drug pyrazinamide in tubercle bacillus.," *Nature Medicine*, vol. 2, no. 6, pp. 662–667, 1996.

[101] K. Hirano, M. Takahashi, Y. Kazumi, Y. Fukasawa, and C. Abe, "Mutation in *pncA* is a major mechanism of pyrazinamide resistance in *Mycobacterium tuberculosis*," *Tubercle and Lung Disease*, vol. 78, no. 2, pp. 117–122, 1998.

[102] N. A. Dillon, N. D. Peterson, B. C. Rosen, and A. D. Baughn, "Pantothenate and pantetheine antagonize the antitubercular activity of pyrazinamide.," *Antimicrobial Agents and Chemotherapy*, vol. 58, no. 12, pp. 7258–7263, 2014.

[103] J. C. Evans, C. Trujillo, Z. Wang, H. Eoh, S. Ehrt, D. Schnappinger, H. I. M. Boshoff, K. Y. Rhee, C. E. Barry, and V. Mizrahi, "Validation of *CoaBC* as a bactericidal target in the coenzyme a pathway of *Mycobacterium tuberculosis*," *ACS Infectious Diseases*, vol. 2, no. 12, pp. 958–968, 2016.

[104] M. A. Zaunbrecher, R. D. Sikes, B. Metchock, T. M. Shinnick, and J. E. Posey, "Overexpression of the chromosomally encoded aminoglycoside acetyltransferase *eis* confers kanamycin resistance in *Mycobacterium tuberculosis*," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 47, pp. 20004–20009, 2009.

[105] H. E. Takiff, L. Salazar, C. Guerrero, W. Philipp, W. M. Huang, B. Kreiswirth, S. T. Cole, W. R. Jacobs, and A. Telenti, "Cloning and nucleotide sequence of *Mycobacterium tuberculosis gyrA* and *gyrB* genes and detection of quinolone resistance mutations.," *Antimicrobial Agents and Chemotherapy*, vol. 38, no. 4, pp. 773–780, 1994.

[106] M. A. DeJesus, E. R. Gerrick, W. Xu, S. W. Park, J. E. Long, C. C. Boutte, E. J. Rubin, D. Schnappinger, S. Ehrt, S. M. Fortune, *et al.*, "Comprehensive essentiality analysis of the *Mycobacterium tuberculosis* genome via saturating transposon mutagenesis," *MBio*, vol. 8, no. 1, pp. e02133–16, 2017.

[107] D. L. Swofford, "PAUP*. Phylogenetic analysis using parsimony (*and other methods).," *Version 4. Sinauer Associates, Sunderland, Massachusetts.*, 2003.

[108] I. Comas, M. Coscolla, T. Luo, S. Borrell, K. E. Holt, M. Kato-Maeda, J. Parkhill, B. Malla, S. Berg, G. Thwaites, D. Yeboah-Manu, G. Bothamley, J. Mei, L. Wei, S. Bentley, S. R. Harris, S. Niemann, R. Diel, A. Aseffa, Q. Gao, D. Young, and S. Gagneux, "Out-of-Africa migration and neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans," *Nature Genetics*, vol. 45, no. 10, pp. 1176–1182, 2013.

[109] S. J. Kim, "Drug-susceptibility testing in tuberculosis: methods and reliability of results," *European Respiratory Journal*, vol. 25, no. 3, pp. 564–569, 2005.

[110] B. van Klingeren, M. Dessens-Kroon, T. van der Laan, K. Kremer, and D. van Soolingen, "Drug susceptibility testing of *Mycobacterium tuberculosis* complex by use of a high-throughput, reproducible, absolute concentration method," *Journal of Clinical Microbiology*, vol. 45, no. 8, pp. 2662–2668, 2007.

[111] K. Cohen, T. El-Hay, K. L. Wyres, O. Weissbrod, V. Munsamy, C. Yanover, R. Aharonov, O. Shaham, T. C. Conway, Y. Goldschmidt, W. R. Bishai, and A. S. Pym, "Paradoxical hypersusceptibility of drug-resistant *Mycobacterium tuberculosis* to beta-lactam antibiotics," *EBioMedicine*, vol. 9, pp. 170–179, 2016.

[112] R. W. Lau, P.-L. Ho, R. Y. Kao, W.-W. Yew, T. C. Lau, V. C. Cheng, K.-Y. Yuen, S. K. Tsui, X. Chen, and W.-C. Yam, "Molecular characterization of fluoroquinolone resistance in *Mycobacterium tuberculosis*: functional analysis of *gyrA* mutation at position 74," *Antimicrobial Agents and Chemotherapy*, vol. 55, no. 2, pp. 608–614, 2011.