Detecting mixed *Mycobacterium tuberculosis* infections and differences in drug susceptibility with WGS data

BY

Arlin Keo

(4257111)

Thesis report

Defense date: January 20, 2015

Bioinformatics, Master of Computer Science

Faculty Electrical Engineering, Mathematics and Computer Science

Delft University of Technology, the Netherlands

Preface

This report has been written in 2015/2016 and describes a one-year thesis project at the University of Technology in Delft under the supervision of Dr. Thomas Abeel. The work has been done during the final year of the master Computer Science and the specialization track Bioinformatics, and has been performed at the Delft Bioinformatics lab.

During the project I developed a tool for detecting mixed tuberculosis infections. The tool uses a large marker set to identify and distinguish multiple strains from one sample and it is applied to whole genome sequencing data. These genomic markers are specific to hierarchical clusters of strains derived from a reconstructed phylogenetic tree of tuberculosis strains. Based on these markers the presence of the hierarchical clusters is determined and the tool finds paths of present clusters in the phylogenetic tree. Paths that lead back to the root of the tree support the presence of the tuberculosis species and also give more details about the infecting strains at a higher resolution. When there are multiple paths, the tool finds that the sample consists of multiple strains and thus is a mixed infection. It also estimates the frequencies of the infecting strains, which were compared with detected drug resistance profiles. In this study, I found that mixed infections can be responsible for multi-drug resistance.

This project will be judged by the thesis committee: Dr. Thomas Abeel, Prof. Marcel Reinders, Drs. Thies Gehrmann, Drs. Jasper Linthorst and Dr. Aljoscha Wahl.

I sincerely thank Dr. Thomas Abeel for his guidance and encouragement in carrying out this project work. Without his assistance this work would not have been completed. I also thank the Broad Institute for supplying the data, all the people from the Delft Bioinformatics lab for their assistance and my family and friends who have supported me. The success and final outcome of this project is achieved thanks to them.

Arlin Keo

January 13, 2015

Abstract

Tuberculosis is a pulmonary disease caused by the pathogen *Mycobacterium tuberculosis* and is the second leading cause of death from an infectious disease. Individuals infected by multiple strains referred to as a mixed infection are associated with poor treatment outcomes when the infecting strains differ in their susceptibility against antibiotics. Studies aimed to detect mixed infections are likely to underestimate the true prevalence of mixed infections, because conventional genotypic methods have limited sensitivity to distinguish TB strains. Tools are needed to distinguish strains at a finer resolution and allow the simultaneous detection of multiple strains. Whole genome sequencing yields more information and therefore provides increased resolution to also distinguish closely related strains.

In this study, I detect mixed infections by using a large number of group-specific SNP markers obtained from sequence data. I associated SNPs to clusters in the MTBC phylogenetic tree to obtain cluster-specific SNP markers that allow detecting and estimating frequencies of present strain types at different levels in the phylogeny. The prevalence of mixed infections was found to be ~10% of which half were mixed at the sub-lineage level. Approximately 15% of the mixed infections were found to have ambiguous SNPs corresponding to locations that are known to cause drug resistance, indicating the presence of MTB populations with differing drug susceptibility profiles. The results show that patient-level multi-drug resistance can be caused by multiple strains each with their own resistance to a particular drug. This work illustrates the dire need for high-resolution molecular diagnostics that can pinpoint the exact nature of the infection.

Index

1. Introduction

Mycobacterium tuberculosis is a bacterial pathogen that causes the pulmonary disease tuberculosis (TB), and it infects and kills millions of people worldwide (WHO, 2014). The study of genetic diversity within the *M. tuberculosis* complex (MTBC) is complicated by mixed TB infections, which happens when a person is infected with more than one distinct strain type of MTBC. Undetected mixed infections affect estimates of transmission events (Eyre *et al.*, 2013) and often results in poor diagnosis and treatment of patients as the bacterial sub-populations may have undetected differences in drug susceptibility (Zetola *et al.*, 2014).

MTBC seems to be a clonal expansion from a progenitor population that arose at least 70,000 years ago and spread with human migration out of Africa. This was confirmed in a whole genome sequencing (WGS) study by Comas *et al.*, 2013 who revealed the congruence between the MTBC phylogeny and a tree constructed from human mitochondrial genomes. Figure 1 shows the phylogenetic tree of MTBC and *M. canettii* that shares the most recent common ancestor (Blouin *et al.*, 2014). Strains of MTBC have a ~4.4 Mbp genome and include human adapted *M. tuberculosis* and *M. africanum*, and animal adapted mycobacteria, such as *M. bovis* (Galagan, 2014). The human adapted strains have evolved into lineage 1-7 of which the more recently diversified modern lineages 2-4 are thought to be more virulent and transmissible than other more geographically restricted lineages (Coscolla and Gagneux, 2014). The most geographically restricted lineages are 5 and 6 confined to West Africa and lineage 7 confined to Ethiopia. Lineage 7 is a recently discovered phylogenetic lineage of MTBC considered intermediate between ancient and modern (Firdessa *et al.*, 2013). On average the human-adapted lineages differ by only 1200 SNPs (Coscolla and Gagneux, 2014).

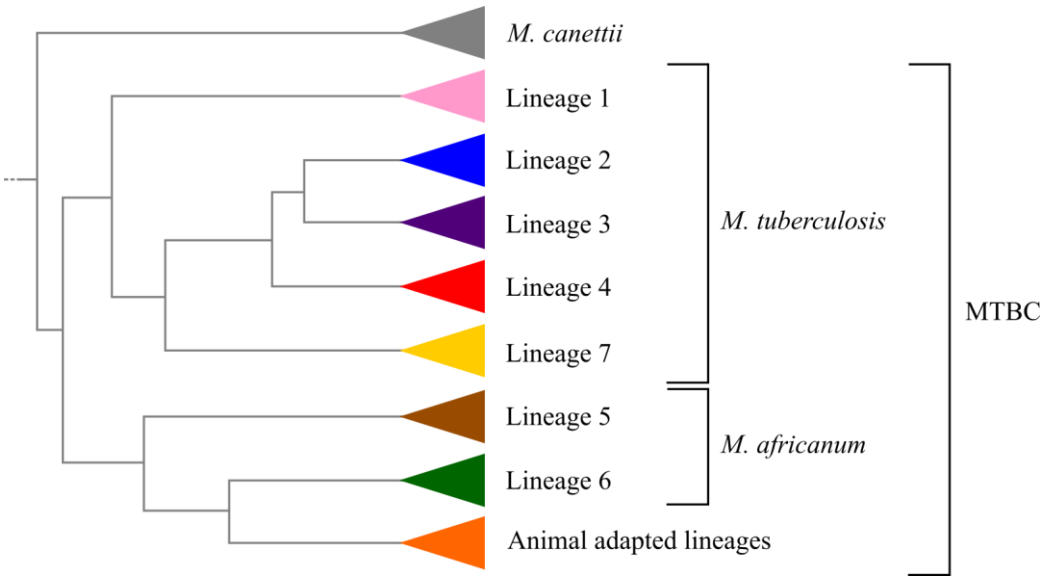

Figure 1 MTBC consists of 7 human adapted lineages, *M.tuberculosis* and *M. africanum*, and animal adapted lineages such as *M. bovis*. It shares the most recent common ancestor with the species *M. canettii* that can be used as an outgroup (Blouin et al., 2014). The more recently diversified modern lineages 2-4 are thought to be more virulent and transmissible than other more geographically restricted lineages (Coscolla and Gagneux, 2014).

MTBC strains have low genetic diversity as the evolution of MTBC is restricted to clonal diversification (Galagan, 2014). Unlike the more distantly related *M. canettii* they have no plasmids or horizontal gene transfer and instead the evolution of MTBC is primarily driven by chromosomal mutations (Warner et al., 2015). A large proportion of acquired mutations in MTBC leads to phenotypic differences and despite the clonality there is evidence that MTBC is highly diverse (Warner *et al.*, 2015). This diversity may be found within the host when a patient is infected with multiple TB strains, complicating TB studies as they have an impact on accurate disease diagnosis, effect treatment of patients, and control of TB in populations (Cohen *et al.*, 2012).

1.1 Challenges in detecting a mixed infection

The detection of a mixed TB infection depends on specimen- collection, handling and the sensitivity of a typing method (Cohen *et al.*, 2012). When a patient is infected by multiple TB strains some variants may not be present in the sputum at the time of sampling. The sample may thus only contain a sub-set of variants and others will be missed during detection. Within the host the different infecting genotypes are assumed to be competing with each other, causing some strains to occur in small numbers compared to the dominant strain (Warner *et al.*, 2015). These minority variants may be lost when samples are cultured to increase the bacterial population for DNA extraction.

Conventionally, MTBC strains were detected using a limited number of markers based on repetitive elements like Clustered Regularly Interspaced Short Palindromic Repeats regions (CRISPRs), and insertion sequences (ISs) that are other important sources of genomic variation to distinguish MTBC strains (Coscolla and Gagneux, 2014). By observing different banding patterns for distinct strains mixed infections could be detected, but requires the selection of multiple colonies for separate culture and detection, otherwise overlapping patterns may appear (Cohen *et al.*, 2012). MIRU-VNTR is based on Variable Number of Tandem Repeats (VNTRs) loci and can detect a mixed infection from a single sample, but the use of 24 loci provides limited information for phylogenetics and strain classification (Cohen *et al.*, 2012; Jagielski *et al.*, 2014). In addition, the power of these methods to differentiate strains and their results do not always agree (Ford et al., 2012).

Because of these difficulties the prevalence of mixed TB infection remains uncertain and estimates from previous studies are likely underestimated (Plazzotta *et al.*, 2014). TB Studies that focused on mixed infections have documented that 10-20% of the patients are infected with multiple strains (Cohen *et al.*, 2012; Hanekom *et al.*, 2013; Huang *et al.*, 2010; Navarro *et al.*, 2011), strongly suggesting that mixed infections are common. To increase chances of detecting a mixed infection a typing method should be able to distinguish closely related strains based on more markers and therefore allow detection at fine resolutions. WGS yields a great number of SNPs that can support accurate classification and will enable better estimates of mixed infections as it offers a higher resolution to differentiate MTBC strains that is not possible with other genotyping methods (Cohen *et al.*, 2012 Coscolla and Gagneux, 2014).

Although WGS allows detecting mixed infections at a finer level, some methods aim to identify only a single strain assuming that each patient is infected by a single strain (Zetola *et al.*, 2014; Benavente *et al.*, 2015; Larkeryd *et al.*, 2014). This requires multiple colonies to be sequenced to detect a mixed infection. To obtain a pure sample it is common that single

colonies are selected for sequencing, so a mixed infection may still be missed even when a method is designed to detect mixed infections (Eyre *et al.*, 2013).

There are studies aimed to detect mixed infection for clonal pathogens with a low mutation frequency like MTBC. Eyre *et al.*, 2013, used WGS data to detect mixed infections of the bacterial pathogen *Clostridium difficile*, but their method assumes that each sample contains only up to two infecting strains. Pulido-Tamayo *et al.*, 2015, developed a bacterial haplotype reconstruction method based on methods designed for viral sequences (that are much shorter in length and have a higher density of SNPs) and can also be used to identify mixed infections. Read overlap information is used, but the key to extend and reconstruct genome-wide haplotypes is to use frequency information of the present bacterial strains supported by mapped reads. *M. tuberculosis* has insufficient diversity to apply this work at high resolution and the frequency based approached requires inbalance between the strains to identify them. Although this method allows detecting more than two strain variants, it is not necessarily needed to reconstruct present genotypes to identify mixed infections. It should be sufficient to determine the presence of markers that are specific to distinct groups of strains in a single sample. Because mixed infections are known to impede drug susceptibility tests, Bradley *et al.*, 2015, developed a method that assumes drug resistance may be caused by a minority or majority strain in a mixed sample. Genotyping and frequency estimates are based on drug resistance mutations, so when different drug resistance profiles have been identified it implies the presence of multiple strains. This means that the tool does not identify a mixed infection when all the infecting strains are drug susceptible or when they are all resistant. In addition, this method relies on a frequency threshold (10%), so resistant minority strains present below this threshold are missed.

Several studies have suggested the use of an minimal set of SNPs that should be effective and sufficient for strain classification (Filliol *et al.*, 2006; Coll *et al.*, 2014; Comas *et al.*, 2009; Homolka *et al.*, 2012). SNPs were selected for their phylogenetic informativeness under strict criteria reducing the number of SNP markers greatly. As a consequence, the small number of markers reduces classification support and detection resolution. The sizes of the minimal SNP sets vary from 45 to 93 SNPs and there is little congruence between these sets. This may be explained by the different approaches used to extract SNPs and the type of SNPs they selected for which they assumed to have a higher contribution in MTBC classification. The selection of (sub-)lineage specific SNPs also depends on how the lineages and sub-lineages are defined, which is also different among these studies. The minimal SNP set of Coll *et al.*, 2014, contains one SNP per (sub-)lineage to differentiate 7 lineages and 55 sub-lineages, while choosing more than one SNP will avoid the inequalities coming from varying SNP-typing technologies (Homolka *et al.*, 2012). A decade ago, Filliol *et al.*, 2006, claimed a minimal SNP set would be advantageous because SNP analysis is relatively expensive, but this claim no longer fits to the current situation. It is more suitable to use a less strict selection on SNPs to collect a wide range of phylogenetically informative SNPs for a robust MTBC classification method.

A first attempt to tackle the challenge was made by a student intern, Joseph Romano, who developed *Macaw*, a method that can detect mixed infections at the lineage level. *Macaw* uses SNP markers that are specific to the 7 main lineages and *M. bovis*. His proof-of-concept work indicated that it was possible to identify mixed infections at least at the lineage level. While it was a step in the right direction, *Macaw* had inherent limitations that needed to be addressed. It uses the Fisher exact test to determine which lineages are significantly present, those with a p-value <0.05, but this method depends of the absence of other lineages, because higher p-

values are obtained if markers from other lineages are absent. When the test would be used to distinguish fewer clusters, the total number of absent markers would also be lower. This means that the test is dependent on the number of groups to distinguish and so the significance is reduced for mixed lineage infections, because there are fewer absent lineages. In addition, frequencies are determined in *Macaw* based on the total number of mapped reads. This results in an estimation bias, as reads mapped to markers with a read depth <5 are included in the calculation, while these markers were first considered absent. One of the goals of my project is to address these problems.

1.2 TB drug resistance

Global efforts to control TB have been challenged by the emergence of drug resistant TB (Zetola *et al.*, 2014). Strains can survive antibiotic treatment when they acquire mutations in genes involved in DNA repair systems, or genes that affect the expression of drug efflux pumps (Fonseca *et al.*, 2015). Stepwise acquisition of mutations associated with drug resistance can lead to resistance to multiple drugs (Perdigão *et al.*, 2013). Multiple drug resistant TB (MDR-TB) is defined as at least resistant to first-line drugs rifampicin (RIF) and isoniazid (INH), the two most powerful anti-TB drugs (WHO, 2014). Strains with additional resistance to any fluoroquinolone (FQ) and a second-line agent are classified as extensively drug resistant (XDR-TB). Resistance to all drugs tested has been described as totally drug resistant (TDR-TB).

A mutational event that leads to antibiotic resistance may have an effect on the fitness of drug resistant strains (Cohen *et al.*, 2003). Drugs typically target genes/proteins required for survival, so drug resistance mutations often come with a high fitness cost and in the absence of treatment drug resistant strains are considered to be less fit than drug susceptible strains (Galagan, 2014). The drug resistance mutations are selected when a patient receives antibiotics that kills all other bacteria, allowing the drug resistance strains to thrive (Warner *et al.*, 2015). This leads to the subsequent fixation of drug resistant strains in MTBC.

Compensatory mutations may also play a prominent role in the spread of drug resistance as they can restore the fitness of drug resistant strains (Fonseca *et al.*, 2015; Cohen *et al.*, 2003). Combinations of drug resistance and compensatory mutations suggest that resistance may be acquired without compromising fitness and transmissibility (Casali *et al.*, 2014). Studies have found mutations in the *rpoA*- and *rpoC*-genes that putatively restore the fitness of RIF-resistant strains that have mutations in the *rpoB*-gene (de Vos *et al.*, 2013; Perdigão *et al.*, 2014; Casali *et al.*, 2014; Cohen *et al.*, 2015). These mutations are widespread in MDR-TB and associated with ongoing transmission (Köser, Ellington, *et al.*, 2014; Warner *et al.*, 2015).

Studies have identified drug resistance mutations by comparing genetic identified mutations from WGS data with phenotypic drug resistance profiles and by finding whether phenotypic resistance can be explained by the detected mutations (Zhang *et al.*, 2013; Walker *et al.*, 2015). Mutations are characterized as resistance determining if it occurs in phenotypic resistant isolates and mutations that occur only in phenotypic resistant strains are characterized as benign. This genetic identification of resistance is desirable because it is faster than phenotypic drug susceptibility testing (DST) which takes at least 7 days and also shows poor reproducibility for some drugs (Feuerriegel *et al.*, 2015). Other WGS studies have used these identified mutations and showed that drug resistance can be assessed based on a library of known drug resistance mutations (Benavente *et al.*, 2015; Bradley *et al.*, 2015; Coll *et al.*, 2015; Feuerriegel *et al.*, 2015; Rosales-Klintz *et al.*, 2012). Because these studies have

only used known mutations they can target only the most frequent resistance-mediating mutations for genotypic DST.

The mechanisms of drug resistance in MTBC are complex and needs better understanding of the emergence and fixation of drug resistant strains. A significant proportion of resistance phenotypes of MTBC isolates cannot solely be explained by these mutations: up to 30% of isolates resistant to INH and approximately 5% of those resistant to RIF do not harbor mutations in the known resistance genes (Fonseca *et al.*, 2015). Mutations in genes that cause resistance to ethambutol (EMB) can confer low-level resistance that is below the threshold for clinically relevant resistance when tested for the minimum inhibitory concentration (MIC) (Fonseca *et al.*, 2015). This low-level resistance to EMB associated with these mutations might be a stepping-stone to higher level resistance, this is supported by the observation that several EMB resistance mutations can be present in the same resistant isolate (Plinke *et al.*, 2010). Two SNPs in *gidB*-gene have been found frequently in strains susceptible and resistant to streptomycin (SM) and were considered to be not related to SM resistance (Spies *et al.*, 2011). These mutations are more likely to be present in certain MTBC lineages and should be classified as lineage-specific SNPs (Fonseca et al., 2015; Feuerriegel et al., 2015).

Mixed infections impede antibiotic resistance tests, because the present strains may have differences in their drug susceptibility (Fonseca *et al.*, 2015; Zetola *et al.*, 2014). Drug resistance, minority strains may not be detected when mixed infections are treated as a single-strain sample and when the patient is given an antibiotic, only the susceptible strains are killed and the resistant strains can take over the in-host population (Figure 2). The patient is then subjected to a more severe infection and because rediagnosis is rare, mixed infections are associated with poor treatment outcomes. The Xpert MTB/RIF assay is a genotypic test to rapidly diagnose resistance to rifampicin (RIF), but Zetola *et al.*, 2014, showed that false-negative results were significantly associated with mixed infections and that it fails to detect RIF resistance in vitro when resistant sub-populations accounted for <90%.

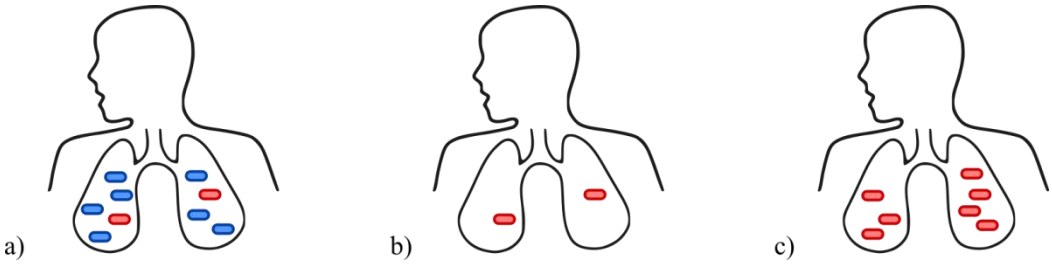

Figure 2 a) Individual infected with two distinct TB strains. The blue type is drug susceptible and dominant; the red type is a minority type that is drug resistant and less fit. b) When a patient has a mixed TB infection and is treated with an antibiotic, the susceptible strains are killed and drug resistant strains survive. c) The bacterial in-host competition is changed and the drug resistant strains can thrive.

1.3 Study aim and overview

This study aims to detect a putative mixed TB infection at different levels in MTBC and to determine the frequencies of the present strains based on established tree paths in the MTBC phylogenetic tree. The project builds upon Romano's proof-of-concept work developing *Macaw*, which only differentiates between the main lineages 1-7 while WGS allows much higher resolution to differentiate MTBC strains. Detected mixed infections were also

examined to confer strains with differing drug resistance profiles. Figure 3 shows an overview of my approach and three specific aims are stated to accomplish the goal:

1. **Define hierarchical clusters and cluster-specific SNPs.**
   Sub-lineages are not clearly defined in literature, and differentiating MTBC strains into the main lineages is too coarse. A hierarchical clustering approach was used to find clusters of strains and allows differentiation at varying levels of the MTBC phylogenetic tree.
2. **Detect mixed infections and estimate frequencies.**
   The tool will allow researchers to detect multiple strains in a single sample and to determine their frequencies by using only raw sequencing data. By mapping reads to cluster-specific SNP markers, paths of present clusters in the MTBC phylogenetic tree can be detected. Full tree paths that lead back to the MTBC root give additional support for the presence of certain strains and paths that split after the root indicates the presence of multiple MTBC strains. Frequencies of present strains are estimated based on the number of reads for which markers have been identified as substrings.
3. **Detect drug susceptibility differences of infecting strains.**
   If a sample is detected to be a mixed TB infection, the present strains may have different drug resistance profiles that need to be determined to improve treatment of TB patients. Using a manually curated library of known drug resistance mutations a sample can be tested to confer genotypic resistance. When there is an ambiguous base call, frequencies of detected nucleotides at the position associated with drug resistance can be matched with frequencies found in aim 2 to find how the infecting strains differ in their drug susceptibility.

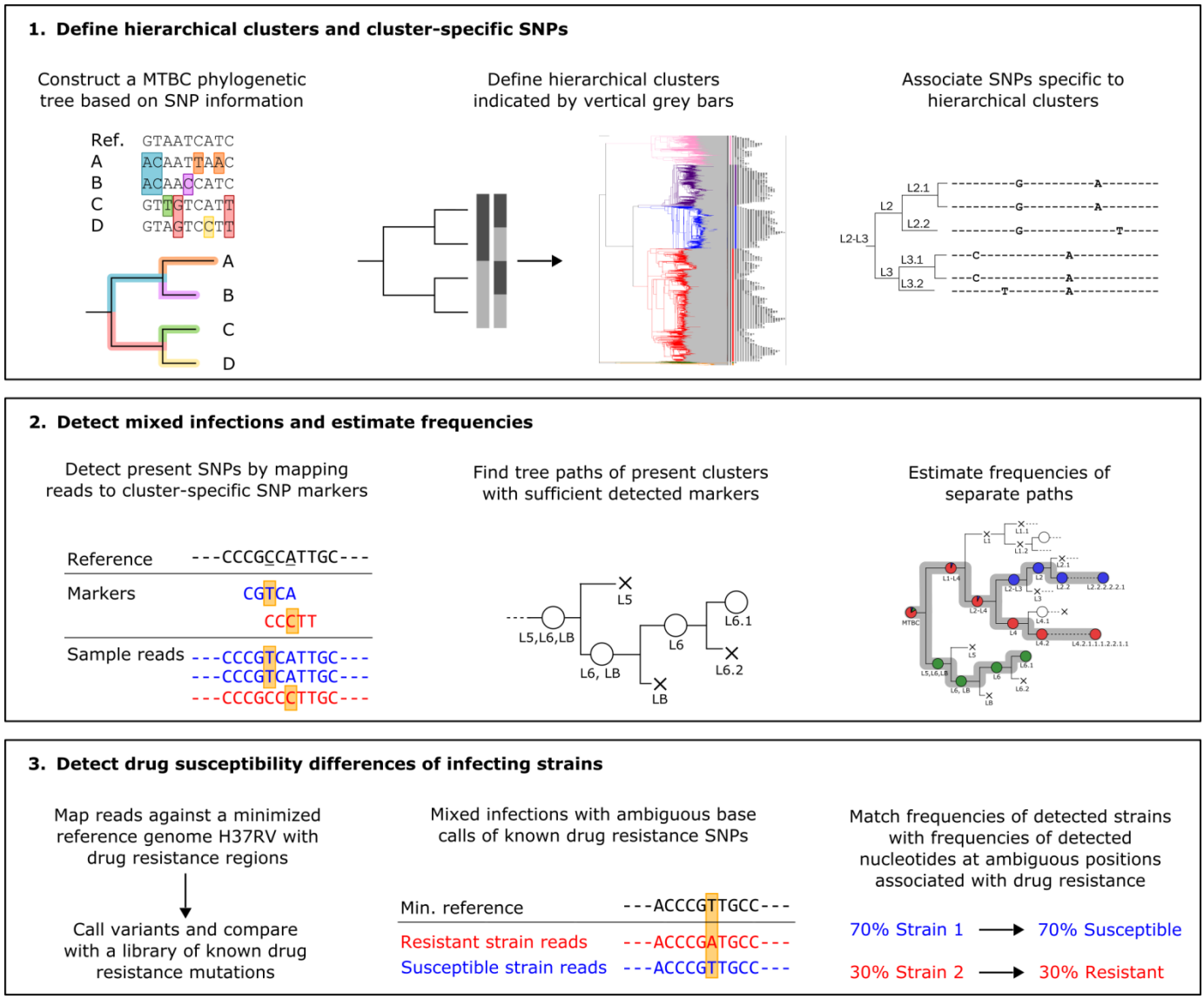

Figure 3 Study overview of the applied approaches to detect mixed TB infections and differences in drug susceptibility of the infecting strains. First hierarchical clusters were defined and SNPs were associated to these clusters. The cluster-specific markers were used to obtain SNP markers to detect the presence of clusters and paths of present clusters. Paths that split after the MTBC root indicate the presence of multiple TB strains for which frequencies are estimated. A minimized reference genome is used to detect known drug resistance SNPs. Frequencies of nucleotide bases at a position associated with drug resistance is compared with the frequencies of detected strains in aim 2.

2. Methods

2.1 Phylogenetic tree of global dataset

A global dataset of 8217 samples was gathered from multiple TB studies with WGS data (Table S1). I selected 6100 samples for training based on results of *Macaw*, those that were not mixed infections and had an unambiguously predicted lineage. The lineages were determined by assessing their position in the MTBC phylogenetic tree (Figure 1). Bam-files and variant call files (VCFs) were available for paired-end read data based on reference genome H37Rv (CP003248.2) obtained with Pilon 1.11 (Walker *et al.*, 2014). There were 76 samples from Comas *et al.*, 2013, that are single-end read data and VCFs for these were unavailable, but I produced them using the same pipeline.

High quality SNPs were extracted from VCFs to construct a MTBC phylogenetic tree. Given each SNP position found in at least one sample SNP sequences were created for each sample. At each position a sample may have an alternative base when a mutation was detected and the reference base when a mutation was not detected. In this way, all genomic invariant positions were left out and there were no positions for which no SNP was found in all samples. The output is a phylip-formatted file that complies to the RAxML required input. Another tool performs the same filtering, but outputs a multi-fasta file. Both methods were found to give identical outputs. SNP sequences may contain unknown nucleotides indicated by 'N' instead of 'A', 'C', 'G', or 'T'. Samples with SNP sequences that had >5% 'N's of the total length were removed from analysis. A smaller prototype trainingset of 421 strains was first used to construct a phylogenetic tree and define sub-lineages (Table S2). The final set of 6100 strains was reduced to 5992 samples when samples were removed that contained >5% 'N's (Table S3).

The phylogenetic tree was constructed with RAxML 7.7.8 (Stamatakis, 2006) under the GTRCAT-model to use a general time reversible model while reducing computational cost and memory consumption. The tree was based on 5992 strains and mutations are relative to the reference genome H37Rv (CP003248.2) (Figure S1). The largest group of strains is from lineage 4, while lineage 7 had only few representatives and was excluded from further analysis (Table S3). Lineage 1, 2 and 3 had representations >12% and lineage 5, 6 and *M. bovis* had limited (<1%), but sufficient strains. Based on phylogenies from literature (Coscolla and Gagneux, 2014) the root was placed manually in Dendroscope (Huson and Scornavacca, 2012) between groups of lineages 1-4, and 5, 6 and *M. bovis*. In phylogenetic trees, bootstrap values indicate the percentage in which a node was recovered in pseudo-replicate trees when resampling the nucleotide characters (Soltis and Soltis, 2003). These bootstrap values were not calculated, because nodes are likely to be recovered when they are close to the root and therefore more clear distinctions can be made between the strains. For nodes that are closer to the leaves bootstrap values logically decrease, because there is less variation between the strains. Considering the large number of samples, no outgroup was used to root the tree, which reduces the number of variable positions, and consequently reduces memory usage and runtime to construct the phylogenetic tree.

2.2 Defining sub-lineages

HierBAPS (Cheng *et al.*, 2013) was initially used to find sub-clusters of the main MTBC lineages. The algorithm defines clusters given a user-defined maximum value of *K* and subsequently defines sub-clusters given the number of hierarchical levels *L*. Clustering was

first performed on a sub-set of 421 strains from lineage 1-7 and *M. bovis* (Table S2) with K = 2, 3, 4, 5, 6, 10, and 15, and L=10 for each run. A vast amount of samples are lineage 4 (66.7%), and lineage 2 (21.6%) strains. This clustering method failed because paraphyletic groups were obtained for which no SNPs could be associated (Table S4). To overcome this problem I switched to another approach to define clusters.

Hierarchical clusters were used based on the reconstruction of the MTBC phylogenetic tree with 5992 taxa. Each branch in this tree splits into two new branches and represents groups of strains that are further divided into two sub-groups of strains. Clusters were defined based on these partitions, so that each cluster is a monophyletic group and there are no paraphyletic groups. Starting at the MTBC root node, two new sub-clusters were defined at each ancestral tree node if both sub-clusters have at least 10 strains. The recursive call for sub-clusters continued while a sub-cluster has at least 20 strains. The size of each leaf cluster is ≥10 and <20 given this algorithm, while internal clusters consist of at least 20 strains.

I modified the tree visualization tool Peacock (*https://github.com/AbeelLab/peacock*) to optionally show clusters alongside the tree (Figure S1: phylogenetic tree of 5992 strains with hierarchical clusters). Vertical colored bars placed next to the leaves/strains indicate which strains are clustered together. Clusters that are defined hierarchically can be visualized using multiple columns, where each subsequent column has bars that indicate sub-clusters (Figure 4).

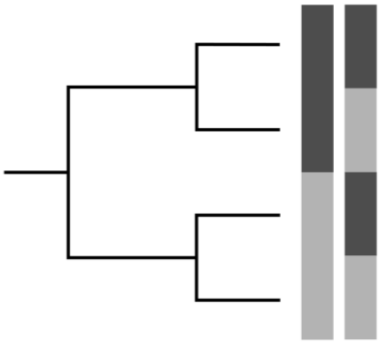

Figure 4 A phylogenetic tree with 6 hierarchical clusters indicated by the vertical dark and light gray bars. The first column shows two clusters of strains based on the root node that splits into two new branches. The second column shows the subsequent division into four sub-clusters (two for each node/cluster).

2.3 Phylogenetically informative SNPs

Phylogenetic clades differentiate from their ancestor group when accumulated mutations become fixed, and therefore these mutations indicate the presence of such clusters and are phylogenetically informative. Figure 5 shows an example of SNPs that can be associated to hierarchical clusters. Here, a cluster that consists of lineage 2 and 3 is split into two groups for each lineage. Lineage 2 has 3 strains that harbor a specific mutation ‘G’ and lineage 3 has 3 strains with a mutation ‘A’, dashes are invariable positions within this example. Each lineage is than divided into sub-lineages for which there are also mutations that are specific to the sub-groups. SNPs are selected to be phylogenetically informative when they can be associated to a defined cluster.

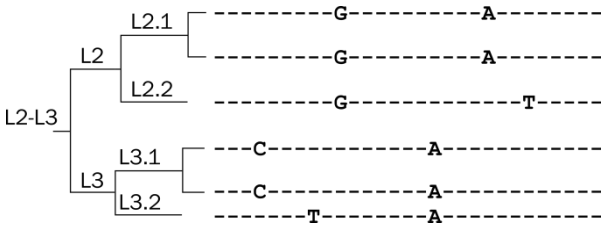

Figure 5 Clades differentiate from their ancestor group when accumulated mutations become fixed, and therefore these mutations indicate the presence of such clusters and are phylogenetically informative. Dashes are invariable positions, nucleotide bases indicate differences in the genomes of lineage 2 (L2) and 3 (L3) and their sub-lineage clusters. L2 has 3 strains with a mutation 'G' and L3 has 3 strains with a mutation 'A', sub-clusters of these lineages also have SNPs that are phylogenetically informative.

Initially, a simple approach was applied to find SNPs specific to the known lineages of MTBC. SNPs were associated to a lineage cluster when >95% of the strains in a lineage harbor a SNP and the SNP was found in <5% of strains in all other clusters. These thresholds allow lineage-specific SNPs to not occur in a small fraction of the lineage and to occur in a small fraction of strains that are not within this lineage. This method failed to find SNPs that allow for strain type detection, because lineage detection based on these SNPs could not recover the known lineages.

To find SNPs that are associated to hierarchical clusters, I used a contingency table for each cluster and each SNP (Table 1). I included only SNPs in the analysis to minimize memory and computation time. Samples were classified to be true/false positives/negatives based on their occurrence in a cluster and whether they harbor a mutation (Table 1). Strains within a cluster are true positives (TP) when they have the specific mutation, and are considered false negatives (FN) if they do not have this mutation; strains outside this cluster are either false positives (FP), with the mutation, or true negatives (TN), without the mutation.

Table 1 Contingency table for a specific cluster and a specific SNP. A SNP was selected when the true positive rate (TPR, sensitivity), true negative rate (TNR, specificity), positive predictive value (PPV), and negative predictive value (NPV) all were >0.95.

	Strains within the cluster	**Strains outside the cluster**	
Strains that harbor the mutation	True positives (TP)	False positives (FP)	Positive predictive value: PPV = TP/(TP + FP)
Strains without the mutation	False negatives (FN)	True negatives (TN)	Negative predictive value: NPV = TN/(TN + FN)
	True positive rate: TPR = TP/(TP + FN)	True negative rate: TNR = TN/(TN + FP)	

Based on these numbers the true positive rate (TPR, sensitivity), true negative rate (TNR, specificity), positive predictive value (PPV), and negative predictive value (NPV) were calculated and when all these values were >0.95 a SNP was associated to the cluster. The minimum true positive value was set to 10, meaning that for each cluster at least 10 strains are required to find an associated mutation. The TPR considers all strains in the clusters and given that the minimum is set to 0.95, no more than 5% strains within this cluster are allowed to not have this SNP. Strains that are outside the cluster are analyzed with the TNR requiring that more than 95% of these strains do not have the SNP and allowing no more than 5% of the strains to have this SNP. The PPV considers all strains that have the specific SNP of which

95% of these strains should be within the cluster and at most 5% of these strains are found outside the cluster. Finally, the NPV ensures that more than 95% of all negatives, strains without the specific SNP, are found outside the cluster and that no more than 5% is allowed within the cluster.

First, the method was applied locally on pairs of clusters c_1 and c_2 that share a direct ancestor. So when using the contingency table for c_1, strains outside the cluster are those strains that are in c_2 and so SNPs are called relatively to cluster c_2, but also vice versa. Using this method SNPs are only unique compared to their sibling cluster, but within the phylogenetic tree SNPs can also be specific to other clusters. This method did not work for lineage detection. Therefore, SNPs need to be associated on a global scale and all strains in the phylogenetic tree need to be considered when using the contingency table.

Global SNP association was done for each hierarchical clustering to find SNPs that are specific to the strains in a cluster and not specific to all other strains in the phylogenetic tree. In this way, the number of false positives and true negatives were based on all other samples in the dataset. For each single cluster *c*, SNPs specific to the samples in cluster *c* were called relatively to all samples that are not in cluster *c*. In total there were 5992 samples and 226570 unique SNPs and for each SNP analysis all samples were taken into account. SNPs that are globally associated to a cluster can indicate the presence of a hierarchical cluster in MTBC.

Clusters that contain the reference genome have few to zero SNPs, so associated SNPs were not found for these clusters. Instead SNPs were inversely associated and selected when found in most strains outside the cluster and rarely found within the cluster. Figure 6 shows an inverse associate SNP 'G' that is found in all clusters that do not contain the reference genome (blue nodes). Any other nucleotide than 'G' then indicates the presence of lineage 4.2.2.

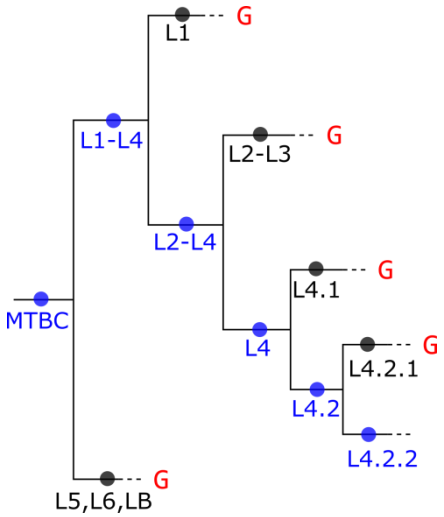

Figure 6 SNPs are inversely associated when clusters contain the reference genome (blue nodes). An inversely associated SNP 'G' is found in clusters that do not have the reference (black nodes) and is rarely found in clusters that do contain the reference. Any other nucleotide than 'G' then indicates the presence of lineage 4.2.2.

2.4 Cluster specific SNP markers

Filtering of SNPs and SNP markers was done within a cluster. SNPs are removed that reside within 10 bp of another SNP within the same cluster; this assures that the SNP marker sequences will not overlap when 21 bp markers are generated centered on the specific SNP. Short sequences have been used earlier to construct phylogenies from WGS data and *l*-mers of this length were found to be in the optimal range for genome comparison (Sims *et al.*, 2009; Sims and Kim, 2011). The selected SNPs were extended with 10 bp on each side based on the reference genome H37Rv (CP003248.2, Figure 7). Clusters that contain the reference genome have inverse associated SNPs and therefore markers are based on the other 3 possible nucleotide bases to obtain complementary markers (Figure 6 and 8). One of these bases is the reference base and the other bases are included because strains can still have a mutation relative to the reference genome, even when they are in the same cluster. These opposite nucleotides indicate the presence of strains in clusters containing the reference genome and allow reads to map for quantification.

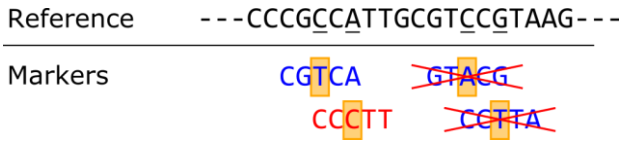

Figure 7 SNPs that indicate the presence of a cluster are extended with 10 bp on each side based on the reference genome, the example shows an extension with 2 bp. SNPs that are within 10 bp of another SNP within the same cluster are removed (blue markers on the right), such that the markers will not overlap at variable positions. This assures that these marker sequences can be detected. Markers are allowed to overlap with markers from other clusters, because then the marker sequences can still be detected in strains (blue and red markers on the left).

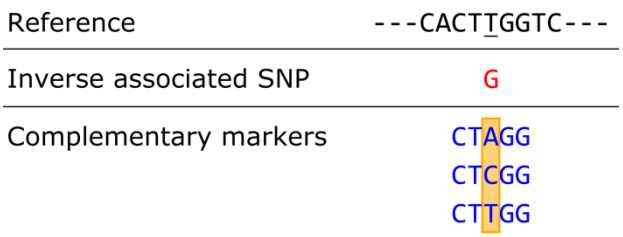

Figure 8 Complementary markers based on an inverse associated SNP ('G') indicating the absence of a cluster. Three other possible nucleotides inversely indicate the presence of the cluster and include the reference base. The other bases are included because strains can still have a mutation relative to the reference genome, even when they are in the same cluster.

Duplicate markers with identical sequences within a cluster are removed such that only unique SNP markers are left, this removed 38 potential markers. Sibling clusters (that share a direct ancestor) were removed for detection when both clusters have 0 SNP markers, and descendants of these nodes were also removed. Samples cannot be classified to these groups because without any SNPs these two groups cannot be distinguished and their descendants will also not be detectable.

The performance of each marker is validated with my detection method on the training set of 5992 strains (see Table S3). Markers that were not detected in any of the strains in the cluster to which it is specific were removed, this removed 1228 markers.

2.5 Curated list of known drug resistance mutations

I started from a curated drug resistance mutation list from Coll *et al.*, 2015, to detect known drug resistance mutations. Codon changes were translated to nucleotide changes if there is a unique possibility, otherwise the nucleotide change is unknown and left out of the analysis; if nucleotide change is known then the coordination of the SNP is known. Identical SNPs that indicate resistance to multiple drugs are merged so that there are only unique SNPs in this list. Genomic positions of mutations relative to H37Rv NC_000962.3 were translated to genomic positions in H37Rv CP003248.2. The *rrs*-gene in reference H37Rv NC_000962.3 has 10 bp extra at the beginning and 7 bp extra at the end compared to reference H37Rv CP003248.2, the rest of the gene sequence is identical. The *rrl*-gene in reference genome H37Rv NC_000962.3 has 2 bp extra at the beginning of the gene sequence compared to reference H37Rv CP003248.2. There is one mutation in the list in gene *thyA* with gene coordination 87, for which the reference bases in H37Rv CP003248.2 do not correspond to the reference bases in H37Rv NC_000962.3 from Coll *et al.*, 2015. This is probably because the wrong chromosome coordination was given by Coll *et al.*, 2015, this mutation is unreliable and left out. There are three mutations that are insertions of unknown nucleotide bases 'N'; all three are from the source MUBII-TB-DB (Flandrois *et al.*, 2014). The library contains 130 indels (insertions and deletions) of which 121 are from sources MUBII-TB-DB and TBDreaMTB (Sandgren *et al.*, 2009).

Additional drug resistance SNPs were gathered from three studies (Walker *et al.*, 2015; Casali *et al.*, 2014; Rosales-Klintz *et al.*, 2012). Drug associated mutations from Walker *et al.*, 2015, and Casali *et al.*, 2014, are indicated by the number of resistance and susceptible phenotypes, and also include mutations only found in drug susceptible strains. Adding mutations with resistant phenotypes with mutations from the study of Coll *et al.*, 2015, was complicated as I found discordance; mutations supported by more susceptible phenotypes than resistant phenotypes in studies of Walker *et al.*, 2015, and Casali *et al.*, 2014 that were already included in the library. Due to the complexities in merging these conflicting lists of mutations, I choose to not use the additional drug resistance SNPs found in other studies.

Drug resistance/susceptibility detection with markers

At first I tried to detect genotypic drug resistance with 21 bp markers that are centered on SNPs associated with drug resistance. Susceptibility markers were also included to detect differing resistances. Markers are only generated from SNPs (indels were excluded) that are supported by at least 5 study references. For each SNP position, a drug susceptibility marker was made based on the nucleotide base in reference H37Rv (CP003248.2). Remaining nucleotide bases that are missing from the library are included as markers with unknown drug susceptibility. So for each genomic position of a drug resistance SNP, all four nucleotides 'A', 'C', 'G', 'T' can be detected and linked to unknown or known drug resistance/susceptibility. When sample reads are mapped against the drug resistance/susceptibility markers, there are 4 cases to distinguish for each SNP position: 1) and 2) only resistance or susceptibility markers are detected for a specific genomic position, 3) both resistance and susceptible markers are present for a specific genome position, 4) both resistance and susceptible markers are absent for a specific genome position. Drug resistance and also different drug resistance profiles could be detected for mixed infection. However, this method failed to detect drug resistance SNPs when they are within 10 bp of another drug resistance SNP that is also present. Because drug resistance SNPs are near or within genes, it

is also likely that strains can have multiple drug resistance SNPs within 10 bp of each other that are in the same gene.

Drug resistance detection through read mapping and calling variants in DR regions

I choose to adapt the approach of Coll *et al.*, 2015, to detect drug resistance by mapping reads to a minimized reference genome. This reference consists of drug resistance associated regions and 1000 bp flanking regions and allows to detect variants that are associated with drug resistance. These genes were obtained from a curated library of drug resistance mutations (Coll *et al.*, 2015) and mixed infection samples detected in this study were analyzed for mutations in these regions. Phenotypic DST results were available for 337 samples derived from the study by Cohen *et al.*, 2015, and used as a reference standard.

Read data is mapped with BWA 0.7.12 against the minimized reference genome and variants in drug resistance associated regions are called with Pilon 1.11 (Walker *et al.*, 2014), and stored in a variant call file (VCF). SAMtools (Li *et al.*, 2009) was used for file conversion.

I determined the presence of SNPs associated with drug resistance, and whether these were ambiguous calls that indicate a mixed infection with differences in drug susceptibility (Figure 9). Frequencies of present genotypes are derived from mapped reads with a phred score >30.

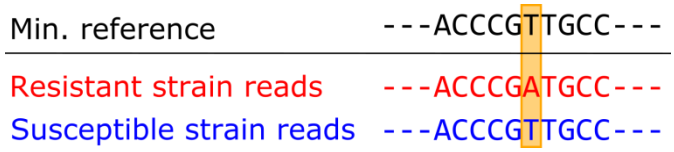

Figure 9 Reads from a mixed infection sample may lead to ambiguous calls suggesting differences in drug susceptibility of present strains. The nucleotide position marked by a yellow block is a position associated with drug resistance of which the resistant strain has the actual mutation 'A'.

3. Results and discussion

3.1 Defining clusters within MTBC

A phylogenetic MTBC tree (see Figure S1) was first constructed to assess how strains are related. Then clusters were defined in this tree.

Extract SNP information for phylogeny construction

Initially, SNP information was extracted from 6100 strains and this resulted in 418,348 variable genome positions. There were 108 samples that contained >5% unknown nucleotides 'N' at the variable genome positions and these were removed from the analysis. Excluding these samples also excluded many variable positions found in the total set of samples, thus the total the number of variable positions reduced to 222,376 positions based on 5992 strains with 226,570 unique SNPs.

The phylogenetic tree was constructed based on the 226,570 SNPs at 222,376 genomic SNP positions found in 5992 strains. Lineage positioning in the constructed phylogenetic tree was compared to phylogenies from literature and was found to be the same (Coscolla and Gagneux, 2014).

Defining sub-lineages with HierBAPS

Casali et al., 2014, clustered their MTBC samples using HierBAPS (Cheng et al., 2013), a hierarchical Bayesian genetic population clustering method often used for population genetics. HierBAPS was initially used to define sub-lineage clusters.

I looked for similarities between the clustering results obtained with different values of *K*, maximum number of clusters at each level, and levels *L*=10. HierBAPS does not necessarily recover the known lineages of MTBC and deeply nested clusters at higher levels are scattered with many clusters consisting of a single strain. I used two clustering results based on 421 samples where only lineage 2 and 4 were split into sub-lineages because of their prevalence in this dataset (Table S4). For these two clustering results I tried to select sub-lineage specific SNPs using a simple approach: SNPs that are present in >95% of the cluster samples, and <5% in all other clusters. Unique markers from this method were tested on three samples for which the lineage information was known from *Macaw*: lineage 2; mixed infection with lineages 2 (0.48) and 4 (0.52); and a mixed infection with lineages 2 (0.04), 4 (0.85), and 6 (0.11). The markers could not correctly identify the present lineage(s) with corresponding frequencies when compared to results of Romano's lineage markers.

The obtained clusters consisted of monophyletic groups and paraphyletic groups (Figure 10). A monophyletic group consists of strains that have a common ancestor and includes all descendants. Paraphyly refers to groups that have a common ancestor, but do not include all descendants. In the process of phylogenetic patterns, paraphyly is a natural transitional stage in the evolution of taxa, a paraphyletic group was monophyletic before younger derivatives arose from this group (Hörandl and Stuessy, 2010). Unlike monophyletic groups, it was not possible to find SNPs associated with paraphyletic groups. SNPs found in strains within paraphyletic groups are not specific to the group, because they are also found outside this group in strains that share the common ancestor.

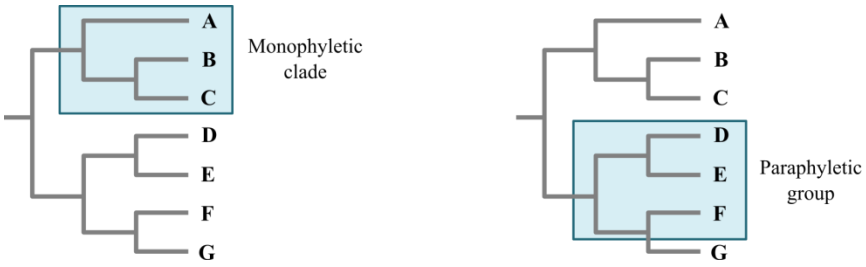

Figure 10 A monophyletic clade consists of descendants that have a common ancestor and includes all descendants. A paraphyletic groups also consists of descendants that have a common ancestor, but does not include all descendants.

I could switch to other clustering methods to define sub-lineages within MTBC as there are many algorithms for populations genetics as described by Excoffier and Heckel, 2006, but this may still lead to the problem of defining clusters that are paraphyletic groups. Any clustering method will find clusters that are statistically relevant, but to find biologically relevant groups is hard. Also, many clustering methods require a user-defined (maximum) optimal number of clusters that is unlikely to be known *a priori*.

Hierarchical clustering

I wanted to use a clustering method that excludes paraphyletic groups, such that each cluster is a monophyletic group (node in phylogenetic tree) for which I can find cluster-specific SNPs (Figure 10). Given that the phylogenetic tree is bifurcating (each branch splits into two branches), hierarchical clusters were obtained as described in the methods section. This approach only returns monophyletic groups, because for each ancestral node all of its descendants are included. There were 496 clusters defined (MTBC root excluded), see Figure S1.

At the root of the MTBC phylogeny, samples are clustered into lineage 1-4 or lineage 5, 6 and *M. bovis*. Subsequently, they are clustered into one of the two clusters according to the phylogenetic tree. Deeper phylogenetic clusters have been found for lineages that were overrepresented in the dataset such as lineage 4 (Table S5). Lineage 5 and *M. bovis* were not split into sub-lineages and only two sub-lineages were defined for lineage 6. This is because each of these lineages is represented by less than 60 strains (<1% in the total training set). In my phylogenetic tree the reference genome is included in lineage 4.2.2.2.2.2.2.2.2.2.

3.2 Associate SNPs to hierarchical clusters

At first SNPs were associated using the contingency table (Table 1) by comparing pairs of sibling clusters that share a direct ancestor. These associated SNPs were meant to distinguish sibling clusters when the ancestor cluster had been detected in the proposed hierarchical detection method. The number of SNPs indicating the presence of specific clusters varies between 1 and 830.

Inverse specific SNPs can be found for clusters that include the reference genome and they indicate the absence of this cluster, but because there are two clusters to distinguish these SNPs are identical to SNPs that indicate the presence of the other cluster. This means that for each cluster that includes the reference genome no SNPs can be found that are unique to that cluster. Using this pair-wise method also yielded SNPs that are only unique compared to their

sibling cluster, but within the phylogenetic tree they can also be specific to other clusters. Due to these reasons SNPs need to be associated on a global scale so that all strains in the phylogenetic tree are considered when using the contingency table (Table 1).

Global SNP association

In this approach SNPs were associated to a cluster by comparing strains within a cluster with all other strains in the total training set that are not within this cluster. This ensures that SNPs are not also associated to other clusters in the phylogenetic tree that do not have the same strains. A cluster that is an ancestor or descendant of the specific cluster may have the same cluster-specific SNPS, because these clusters contain sub-sets of strains in the specific cluster. Between 1 and 618 SNPs were found that can indicates the presence of SNPs including duplicates. In total there are 13633 associated SNPs, 13337 SNPs that do not occur within 10 bp of another SNP within the same cluster.

Associated SNPs could not be found for clusters that contain the reference genome, likely because little variant information is obtained for strains that are closely related to the reference genome. This makes it difficult to obtain read depths for these groups. As a solution inverse-associated SNPs were used to obtain complementary markers (Figure 8). When variant calling is performed relative to a reference genome outside MTBC, e.g. *M. canettii*, it would be possible to only find presence-indicating SNPs for every cluster within MTBC and omit the use of complementary markers. This would also allow finding SNPs specific to the MTBC species as this group would not contain the reference genome. Two sibling clusters that both have 0 SNP markers create 'gaps' in the phylogenetic tree for which no paths can be detected, and were removed together with its descendants.

3.3 Cluster-specific SNP markers

Markers of 21 bp were generated from the filtered SNPs by extending them with 10 bp sequences on each side based on the reference genome (Figure 7). Based on these SNPs, including complementary SNPs (Figure 8), 14861 markers were generated of which 14823 markers are unique within the cluster. Given these markers 308 clusters can be distinguished (MTBC root excluded) of which 261 clusters have associated SNPs (Table S5).

Marker validation

Each marker was validated based on the number of times it has been detected successfully in strains that belong to the cluster to which the marker is specific. The 14823 markers were used for strain-type detection in the trainingset of 5992 strains. There were 1228 markers that were not detected in strains specific to the cluster. Those were removed from the marker set because they cannot be used to detect the specific cluster; 13595 markers were left that are specific to 261 clusters. There were 9919 markers that were detected in strains in unrelated clusters (clusters that are not ancestors or descendants of the cluster to which the marker is specific). If I removed 10970 markers for which both constraints were true, 3853 markers would be left that are suitable for strain type detection. If markers are removed because they are detected in at least 10 strains in unrelated clusters, 6608 markers would be left. If I use a softer threshold and allow SNPs to be detected in at least 20 strains in unrelated clusters, 8774 markers would be left. This filter is too stringent to leave a sufficient number of markers left

for detection. Also, I want to maintain the number of clusters that have cluster-specific markers. If many markers are removed some clusters cannot be detected anymore, because all their markers were removed.

3.4 Detecting samples with multiple strains and estimate frequencies

My method allows detecting a mixed infection at each level in the MTBC hierarchy using SNP markers specific to hierarchical clusters. By finding path(s) of present clusters that start at the MTBC root, the lineage can be determined and refined to sub-lineages based on hierarchical splits.

The markers can be used to determine the presence of cluster-specific SNPs in a sample. Markers are detected as substrings in reads and need to match exactly. A marker is considered present if the read depth of the marker is ≥5, reads mapped to markers below this threshold are considered sequencing errors. A cluster that has a set of markers is present if more than 1/8 of its markers are detected such that a sufficient number of markers is needed, but minority strains can still be detected (Figure 11).

Figure 11 Cluster-specific SNPs (yellow blocks) are extended with 10 bp on each side based on the reference genome to generate 21 bp markers, in this example they are extended with 2 bp. Genomic SNP positions in the reference are underlined. SNP markers are detected in sample reads as substring to determine the SNPs that are present in the sample.

The method takes into account that clusters with sufficient detected markers may be falsely detected (this includes clusters that have 0 markers) by finding paths of possible present clusters that trace back to the MTBC root. In this way, the presence of a cluster is supported by the presence of ancestral clusters (Figure 12a). Because there are clusters with 0 markers for which no reads can be mapped, they are considered to be possibly present (Figure 12b). Clusters with 0 markers complicate the detection of present paths, because the algorithm needs to assume these clusters may be present although there are no mapped reads. For some of these clusters their presence can still be inferred. By comparing the coverage detected at the ancestor and sibling cluster, we can observe whether there is missing read depth between these clusters. Only when read depth is missing the paths should split and thus reveal that both siblings are present. Present clusters that do not have a path of ancestral clusters leading back to the MTBC root are discarded, so gaps are not allowed (Figure 12c).

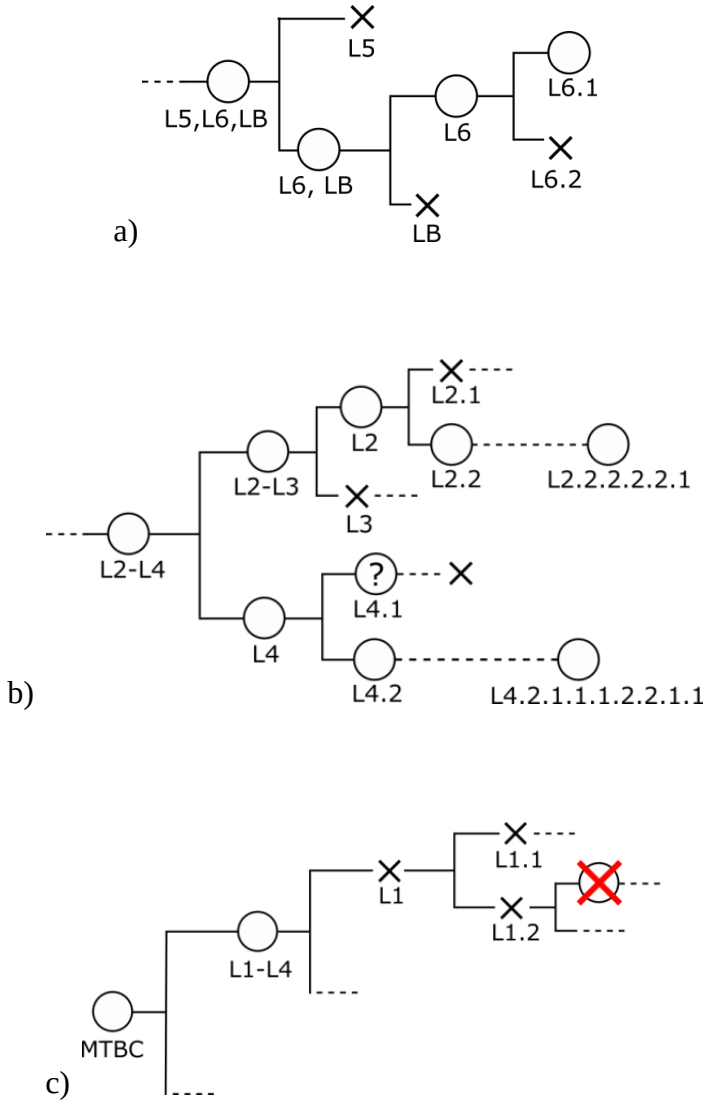

Figure 12 Detection of present clusters and present paths in MTBC. a) The presence of ancestral clusters supports the presence of a strain. Here, lineage 6.1 is present because its ancestors have also been detected. b) There were clusters that had 0 markers, such as lineage 4.1. These had to be considered as possibly present and its presence is only confirmed when both the ancestor and sibling cluster has been detected, but there was missing read depth. c) Present clusters are discarded when they do not have a path that leads back to the MTBC root, so gaps are not allowed.

Figure 13a shows the MTBC tree where all sub-trees from Figure 12 are merged and paths of clusters are found as described. The presence of multiple paths indicates the presence of multiple strains and thus a mixed infection. This example yields 3 paths that all lead back to the MTBC root. For each cluster I take the median read depth of present markers. Paths can now be separated based on their overlap indicated by different colored nodes (Figure 13b). Clusters that have same colors also have similar read depth and this read depth is divided when a path splits. The total read depth is found at the MTBC root node and is divided over the paths that contain purple and green nodes. Then the read depth of the purples nodes is divided over the paths with blue and red nodes. The detected read depths are used to estimate the frequencies of the present strains (Figure 13c). In each hierarchical level, the frequencies of detected phylogenetic clusters will add up to 1.

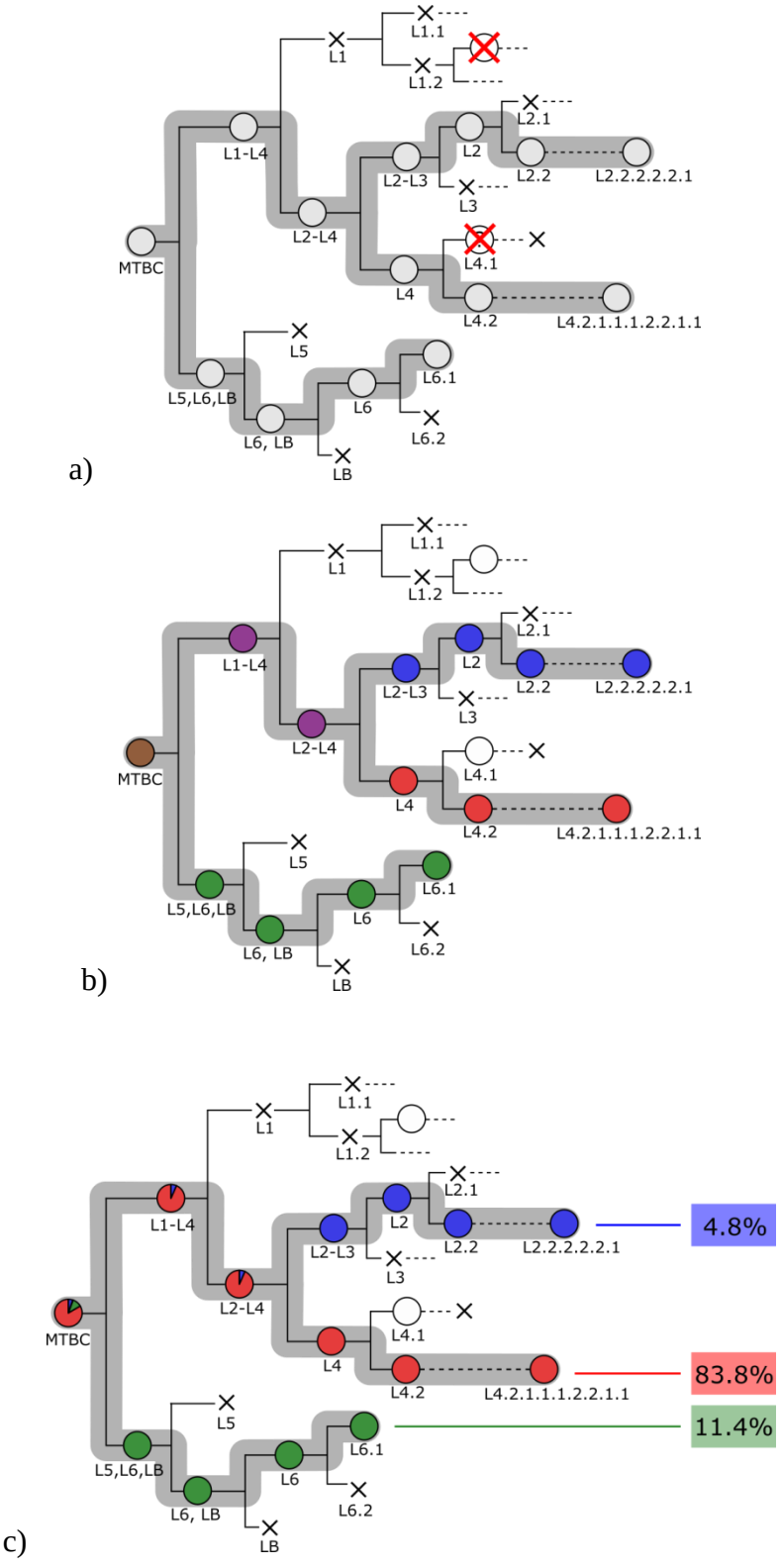

Figure 13 Frequency estimation based on detected coverage in separate paths. a) Clusters are present (circles) based on the number of markers that have been mapped for each clusters. Cluster with less than 0.125 of its markers detected (crosses) are excluded from analysis. In this example, there are paths of present clusters that start at the MTBC root node and split into three individual paths. b) Colors indicate separate paths with their own read coverage before and after they split. The read depth at the MTBC root node (brown) is the total read depth of all present MTBC strains. c) Frequencies are estimated for separate paths based on the mean read depth of clusters in these paths.

Unbalanced clusters

An ancestral group that splits into one group containing most of its descendants and one much smaller group resulted in unbalanced clusters, clusters that share a direct ancestor, and for which these sibling clusters have a large difference in the number strains they hold (Figure 14).

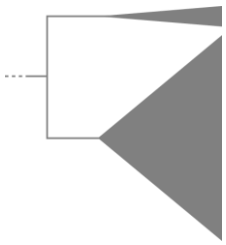

Figure 14 Unbalanced clusters are clusters that share a direct ancestor and for which these sibling clusters have a large difference in the number of strains they hold. The size of the triangles indicates the relative number of strains in these clusters.

Unbalanced sibling clusters leads to the detection of both clusters while the read coverage of the ancestor is not divided over the siblings, instead the read depth of both descendent clusters is similar to the read depth of the ancestor cluster. This is because the dominating sub-cluster is similar to the ancestral group and therefore the set of SNPs associated with this cluster will largely overlap with the SNP set of the ancestral group. This means that whenever the ancestral group is identified based on the number of markers detected, the dominant cluster will also be detected although this group may not be present. Cases where the read depth of an ancestral nodes is not divided over both descendent nodes indicate that the dominant cluster is falsely present. The smaller sibling cluster contains a small sub-set of strains from the ancestor group, therefore SNPs associated to this group are not similar to the set of SNPs from its ancestor group. So when both sibling clusters have been detected, the smaller group is actual present.

To find these unbalanced clusters, we have to find sibling clusters for which the sum of the read depth is approximately twice the read depth of the ancestor cluster (Figure 15). If the total coverage of siblings is at least 1.8 times greater the coverage of the ancestor the detection method excludes the presence of the larger cluster.

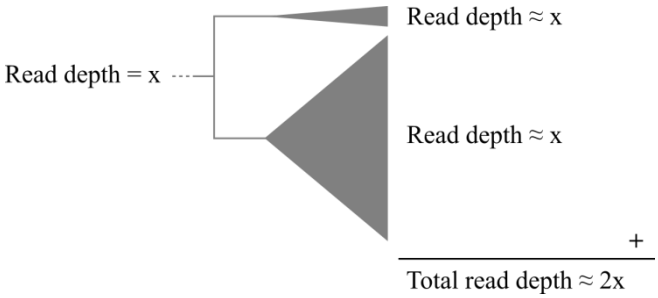

Figure 15 Unbalanced clusters for which the read depth compared to their ancestor remains similar and is not divided over the sibling clusters. The total read depth of the sibling clusters is approximately twice as high. For these clusters I exclude the presence of the larger cluster that is most similar to its ancestor cluster.

Validation of lineage detection

I validated 87 test samples that were not included in the trainingset and for which lineages were previously predicted using *Macaw*. There were 77 single lineage samples for which 75 samples my prediction agreed with *Macaw* predictions. The remaining two samples were classified as non-TB samples, because paths of present clusters could not be found. Even though *Macaw* outputs lineage designations, they were supported by very few markers (2 out of 139 and 29 out of 879 lineage markers) and therefore likely incorrect. For 9 out of 10 samples that contain mixed lineages *Macaw* and my predictions agreed completely. For the one remaining sample, lineages 2 and 4 were previously identified, which I confirmed. Furthermore, I identified two sub-lineages of lineage 4 to be present, illustrating the increased resolution.

In the total set of 8217 training and testing samples the lineage could not be determined for 21 samples, because a path was detected for which the end cluster in the path is a superlineage cluster or cluster of lineages. For some cases the cluster consisting of lineage 2 and 3 was detected. Markers from lineage 2 were detected, but the read depth was insufficient to determine its presence.

Samples for which the cluster consisting of lineage 2, 3 and 4 was detected and for which the coverage was sufficient, the phylogenetic location of the strain indicates them to be lineage 7 strains. SNPs specific to the group of modern lineages 2-4 also includes SNPs specific to lineage 7, because it is more closely related to this group than the ancient lineages 5 and 6 (*M. africanum*) and *M. bovis* (Figure 1). Lineage 7 was excluded from the SNP marker discovery so we have no specific markers, but the classification as lineage 2-4 supports that my method is robust to deal with unseen lineages and further supports the idea that there may be another undiscovered lineage at the intersection of lineages 2 and 3 when samples have sufficient coverage.

3.5 Validation of frequency estimates

Frequency estimates of the hierarchical detection method were validated to examine the accuracy. I tested this method on a set of 9 samples that were made by computationally mixing reads from pure samples belonging to lineage 2 and 4 in varying proportions. The total recovered read depth of these samples was between 167 and 174. The frequencies of lineage 2 and 4 were predicted with a maximum deviation of 0.02 (Table 2).

Table 2 Validation of frequency estimates of mixed infections test samples that were made by computationally mixing varying proportions of a pure lineage 2 and 4 strains. Frequencies were predicted with a maximum deviation of 0.02.

	True frequencies		Predicted frequencies		
Test sample	**L2**	**L4**	**L2**	**L4**	**Deviation**
1	0.90	0.10	0.91	0.09	0.01
2	0.80	0.20	0.80	0.20	0.00
3	0.70	0.30	0.71	0.29	0.01
4	0.60	0.40	0.62	0.38	0.02
5	0.50	0.50	0.51	0.49	0.01
6	0.40	0.60	0.41	0.59	0.01
7	0.30	0.70	0.31	0.69	0.01
8	0.20	0.80	0.22	0.78	0.02
9	0.10	0.90	0.10	0.90	0.00

Prevalence of mixed infections

I tested my method on 8217 samples from the total dataset and for 8028 samples TB was detected. Of these TB samples 802 were found to be mixed infections and 430 were mixed at the sub-lineage level. In this study, the prevalence of mixed infections in TB samples is estimated to be ~10% of which ~50% are mixed sub-lineage infections. Most (89.2%) of the mixed infection samples consisted of two sub-populations and for some 3 or more strains were detected (2.1%) (see Table 3). There were 194 (2.4%) samples for which no MTBC strains could be detected at all of which 17 samples were known to be *M. canettii*. The other samples may have insufficient read coverage to determine the presence of a MTBC strain type.

Table 3 Detected number of bacterial TB populations in 8217 samples. In 8028 samples TB was detected of which 802 samples contain a mixed TB infection.

Number of populations	0	1	2	3	>3
Number of samples	194	7226	715	70	17

3.6 Drug resistance detection

Once a sample is found to be a mixed infection, the question arises whether the distinct TB populations have differences in their drug susceptibility. Drug resistance can be detected by comparing observed mutations with known drug resistance mutations associated in other studies. The mutation library from Coll *et al.*, 2015, was examined and I first tried to detect drug resistance by using 21 bp markers created from drug resistance SNPs. This method failed because a drug resistance SNP could not be detected in samples that have two drug resistance SNPs that are within 10 bp of each other. Drug resistance SNPs are more likely to occur within 10 bp because they are found within genes that confer resistance. I then used a minimized reference genome to detect SNPs in drug resistance associated regions and compared them with the drug resistance mutation library of Coll *et al.*, 2015.

Compare detected drug resistances with DST phenotypes

Phenotypic DST results were available for 337 samples from Cohen et al., 2015, and I compared them to results obtained from the minimized reference genome (Table S6). Phenotypic results for some drugs were not available for all strains, in particular pyrazinamide (PZA), ethambutol (EMB), ethionamide (ETH), and capreomycin (CAP) were lacking. Table 4 shows the number of genotypic results in this study compared to the phenotypes obtained by Cohen et al., 2015. It should be taken into account that phenotypic drug susceptibility tests show poor reproducibility for some drugs. Phenotypic resistance to PZA was known for 143 samples of which 65 were misclassified as being susceptible, while the drug resistance library contained many known SNPs associated with resistance to PZA. False negative genotypes are a known problem for PZA, because unknown resistance mechanisms are responsible for the majority of the unexplained phenotypic resistance (Köser, Comas, *et al.*, 2014). For the remainder, I focused on isoniazid (INH), rifampicin (RIF), and fluoroquinolones (FQ), because they are the best studied antibiotics and their resistance mutations are relatively well understood as reflected in Table 4. For most samples all drug resistance predictions to RIF, INH, and FQ were in concordance with the phenotypic test results (86%).

Table 4 Number of strains that have concordant phenotypic and genotypic results, false negatives (phenotypic resistant/genotypic susceptible) and false positives (phenotypic susceptible/genotypic resistant) when detection was applied to 337 samples using phenotypic DST results from Cohen et al., 2015, as a reference standard. Phenotypic DST results were available for drugs rifampicin (RIF), isoniazid (INH), kanamycin (KAN), pyrazinamide (PZA), ethambutol (EMB), streptomycin (SM), ethionamide (ETH), capreomycin (CAP), and fluoroquinolones (FQ). DST results for FQ only includes oxafloxacin. Some strains have an unknown phenotype, because not all strains were tested to all drugs.

	RIF	**INH**	**KAN**	**PZA**	**EMB**	**SM**	**ETH**	**CAP**	**FQ**
Phenotypic and genotypic susceptible	99	91	212	60	108	128	66	78	223
Phenotypic and genotypic resistant	214	225	75	7	95	131	53	40	103
Phenotypic resistant / Genotypic susceptible	6	8	8	65	2	42	13	4	5
Phenotypic susceptible / Genotypic resistant	18	13	42	10	28	36	11	21	6
Unknown phenotype/ Genotypic susceptible	0	0	0	166	17	0	131	142	0
Unknown phenotype/ Genotypic resistant	0	0	0	29	87	0	63	52	0

Drug susceptibility differences in mixed infections

Mixed infection samples detected in this study were analyzed to detect differences in drug susceptibility of the present strains. Out of 802 mixed infections, there were 463 samples for which no drug resistance was detected, therefore these samples have strains that are all drug susceptible. MTBC genomes are haploid, so heterozygous or ambiguous base calls suggest there is more than one strain present in the sample. There were 118 (~15%) mixed infection samples that contain known drug resistance SNPs that are ambiguous calls (Table S7). Nucleotide counts of ‘A’s, ‘C’s, ‘G’s, and ‘T’s based on the read depth were used to calculate frequencies of the detected nucleotides. The variant caller only flags samples with ambiguous base calls if nucleotide frequencies are in the range of 25-75%. So 118 is a conservative estimate as only mixed infections with a ratio in the 25-75% range will show up in the analysis.

I compared the estimated frequencies of present strains with frequencies obtained from the nucleotide counts at ambiguous positions. I selected those mixed infections that consist of two strains, samples consisting of 3 strains are more difficult to solve. Also when the infecting strains occur in a 50/50 ratio, the detected frequencies cannot be matched. Samples for which the estimated frequencies are close to a 50/50 ratio can be safely match when there is sufficient coverage. There are 57 mixed infection samples that were manually compared with drug resistance frequencies for which 3 examples are given:

Example 1:

> In sample PRJNA183522 (Table S7, row 96) a sub-lineage infection of two lineage 4 strains was found. Strains 4A and 4B were in a 0.56/0.44 ratio. In this sample, I found an ambiguous call for the well-known mutation S315T in gene *katG* which had similar frequencies of nucleotide base calls. The nucleotide change from 'G' (susceptible variant) to 'C' (resistant variant) was detected with frequencies of respectively 0.45 and 0.54, thus the strain from lineage 4A carries a mutation making it resistant to isoniazid (INH).

Example 2:

> In sample ERR234611 (Table S7, row 15) a mixed infection was found to have two drug resistant strains that together cause MDR-TB at the patient-level. The sample consisted of a lineage 2 and 4 strain in a ratio of 0.61/0.39. The lineage 2 strain is resistant to RIF caused by a mutation L430P in the *rpoB*-gene, nucleotide bases were detected at a frequency of 0.57 ('C') for the resistant variant and 0.41 ('T') for the susceptible variant. The lineage 4 strain is resistant to INH caused by mutation S315T in the *katG*-gene, the mutation was found in frequency 0.35 ('C') for the resistant variant and 0.65 ('G') for the susceptible variant.

Example 3:

> In sample ERR067701 (Table S7, row 23), I found strains from lineage 2 and 4 in a ratio of 0.77/0.33. In this sample XDR-TB at the patient-level is caused by two MDR-TB strains. Mutations that cause resistance to RIF and INH were non-ambiguous, so these mutations are found in both strains. Then there were two ambiguous SNPs for which the detected frequencies can be matched to the present strains. The lineage 2 strain is resistant to kanamycin caused by G-10A in the *eis*-promoter, the mutation to 'A' was detected at frequency of 0.87 and the susceptible variant 'G' was detected at 0.13. The lineage 4 strain is resistant to a fluoroquinolone (FQ) caused by mutation A90V in the *gyrA*-gene, the mutation was found in a frequency of 0.29 ('T') for the resistant variant and 0.71 ('C') for the susceptible variant. XDR is defined as resistance to all these four drugs and so for this mixed infection the strains together from XDR-TB at the patient-level.

In this way, I found 22 mixed infection samples that contain a MDR strain and 3 samples with an XDR strain. There were 28 samples that consisted of one drug susceptible and one drug resistant strain and 29 samples consisted of two drug resistant strains.

4. Conclusion

I found that it is possible to rapidly detect mixed TB infections by using SNP markers that are specific to hierarchical clusters in the MTBC phylogenetic tree. The presence of multiple paths with sufficient read coverage supports the presence of multiple strain types indicating a mixed infection. Confidence for determining their presence is based on read support from multiple SNP markers that are specific to the detected 'end' group and to all their ancestor clusters leading back to the MTBC root.

A sizeable fraction of the mixed infections detected in this study was found to contain ambiguous SNPs associated with drug resistance in literature (Coll *et al.*, 2014). These mixed infections have bacterial populations that differ in their drug susceptibility.

4.1 Prevalence of mixed infections

In this study the prevalence of mixed infection is estimated to be ~10% based on the global dataset of 8217 samples. In literature, mixed infections were often detected in as many as 10-20% (Cohen *et al.*, 2012; Hanekom *et al.*, 2013; Huang *et al.*, 2010; Navarro *et al.*, 2011). Studies aimed to detect mixed infections are likely to underestimate the prevalence of mixed infections due to sample handling or the genotyping sensitivity and therefore these estimates serve as a lower bound (Plazzotta *et al.*, 2014; Cohen *et al.*, 2012). Conventionally, mixed infections were detected using genotypic methods that have limited sensitivity because of the limited numbers of markers they use to distinguish strains (Ford *et al.*, 2012). WGS enables improved estimates of mixed infections as it offers an higher resolution to differentiate MTBC strains that is not possible with conventional genotyping methods (Walker *et al.*, 2012). Although multiple strains can be detected simultaneously, mixed infections may have been missed because it is common practice to isolate single colonies, reducing the chance of detecting a putative mixed infection (Galagan, 2014; Ford *et al.*, 2012; Cohen *et al.*, 2012). The ability to detect mixed infections is also dependent of the within-host diversity, the sequencing technology and sequencing coverage (Eyre *et al.*, 2013; Pulido-Tamayo *et al.*, 2015; Köser, Ellington, *et al.*, 2014).

4.2 Detection sensitivity

The sensitivity of my method and thus the ability to detect minority strains in a mixed infection depends on thresholds that were used. Distinguishing a sequence error from a true SNP is more difficult for SNPs present in bacterial strains that occur in low frequencies. Here, a minimum depth of 5 reads was chosen to support the presence of a marker, such as to reduce the effect of sequencing errors. This means that sufficient sequencing depth is required to identify minority strains and to distinguish them from other present strains.

Each hierarchical cluster needs a large number of SNP markers to support its presence. A cluster is considered to be present if more than 1/8 of its markers are detected, such that a reasonable fraction of markers is at least needed to detect clusters and the detection of minority strains is still possible. This fraction threshold will easily include unwanted clusters that have few SNP markers, e.g. if a cluster has at most 8 markers only one marker needs to be detected to determine its presence. So when clusters are defined and the SNP markers have been obtained it is preferable to only keep clusters for detection that have a large set of markers to give more confidence for determining the presence of cluster.

The accuracy of frequency estimates were validated using a test set of samples with reads from a lineage 2 and 4 sample mixed in different proportions (Table 2). The results show that predicted frequencies of present sub-populations are very close to the true frequencies and that they could be estimated with a low maximum deviation of 0.02.

4.3 Defining hierarchical clades as clusters

I choose to define hierarchical clades in MTBC as clusters for lineage identification when these clades consisted of at least 10 strains. The question arises whether these groups are relevant and will reoccur when reconstructing the phylogenetic tree, especially smaller groups that are close to the leaves. To define hierarchical clusters it may be preferable to select clusters by bootstrap values e.g. >95% instead of a threshold based on a cluster size of at least 10 strains. A bootstrap value is the proportion of pseudo-replicate trees in which a clade is recovered, by resampling across the nucleotide characters the tree is rebuild. Bootstrap confidence levels apply to single tree nodes and are usually lower near leaf nodes (Soltis and Soltis, 2003).

4.4 Detection of drug resistance SNPs

Detection of drug resistances depends on our current knowledge of mutations known to cause resistance to particular drugs. Known mutations from Coll *et al.*, 2015, have been detected in data from Cohen *et al.*, 2015, for which phenotypic resistance profiles of 9 anti-TB drugs were available. I found large concordance between phenotypic and genotypic detected resistance. False negatives may results from resistance phenotypes of MTBC isolates that do not harbor mutations in the known resistance genes (Fonseca *et al.*, 2015). This may negatively affect the treatment of mixed infections (Zetola *et al.*, 2014). Genotypic resistant strains that are phenotypic susceptible may confer low-level resistance that is below the threshold for clinically relevant resistance when tested for the MIC of a particular drug (Zetola *et al.*, 2014). When mutations in drug resistance genes are associated with certain (sub-)lineages genotypic testing can result in false positives (Köser, Ellington, *et al.*, 2014). Further, it should be taken into account that phenotypic DST tests show poor reproducibility for some drugs such (Feuerriegel *et al.*, 2015).

In the manually curated drug resistance list, some codon changes cannot be translated to nucleotide changes, so the data lacks information about the specific nucleotide change that has been detected in their study.

4.5 Mixed infections with differences in drug susceptibility

Merging results from the genotyping method and drug resistance detection showed that a mixed infection can contain a drug susceptible and resistant strain. Mixed infections have been linked with poor treatment outcome when the infecting strains differ with respect to drug susceptibility, thus when a patient is treated it is important to identify the presence of multiple resistance profiles to avoid positive selective pressure of drug resistant strains (Hingley-Wilson *et al.*, 2013; van Rie *et al.*, 2005). Cases of resistant minority strains may be missed if genotyping methods are designed to detect a single strain, therefore it is crucial to have mixed infection detection methods. The Xpert MTB/RIF assay is a genotypic test to rapidly diagnose resistance to RIF, but it fails to detect RIF resistance in vitro when resistant sub-populations accounted for <90% (Zetola *et al.*, 2014). I have observed that RIF resistance can also be detected for minority strains present in 20% of the population with matching frequencies.

This method can be applied to raw sequence reads to determine the presence of a single or multiple MTBC strain type(s) simultaneously. There is no need to map the sample reads against a reference genome to reconstruct the whole genome and call variants. Sample reads can be used directly to identify cluster-specific SNP markers for strain type detection.

Mixed infections and frequencies of present populations can be detected by using a large set of SNP markers and constructing tree paths of present clusters in MTBC. In addition, these findings show that some mixed infections indeed have differences in drug resistance that may contribute to the poor treatment of mixed infections.

The results suggest that resistance to multiple strains at the patient-level may be explained by the presence of multiple strains with individual resistance to particular drugs. Extrapolating from my findings I estimate that at least 135000 patients each year are affected by mixed infections for which the strains have different drug resistance profiles. This is when 9 million cases of TB infections consists of 10% mixed infections and for which 15% of these mixed infections the strains differ in their drug susceptibility. The infecting strains together form a more severe infection that is more difficult to treat. Using the hierarchical clustering method allows to detect putative mixed infections to improve diagnostics and therefore also enable improved treatment of TB patients.

5. References

Benavente,E.D. *et al.* (2015) PhyTB: Phylogenetic tree visualisation and sample positioning for M. tuberculosis. *BMC Bioinformatics*, **16**, 155.

Blouin,Y. *et al.* (2014) Progenitor 'Mycobacterium canettii' clone responsible for lymph node tuberculosis epidemic, Djibouti. *Emerg. Infect. Dis.*, **20**, 21–8.

Bradley,P. *et al.* (2015) Rapid antibiotic resistance predictions from genome sequence data for S. aureus and M. tuberculosis. *bioRxiv*, **6**, 018564.

Casali,N. *et al.* (2014) Evolution and transmission of drug-resistant tuberculosis in a Russian population. *Nat. Genet.*, **46**, 279–286.

Cheng,L. *et al.* (2013) Hierarchical and Spatially Explicit Clustering of DNA Sequences with BAPS Software. *Mol. Biol. Evol.*, **30**, 1224–1228.

Cohen,K.A. *et al.* (2015) Evolution of Extensively Drug-Resistant Tuberculosis over Four Decades: Whole Genome Sequencing and Dating Analysis of Mycobacterium tuberculosis Isolates from KwaZulu-Natal. *PLOS Med.*, **12**, e1001880.

Cohen,T. *et al.* (2012) Mixed-strain Mycobacterium tuberculosis infections and the implications for tuberculosis treatment and control. *Clin. Microbiol. Rev.*, **25**, 708–719.

Cohen,T. *et al.* (2003) The effect of drug resistance on the fitness of Mycobacterium tuberculosis. *Lancet Infect. Dis.*, **3**, 13–21.

Coll,F. *et al.* (2014) A robust SNP barcode for typing Mycobacterium tuberculosis complex strains. *Under Rev.*, **5**, 4812.

Coll,F. *et al.* (2015) Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. *Genome Med.*, **7**, 51.

Comas,I. *et al.* (2009) Genotyping of genetically monomorphic bacteria: DNA sequencing in Mycobacterium tuberculosis highlights the limitations of current methodologies. *PLoS One*, **4**, e7815.

Comas,I. *et al.* (2013) Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans. *Nat. Genet.*, **45**, 1176–1182.

Coscolla,M. and Gagneux,S. (2014) Consequences of genomic diversity in Mycobacterium tuberculosis. *Semin. Immunol.*, **26**, 431–444.

Excoffier,L. and Heckel,G. (2006) Computer programs for population genetics data analysis: a survival guide. *Nat. Rev. Genet.*, **7**, 745–758.

Eyre,D.W. *et al.* (2013) Detection of Mixed Infection from Bacterial Whole Genome Sequence Data Allows Assessment of Its Role in Clostridium difficile Transmission. *PLoS Comput. Biol.*, **9**, e1003059.

Feuerriegel,S. *et al.* (2015) PhyResSE: a Web Tool Delineating Mycobacterium tuberculosis Antibiotic Resistance and Lineage from Whole-Genome Sequencing Data. *J. Clin. Microbiol.*, **53**, 1908–14.

Filliol,I. *et al.* (2006) Global Phylogeny of. **188**, 759–772.

Firdessa,R. *et al.* (2013) Mycobacterial lineages causing pulmonary and extrapulmonary Tuberculosis, Ethiopia. *Emerg. Infect. Dis.*, **19**, 460–463.

Flandrois,J.-P. *et al.* (2014) MUBII-TB-DB: a database of mutations associated with antibiotic resistance in Mycobacterium tuberculosis. *BMC Bioinformatics*, **15**, 107.

Fonseca,J.D. *et al.* (2015) The complex evolution of antibiotic resistance in Mycobacterium tuberculosis. *Int. J. Infect. Dis.*, **32**, 94–100.

Ford,C. *et al.* (2012) Mycobacterium tuberculosis--heterogeneity revealed through whole genome sequencing. *Tuberc.*, **92**, 194–201.

Galagan,J.E. (2014) Genomic insights into tuberculosis. *Nat. Rev. Genet.*, **15**, 307–320.

Hanekom,M. *et al.* (2013) Population Structure of Mixed Mycobacterium tuberculosis Infection Is Strain Genotype and Culture Medium Dependent. *PLoS One*, **8**, e70178.

Hingley-Wilson,S.M. *et al.* (2013) Undetected multidrug-resistant tuberculosis amplified by first-line therapy in mixed infection. *Emerg. Infect. Dis.*, **19**, 1138–1141.

Homolka,S. *et al.* (2012) High resolution discrimination of clinical Mycobacterium tuberculosis complex strains based on single nucleotide polymorphisms. *PLoS One*, **7**, e39855.

Hörandl,E. and Stuessy,T.F. (2010) Paraphyletic groups as natural units of biological classification. *Taxon*, **59**, 1641–1653.

Huang,H.Y. *et al.* (2010) Mixed infection with Beijing and non-Beijing strains and drug resistance pattern of Mycobacterium tuberculosis. *J. Clin. Microbiol.*, **48**, 4474–4480.

Huson,D.H. and Scornavacca,C. (2012) Dendroscope 3: An Interactive Tool for Rooted Phylogenetic Trees and Networks. *Syst. Biol.*, **61**, 1061–1067.

Jagielski,T. *et al.* (2014) Current methods in the molecular typing of mycobacterium tuberculosis and other Mycobacteria. *Biomed Res Int*, **2014**, 645802.

Köser,C.U., Comas,I., *et al.* (2014) Genetic diversity within Mycobacterium tuberculosis complex impacts on the accuracy of genotypic pyrazinamide drug-susceptibility assay. *Tuberculosis*, **94**, 451–453.

Köser,C.U., Ellington,M.J., *et al.* (2014) Whole-genome sequencing to control antimicrobial resistance. *Trends Genet.*, **30**, 401–407.

Larkeryd,A. *et al.* (2014) CanSNPer: a hierarchical genotype classifier of clonal pathogens. *Bioinformatics*, **30**, 1762–1764.

Li,H. *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Navarro,Y. *et al.* (2011) Systematic survey of clonal complexity in tuberculosis at a populational level and detailed characterization of the isolates involved. *J. Clin. Microbiol.*, **49**, 4131–4137.

Perdigão,J. *et al.* (2013) From multidrug-resistant to extensively drug-resistant tuberculosis in Lisbon, Portugal: The stepwise mode of resistance acquisition. *J. Antimicrob. Chemother.*, **68**, 27–33.

Perdigão,J. *et al.* (2014) Unraveling Mycobacterium tuberculosis genomic diversity and evolution in Lisbon, Portugal, a highly drug resistant setting. *BMC Genomics*, **15**, 991.

Plazzotta,G. *et al.* (2014) Magnitude and Sources of Bias in the Detection of Mixed Strain M. tuberculosis Infection. *J. Theor. Biol.*, **368**, 1–7.

Plinke,C. *et al.* (2010) embCAB sequence variation among ethambutol-resistant Mycobacterium tuberculosis isolates without embB306 mutation. *J. Antimicrob. Chemother.*, **65**, 1359–67.

Pulido-Tamayo,S. *et al.* (2015) Frequency-based haplotype reconstruction from deep

sequencing data of bacterial populations. *Nucleic Acids Res.*, **43**, e105–e105.

van Rie,A. *et al.* (2005) Reinfection and Mixed Infection Cause Changing *Mycobacterium tuberculosis* Drug-Resistance Patterns. *Am. J. Respir. Crit. Care Med.*, **172**, 636–642.

Rosales-Klintz,S. *et al.* (2012) Drug resistance-related mutations in multidrug-resistant Mycobacterium tuberculosis isolates from diverse geographical regions. *Int. J. Mycobacteriology*, **1**, 124–130.

Sandgren,A. *et al.* (2009) Tuberculosis drug resistance mutation database. *PLoS Med.*, **6**, 0132–0136.

Sims,G.E. *et al.* (2009) Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proc. Natl. Acad. Sci. U. S. A.*, **106**, 2677–2682.

Sims,G.E. and Kim,S.-H. (2011) Whole-genome phylogeny of Escherichia coli/Shigella group by feature frequency profiles (FFPs). *Proc. Natl. Acad. Sci. U. S. A.*, **108**, 8329–8334.

Soltis,D.E. and Soltis,P.S. (2003) Applying the Bootstrap in Phylogeny Reconstruction. *Stat. Sci.*, **18**, 256–267.

Spies,F.S. *et al.* (2011) Streptomycin resistance and lineage-specific polymorphisms in Mycobacterium tuberculosis gidB gene. *J. Clin. Microbiol.*, **49**, 2625–2630.

Stamatakis,A. (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, **22**, 2688–2690.

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

Walker,B.J. *et al.* (2014) Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*, **9**, e112963.

Walker,T. *et al.* (2012) A whole-genome sequencing approach to targeting Mycobacterium tuberculosis outbreak management. *Lancet*, **380**, S77.

Walker,T.M. *et al.* (2015) Whole-genome sequencing for prediction of Mycobacterium tuberculosis drug susceptibility and resistance: a retrospective cohort study. *Lancet. Infect. Dis.*, **3099**, 1–10.

Warner,D.F. *et al.* (2015) Diversity and disease pathogenesis in Mycobacterium tuberculosis. *Trends Microbiol.*, **23**, 14–21.

WHO (2014) Global tuberculosis report 2014 (WHO). 171.

Zetola,N.M. *et al.* (2014) Mixed Mycobacterium tuberculosis Complex Infections and False-Negative Results for Rifampin Resistance by GeneXpert MTB/RIF Are Associated with Poor Clinical Outcomes. *J. Clin. Microbiol.*, **52**, 2422–2429.

Zhang,H. *et al.* (2013) Genome sequencing of 161 Mycobacterium tuberculosis isolates from China identifies genes and intergenic regions associated with drug resistance. *Nat. Genet.*, **45**, 1255–1260.