Challenges to identify mixed *Mycobacterium tuberculosis* infections in whole genome sequencing data

Master Bioinformatics Research
Arlin Keo, 4257111

April 8, 2015

Abstract

Strains of the *Mycobacterium tuberculosis* complex are known to cause the infectious disease tuberculosis killing millions of people worldwide. Insights into the phylogenetic relations of the varying tuberculosis strains that have varying disease outcomes helps to improve our understanding of this bacterial pathogen, and therefore enable better diagnostics and treatment of tuberculosis. Thus, there is a need to differentiate the strains of the *Mycobacterium tuberculosis* complex with high resolution to also distinguish very closely related strains. A typing method that differentiates with high resolution is also desired for the detection of mixed infections, infections where the bacterial in-host population consists of multiple different genotypes that are responsible for different disease outcomes, in example differences in drug susceptibility. Mixed infections impede the treatment of tuberculosis as the detection of the differing strains is hampered by low resolution genotyping methods. Conventional typing methods are limited in the detection of more closely related strains and tend to homoplasy, because the discriminative markers are based on repetitive or mobile genetic elements that have partly caused the bacterial diversification, but only cover a small part of the whole genome.

More recent approaches have applied whole genome sequencing (WGS) to differentiate strains based on single nucleotide polymorphisms (SNPs) that allow a finer resolution. Although WGS studies yielded many new important insights into the pathogen, there are several limitations to overcome with whole sequence data. Most often variants are called relative to a reference genome that is genetically closely related, but to compare strains that are more closely related to each other than the reference strain may give redundant variants between the strains. While application of an assembly-based approach allows to call variants in multiple strains simultaneously relative to each other, the method seems to be restricted to the order in which assembly graphs, that contain variant information, are pairwise merged. Additionally, both mapping-based and assembly-based variant callers are restricted to the use of short reads to resolve repetitive regions and genomic duplications leaving ambiguous calls. Furthermore, studies differ in their approach to construct SNP-based phylogenies and defining phylogenetically informative SNPs. Large amounts of data and less stringent SNP filters are preferred to differentiate tuberculosis strains at a higher level. Finally, SNP markers may be utilized for the detection of mixed infections by applying an alignment-free approach using only raw sequencing data.

Contents

1 Introduction

Tuberculosis (TB) is ranked as the second leading cause of death from an infectious disease worldwide, after HIV.[1] In 2013, an estimated 9 million people developed TB and 1.5 million died from the disease, improved diagnosis and treatment of TB are needed to reduce the global high morbidity and mortality rates.[1] The bacterial pathogen has been coevolving with humans for tens of thousand years and since the use of drugs in the last century drug resistant strains have been observed.[2,3] To control and fight the ongoing TB epidemic it is important to understand the underlying characteristics of this human obligate pathogen.[4] Earlier studies on genotype-phenotype associations were based on repetitive or mobile genetic elements which are thought to have a higher mutation frequency.[5] However, these conventional strain typing methods have limitations and cannot distinguish between genetically closely related strains of TB.[4,5]

With the advent of whole genome sequencing (WGS) a higher resolution to differentiate TB strains is achieved.[4] Advances in sequencing technologies have made available many whole genome sequences of TB from around the world and new genomic insights into the origin of TB has improved our understanding of the pathogen.[3,5] Comparative WGS studies led to the reconstruction of a robust phylogenetic tree and revealed that TB strain diversity was much higher than previously considered.[5,6,7] Also, strains of TB have been linked to geographical regions, and the transmissibility of TB lineages is associated with differences in disease severity.[5]

In this paper I review the current knowledge about TB, the challenges in detecting different TB strains in mixed infections, and the opportunities of applying sequence technologies to enrich our view about this pathogen. Finally, a study is proposed to develop an alignment-free tool that is capable of differentiating TB strains at an unprecedented resolution in WGS samples with mixed infection.

2 Pathogen-host interaction

The causal pathogen of TB is the bacterium *Mycobacterium tuberculosis* and is transmitted by inhalation.[3] Typically, the disease affects the lungs causing pulmonary TB, but can also affect other sites referred to as extrapulmonary TB.[1] The success of this pathogen is partly due to its ability to survive in macrophages.[3] After it is inhaled via aerosols, *M. tuberculosis* is ingested by the alveolar macrophages[8]. In contrast to pathogens that avoid uptake into host phagocytic cells, *M. tuberculosis* takes advantage of this mechanism to intrude and exploit the host phagocytic cells.[8] Within the macrophage in granuloma the bacterium can stay in a dormant, asymptomatic state in which the replication rate is thought to be slower.[3] This latent infection is thought to exist in approximately one-third of the world's population. In a small fraction (5-10%) of the infected individuals, granuloma are broken down and active disease ensues.[8] This causes the bacilli to be released into the airways whereafter respiratory transmission to the next host can occur.[3,8] The exact dose of TB transmission that results in infection or active disease is still unclear.[9]

Coinfection with HIV alters the hosts immune response to TB and increases the risk of active disease, compounding the global health problem.[8,9] The immune deficiency in HIV-infected individuals also correlates with an increase in the likelihood of extrapulmonary dissemination.[8,9] Although there is an increased risk of active disease, disease occurs is a host with compromised immunity that reduces the risk of TB transmission.[10,11] This compromised immunity threatens to disrupt the TB infection cycle and therefore HIV-positive patients are considered poor TB transmitters.[10,11] So, for TB to complete its life cycle of infection, disease, and transmission it depends on the hosts functional adaptive immune response. This also suggests that the ability to cause active disease will directly

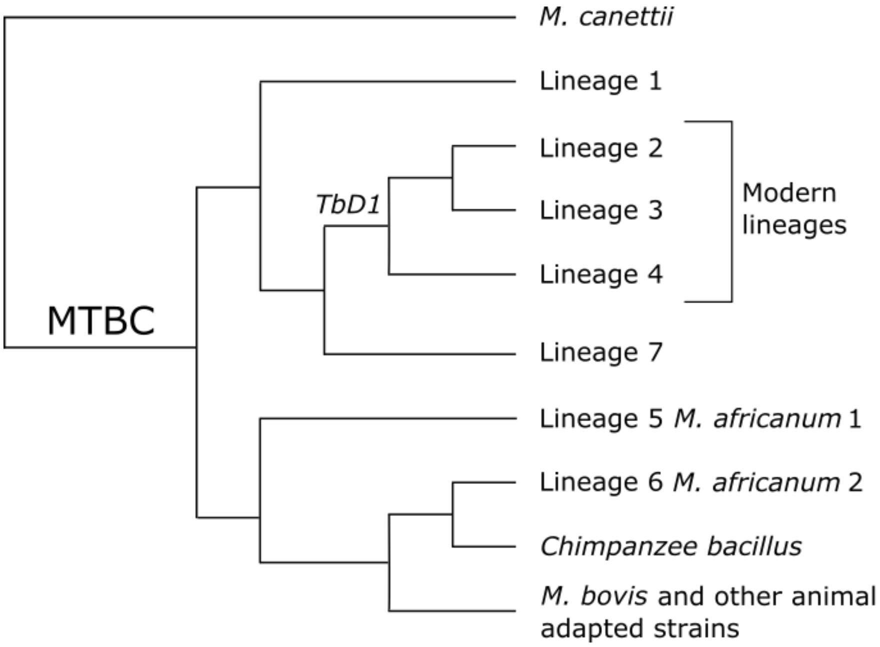

Figure 1: The *Mycobacterium tuberculosis* complex (MTBC) consists of human and animal adapted strains that share a recent common ancestor with *M. canettii*. The modern monophyletic lineages 2-4 are characterized by a tuberculosis specific TbD1 deletion that are considered more virulent and transmissible than other lineages.[5] The animal adapted strains branche from a presumed human adapted *M. africanum* that is currently restricted to West Africa.[3]

impact transmissibility, and genetic selective pressure for transmission is likely to be associated with an increase in strain virulence.[11]

3 Genomic diversity in MTBC

M. tuberculosis is a member of a group of organisms known as the *M. tuberculosis* complex (MTBC).[3] This group seems to be a clonal expansion from a progenitor population that arose at least 70000 years ago and spread with human migration out of Africa.[2,3] This was confirmed by a WGS study by Comas et al.[2] who revealed the congruence between the MTBC phylogeny and a tree constructed from human mitochondrial genomes. MTBC includes strains of human adapted *M. tuberculosis* and *M. africanum*, and animal adapted mycobacteria, such as *M. bovis* (Figure 1).[3] The human adapted strains are split 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.[5] This modern monophyletic group is characterized by a genomic deletion TbD1 and is globally more succesful than the ancestral paraphyletic group that do not have this deletion and is locally restricted.[5] MTBC strains have low genetic diversity and there is no unambiguous evidence of horizontal gene transfer (or lateral gene transfer).[3] Unlike the more distantly related *M. canettii* they have no plasmids and MTBC is characterized by clonality.[3,5] Thus, the evolution of *M. tuberculosis* seems to be more restricted as diversification is primarily driven by chromosomal mutations.[11,12] Except for mutations involved in drug resistance, the presence of convergent evolution events in MTBC is rare.[5] A large proportion of acquired mutations in MTBC lineages lead to phenotypic differences and despite the clonality there is evidence that MTBC is highly diverse.[11]

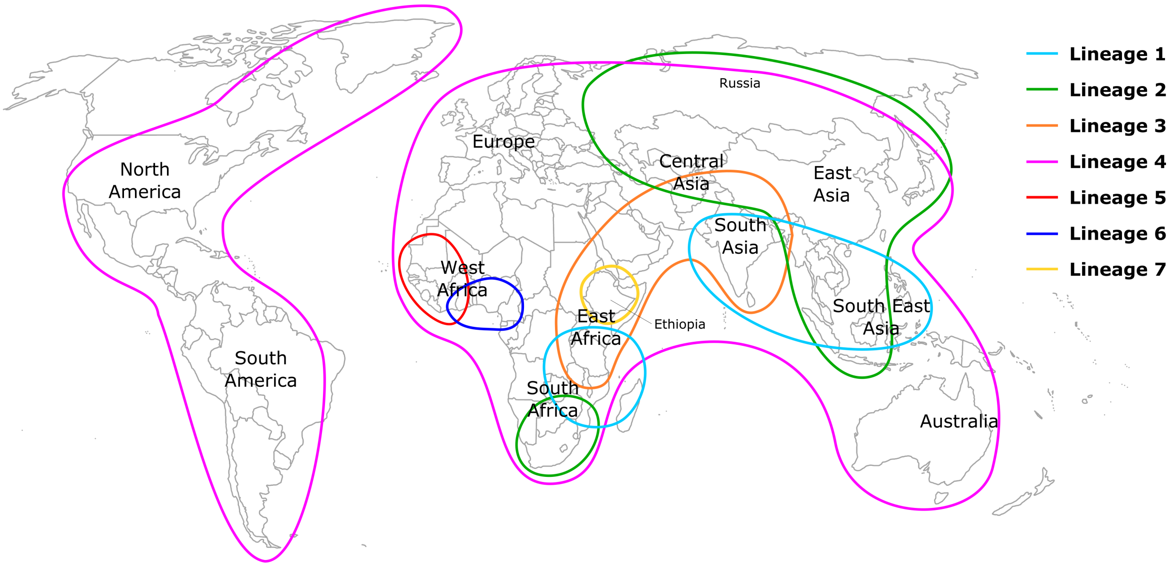

Figure 2: Phylogeographic distribution of MTBC lineages 1-7. Each lineage is represented by a different color that covers an area on the map that indicates the distribution of the lineage, areas are based on sample isolates obtained in these regions from different studies.[5,13,14] Transmissibility of the different lineages is linked to their virulence, the modern lineages are generally more virulent and globally succesful. Lineage 2 and lineage 4 are the most widely distributed groups, while other lineages are more locally restricted.

3.1 Phylogeographic distribution

Several studies have shown that the human adapted lineages have a strong phylogeographical population structure, with different lineages associated with distinct geograpical regions (Figure 2).[2,5,6,7] MTBC lineages differ in their severity of disease and their virulence is directly linked to transmissibility. Hypothetically, human population expension during Neolithic- and Industrial revolutions have been linked to an increase in virulence of some MTBC lineages, the highly transmissible chains tend to cause more severe disease. Compared to other more geographically restricted lineages, the modern lineages are generally more virulent and globally successful.

East-Asian lineage 2 and Euro-American lineage 4 are the most widely distributed groups (Figure 2).[5,13,14] Lineage 2 includes the Beijing family and mainly occurs in East Asia, but is also present in Central-, South East Asia, Russia, and South Africa.[5] Although lineage 2 strains have also been identified in isolates sampled in North America, these were found in patients that emigrated from Asia.[2] Lineage 4 is found in populations from Asia, Europe, Africa, and America. Lineage 1 and 3 are more geographically restricted, limited to East Africa, Central-, South-, and South East Asia. The most geographically restricted lineages are 5-7. Lineage 5 and 6 are known respectively as *M. africanum* West Africa 1, occuring in the Eastern part of West Africa, and West Africa 2, present in de Western part of West Africa.[5] A novel phylogenetic lineage of MTBC, lineage 7, appears to be intermediate between ancient and modern, and is confined to Ethiopia and immigrants from this region.[5,15]

3.2 Genomic distances

To provide a first indication of the relative genomic distances between the lineages of MTBC, Coscolla and Gagneux[5] quantified the genomic diversity of 217 human adapted MTBC clinical strains previously published. By calculating the number of single nucleotide polymorphisms (SNPs) between any pair of strains, they found the within diversity of two human adapted strains to differ by about 1200 SNPs (about 0.03% of the genome, repetitive sequences excluded), and 2.7% between *M. canettii* and any MTBC strain. Lineage 1 harbors the largest genetic within diversity with an average of 730 SNPs between any two strains in this phylogenetic group. The lowest average distance was 230 SNPs between any two strains in lineage 7. The between lineage diversity of the modern lineages 2-4 was found to be 970 SNPs on average, the ancient lineages 1, 5 and 6 differ by 1500 SNPs. Between lineage 7 and either lineage 1, 5 and 6 the maximum distance is 1800 SNPs.

4 The emergence of drug resistant strains

Horizontal gene transfer plays no significant role in the emergence of drug resistance in MTBC and diversification in MTBC evolution is primarily driven by chromosomal mutations, but nevertheless multi-drug resistant strains emerge rapidly.[3,16] Strains of MTBC may evolve from drug susceptible tuberculosis, having no resistance to any drugs, to totally drug-resistant tuberculosis (TDR-TB), strains not susceptible to any tested drugs.[17] In 2013 on average, an estimated 9.0% of patients with multi-drug resistant tuberculosis (MDR-TB, defined as at least resistant to first-line drugs rifampicin and isoniazid) had extensively drug resistant tuberculosis (XDR-TB, MDR with additional resistance to any fluoroquinoline and at least one second-line agent, e.g. kanamycin, amikacin, capreomycin).[1,12,18] In some parts of the world the higher resistance levels and poor treatment outcomes are of major concern.[1]

Drug pressure in recent evolution of MTBC is thought to cause the selection of drug resistant strains.[5] Through the sequential fixation of mutations, drug resistant strains develop and persist to exist under positive selective pressure during poor treatment.[17,18] A signature of positive selection is the observation that resistance mutations to many drugs are often independently acquired by different isolates.[3,5] For example in the Western Cape and KwaZulu-Natal regions of South Africa, highly drug resistant strains emerge independently in the same geographic area from monoresistant isolates with distinct resistance mutations.[6]

In the absence of treatment drug resistant strains are assumed to suffer a fitness cost relative to the susceptible strains in the host,[3] while in the presence of treatment selection of resistance mutations become exacerbated as mutations in low-fitness strains become fixed when it outcompetes the drug susceptible strains.[11] The fitness cost may also be affected by epistasis where the phenotypic effect of a mutation depends on the presence or absence of other mutations in the genome. The fact that fitness cost can be offset by compensatory mutations implies that resistant strains may persist even in the absence of drug treatment.[3,12] Because drug resistance effects are attributed to a set of discrete SNPs, drug resistance could be measured by observing the genetic change.[17] Compensatory mutations were investigated by Casali et al.[12] to determine the transmissibility and prevalence of drug resistance genotypes, they found that drug resistance may be more multi-factorial than previously appreciated and resistance-conferring mutations are acquired without compromising fitness and transmissibility. This shows that in addition to weaknesses in tuberculosis control programs, persistence and spread of drug resistant tuberculosis is driven by biological factors.[12]

5 Prevalence and effects of mixed infections

While MTBC is considered largely clonal, there is more strain diversity within the infected hosts than previously expected.[6] This type of infection with more than one distinct strain of MTBC is referred to as mixed infection and has been detected in 10-20% of the cases in areas where the incidence of TB is high.[19] MTBC in-host diversity have been found both in extrapulmonary and respiratory sites, meaning that this variability can be transmitted and impact the inference on transmission events and subsequentially insights into the epidemiology of TB.[19,20]

Microdiversity within a single host may exist through: 1) the simultaneous infection by multiple strains, 2) superinfection, reinfection with a new strain, or 3) in-host evolution, that is genetic diversity may arise during the course of infection.[6,21] In the latter case, microdiversity within the host develop from a single infecting strain.[11] Mixed infections have been linked with poor treatment outcomes when the infecting strains differ with respect to drug susceptibility.[19] These differences in drug susceptibility may arise through in-host evolution, as mentioned in the previous section drug resistance is acquired through the sequential fixation of mutations.[6,21] Hence it is of great importance to determine how mixed infections can influence disease outcome and treatment efficacy.[17]

5.1 Challenges to detect mixed infections

There are several complications in the detection of mixed infections that can occur during specimen collection and selection.[21] Clinical isolates are usually cultured from sputum samples that may not represent the true bacterial heterogeneity in the host, as it is likely that not all cavitary lesions are open to the airways at a given time and lesions producing sputa may change during the course of infection.[6,11] Consequently, heterogeneity might not be adequately reflected and instead isolates might only contain a subset of the variants.[11,21] Most often a single sputum sample is collected, while it is suggested that increasing the number of sputum samples increases the likelihood of detecting mixed infections. Also, using specimen from multiple sites is assumed to increase the sensitivity of detection of mixed infections.[21]

Detection of in-host heterogeneity also depends on how specimen are handled, there are different approaches for collecting DNA material from culture that may contribute to the sensitivity of the assay.[21,22] A culture step is employed to increase the bacterial population for extraction of bacterial DNA, but during culture the clonal composition may be affected as MTBC strains may have different abilities to grow and divide in different medium types.[21,22] This was shown in a study by Hanekom et al. (2013)[23] in which Lowenstein-Jensen media was found to be more sensitive in detecting mixed infections than Mycobacterial Growth Indicator Tubes media. As the clonal composition changes, minority variants may be lost, leading to underestimates of mixed infections.[21] Moreover, delays from collection to laboratory and the risk of cross-contamination will likely decrease the sensitivity of methods to detect mixed infections, influence of delays and decontamination are not known.[21] To completely cover the bacterial diversity and detect heterogeneity within a sample, it is preferred to use DNA that was extracted from all bacterial colonies from a culture instead of a single colony.[3,17]

The actual prevalence of mixed infections in human populations may be underestimated given the typing methods that are not sufficiently sensitive to differentiate MTBC strains and detect mixed infections.[19,21,22] Only methods that differentiate with high resolution can quantify the true proportion of mixed infections.[23]

6 Limitations of conventional typing methods

An ideal typing method is not only characterized by its power to differentiate strains, the procedure must also be suitable for standardization, enable a high inter- and intralaboratory reproducibility, and the turn-around time of a method must be as short as possible.[4]

The first attempts to differentiate strains were based on phage typing, a phenotypic method that exposes MTBC isolates to a predefined selection of bacteriophages, viruses that can infect and lyse specific types of mycobacteria.[15,21] A mixed infection is detected when at least two mycobacteriophages with nonoverlapping specificities infect mycobacteria from a single sample.[4,21] However, most available mycobacteriophages are nonspecific and can infect a wide range of MTBC.[21] Therefore, phage typing can only differentiate strains at a coarse resolution and exhibits poor sensitivity for the detection of mixed infections.[21] It provides no quantification of strains and has been displaced by genetic methods.[15,21]

Some studies have utilized a PCR-based approach to detect the presence or absence of different MTBC strains by using strain specific PCR probes to amplify the mycobacterial DNA.[22,23] The PCR primers are designed to only target a specific family of strains complemented with primers that are specifically not identical to the family of strains to be detected, for example Beijing and non-Beijing primers.[22,23] Strain specific PCR amplification may be applied to DNA directly extracted from sputum without the need to culture, but due to the limited sensitivity a mixed infection cannot be detected if it harbors different strains of a single lineage or sublineage, and lineages for which primers are not included would be missed.[21]

The lack of horizontal gene transfer precludes MTBC strains from reacquiring genomic regions that have been lost and therefore genomic deletions or regions of difference (RDs) have marked the evolutionary history of MTBC.[14,16,24] Therefore, the presence or absence of RDs is phylogenetically informative and lineages defined by RDs were shown to be associated with different geographical regions.[4,16] Determination of RDs in laboratory is easy and straightforward, but the power to differentiate is limited for lineages that do not exhibit the assayed RDs.[4]

6.1 Genotyping based on repetitive elements

Most widely used conventional genotyping methods are based on repetitive elements like Clustered Regularly Interspaced Short Palindromic Repeats regions (CRISPRs), Variable Number of Tandem Repeats (VNTRs) loci and insertion sequences (ISs) that are other important sources of genomic variation.[5]

6.1.1 IS6110-RFLP typing

Restriction Fragment Length Polymorphisms (RFLP) analysis of the repetitive genetic mobile element IS6110 involves cleaving the genomic DNA with a restriction enzyme, followed by separating the restriction fragments in gel electrophoresis, blotting the fragments to a nylon membrane, and labeling with a probe that contains the IS6110-sequence.[3,4,5] Based on differences in both copy numbers and the genomic location of IS6110 different banding patterns are obtained for different isolates.[4,5] The presence of a mixed infection is indicated by different banding patterns after subculturing the multiple colonies from a single sample, or the presence of ”low intensity bands” in a single gel.[21] IS6110-RFLP is considered the most sensitive of all traditional typing methods.[25] However, IS6110-RFLP typing is difficult to reproduce and requires a large amount of good quality data, which makes the typing method time-consuming as it can takes weeks to grow a sufficient amount of bacteria to extract DNA from.[4,5] On the other hand, PCR-based methods such as spacer oligonucleotide typing (spoligotyping) and Mycobacterial Interspersed Repetitive Unit - Variable Number of Tandem Repeats (MIRU-VNTR) analysis require little DNA.[5]

6.1.2 Spoligotyping

Spoligotyping is based on the presence or absence of 43 spacer sequences (35-41 bp) between direct repeats (36 bp) in the Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) locus of the MTBC genome.[3,4,5] The power to differentiate MTBC strains of this technique is low as it targets only a single locus that accounts for 0.1% of the genome,[3,25] but spoligotyping is fast and inexpensive with a good portability between laboratories.[4,21] This typing approach has been used regularly for strain differentiation and dated phylogenies were derived from spoligotyping patterns.[7] The presence of a mixed strain infection is indicated by different spoligotype patterns observed after subculturing multiple colonies from a single culture sample, but when typing multiple strains in a single sample the spoligotyping pattern may appear as an overlapping pattern of all strains present or show only spacers from a dominant strain.[21,26]

6.1.3 MIRU-VNTR

MIRU-VNTR typing is based on the number of repeats at 24 different VNTR loci that is determined by PCR.[4,21] Primers specific to each locus are used to amplify an unknown number of repeats, resulting in varying amplicon lengths that are determined in gel electrophoresis from which the number of tandem repeats present is determined by extrapolation.[4] A mixed infection may be detected from a single sample and is indicated by the presence of multiple bands (amplicons of varying lengths, and thus variation in copy number) at a single MIRU-VNTR locus.[21,26] The number of VNTR loci can exhibit different discriminatory power in different lineages, MIRU-VNTR based on 15 loci was shown to be less discriminatory than with an additional 9 loci (the current 24 loci-based method), and also tends more to homoplasy.[27] However, the current 24 loci-based approach is still limited in strain differentiation within certain lineages.[21,27] In spite of that, easy implementation and portability of this typing tool allowed the simultaneous investigation of strain transmission in epidemiological studies as well as the phylogenetic lineage classification of clinical isolates,[4,28] but it is suggested that MIRU-VNTR should only be used in a lineage dependent manner.[27]

6.2 Detection limits in mixed infections

There is a limited sensitivity in detecting the prevalence of mixed infections when using the typing methods described above. Phage typing, strain specific PCR, and determination of RDs can only make crude distinctions between MTBC strains and may greatly underestimate the frequency of mixed infections.[21] While genotyping methods based on repetitive elements permit differentiation with much greater resolution, the corresponding molecular markers still provide limited functional information and are highly prone to convergent evolution.[5,6,21] Cohen et al.[26] have used MIRU-VNTR typing to detect mixed infections and clonal heterogeneity in patients in KwaZulu-Natal and suggest that due to the limited sensitivity of MIRU-VNTR, the frequencies of mixed infections may by higher than they found in their study. Same patterns may occur for phylogenetically unrelated strains and therefore the genotyping methods are limited for phylogenetics and strain classification.[5,6] In addition, the power of these methods to differentiate strains and results do not always agree.[6] For example, isolates of the TB reference genome H37Rv and the avirulent sister strain H37Ra were indistinguishable using spoligotyping, but could be detected using IS6110-RFLP analysis.[25]

Despite the limitations, first insights into the global population structure and the phylogenetic lineage composition have been gained by IS6110-RFLP and spoligotyping.[28] Moreover, spoligotyping and MIRU-VNTR have been routinely used to trace ongoing chains of transmission, detect laboratory cross-contamination, and have been successfully applied to address a variety of epidemiological questions.[27] Apart from the earlier successful applications of the conventional typing methods it must be taken into account that the current estimates of the prevalence of mixed infections serve as a lower bound and the actual proportion of cases of mixed infections may be significantly larger.[21] WGS on the other hand, will enable better estimates of mixed infections as it offers a higher resolution to differentiate MTBC strains that is not possible with other typing methods.[21,23]

7 The advent of WGS

MTBC strains differ in their content of SNPs, small insertions and deletions, large genomic deletions, large duplications, and insertion sequences. As WGS theoretically reveals all types of mutations, it has the highest power to differentiate strains that is possible on DNA level.[5] The relatively low amount of genetic diversity of MTBC makes WGS particularly powerful and enables phylogenetic analysis at a high resolution.[4,6] During the past years, WGS studies have demonstrated the use of SNPs to detect lower levels of genotypic variation and the phylogenies were shown to more robust than nonsequence-based phylogenies that had low bootstrap support and high rates of homoplasy.[2,7,12,15,18,27,28,29] Thus, WGS sets a new gold standard for phylogenetic classification that allows to define deep phylogenetic groupings with very high confidence, and to answers questions about evolution, transmission chains, and drug resistance.[6,27,28]

7.1 Sequencing technologies

WGS uses deep sequencing to sample the genome that allows the detection of low-frequencies mutations, which is an important feature for the detection of drug resistance mutations in mixed infections.[3,30,31] Bacterial DNA is subjected to random shearing and the variable-sized fragments are amplified and sequenced to obtain reads. Reads are then regularly mapped to a reference genome that is genetically closely related and the coverage of reads is used to assemble the original sequence.[31]

The fastest and most cost-effective method is shotgun sequencing to generate short reads (40-250 bp), Illumina instruments are currently most widely used generating hundreds of millions short reads per run.[3,31] Typically, long repeat regions are not captured when using short reads and genome arrangements are likely to be undersampled, pair-end sequencing and jumping libraries can mitigate these issues.[3] Pacific Biosciences (PacBio) sequencers generate longer reads that can be combined with short reads to improve assembly.[3] Unfortunately, they are limited by cost and generate fewer reads in a single run that seems to be a trade-off for depth coverage.[31,32] The lower depth coverage makes it difficult to distinguish true variants from sequencing errors and consequently has an impact on the detection of variants.[31] Error rates also increase towards the end of longer reads, but with the rapidly advancing sequence technologies such current technical limitations may be overcome.[3,31]

7.2 Construction of phylogenies

Different methods and computational tools have been employed in recent studies to construct a robust WGS phylogenetic tree of MTBC. The primary steps in these comparative studies involve variant calling, variant selection, phylogenetic computations, and cluster analysis. In Table 1 an overview of the applied methodologies of eight WGS studies are summarized.

The construction of a phylogenetic tree is mostly based on a rooted outlier *M. canettii* that represent an early branching lineage of human infecting mycobacteria and shares the most recent common ancestor with MTBC (Figure 1).[3] Based on the genetic distances between the strains a phylogenetic tree may be constructed on the basis of the best-scoring maximum likelihood tree,[2,12,15,18,27,28,29] other

Table 1: Overview of methodologies applied by eight WGS studies to construct MTBC phylogenies. Different approaches have been applied to construct SNP-based phylogenies using different tools to analyze the data. Data set size indicates the number of samples used to contruct the phylogeny. Statistical support in the study by Zhang et al.[29] could not clearly be derived from the paper and is indicated with a ”?”.

Study	Data set size	Rooted outlier	Variant caller	SNP selection	SNP set size	Phylogeny construction (software tool)	Statistical support	Phylogeny comparison
Filliol et al. (2006)[7]	219	Unrooted	Pairwise comparison of H37Rv, CDC1551, strain 210, and *M. bovis*.	SNPs that are well-distributed across the genome and unique to one of the four strains.	212	Consensus parsimony tree from 500 bootstrap replicates (MEGA), distance-based neighbor-joining algorithm (MEGA), model-based clustering analysis (STRUCTURE).	Bootstrap values $> 80\%$.	Spoligotyping, MIRU-VNTR.
Comas et al. (2009)[27]	97	*M. canettii*	*De novo* DNA sequence data.	SNPs discovered in 89 genes.	339	Neighbor-joining analysis (MEGA), max. likelihood analysis (PHYLM), and MCMC-based Bayesian analysis (MrBayes).	1000 Bootstrap replicates, and Bayesian *a posteriori* support values.	Spoligotyping, MIRU-VNTR.
Homolka et al. (2012)[28]	68	*M. canettii*	SeqScape	Variants discovered in 26 genes with an overall higher mutation rate.	155 + 6 deletions	Max. likelihood analysis (Treefinder)	Bootstrap values $> 90\%$, 1000 Bootstrap replicates.	Spoligotyping, MIRU-VNTR.
Comas et al. (2013)[2]	220	*M. canettii*	MAQ SNP caller and SAMtools.	SNPs called by both variant calling approaches.	34167	Max. likelihood analysis (RAxML), neighbor-joining analysis (MEGA).	1000 Bootstrap replicates.	Human mitochondrial DNA.
Zhang et al. (2013)[29]	183	*M. canettii*	Bases with a quality score ¡20 assessed by an inhouse program.	Number of most abundant (n1) and second most abundant (n2) nucleotides were examined: i) n1 base was different than in H37Rv, II) $n1 + n2 \geq$ 10, II) $n1/n2 \geq 5$. SNPs in repetitive regions were excluded, high quality SNPs filtered.	18970	Max. likelihood analysis (TreeBeST)	?	Spoligotypes.
Clark et al. (2013)[18]	51	Unrooted	SAMtools/BCFtools, Breakdancer, CREST, Pindel, Delly, Velvet.	High quality (1 error/1000) variants in highly variable gene families and nonunique regions supported by both directional reads, excluding polymorphisms with two or more missing genotypes.	6857	Clustering dendogram (R), max. likelihood analysis (RAxML)	100 Bootstrap replicates.	Spoligotypes.
Casali et al. (2014)[12]	1035	Unrooted	SAMtools, Pindel.	SNPs within repetitive regions were excluded.	32445	Max. likelihood analysis (RAxML)	100 Bootstrap replicates.	Spoligotypes.
Coll et. al. (2014)[15]	1601	*M. canettii*	SAMtools/BCFtools, GATK.	Filtered out nonunique SNP sites.	91648	Bootstrapping combined with max. likelihood search (RAxML).	Bootstrap values.	Lineage specific RDs, spoligotypes.

methods include neighbor-joining analysis and maximum parsimony analysis.[2,6,7,27] To assess the robustness of a constructed phylogeny, bootstrapping is applied to provide statistical support. In this method, replicate bootstrap data sets are produced by resampling across the nucleotide characters for which the tree is rebuild. This is done in many iterations so as to determine the proportion of pseudoreplicate trees in which a clade is recovered, presented as the bootstrap value or bootstrap support of a node or clade.[33] The resulting phylogenies are often compared to spoligotype-based phylogenies to investigate the population structure of the samples.

8 Variant calling for strain differentiation

Mapping assembly and annotation of unfinished genomes often rely on a finished reference genome, where reads are mapped to reconstruct the original genome.[32] In MTBC studies, the genome sequence of *M. tuberculosis* H37Rv laboratory strain has been widely used to map short reads and perform variant calling, but genome assembly with a reference genome may give obscure SNPs present at low frequencies.[11,25] In addition, detection of structural variants is limited by the use of short reads, which cannot resolve large-scale structural mutations mediated by repetitive transposable elements.[32]

8.1 Mapping-based variant calling

Improved variant calling methods to detect larger genomic deletions and alterations are required to provide better insights into MTBC diversity.[11] Popular variant callers like GATK Unified Genotyper[34] and SAMtool/BCFtools[35] are frequently used tools that rely on a reference genome to detect variants in MTBC. More recently, a novel mapping-based variant caller Pilon[36] has been developed that outperforms the state-of-the-art tools and makes fewer mistakes. Other than traditional variant calling tools, Pilon has the ability to find insertions and deletions that are considerably larger than sequence read length.[36] The variant caller is particularly powerful when both fragment insert and long insert libraries are used, then it is also able to identify large-scale events including large-scale genome duplications. When using only short fragment reads, Walker et al.[36] showed that Pilon was unable to span the entire length of the IS6110-element when calling variants in the *M. tuberculosis* F11 genome against *M. tuberculosis* H37Rv , adding long insert data improved the ability to detect large insertions. For calling SNPs the added value of long insert libraries is small, but for SNPs in repetitive regions long insert data is very helpful in disambiguating these events. Pilon is easy to apply as it is able to operate immediately, no parameters has to be set first, and no separate filtering criteria needs to be specified to determine which calls are of high confidence.

In mapping-based approaches comparison to a reference genome makes it difficult to detect variable regions specific to the sequenced strain, therefore the reference genome should be closely related to the species under investigation.[3] Also, the use of different references can complicate analysis[3] and to avoid this, H37Rv annotated by the Sanger institute is suggested to be the starting point for future studies, because it has undergone extensive manual curation.[25]

8.2 Assembly-based variant calling

A whole genome may be aligned against another whole genome to detect variants relative to each other without the need of a reference genome and the result can be visualized in an assembly graph that allows to observe the detected variants between the strains.[37] Such an approach allows the simultaneous assembly of more than two samples and such a variant calling method was developed by Linthorst et al.[37], called Probubble, that combines the use of string graphs for variant calling and a scalable method to enable the construction of a multi-sample variation graph. Sequences that are common in both strains are merged into one node and when variation is detected a bubble structure in the multi-assembly graph can be seen with nodes containing alternative sequences (Figure 4).

Probubble its ability to call variants in multiple genomes simultaneously was tested on strains of MTBC combined with Cytoscape[38] to visualize the assembly graphs. A multi-assembly graph was constructed for inter- and intralineage variant calling, respectively: (1) a set containing one strain from each lineage 1-6 and the reference genome H37Rv (identity of 80.56%), and (2) a set of eight strains that are more closely related from lineage 4 only (identity of 95.16%, including H37Rv to the multi-assembly resulted in a identity of 94.73%). A multi-assembly graph was constructed by starting with the alignment of two strains followed by merging two of the resulting assembly graphs until all samples were merged (Figure 3a and 3b). All parameters were set to default, except for the global merging threshold, that is the minimal length of common substrings, was set to k=2500. Incorrect calls were not analyzed through the whole graph, the following examples just show there is some inconsistency in the method that result from multiple graphs being merged.

Results show that the final multi-assembly is dependent of the assembly order (Figure 3), this can be seen in the bubble structures in Figure 4 where part of the graphs are given (the reference strain H37Rv is abbreviated to ref., L1-6 indicates the lineage that is represented by a strain, and S1-8 are identifiers for different strains from lineage 4). In Figure 4a the alternative sequence ”CG” in L4 (purple node) is a subsequence of ”CGC” in ref. and L1-3 (red node) at the same genomic position and these variants were not merged. This is because the assembly graph for ref. and L1-3, that contained no bubble at this part of the graph, was merged with the graph for L4-6, that did contain a bubble at this position in the sequence, and variants from one graph were not identified as subsequences of variants from the other graph. This shows how the multi-assembly graph is dependent of the assembly order and the same problem can be observed in Figure 4b, ”C” in S5-6 and S7-8 (purple and light blue nodes) is a subsequence of ”GCT” in S1-4. Besides, this pair of nodes should have been merged earlier during assembly and this also applies to the green and light green nodes that contain ”A”, and the orange and yellow nodes that contain ”G”. Thus, duplicate nodes that represent the same base or sequence can be observed at a single position and should have been merged by Probubble.

As Probubble did not properly merged the individual assembly graphs together, the bubble structures are more complex then necessary. An optimal structure is expected to be represented as in the manually reconstructed multi-assembly graphs in Figure 5, where nodes that are parallel and represent the same alternative sequence are merged. In these graphs, there are no variants that are subsequences of other variants and that have the same position in the genome. For example in Figure 5a, ”G” in the pink node is a subsequence of ”CG” in the orange node, but the identical bases do not share the same genomic position and so it is a mutation in L5-6. In Figure 5b, the middle position of the varying 3 bp sequence is a ”C” in all strains S1-8 indicated by a blue small node, that makes the neighboring nodes (pink, red, grey, and brown nodes) single base mutations or SNPs.

Thus, in the current version of Probubble the order of alignment may result in different assembly graphs and one should consider the order before comparing multiple samples, which in turn may be too demanding for researchers. If more is known about the sample genomes being studied, phylogenetically closely related strains could be compared first before merging to a larger assembly graph. For example when comparing lineages of MTBC and considering their phylogenetic relations, the pairs L2 and L3, L4 and reference H37Rv, L5 and L6 should be first aligned, where upon the first two pairs can be merged into an assembly graph for L2-4 and H37Rv, followed by adding L1, and finally the pair L5 and 6 are added to obtain the final graph. However, for more closely related strains the phylogenetic relations are often not known.

The current implementation is not capable of handling complex bubbles that arise through repeti-

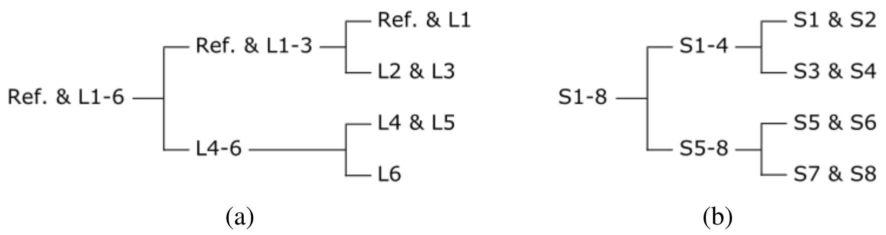

Figure 3: Multi-assembly order for (a) interlineage variant calling of six strains of each lineage L1-6 and reference strain H37Rv, and (b) intralineage variant calling of eight strains S1-8 from lineage 4 only. Multi-assembly graphs from Probubble are dependent of the assembly order 4.

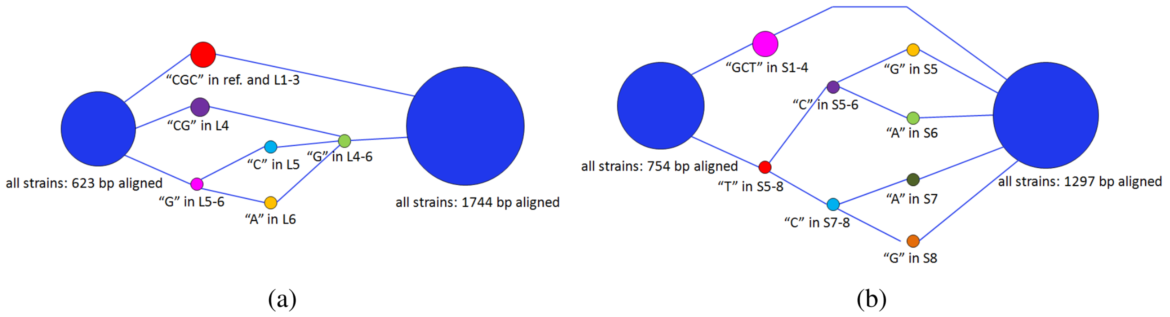

Figure 4: Part of the multi-assembly graph from Probubble for **(a)** interlineage variant calling of six strains of each lineage L1-6 and reference strain H37Rv (ref.), and **(b)** intralineage variant calling of eight strains S1-8 from lineage 4 only. The nodes vary in size according to the size of the sequence they represent and the color of a node indicates the strain(s) that contain the sequence, node information is given below the nodes. The large blue nodes represent a matching sequence found in all strains for which the size of the sequence is given, and in between a bubble structure shows variation found between the strains in a 3 bp sequence, the small nodes represent alternative sequences that is given below the nodes (colors in **(b)** represent other strains). It can be observed that the graph structure is more complex than necessary, parallel nodes that contain the same alternative sequence should have been merged. Ideally the structure is represented as in Figure 5.

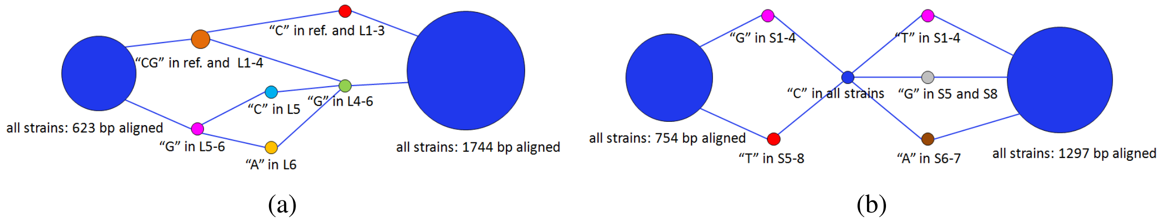

Figure 5: Optimal manually reconstructed graph structure suggested for the graph obtained from Probubble in Figure 4 for **(a)** interlineage variant calling of six strains of each lineage L1-6 and reference strain H37Rv (ref.), and **(b)** intralineage variant calling of eight strains S1-8 from lineage 4 only. **(a)** In this manually reconstructed graph the orange node merges L4 with ref. and L1-3 that all have the alternative sequence ”GC” at that position and then splits into two nodes: red for ”C” in the reference strain and L1-3, and green for ”G” in L4 merged with ”G” in L5-6. **(b)** Between the strains S1-8, there is no mutation at the middle position between the two large blue nodes, because all strains have a ”C” at that position (small blue node), Probubble did not recognized this identical base (pink, purple, and light blue nodes in Figure 4b). At the next position, duplicate nodes indicating the same base in the graph of Figure 4b are removed in the reconstructed graph here and there are three varying bases ”T”, ”G”, and ”A” at this position for all strains S1-8.

tive sequences which may result in loops in the alignment or tip nodes (containing unaligned flanked regions of reads) due to limited coverage. Authors of Probubble are currently working on the needed improvements for this recently developed tool and towards an implementation that can deal with repeats and determine copy number.[37]

8.3 Assembly- versus mapping-based variant calling

The VCF-file obtained for interlineage comparison was analyzed to count the number of variants called by Probubble and only contains variants that passed the criteria filter. Since this involves a multi-alignment, multiple alternatives are given for a specific variant position, and one possible variant may be found in more than one genome. Considering the multiple alternative sequences and excluding the possibility they might be shared by multiple genomes, 1113 mutations (if length of the reference sequence is equal to the length of the alternative sequence), 7996 deletions (if reference length is larger than alternative length), and 202 insertions (if alternative length is larger than reference length) were counted, see also Table 2. It should be mentioned that Probubble's output in the VCF-file was inconsistent and was not yet solved in the implementation at the time. To compare these numbers with Pilon's mapping-based approach, variants were also counted from a concatenated list of variants of the same genomes representing six lineages that were individually mapped against H37Rv. For fair comparison multiple alternatives at a specific base position were taken into account and the fact that an alternative sequence may be shared by more than one genome was again excluded. For variants that passed all filter criteria, this resulted in 6431 mutations, 556 deletions, and 564 insertions (Table 2). The difference in the number of mutations, deletions, and insertions between Probubble and Pilon may be explained by the inconsistent merging of pairwise alignments in Probubble. For example in Figure 4a the reference sequence is "CGC" having length 3 with four optional alternatives "G", "C", "CG", and "A" that all have a smaller sequence length and were counted as 4 deletions based on the VCF-file. Under close observation one can conclude these are actually mutations, because each strain in this bubble structure has a sequence of length 3 with varying bases. How to properly write variants resulting from a multi-assembly in a VCF-file is still debated.

Comparing an assembly-based variant caller to a mapping-based variant caller leaves us with an important question to address about assembly-based performance. If genomic distance from the reference genome increases for two closely related strains under study, does an assembly-based approach call less false positives than a mapping-based approach? Thus, intralineage variants called by Probubble and Pilon were also counted and the number of variants called by Pilon is significantly larger. Probubble called 413 mutations, 326 deletions, and 51 insertions, and Pilon called 2032 mutations, 224 deletions, and 305 insertions, see also Table 2. Here, calling variants relative to each other instead of a reference genome resulted in less calls for closely related strains.

In comparison to mapping-based approaches, assembly-based methods needs higher read coverage to produce a high quality assembly.[37] Independent of using a reference genome, the main challenge for variant callers is to solve repetitive regions.[6] Tandem repeats remains challenging for both Pilon and Probubbble as assembly ambiguities in these regions make it unable to determine the copy number.[36,37] These genomic duplications are difficult to detect with short reads, because there are multiple ways to build contigs.[6]

Table 2: Comparison of inter- and intralineage variant calls by an assembly-based variant caller Probubble[37] and mapping-based variant caller Pilon[36]. A variant was considered as a: (1) mutation, if the length of the reference sequence is equal to the length of the alternative sequence, (2) deletion, if the reference length is greater than the alternative length, or (3) insertion if the alternative length is greater than the reference length.

Comparison	**Variant calling method**	**Mutations**	**Deletions**	**Insertions**	**Total number of variants**
Interlineage	Probubble[37]	1113	7996	202	9311
	Pilon[36]	6431	556	564	7551
Intralineage	Probubble[37]	413	326	51	790
	Pilon[36]	2032	224	305	2561

9 Phylogenetically informative SNPs

Despite the evidence of large-scale genome rearrangements causing gene loss or gene duplication, genome evolution in MTBC is thought to be mostly driven by sequential chromosomal nucleotide substitutions.[3] SNPs are unlikely to converge, as can be the case with spoligotype and MIRU-markers,[7] and due to the low-level homoplasy in MTBC, SNPs can be usefully applied to measure evolutionary distances between isolates.[5,11]

9.1 Types of SNPs and their added value

SNPs can be classified into two major groups: synonymous and nonsynonymous, and additionally they can be classified into intergenic, and nonsense SNPs.[5,24] The type of SNPs differ in their phenotypic consequences.

Nonsynonymous SNPs in coding regions introduce amino acid changes to the protein that may influence the phenotype, and hence are thought to be under selective pressure.[2,15,24,29] In previous findings the majority of the SNPs located in coding regions are nonsynonymous polymorphisms present at approximately 60% of all detected SNPs.[2,15,18,28] Occasionally, the dN/dS ratio (nonsynonymous-/synonymous SNPs ratio) is used as a measure of individual protein evolution as well as of the impact of selection on the genome,[29] a ratio >1 reflects positive selection for increased diversity.[6] Zhang et al.[29] found that dN/dS ratios for drug resistant and drug susceptible isolates were similar, but found an increased ratio in drug resistance associated genes indicating a relatively large positive selective effect.

Although, synonymous SNPs are considered evolutionary neutral, because they do not alter the amino acid profile, they can influence function and do have important phenotypic effects.[5,11] Synonymous SNPs have been shown to alter transcriptional start sites (TSS) increasing expression of genes essential to TB.[5] Thus, synonymous SNPs still can act as unambiguous markers for certain lineages.[7]

Intergenic SNPs may also affect gene expression if they are within regions that encode small RNAs, that can regulate gene expression in response to environmental changes, or within promoters of genes, that may lead to overexpression of downstream genes in MTBC.[29] Intergenic SNPs were found in intergenic regions that have a strong association with drug resistance and they may also have a compensatory role in drug resistance.[29]

A nonsense mutation is a mutation that results in a premature stop codon and a truncated nonfunctional peptide, the structurally disturbed gene is referred to as a pseudogene.[5] These nonsense mutations are thought to have a minor impact on phenotype, because many of these mutations are in genes that have analogous genes or pathways.[5] In despite, this type of SNP should not be disregarded as it still may have a phenotypic impact. This was shown in a study by Casali et al.[12], a nonsense

SNP was identified in the KatG locus that mediated resistance. KatG is a frequently mutated drug resistance locus that encodes an enzyme required for the activation of the prodrug isoniazid.[39]

Independent of whether SNPs are synonymous, nonsynonymous, located in genes, or intergenic, all SNPs types might have an impact on the phenotype. To avoid loss of informative SNPs in comparative studies it is suggested to relax the stringency of SNP filters applied.[17]

9.2 Minimal SNP marker sets

Several studies have suggested the use of a minimal SNP set that is sufficient for strain classification.[7,15,27,28] Recently, Coll et al.[15] identified a panel of 62 robust lineage informative SNPs, that are nonsynonymous and occur in essential genes, to construct high resolution phylogenies. The main lineages and sublineages were initially identified from the phylogenetic tree, based on the spoligotype and the RDs present in each clade. They refer to this minimal set as a SNP barcode and benchmarked their minimal set with minimal SNP sets proposed by three other studies. One of them is a study by Filliol et al.[7] that proposed 45 SNPs including synonymous, nonsynonymous, and intergenic SNPs to distinguish MTBC lineages and *M. bovis*. Although this set can be used for initial characterization of isolates, WGS allows for a finer resolution.[6] In another study by Comas et al.[27] 89 genes were selected to identify SNPs of which 93 are (sub)lineage specific. 72 SNPs were proposed by Homolka et al.[28] identified from 26 genes selected to be highly variable to distinguish 17 (sub)lineages. The methodologies applied to construct the WGS phylogenies is included in Table 1. In the benchmark procedure, Coll et al.[15] tested all four minimal SNP sets to resolve 7 lineages and 55 sublineages that they defined in their large global data collection of 1601 genomes. Unsurprisingly, they found their 62 SNPs to perform better in classifying MTBC strains than those proposed by the three other studies.

In all four studies the SNPs were selected for their phylogenetic informativeness, but there is litte congruence between the minimal SNP sets. A Venn diagram (in Figure S1) shows the little overlap of the minimal SNP sets that is determined from analyzing the SNP positions in the reference genome H37Rv (Table S1). The 45 SNPs suggested by Filliol et al.[7] was shared by none of the other studies, and the minimal SNP sets of Comas et al.[27] and Coll et al.[15] only shared one SNP. The largest concordance was found between the SNP sets of Comas et al.[27] and Homolka et al.[28] sharing 22 SNPs, this is explained by their choice to select SNPs only from genes of which 10 genes were analyzed in both studies. The little similarity between all four studies may be the result of the different approaches employed to extract SNPs (Table 1) 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 sublineages are defined, which is different among these studies.

The minimal SNP set of Coll et al.[15] contains one SNP per (sub)lineage to differentiate MTBC strains, while choosing more than one SNP will avoid the inequalities coming from varying SNP-typing technologies.[28] Almost a decade ago, Filliol et al.[7] 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 should be more suitable to use a wide range of SNPs, to collect as much as possible phylogenetically informative SNPs for a robust MTBC classification method.

9.3 Lineage specific SNPs to detect mixed infections

Genotyping methods are limited in their power to differentiate and detect mixed infections at the strain level. In the next section, we propose an alignment-free method that enables the detection of MTBC mixed infections at sublineage level. In addition, the tool also determines the frequencies of the strains present in a mixed infection sample. The proposal is based upon the tool Macaw that detects strains at lineage level without alignment to a reference genome, in samples with mixed infection it can determine the frequencies of the present lineages with a detection limit of 0.05. For this tool 3722 lineage specific markers were identified from a global set of 1297 samples and was used for strain identification, lineage specific markers were identified using Bonferroni corrected Fisher's exact test. Because SNPs were called relative to H37Rv that belongs to lineage 4, inverse lineage specific SNPs were selected for lineage 4 (SNPs present in all other lineages except lineage 4).

To get acquainted with Macaw, the tool was tested on three types of samples: one sample did not contain a mixed infection and only lineage 2 was detected, and the other two samples both contained a mixed infection with two strains (lineage 2 and 4) and three strains (lineage 2, 4, and 6). For each lineage specific marker the presence or absence in a sample is determined statistically based on the read depth, if coverage is sufficient then the marker is present otherwise absent. Macaw outputs a list of all their markers and whether they were found to be present or absent in the tested sample. Also, a raw marker total is given for each lineage that indicates the number of sample reads that have been aligned to the lineage specific markers. Based on this number the lineage fraction present in the sample is given, this fraction is corrected at the end to eliminate noise. The sample containing a mixed infection with two strains, lineage 2 and 4, were found to be present at 0.48 and 0.52 respectively. The sample containing three strains, lineage 2, 4, and 6, were detected at 0.04, 0.85, and 0.11 respectively. Macaw classifies MTBC strains into one of the 7 main lineages or *M. bovis*, but WGS allows a higher resolution to differentiate MTBC strains and has the potential to distinguish drug resistant strains from drug susceptible strains. A study is proposed in the next section to address an important question stated by Ford et al.[6]: ”Can we look at key SNPs from pooled sweeps of colonies to estimate their frequencies in mixed cultures and rapidly detect low frequencies of drug resistance mutations in an individual?”

10 Project Proposal: Alignment-free identification and quantification of mixed *Mycobacterium tuberculosis* infections at the sublineage resolution

MTBC is highly diverse and bacterial diversity may even exist in a single host, referred to as a mixed infection, complicating diagnosis and treatment. The advance of whole genome sequencing (WGS) has improved our understanding of the pathogen and holds promise to be further exploited. To our knowledge, there exists no tools that uses WGS data to detect and quantify the strain variation in a single sample.

10.1 Specific aims

The goal is to develop a tool to detect strain variation and determine their frequencies in mixed infections at sublineage level. The first step to get to the goal is defining sublineages. The project will be build upon Joe Romano's work who developed Macaw, which only differentiates between the main lineages 1-7 while WGS allows a higher resolution to differentiate MTBC strains. A mapping-free-based approach will be utilized that allows researchers to study mixed infections more easily and faster. Using a set of sublineage specific SNPs and raw sequencing data from patient isolates, mixed infections can be easily detected and the presence of strains at sublineage level can be determined, as wel as their frequencies with a low detection limit. In addition, individuals infected with closely related species of MTBC, like *M. canettii* show similar symptoms as TB and the tool may also be utilized to detect the infecting species. Three specific aims are stated that describe how the goal will be accomplished:

- **Aim 1: Define sublineages and sublineage specific SNPs**
  Sublineages are not clearly defined in literature, and differentiating MTBC strains into the main lineages is too coarse. A set of SNPs will be retrieved from VCF-files produced by Pilon that will be used to build a phylogenetic tree. In combination with a clustering algorithm, Bayesian Analysis of Population Structure (BAPS)[40,41,42], clusters of the tree will be defined as sublineages. The initial set of SNPs will then be filtered to only contain SNPs that are unique to one of these sublineages.
  Expected outcome: A list of SNPs to differentiate MTBC strains at sublineage level.
  Benchmark: Evaluate whether the list of sublineage specific SNPs is indeed capable of detecting the defined sublineages. The tool that uses these SNPs to detect the sublineages needs to be developed first and is part of aim 2.

- **Aim 2: Develop a tool to detect strains in mixed infections and their frequencies**
  The tool will allow researchers to detect mixed infections and their frequencies by using only raw sequencing data. The sublineage specific SNPs from aim 1 will be extended to k-mers of 21 bp that will be used as markers in the tool. Reads are aligned to these markers, whereas sufficient coverage wil reveal the presence or absence of SNPs in the sample being analyzed.
  Expected outcome: A tool to detect mixed infections and their frequencies using only raw sequencing data.
  Benchmark: 1) Detect sublineages of aim 1, 2) *in silico* data set of mixed infections with gradually changing proportions of 2 (or more) strains to determine the detection limit, 3) and evaluate the tool using real mixed infection data.

- **Aim 3: Extend the tool to detect genotypes associated with drug resistance.**
  Generate a new set of markers from SNPs associated with resistance to specific tuberculosis drugs, based on literature, and extend the tool to detect resistance genotypes aside from sublineage identification. In case of mixed infections it is of great interest to know which drug resistance genotypes are present in a sample, this knowledge can greatly improve treatment of TB patients. Other than the sublineage identification method, each drug resistance specific marker needs to be complemented with an alternative marker that does not contain the specific SNP to identify drug susceptible genotypes.
  Expected outcome: Next to sublineage identification, the tool can be used to determine drug resistance genotypes present in the sample.
  Benchmark: 1) Detect drug resistance genotypes in drug resistant isolates, 2) determine frequencies of drug- resistance and susceptible genotypes in *in silico* mixed infection data and the detection limit, 3) and apply tool to detect real mixed infections containing both drug susceptible and drug resistant strains.

10.2 Approach

The following describes how the aims specified above will be achieved. A time plan is given in Table 3 that shows the tasks that will be done in nine months for the proposed project.

Literature research will be done about existing methods that detect mixed TB infections with WGS data. This will reveal the state-of-the-art tools that are available and their potential shortcomings when it comes to detecting mixed infections. If tools exists that are capable of detecting and quantifying mixed infections in WGS data, they can be used in the benchmarking step to evaluate and compare it with our own tool. Genome data will be collected from multiple large WGS studies to attain a large global comprehensive data set, mainly to find the phylogenetically informative SNPs that will be used as markers in the detection of mixed infections. A training and test set will be composed, to develop the tool and evaluate the tool respectively. For this MTBC genome data collection, variant calling has been performed for each genome with the tool Pilon[36] by assembling the sample reads to the finished reference genome H37Rv. From these VCF-files a list of all SNPs found in the whole data collection will be extracted, whereupon a phylogenetic tree may be constructed. This phylogeny will be compared to other phylogenies from other studies, including trees derived from conventional strain typing methods.

The main lineages 1-7 in MTBC are clearly defined in literature, but the sublineages are not well defined. Casali et al.[12] have performed clustering of their samples based on Bayesian genetic population clustering (BAPS), which will be adopted to identify sublineages in our dataset. Literature research could reveal additional sublineage clustering methods, and their underlying reason of applying the specific method. Possibly, another method may be preferred, given the information obtained from literature, and may be applied to define the sublineages in our data. Evaluation can not be performed until aim 2 is accomplished.

When sublineages are defined satisfactorily, the SNPs called by Pilon[36] can now be classified specific to the defined sublineages. A SNP is sublineage specific if it is only found in strains from this cluster and not in any other cluster. The Fisher's exact test is a possible measure to identify the sublineage specific SNPs, but literature could reveal potential other approaches that we have to search for. Markers will be generated by extending the sublineage specific SNPs to 21-mers with 10 bp sequence on both sides derived from the reference genome H37Rv. There is a possibility that SNPs resides within 10 bp of other SNPs, these have to be discarded. Also, nonunique marker sequences have to be removed to avoid complications when read coverage is used to calculate strain frequencies.

The final set of SNP markers needs to be aligned to the raw sequencing sample reads, and read coverage will be determined. Incorporating statistics will yield confidence for read coverage and enable a reliable output of the tool. The finished tool can be evaluated and tested whether the earlier defined sublineages are indeed identified (aim 1). An in silico data set of mixed infections with gradually changing proportions of the present strains will be generated to determine the detection limit in mixed infections. The tool will also be evaluated on real mixed infection data.

Next to strain identification, there is a need to detect drug resistance genotypes present in a single sample. So the tool will be extended such that it can also be used to detect specific drug resistance genotypes and again determine the frequencies. Therefore, literature research will reveal the SNPs that are associated with resistance to specific TB drugs. Instead of selecting sublineage specific markers, the list of SNPs derived from the Pilon VCF-files will now be filtered to only contain drug resistance specific SNPs from which 21 bp markers will be generated. Above that, each marker needs to be complemented with an alternative marker to identify strains susceptible to the specific drug. These alternative markers contain the same 10 bp sequence surrounding the specific SNP except that it does not contain the SNP it self, but the base present in the pansusceptible reference genome H37Rv. The

Table 3: Proposal time plan in 9 months

Aim	Task	Subject	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
1	1	MTBC genome data collection	x	x	x						
	2	Literature research into detection of mixed infections	x	x							
	3	Generate list of SNPs and their coordination in ref. H37Rv from Pilon VCF-files		x	x						
	4	Build phylogeny and compare with phylogenies from other studies			x						
	5	Define sublineages with BAPS			x						
	6	Test whether the tool is capable of detecting the defined sublineages						x			
2	7	Build a prototype of the tool for a small (sub)dataset	x	x							
	8	Literature research on clustering methods to define sublineages		x							
	9	Select sublineage specific SNPs			x	x					
	10	Generate sublineage/species specific markers				x					
	11	Filter redundant markers					x				
	12	Implement tool					x	x			
	13	Determine detection limit of the tool with in silico data						x			
	14	Evaluate tool with real mixed infection data							x		
3	15	Find drug resistance associated SNPs in literature			x						
	16	Associate SNPs from task 3 with drug resistance genotypes found in literature				x					
	17	Generate drug resistance markers				x					
	18	Filter redundant markers					x				
	19	Extend tool to detect drug resistance genotypes						x	x		
	20	Evaluate whether drug resistance genotypes are detected in real data								x	
	21	Write thesis								x	x

extension of the tool will be tested on whether it is capable of detecting drug resistance genotypes in our own genome data set. Also, the detection limit will be determined using *in silico* generated samples with both drug susceptible and drug resistant strains. Finally, the performance on real samples with drug resistance genotypes will be evaluated.

10.3 Potential pitfalls and backup plan

A prototype tool to detect mixed infections and drug resistance associations will be implemented prior to the final implementation, that will only be based on a small subset of the MTBC genome data. This will help to prevent mistakes in the final implementation and to learn from implications that arise during the study.

We expect to attain a list of sublineage SNPs to diffferentiate strains of MTBC, whereas the sublineages will be defined with BAPS software. In case BAPS cannot be applied as desired or sublineages cannot be identified by our tool afterwards, sublineages still have to be defined so as to reach the overall aim of this project. Another way of defining sublineages is described by Coll et al.[15] where they identify clades in the SNP-based phylogeny based on *in silico* spoligotypes, using SpolPred software, and RDs deletions, identified by mapping contigs that cover RDs to the reference genome.

The Fisher's exact test may be used to find sublineage specific SNPs, as was done for the development of the tool Macaw, but literature research will give insights into other possibilities. If it turns out that a selected method fails to properly identify sublineage specific SNPs, other methods may be preferred based on information gained from literature.

11 Acknowledgements

I thank Jasper Linthorst for helping in the application of Probubble and to clear out the underlying methodology. Also, I thank Thomas Abeel for supplying the genome files of the MTBC strains that were used for assembly-based variant calling and detecting mixed infections, and for his guidance to write this paper.

References

1. World Health Organization. *Global Tuberculosis Report*. World Health Organization, Geneva, Switzerland, 2014.

2. I. Comas, M. Coscolla, T. Luo, S. Borrell, K. E. Holt, M. Kato-Maeda, J. Parkhill, B. Malla, S. Berg, G. Thwaites, D. Yeboah-Manu, G. Bothamley, J. Mei, L. Wei, S. Bentley, S. R. Harris, S. Niemann, R. Diel, A. Aseffa, Q. Gao, D. Young, and S. Gagneux. Out-of-Africa migration and neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans. *Nat. genet.*, 10:1176–1182, 2013.

3. J. E. Galagan. Genomic insights into tuberculosis. *Nature Rev. Genet.*, 15:307–320, 2014.

4. A. C. Schürch and D. van Soolingen. DNA fingerprinting of *Mycobacterium tuberculosis*: From phage typing to whole-genome sequencing. *Infect., Genet. and Evol.*, 12:602–609, 2012.

5. M. Coscolla and S. Gagneux. Consequences of genomic diversity in *Mycobacterium tuberculosis*. *Seminars in Immunol.*, 26:431–444, 2014.

6. C. Ford, K. Yusim, T. Loerger, S. Feng, M. Chase, M. Green, B. Korber, and S. Fortune. *Mycobacterium tuberculosis* - Heterogeneity revealed through whole genome sequencing. *Tuberculosis*, 92:194–201, 2012.

7. I. Filliol, A. S. Motiwala, M. Cavatore, W. Qi, M. H. Hazbón, M. Bobadilla del Valle, J. Fyfe, L. García-García, N. Rastogi, C. Sola, T. Zozio, M. I. Guerrero, C. I. León, J. Crabtree, S. Angiuoli, K. D. Eisenach, R. Durmaz, M. L. Joloba, A. Rendón, J. Sifuentes-Osornio, A. Ponce de León, M. D. Cave, R. Fleischmann, T. S. Whittam, and D. Alland. Global phylogeny of *Mycobacterium tuberculosis* based on single nucleotide polymorphism (SNP) analysis: Insights into tuberculosis evolution, phylogenetic accuracy of other DNA fingerprinting systems, and recommendations for a minimal standard SNP set. *Journal of Bacteriology*, 188(2):759–772, 2006.

8. J. A. Philips and J. D. Ernst. Tuberculosis pathogenesis and immunity. *Annu. Rev. Pathol. Mech. Dis.*, 7:353–384, 2012.

9. A. O'Garra, P. S. Redford, F. W. McNab, C. I. Bloom, R. J. Wilkinson, and M. P. Berry. The immune response in tuberculosis. *Annual review of immunology*, 31:475–527, 2013.

10. C.-C. Huang, E. T. Tchetgen, M. C. Becerra, T. Cohen, K. C. Hughes, Z. Zhang, R. Calderon, R. Yataco, C. Contreras, J. Galea, et al. The effect of HIV-related immunosuppression on the risk of tuberculosis transmission to household contacts. *Clinical infectious diseases*, 58(6):765–774, 2014.

11. D. F. Warner, A. Koch, and V. Mizrahi. Diversity and disease pathogenesis in *Mycobacterium tuberculosis*. *Trends in Microbiology*, 23:14–21, 2015.

12. N. Casali, V. Nikolayevskyy, Y. Balabanova, S. R. Harris, O. Ignatyeva, I. Kontsevaya, J. Corander, J. Bryant, J. Parkhill, S. Nejentsev, R. D. Horstmann, T. Brown, and F. Drobniewsky. Evolution and trasnmission of drug-resistant tuberculosis in a Russian population. *Nat. genet.*, 46:279–286, 2014.

13. S. Gagneux, K. DeRiemer, T. Van, M. Kato-Maeda, B. C. de Jong, S. Narayanan, M. Nicol, S. Niemann, K. Kremer, M. C. Gutierrez, et al. Variable host-pathogen compatibility in *Mycobacterium tuberculosis*. *Proceedings of the National academy of Sciences of the United States of America*, 103(8):2869–2873, 2006.

14. S. Gagneux and P. M. Small. Global phylogeography of *Mycobacterium tuberculosis* and implications for tuberculosis product development. *The Lancet infectious diseases*, 7(5):328–337, 2007.

15. F. Coll, R. McNerney, J. A. Guerra-Assunção, J. R. Glynn, J. Perdigão, M. Viveiros, I. Portugal, A. Pain, N. Martin, and T. G. Clark. A robust snp barcode for typing mycobacterium tuberculosis complex strains. *Nature communications*, 5, 2014.

16. S. Gagneux. Genetic diversity in *Mycobacterium tuberculosis*. In *Pathogenesis of Mycobacterium tuberculosis and its Interaction with the Host Organism*, pages 1–25. Springer, 2013.

17. A. Koch and R. J. Wilkinson. The road to drug resistance in *Mycobacterium tuberculosis*. *Genome Biology*, 15:520, 2014.

18. T. G. Clark, K. Mallard, F. Coll, M. Preston, S. Assefa, D. Harris, S. Ogwang, F. Mumbowa, B. Kirenga, D. M. OSullivan, et al. Elucidating emergence and transmission of multidrug-resistant tuberculosis in treatment experienced patients by whole genome sequencing. *PloS one*, 8:e83012, 2013.

19. G. Plazzotta, T. Cohen, and C. Colijn. Magnitude and sources of bias in the detection of mixed strain *M. tuberculosis* infection. *Journal of theoretical biology*, 368:67–73, 2015.

20. L. Pérez-Lago, I. Comas, Y. Navarro, F. González-Candelas, M. Herranz, E. Bouza, and D. García-de Viedma. Whole genome sequencing analysis of intrapatient microevolution in *Mycobacterium tuberculosis*: potential impact on the inference of tuberculosis transmission. *Journal of Infectious Diseases*, page jit439, 2013.

21. T. Cohen, P. D. van Helden, D. Wilson, C. Colijn, M. M. McLaughlin, I. Abubakar, and R. M. Warren. Mixed-strain *Mycobacterium tuberculosis* infections and the implications for tuberculosis treatment and control. *Clinical Microbiology Reviews*, 25:708–719, 2012.

22. K. Mallard, R. McNerney, A. C. Crampin, R. Houben, R. Ndlovu, L. Munthali, R. M. Warren, N. French, and J. R. Glynn. Molecular detection of mixed infections of *Mycobacterium tuberculosis* strains in sputum samples from patients in Karonga district, Malawi. *Journal of clinical microbiology*, 48(12):4512–4518, 2010.

23. M. Hanekom, E. M. Streicher, D. Van de Berg, H. Cox, C. McDermid, M. Bosman, N. C. G. van Pittius, T. C. Victor, M. Kidd, D. van Soolingen, et al. Population structure of mixed *Mycobacterium tuberculosis* infection is strain genotype and culture medium dependent. *PloS one*, 8(7):e70178, 2013.

24. T. Jagielski, J. van Ingen, N. Rastogi, J. Dziadek, P. K. Mazur, and J. Bielecki. Current methods in the molecular typing of *Mycobacterium tuberculosis* and other mycobacteria. *BioMed research international*, 2014, 2014.

25. C. U. Köser, S. Niemann, D. K. Summers, and J. A. Archer. Overview of errors in the reference sequence and annotation of *Mycobacterium tuberculosis* H37Rv, and variation amongst its isolates. *Infection, Genetics and Evolution*, 12(4):807–810, 2012.

26. T. Cohen, D. Wilson, K. Wallengren, E. Y. Samuel, and M. Murray. Mixed-strain mycobacterium tuberculosis infections among patients dying in a hospital in kwazulu-natal, south africa. *Journal of clinical microbiology*, 49(1):385–388, 2011.

27. I. Comas, S. Homolka, S. Niemann, and S. Gagneux. Genotyping of genetically monomorphic bacteria: Dna sequencing in mycobacterium tuberculosis highlights the limitations of current methodologies. *PLoS One*, 4(11):e7815, 2009.

28. S. Homolka, M. Projahn, S. Feuerriegel, T. Ubben, R. Diel, U. Nübel, and S. Niemann. High resolution discrimination of clinical mycobacterium tuberculosis complex strains based on single nucleotide polymorphisms. *PLoS One*, 7(7):e39855, 2012.

29. H. Zhang, D. Li, L. Zhao, J. Fleming, N. Lin, T. Wang, Z. Liu, C. Li, N. Galwey, J. Deng, et al. Genome sequencing of 161 *Mycobacterium tuberculosis* isolates from China identifies genes and intergenic regions associated with drug resistance. *Nature genetics*, 45(10):1255–1260, 2013.

30. C. U. Köser, M. J. Ellington, and S. J. Peacock. Whole-genome sequencing to control antimicrobial resistance. *Trends in genetics*, 30(9):401–407, 2014.

31. K. McElroy, T. Thomas, and F. Luciani. Deep sequencing of evolving pathogen populations: applications, errors, and bioinformatic solutions. *Microbial informatics and experimentation*, 4(1):1, 2014.

32. S. Koren and A. M. Phillippy. One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. *Current opinion in microbiology*, 23:110–120, 2015.

33. P. S. Soltis, D. E. Soltis, et al. Applying the bootstrap in phylogeny reconstruction. *Statistical Science*, 18(2):256–267, 2003.

34. A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, et al. The Genome Analysis Toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data. *Genome research*, 20(9):1297–1303, 2010.

35. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, et al. The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009.

36. B. J. Walker, T. Abeel, T. Shea, M. Priest, A. Abouelliel, S. Sakthikumar, C. A. Cuomo, Q. Zeng, J. Wortman, S. K. Young, et al. Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS one*, 9(11):e112963, 2014.

37. J. Linthorst. Probubble: variant calling through the alignment of overlap-based de-novo assembly graphs. Master's thesis, Delft University of Technology and Leiden University, 2014.

38. C. Consortium. Cytoscape: Network data integration, analysis, and visualization in a box, 2001-2015.

39. S. Shekar, Z. X. Yeo, J. C. Wong, M. K. Chan, D. C. Ong, P. Tongyoo, S.-Y. Wong, and A. S. Lee. Detecting novel genetic variants associated with isoniazid-resistant *Mycobacterium tuberculosis*. *PloS one*, 9(7):e102383, 2014.

40. J. Corander, P. Waldmann, and M. J. Sillanpää. Bayesian analysis of genetic differentiation between populations. *Genetics*, 163(1):367–374, 2003.

41. J. Corander, P. Marttinen, J. Sirén, and J. Tang. Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. *BMC bioinformatics*, 9(1):539, 2008.

42. L. Cheng, T. R. Connor, J. Sirén, D. M. Aanensen, and J. Corander. Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. *Molecular biology and evolution*, 30(5):1224–1228, 2013.

Supplementary information

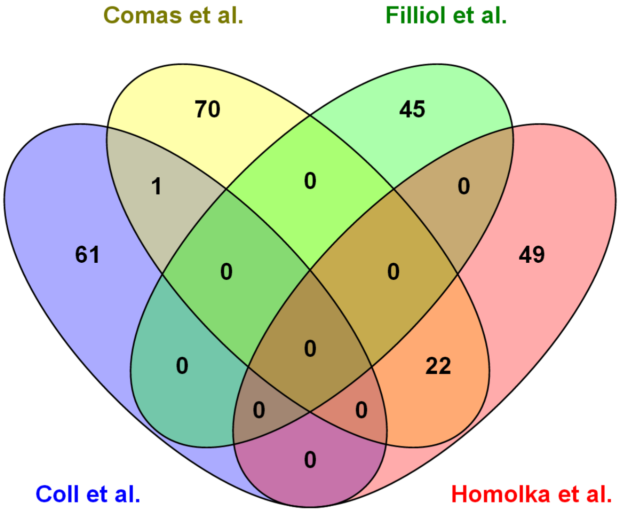

Figure S1: The Venn diagram of minimal SNP sets from four different studies shows that overlap between the sets of SNPs that are selected to be phylogenetically informative is sparse. The SNPs selected by Filliol et al.[7] was shared by none of the other studies, and the minimal SNP sets of Comas et al.[27] and Coll et al.[15] only shared one SNP. The largest concordance was found between the SNP sets of Comas et al.[27] and Homolka et al.[28] sharing 22 SNPs. Overlap of minimal SNP sets was determined from SNP positions in in the reference genome H37Rv (Table S1). Diagram source: http://bioinfogp.cnb.csic.es/tools/venny/.

Table S1: SNP positions in the reference genome H37Rv (version 2) of minimal SNP sets suggested by four different studies.[7,15,27,28] The genomic positions in each study were analyzed and corrected for their position in H37Rv version 2 and this information was used to determine the overlap of the minimal SNP sets with a Venn diagram (Figure S1).

Study	SNP positions					
Coll et al. (2014)[15]	62657	783601	1491275	2505085	3470377	4248115
	107794	797736	1501468	2622402	3479545	4249732
	346693	874787	1502120	2694560	3570528	4260268
	355181	891756	1719757	2831482	3722702	4307886
	403364	931123	1759252	2874344	3836274	4316114
	497491	1084911	1799921	2875883	3836739	4398141
	514245	1132368	1816587	3021283	3977226	4404247
	541048	1137518	1850119	3216553	4125058	
	615614	1237818	1881090	3273107	4151558	
	615938	1455780	1882180	3388166	4229087	
	764995	1487796	2411730	3466426	4246508	
Filliol et al. (2006)[7]	37031	918316	1548149	2376135	3438386	4119246
	43945	923065	1692069	2462871	3440464	4137832
	92199	949221	1692685	2532616	3440542	4254006
	220050	1068151	1884697	2627948	3450725	4255922
	311613	1163134	1892017	2825581	3455686	4280711
	519806	1191861	1952601	2891267	3544710	
	797738	1294398	2158582	2990040	3783058	
	909166	1477598	2223682	3207250	4024273	
Comas et al. (2009)[27]	6112	352058	1043169	2278507	3057309	3597737
	6446	491591	1128814	2280081	3187090	3597682
	7539	491668	1128825	2447426	3187539	3640557
	7585	491742	1128935	2447150	3300479	3769445
	8285	497108	1304443	2603797	3300196	3986987
	8452	495473	1324565	2603736	3300104	3987111
	9143	495322	1461251	2752132	3304966	3987180
	9260	495198	1477588	2752122	3309880	4031202
	9304	495108	1960284	2764939	3309916	4266647
	9566	500223	2223902	2764206	3312728	4352475
	42747	518166	2239055	2952922	3312620	4374228
	157292	558750	2239156	2955233	3314412	4390466
	157129	749176	2239160	2955305	3499247	4393838
	194040	891756	2278426	2955343	3498418	
	194057	1014815	2277350	3004427	3497586	
	351875	1043136	2276918	3057375	3499497	
Homolka et al. (2012)[28]	157292	648756	1129124	2154724	2939716	2955233
	157129	648856	1129128	2223902	2939657	2955305
	156593	648990	1129160	2223896	2939577	2955310
	352058	648992	1129165	2279228	2939373	2955343
	467577	649067	1816848	2279314	2939177	2955957
	491591	649345	2053454	2280081	2940608	2956731
	491668	649446	2053487	2726105	2940562	3986987
	491742	649585	2053682	2752132	2940461	3987050
	497108	649601	2053726	2752122	2941179	3987111
	495473	1128825	2053762	2751866	2940930	3987180
	495322	1128915	2053987	2751764	2955135	3987287
	495108	1128935	2134258	2936944	2955202	