Freie Universität Berlin

Computational Pan-genomics for Detection of Transmission Clusters in Molecular Surveillance with Application in the Epidemiology of Tuberculosis

Dissertation zur Erlangung des Grades
eines Doktors der Naturwissenschaften (Dr. rer. nat.)
am Fachbereich Mathematik und Informatik
der Freien Universität Berlin

von

Christine Jandrasits

Berlin
September 2019

Betreuer: Prof. Dr. Bernhard Renard

Erstgutachter: Prof. Dr. Bernhard Renard
Zweitgutachter: Prof. Dr. Thomas Abeel

Tag der Disputation: 18.11.2019

Abstract

Tuberculosis is a major threat to global health responsible for over a million deaths worldwide every year. It is essential to detect and interrupt transmissions to stop the spread of this infectious disease. With the rising use of next-generation sequencing, its application in the surveillance of *Mycobacterium tuberculosis* has become increasingly important in the last years. The main goal of molecular surveillance is the identification of patient-patient transmission and cluster detection. Whole genome sequencing based base-by-base distance measures have become an integral complement to epidemiological investigation of infectious disease outbreaks. Current approaches are based on single reference sequences and therefore cannot make use of the full diversity of available *M. tuberculosis* genome data and introduce bias towards the chosen reference. Furthermore, they provide inadequate results for comparative analysis of isolates since their resolution is too limited.

In this thesis I present bioinformatic methods for the improvement of molecular surveillance of *M. tuberculosis*. I introduce seq-seq-pan, a framework for adding or removing new genomes from a set of aligned genomes and using these to construct a computational pan-genome. This method is based on sequential whole genome alignment and is optimized for generating a representative linear presentation of the aligned set of genomes, that enables its usage for annotation and in downstream analyses. I present PANPASCO, a pan-genome mapping based distance method that compares high quality variants for each individual pair of samples. It is highly sensitive to differences between cases including variants located in regions of lineage specific reference genomes. This approach allows the comparison of a high number of diverse samples in one analysis. I apply these methods to a large international dataset of drug-resistant *M. tuberculosis* for the detection of transmission clusters. I show their capability of improving surveillance and detection of international transmission and the benefits of including publicly available whole genome sequencing of *M. tuberculosis* for interpretation of national surveillance results. Furthermore, I compare transmission inference methods to answer the important question of 'who infected whom' in *M. tuberculosis* outbreaks.

Acknowledgements

First of all I want to thank my supervisor Bernhard Renard for giving me the opportunity to work in his amazing research group and his continuous advice and support. I am grateful for the freedom given to me for my research and our discussions that were indispensable for my progress. I would like to thank Walter Haas for his supervision and sharing his knowledge about tuberculosis and disease surveillance. Bernhard's and Walter's insight and encouragement were invaluable for my work. Further I would like to thank Thomas Abeel for agreeing to review my thesis. I would like to express my gratitude to all my co-authors for their excellent contributions and feedback. Special thanks to Stefan Kröger for the great team work and support in our collaborations. I want to thank my fellow PhD students and other colleagues for the great atmosphere and working environment. The spirit in our group is unique and exceptional. Many thanks in particular to Tobias, Kathrin, Jakub, Vitor, Martina, Simon, Andreas, Aileen and Elizabeth for inspiring conversations, funny moments and laughs, and extraordinary support and encouragement. I enjoyed spending time with you inside and outside working hours, during coffee breaks, and on conference trips. Many thanks to Pascal and Julius for being awesome students and their great work in our projects. I also want to thank Lei Mao for her perpetual positive attitude and always making me smile. I want to give my warmest thanks to my parents, Inge and Alfred, and my siblings, Edith and Robert, for continuously encouraging me in my path even if it leads me further away. I am thankful for your support and the time we spend together. Last but not least, I am deeply grateful to Claus for his continuous love and encouragement. Thank you for always being there.

Contents

Chapter 1

Introduction

Tuberculosis

Tuberculosis (TB) is an infectious disease caused by the bacterium *M. tuberculosis*, which is transmitted through the respiratory route and manifests most commonly in the lungs (Bloom et al., 2017). Aerosols, a collection of pathogen-laden particles, are formed within an infected person, spread to the air by coughing and may are inhaled by another susceptible person. Exposed individuals either develop a asymptomatic latent infection or an active TB disease (Shiloh, 2016). In most cases the infection remains latent and only 5-10% of infected individuals will develop the disease, with higher rates among immune-compromised patients (Shiloh, 2016; Bloom et al., 2017). Other risk factors for developing active TB disease are co-infection with HIV/AIDS, smoking, indoor air pollution, low body mass index, alcohol use disorder, and diabetes mellitus. Symptoms associated with active TB are: include coughing, chest pains, weakness, weight loss, and fever and in later stages with illness, including wasting and inflammation with tissue damage (Bloom et al., 2017).

Treatment of TB requires taking multiple antibiotics for many months, with prolonged duration in case of drug-resistance. If untreated, the disease shows a 5-year mortality rage of about 60% (Bloom et al., 2017).

Prevention of TB infection mainly consists of vaccinating children and fast detection and treatment of new infections. The only available vaccine is Bacille Calmette–Guérin (BCG), which is effective in children and prevents 20% of infection and 60% of active disease. However, protective effects decrease after 10 years and re-vaccination is not beneficial (Al-Humadi et al., 2017).

The World Health Organization (WHO) reported that in 2015 10.4 million people worldwide fell ill with active TB and 1.7 million died from the disease. This ranks TB among the top 10 causes of death worldwide (WHO, 2018) and makes it the most deadly infectious disease (Koch et al., 2018). Furthermore, about a third of the world's population is estimated to carry a latent infection with TB (Bloom et al., 2017).

Drug Resistance

Antibiotic resistance is a rapidly evolving global health threat, which also holds true for TB. While TB incidences are generally slowly declining (with discouragingly high exceptions in some countries in South-East Asia and Sub-Saharan Africa), an increase in drug-resistant cases has been reported in many countries worldwide (Bloom et al., 2017). The rate of multi-drug resistant tuberculosis (MDR-TB), i.e. cases with resistance against the first-line drugs Rifampicin (RIF) and Isoniazid (INH), has been on the rise over the last decade. An even more dangerous threat developed more recently: extensive-drug resistant tuberculosis (XDR-TB) with resistance against RIF, INH, one Fluroquinolone and one second-line injectable drug (Table 1.1), and totally drug-resistant TB (Dheda et al., 2014). In 2016, almost half a million patients were diagnosed with MDR-TB and 6.2% of those were XDR-TB (Koch et al., 2018).

Category	Drugs
First-line oral drugs	Isoniazid, Pyrazinamide, Ethambutol, Rifampicin, Rifabutin
Fluroquinolones	Levofloxacin, Moxifloxacin, Ofloxacin, Gatifloxacin
Second-line injectable drugs	Kanamycin, Amikacin, Capreomycin, Streptomycin

Table 1.1: First-line, second-line and Fluroquinolone drugs for tuberculosis (TB) (Group 1-3) (Dheda et al., 2014)

Treatment of drug resistant TB takes longer and is more challenging and complex than with susceptible cases. Second-line drugs are more expensive and have more side-effects than the commonly used first-line drugs RIF and INH (Bloom et al., 2017). Together with limited access to expert clinicians, diagnostic tools and availability of second-line drugs, this leads to poor treatment outcome in MDR-TB and XDR-TB patients (Koch et al., 2018).

Lineages of *M. tuberculosis*

Strains of *M. tuberculosis* show remarkable divergence and can be separated into different lineages. Currently, *M. tuberculosis* strains are grouped into seven primary lineages associated with geographically separable human populations (Ford et al., 2013; Wiens et al., 2018) (for geographical distribution see Figure 1.1).

Naming of lineages developed historically and was influenced by early observations of lineage representatives (Gagneux and Small, 2007): 1 - Indo-Oceanic (EAI), 2 -

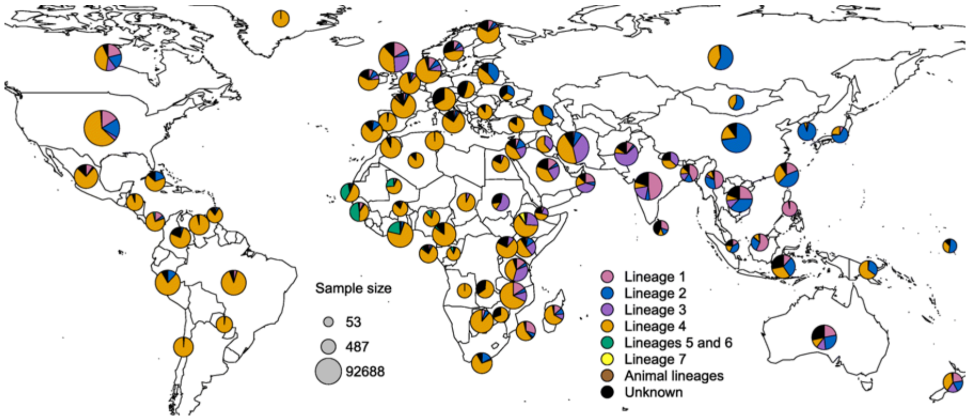

Figure 1.1: **Global spread of *M. tuberculosis* lineages.** Pie charts show the proportion of each lineage where data was available with the radius indicating the amount of available data. (Figure taken from Wiens et al. (2018). Figure is distributed under CC4 `http://creativecommons.org/licenses/by/4.0`)

East-Asian and Beijing, 3 - East-African-Indian (CAS), 4 - Euro-American, 5 - West Africa or *Mycobacterium africanum* I, 6 - West Africa or *M. africanum* II and 7 - Lineage 7 (Ethiopean Lineage). There are three 'ancient' lineages (1, 5, 6) and three 'modern' ones (2, 3, 4) while Lineage 7 seems to be intermediate (Coll et al., 2014). *M. tuberculosis* is part of the *M. tuberculosis* complex (MTBC) which includes species that mainly infect animals such as *Mycobacterium bovis*, *Mycobacterium microti* and *Mycobacterium caprae*. Lineage 2 is a heterogeneous group with subgroups having distinct traits (Merker et al., 2015). Figure 1.2 shows the phylogenetic relationship between all seven primary lineages.

Species of the MTBC have infected and coevolved with humans and animals for around 40,000 years and are probably derived from a collection of ancestral *Mycobacterium protuberculosis* bacteria. Expansion and differential evolution of lineages and species within the MTBC coincides with the expansion and "explosion" of human populations all over the world (Wirth et al., 2008).

Lineages show differing frequency of drug resistance occurrence: Lineage 2 strains are more often reported being drug resistant than Lineage 4 strains and have been shown to have a higher mutation rate (Ford et al., 2013). *M. tuberculosis* lineages are also associated with differential clinical presentation and survival rates (Thwaites et al., 2008) and interact differently with the host's innate immune response and host's genotype (Gagneux and Small, 2007; Caws et al., 2008). Furthermore, the lineages have implications for treatment and control of disease as they spread at different rates due to increased virulence (Gagneux and Small, 2007) and varied response to vaccination prior to infection (Lopez et al., 2003).

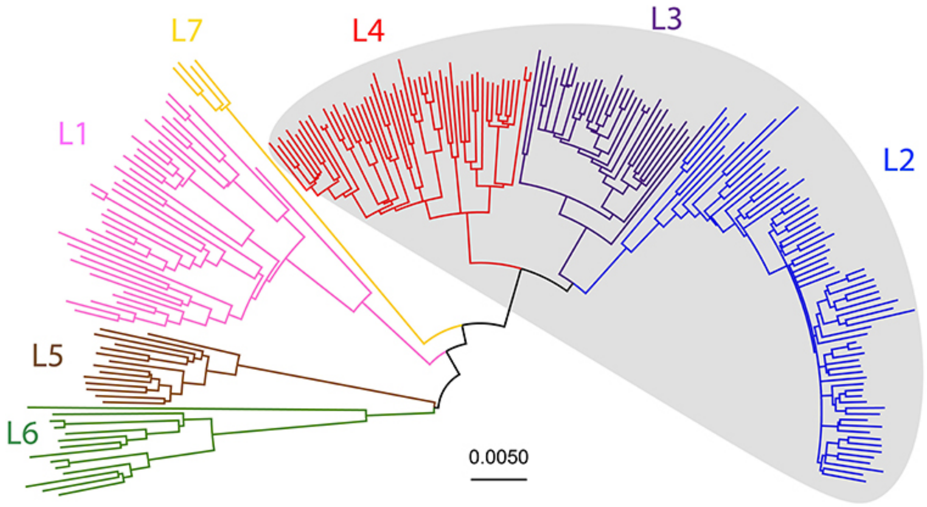

Figure 1.2: **Phylogenetic tree of *M. tuberculosis* lineages (L1-L7).** Lineages are colored and modern lineages marked with grey. (Figure adapted from (Saelens et al., 2019). Figure is distributed under CC4 `http://creativecommons.org/licenses/by/4.0`. Lower part of the original figure was omitted.)

Tuberculosis Control and Surveillance

As transmission and reinfection are key drivers of the global epidemic of TB, transmission control and surveillance has been promoted as effective intervention next to prevention and effective treatment. Transmission control involves the rapid diagnosis and detection of potential infection and drug resistance to promptly start treatment (Bloom et al., 2017). The current WHO recommended strategy of Directly Observed Treatment, Short-course (DOTS) describes treatment of passively found cases and does not appear to be as effective as expected.

The practice of active case finding by contact tracing can play a crucial role in TB control as it reduces the time until detection and treatment of new TB cases and therefore can prevent transmission to further individuals (Shrivastava et al., 2014). Contact tracing involves the identification of relevant contacts, considering the period and proximity of interaction and relationships, and active recruitment of contacts for evaluation. This data is used directly for intervention and potentially initiation of treatment. In the long term surveillance data can be used to understand the epidemiology of the disease to inform on future policies and studies (Begun et al., 2013). Continuous surveillance and modelling of the spread of the disease can help identifying sources of infection and transmission routes, which enables targeted intervention measures for transmission hot-spots and ongoing outbreaks (Jagielski et al., 2014).

While contact tracing currently is the state-of-the-art method for surveillance of TB, it has been shown that acquired information does not always match transmission patterns and many connections are missed (Bjorn-Mortensen et al., 2017). This is because it relies on information on social contacts of patients which can be obscured by recall bias, mobility of patients and reluctance to report contacts. Several molecular typing methods have been developed for the study of TB and are increasingly used in high income countries to improve surveillance (Andrés et al., 2017).

Molecular Typing Methods

Molecular typing methods for infectious diseases aim to support the understanding of the biology and epidemiology of the underlying pathogen, including the definition of infection source and transmission dynamics. For this reason the method of choice must enable the description and discrimination of individual strains. The genome of *M. tuberculosis* is highly homogeneous and conserved among strains. A slow mutation rate of 0.3-0.5 mutations per genome per year was estimated. Additionally, there is no evidence for horizontal gene transfer. These factors render molecular typing and discrimination of strains a challenging task (Jagielski et al., 2014).

Classical genotyping methods that were most frequently used include IS6110 DNA fingerprinting, IS6110 restriction fragment length polymorphism (RFLP), Spoligotyping and mycobacterial interspersed repetitive units variable number of tandem repeats (MIRU-VNTR) (Ei et al., 2016; Meehan et al., 2018). Conventional multilocus sequence typing (MLST), targeting a small number of housekeeping genes at the sequence level, is inefficient for the discriminatory analysis of *M. tuberculosis* typing due to the genome's low degree of sequence polymorphisms (Jagielski et al., 2014).

The *M. tuberculosis* genome includes a high number of repetitive sequences, classified as tandem repeats - repetitive sequences in direct succession - and interspersed repeats scattered across the whole genome. A subclass of interspersed repeats are insertion sequences, among which IS6110 is the most studied one for *M. tuberculosis* (Thierry et al., 1990). The copy number of IS6110 ranges from zero to 25 and the detection of patterns of occurrences within the genome is done with standardized and published methods such as IS6110 DNA finger printing or IS6100 RFLP (Jagielski et al., 2014). However, this method has low discriminatory power in strains with less than six copies of IS6110 which is the case for a large part of lineage 2 strains (Beijing). Additionally the method requires a high amount of high-quality DNA and is technically demanding. Nevertheless IS6110 typing methods are widely used in epidemiological investigations of TB (Ei et al., 2016).

One of the most frequently used polymerase chain reaction (PCR) based molecular typing method for *M. tuberculosis* is spoligotyping. It is based on a single repeat region called the direct repeat (DR) locus. This locus belongs to the family of clustered regularly interspersed short palindromic repeats (CRISPR) and incorporates a 43-spacer set. Individual strains are discriminated by the number of missing spacers of the whole set. This method is fast, simple, cost-effective, and culture-free, which makes it one of the methods of choice for *M. tuberculosis* molecular typing. An international spoligotyping pattern database was established for analysis and comparison of typed probes (Couvin et al., 2019). Spoligotyping has shown weaker discriminatory power than IS6110 typing methods and should therefore only be the first step in epidemiological analyses (Ei et al., 2016). However, it could be used for fast, low-cost sorting of strains into the seven primary lineages (Meehan et al., 2018).

Another type of repeats that is used for *M. tuberculosis* probe differentiation is variable number of tandem repeats (VNTR) with mycobacterial interspersed repetitive units (MIRU) among them. They were initially described as tandem repeats scattered at 41 loci of the *M. tuberculosis* genome with a length of 46-101 base pairs. Their number is determined by the size of PCR-amplicons in relation to the known size of the analyzed repeat unit. While the initially defined 12 locus MIRU-VNTR coding system was less discriminatory than the IS6110 RFLP method, with the extension to more loci, 24 loci MIRU-VNTR became the gold standard method among the molecular typing methods. Still, there are additional tests necessary to discriminate strains of the Beijing lineage (Jagielski et al., 2014).

Other variants such as SNPs and deletions were also considered as a valuable source for information that can be used for typing *M. tuberculosis*. Different methods, including Sanger sequencing (Sanger and Coulson, 1975; Sanger et al., 1977; Homolka et al., 2012) or DNA microarrays (Chee et al., 1996; Salmonière et al., 2004) have been developed, but these methods were superseded by the introduction of whole genome sequencing (WGS). As WGS enabled and influenced many analyses for *M. tuberculosis*, the method is described separately in a following section.

Drug Resistance Diagnosis

Conventionally, drug resistance diagnosis in *M. tuberculosis* has been done with phenotypic culture-based drug susceptibility test (DST) (Koch et al., 2018). This method is based on population growth in culture. Drug resistance is defined on the ability of isolates to grow at or above 'critical' drug concentrations, with those that can grow classified as 'resistant' and those that cannot as 'sensitive'. Different methods are applied that consider resistance ratio or the proportion of resistance (Schön et al., 2017). One of the main problems with these methods is the binary resistance classification of isolates where low levels of resistance can be overlooked and cause extension of treatment (Koch et al., 2018). Additionally, several culture steps are needed and the slow growth of *M. tuberculosis* culture of two to four weeks in average (Ghodbane et al., 2014) renders this process too slow (Koch et al., 2018). Enhancement of culture methods can speed up this process (Ghodbane et al., 2014), however different methods are now preferred over conventional DST (Schön et al., 2017).

In contrast to many bacterial pathogens in *M. tuberculosis* drug resistance is mediated by SNPs rather than horizontal gene transfer (Koch et al., 2018). Several methods have been proposed for molecular assays for drug resistance mutations, among which two are recommended by the WHO: Hain line probe assays (LiPAs) and the Xpert MTB/RIF assay (Schön et al., 2017). Recently a new version of the Xpert MTB/RIF assay was developed that can detect resistance to RIF, INH, fluoroquinolones and aminoglycosides. Two LiPAs are available - one for resistance to INH and RIF and another one for fluroquinolones and the second-line injectable drugs (Table 1.1). The resistance genetics of *M. tuberculosis* are very complex, therefore single molecular diagnostics will not suffice for TB resistance diagnosis and existing methods have to be extended to cover all mutations related to all drugs used for TB treatment (Koch et al., 2018). This challenge can be met using WGS as it can provide information for the majority of mutations related to drug resistance and results can be updated with new data.

Whole Genome Sequencing

WGS allows for the analysis of the whole genome of an organism at once by producing vast amounts of data at low cost (Shendure and Ji, 2008; Mardis, 2008; Wetterstrand, 2019). Many different technologies and platforms have been developed, used and comprehensively described before (Shendure and Ji, 2008; Mardis, 2008; Metzker, 2010; Liu et al., 2012; Tagini and Greub, 2017; Quainoo et al., 2017). In brief, a typical workflow with commonly used WGS *second generation* technologies is as follows: In the first step DNA is digested into several thousands small single stranded DNA fragments. Fragments are bound to a so-called "flowcell" for PCR enrichment to generate DNA fragment clusters. Those clusters are sequenced by synthesis with fluorescent nucleotides and bases are called using the fluorescence information at each step (Metzker, 2010). For detecting small variants, e.g. SNPs, deletions and insertions, reads are typically aligned to a reference genome. Variants are determined from the differences between aligned reads and the reference genome incorporating information about qualities of base calling and alignment (Nielsen et al., 2011).

An alternative approach to alignment and variant detection is assembly. For this, a draft genome is assembled using only the sequence information of the reads. This is done by identifying overlapping reads and combining them to longer continuous sequences (contigs). The result typically comprises a set of contigs, as complete assembly is rarely achieved. This approach is useful in cases were no reference genome is available or to analyse new genomic content, e.g.plasmids, or structural variation. Assembly methods benefit greatly from long reads (several kilobases (kb)) generated by so-called *third generation sequencing* technologies (Carriço et al., 2018).

In recent years WGS based analyses have been successfully used for molecular typing of *M. tuberculosis*. Most WGS based analyses for *M. tuberculosis* focus on SNP analyses and offer the highest discriminatory power in epidemiological investigations (Meehan et al., 2018, 2019). In addition to genotyping, a series of analyses, including drug resistance prediction, lineage differentiation, detection of mixed infection, and distinguishing between relapse and reinfection can be done with WGS. This is the reason for the integration of WGS in surveillance programs of several high-income countries such as Denmark, England, the Netherlands, and the United States (Sanchini et al., 2019).

Lineage Classification and Drug Resistance Prediction

Several studies provide SNP lists for the differentation of *M. tuberculosis* lineages (Comas et al., 2009; Homolka et al., 2012; Stucki et al., 2012; Coll et al., 2014; Feuerriegel et al., 2014; Merker et al., 2015; Lipworth et al., 2019). Among those lists, the one provided by Coll et al. (2014) includes all seven primary lineages of *M. tuberculosis* and was built from a large number of geographically widespread genomes. Therefore, it is widely accepted and used (Coll et al., 2015; Bjorn-Mortensen et al., 2017; Mazariegos-Canellas et al., 2017; Schleusener et al., 2017; Witney et al., 2017; Kohl et al., 2018b; Martin et al., 2018). Other studies aim to extend this list with additional SNPs discriminating sub-lineages of lineage 2 (Merker et al., 2015) or animal strains within the MTBC (Lipworth et al., 2019).

Drug resistance associated mutations of *M. tuberculosis* are extensively researched in an effort to replace phenotypic DSTs for resistance diagnosis. Several lists of drug resistance associated variants have been curated from several hundreds available studies or established from comparative analyses (Koch et al., 2018; Sandgren et al., 2009; Feuerriegel et al., 2015). Patterns indicating the emergence of drug resistance have been identified and can be used to prevent drug resistance in TB patients (Manson et al., 2017).

Various databases, bioinformatic tools and web services for automatically annotating lineages and drug resistance prediction based on SNPs have been developed, e.g. TBDreamDB (Sandgren et al., 2009), KvarQ (Steiner et al., 2014), Mykrobe Predictor TB (Bradley et al., 2015), TB Profiler (Coll et al., 2015), PhyResSe (Feuerriegel et al., 2015), ReSeqTB (Starks et al., 2015) and MTBSeq (Kohl et al., 2018b).

ReSeqTB is a collaborative data-sharing platform recently adopted by the WHO for global surveillance of drug-resistant TB (Critical Path Institute, 2018).

The most commonly used tools - KvarQ, MyKrobe Predictor TB, TB Profiler and Phyresse - have been compared and benchmarked (Schleusener et al., 2017; Ngo and Teo, 2019). Each of them uses raw WGS data as input but applies differing methods of resistance prediction by

- calling mutations directly from sequencing reads without using a reference (KvarQ, Steiner et al. (2014))

- building a de Bruijn graph from resistance information and raw data (Mykrobe Predictor TB, Bradley et al. (2015))

- using only parts of the reference genome (TB Profiler, Coll et al. (2015))

- comparing resistance markers with variants called with a standard WGS workflow (Phyresse, Feuerriegel et al. (2015)).

The applied resistance mutation catalogues differ between the tools, but can be updated with new data with each of them (Schleusener et al., 2017). However, drug resistance prediction does not work equally well for all TB drugs (Ngo and Teo, 2019).

Recurrent Infection and Within-Host Diversity

Recurrent infection describes the event of a person falling ill with TB after being declared cured beforehand. This can have two causes: relapse - the person was not cured, but a undetectable amount of bacteria survived treatment and reinfected the person after some time - or reinfection - the person was infected by another strain after being completely cured from the first infection. Those two cases have differing implication for public health related to selection of drugs and detection of ongoing transmission. Relapse and re-infection can be distinguished by differential analysis of samples taken of both episodes, whereas re-infection with an identical strain can not be differentiated from relapse using genetic information (Hatherell et al., 2016).

Within-host diversity, where differing bacterial populations occur within one person, has two causes: either sub-populations of the infecting strain have evolved (microevolution) or a person was co-infected by two separate strains (mixed infection). Within-host diversity can be identified by analysing mixed basecalls, where differing variants are present in sequencing reads mapped to the same position in the genome. These mixed - or "heterozygous" - SNP calls can be used to separate the differing clones with a sample. Due to the slow mutation rate of *M. tuberculosis*, co-infection and microevolution can be distinguished by the number of heterozygous SNPs, which is much smaller in cases with microevolution (Hatherell et al., 2016). Detecting mixed infection is complicated in cases where infecting strains are very similar or show very different population sizes. This poses a great challenge for treatment drug-resistanc, in cases of undetected drug-resistant co-infecting strains. It also has great implications for transmission analysis and analyzing transmission, as further infections can be caused by the undetected strain (Cohen et al., 2019).

Transmission Analysis

In many studies SNPs analyses were used for differential analyses of *M. tuberculosis* samples. Typically SNPs are called with a standard workflow as described above. Several filters for quality of sequencing reads and detected variations are applied and differential SNPs are counted and compared between samples. Samples with a small SNP distance are grouped into clusters. Cutoffs of 1-12 SNPs are used to infer potential transmission (Pérez-Lago et al., 2013; Roetzer et al., 2013; Walker et al., 2013; Guerra-Assunção et al., 2015; Hatherell et al., 2016; Fiebig et al., 2017). Alternative approaches use identified SNPs in concatenation for phylogenetic analyses (Gurjav et al., 2016; Hatherell et al., 2016). In an effort to standardize WGS based distance analyses, Kohl et al. (2018a) developed a cgMLST scheme for *M. tuberculosis* from 251 genomes including a set of 2891 genes. It was recently updated from a scheme including 3257 genes that was defined on a smaller set of genomes, that did not represent all *M. tuberculosis* lineages well. This method is based on a gene-by-gene allele calling scheme, where individual SNPs within the predefined gene set are transferred into a allele numbering system (Kohl et al., 2018a). Naturally, SNPs located in regions other than this set of genes are excluded from the analysis. Transmission cluster analyses give information about groups of persons that probably infected one another. To determine direction of transmission, often genetic information and epidemiological data, such as the sampling times are combined, but only few methods have been proposed for *M. tuberculosis*. Genetic information is incorporated into the analyses in the form of accumulation of SNPs - higher difference from a reference indicates later infection. However, inference of direction of transmission was defined specifically for analyzed groups of cases only and no standardized method has been proposed yet (Hatherell et al., 2016).

WGS based SNP distance methods offer higher resolution than all of the molecular typing methods described previously (Wyllie et al., 2018), with cgMLST being slightly less discriminatory than SNP distance methods (Kohl et al., 2018a). Meehan et al. (2018) investigated the time period of transmission events that can be detected with spoligotyping, MIRU and WGS based approaches. Small SNP cutoffs can be used to determine transmission events of a few years prior to sampling, whereas with MIRU transmission events were often dated 30 years back and relations found with spoligotyping represented up to 200 years of transmission.

There is room for further improvement of WGS based methods - with the common variant detection workflow sequencing reads are mapped to one reference genome. However, the choice of the reference sequence can greatly influence and bias analysis results (Computational Pan-Genomics Consortium, 2018).

One way to avoid bias in differential analysis and also consider differences between samples in non-reference regions, are reference-free variant calling methods. Most of these methods are based on de Bruijn Graphs that are built from smaller parts – k-mers – of raw sequencing reads (Leggett and MacLean, 2014). Examples of available tools are Cortex (Iqbal et al., 2012, 2013), DIAL (Ratan et al., 2010) and discoSNP++ (Peterlongo et al., 2017). Among those, Cortex offers the most feasible solution as it is fast, memory-efficient and flexible, allowing the integration of raw sequence reads and/or additional annotated sequences (Iqbal et al., 2013; Leggett and MacLean, 2014).

Another way to avoid reference sequence bias is to include not one but several reference sequences into WGS analyses. Combinations of sets of sequences are referred to as pan-genomes. The first definition of a microbial pan-genome included all genes present in all included reference sequences (Tettelin et al., 2005). Later studies and definitions extended this term to include all parts of any included sequence in the pan-genome. The term 'Computational pan-genomics' was introduced and includes the development of efficient data structure and algorithms for the collective analysis of a set of related sequences (Computational Pan-Genomics Consortium, 2018). Many approaches for mapping sequencing reads to a pan-genome are based on a graph representation of reference sequences. Examples for implementations that focus on variant detection using information of multiple reference sequences are MHC-PRG (Dilthey et al., 2015), panVC (Valenzuela et al., 2015) or vg (variation graph, Garrison et al. (2018)).

To date, these approaches are not used in WGS based *M. tuberculosis* surveillance yet, as the interpretation of results and annotation with existing knowledge is challenging.

Thesis outline

In this thesis I present new computational methods for the analysis of *M. tuberculosis* to aid the surveillance of TB. The aim is the improvement of molecular surveillance approaches to make use of the vast amount of publicly available data - in the form of reference genomes, genomic variation annotation and raw sequencing data with metadata. This thesis is compiled of three contributions that are presented in Chapters 2, 3, and 4. I introduce a new pan-genome data structure that can be used for WGS based transmission cluster detection. The presented pan-genome can be united with all SNP based approaches and existing annotation can be fully utilized. I build on prior knowledge for transmission cluster detection and improve the distance measure by including more of the available genomic data from each sample and by using the new pan-genome structure. Furthermore, I show how these new approaches can be used for *M. tuberculosis* surveillance. All contributions were developed with the help and guidance of Bernhard Renard, who participated in method design and conceptualization as well as drafting manuscripts for publication.

Chapter 2 introduces seq-seq-pan (Jandrasits et al., 2018), a computational pan-genome based on whole genome alignment. seq-seq-pan iteratively adds sequences from a pool of reference genome to a pan-genome data structure. All genomic sequences are part of the final pan-genome and can be represented as a linear consensus sequence that can be used in standard WGS workflows as a reference sequence. Positions of SNPs and other small variants with annotations such as drug resistance mutation can be mapped from commonly used *M. tuberculosis* reference genomes to the pan-genome for subsequent analyses. In this contribution I came up with the idea and I designed the concept together with all co-authors. I developed the software with the help of Piotr Woijtek Dabrowski and executed all experiments. I wrote the manuscript with the aid of helpful contributions by all co-authors.

Jandrasits C., Dabrowski P. W., Fuchs S., and Renard B. Y. seq-seq-pan: Building a computational pan-genome data structure on whole genome alignment. *BMC Genomics*, 19(1):47, 2018 `https://doi.org/10.1186/s12864-017-4401-3`

In Chapter 3, I describe PANPASCO, an improved SNP-distance method for the detection of transmission networks. It concerns with the handling of missing information in comparative SNP analyses and introduces a pairwise approach that takes into account all aligned genetic information for each pair individually. Combined with the pan-genome presented in Chapter 2 it improves distance measures between samples that are part of large and diverse datasets. I came up with the idea after fruitful discussions with all co-authors. Stefan Kröger and Walter Haas provided valuable insight into *M. tuberculosis* and TB surveillance. I conceptualized the study and developed PANPASCO. I drafted the manuscript with helpful commentaries from all co-authors.

Jandrasits C., Kröger S., Haas W., and Renard B. Y. Computational Pan-genome Mapping and pairwise SNP-distance improve Detection of Mycobacterium tuberculosis Transmission Clusters. *PLOS Computational Biology.* (in revision)

Chapter 4 addresses the value of including publicly available WGS and meta-data into national surveillance analyses. In this contribution we used PANPASCO to analyse a large set of *M. tuberculosis* samples from all over the world together with a small set of German samples. Using an agglomerative clustering approach designed for detecting clusters of closely related samples, we identified several cross-border transmission networks. We were able to set national samples into relation via samples from different countries. Andrea Sanchini, Stefan Kröger, Walter Haas, Bernhard Renard, and I designed the concept of the study. Lena Fiebig and Marta Andrés were involved in the initial design and analysis process. I set up and used pipelines for analysis and developed most software used in the study with the help of Julius Tembrockhaus. I interpreted and visualized the results of the experiments and provided drug resistance classifications and transmission clusters. Andrea Sanchini collected and aggregated all epidemiological data for both datasets including phenotypic DSTs for the German dataset. He summarized properties of determined clusters and identified cross-border transmission links. Thomas Andreas Kohl and Christian Utpatel provided software for lineage classification, while genomic data and parts of epidemiological data was acquired by the co-authors from the Research Center Borstel. Andrea Sanchini and I wrote the manuscript and Stefan Kröger, Walter Haas, and Bernhard Renard assisted with writing.

Sanchini* A., Jandrasits* C., Tembrockhaus J., Kohl T. A., Utpatel C., Maurer F., Niemann S., Haas W., Renard B. Y., and Kröger S. Improving tuberculosis surveillance by detecting international transmission using publicly available whole genome sequencing data. submission in preparation
*joint first authors

In Chapter 5, I describe a comparison of several methods for inferring 'who infected whom' in four *M. tuberculosis* outbreak datasets. Predictions are evaluated based on available epidemiological data provided for each group of patients. I identified the best performing methods and the minimum amount of epidemiological data necessary for accurate transmission inference. Bernhard Renard and I came up with the concept for this study and wrote the manuscript. I selected methods and datasets, ran the experiments and interpreted results.

Jandrasits C. and Renard B. Y. Inferring transmission chains of tuberculosis from genetic and epidemiological data. manuscript in preparation

Chapter 6 summarizes the contributions of this thesis and will give an outlook for future developments.

Terminology and Abbrevations

Terminology

The term "coverage" is frequently used in WGS based methods and throughout this thesis. However, its meaning is ambigous. There are two potential meanings: Sequencing coverage, closely related to sequencing depth, describes the average number of reads generated for each position on a genome. Desired sequencing depth is defined before sequencing to determine the required number of fragments cycles. Sequencing coverage describes the actual average number of reads that was mapped to a genome. Sequencing depth and coverage depth typically differ from each other. Reasons for this difference are errors in reference sequences, repeats, structural variations and technical errors that complicate mapping of reads to the reference genome. Sequencing coverage for a single position in the genome is defined to be the number of reads mapped to this position. Genome coverage however is defined to be the fraction of regions of a genome where a minimum number of reads was mapped to and describes how well the sequencing results represents the reference genome (or how well the reference genome fits for the analysis of the sequencing experiment). In the context of this thesis the term coverage refers to the coverage depth.

Abbrevations

CRISPR clustered regularly interspersed short palindromic repeats

EEA European Economic Assocation

EU European Union

DNA deoxyribonucleic acid

DOTS Directly Observed Treatment, Short-course

DR direct repeat

DST drug susceptibility test

ECDC European Centre for Disease Prevention and Control

ENA Euroean Nucleotide Archive

INH Isoniazid

kb kilobase

LCB locally collinear block

LiPA line probe assay

MDR-TB multi-drug resistant tuberculosis

MIRU mycobacterial interspersed repetitive units

MIRU-VNTR mycobacterial interspersed repetitive units variable number of tandem repeats

MLST multilocus sequence typing

MTBC *M. tuberculosis* complex

NCBI National Center for Biotechnology

NGS next generation sequencing

PCR polymerase chain reaction

RIF Rifampicin

RFLP restriction fragment length polymorphism

SRA Sequence Read Archive

SNP single nucleotide polymorphism

TB tuberculosis

VNTR variable number of tandem repeats

WGA whole genome alignment

WGS whole genome sequencing

WHO World Health Organization

XDR-TB extensive-drug resistant tuberculosis

Chapter 2

seq-seq-pan: building a computational pan-genome data structure on whole genome alignment

This chapter is based on a published article:
Jandrasits C., Dabrowski P. W., Fuchs S., and Renard B. Y. seq-seq-pan: Building a computational pan-genome data structure on whole genome alignment. *BMC Genomics*, 19(1):47, 2018 `https://doi.org/10.1186/s12864-017-4401-3`

Background

Thanks to the continuous advances in next generation sequencing (NGS) technologies the number of sequenced whole genomes is also continuously increasing. This has led to a 10,000 fold increase in available bacterial genomes over the past 20 years (Land et al., 2015). As a result complete sequence information for many species and phylogenetic clades has become available. The current approach to handle the diversity of sequences within a single population is to define a single reference genome with an accompanying comprehensive catalog of variants and other variable genome elements present within that population (Baier et al., 2016). Unfortunately, this representation is limited, as complex genetic differences such as large deletions, insertions or rearrangements cannot easily be expressed in relation to a single reference genome (Herbig et al., 2012). This presents a significant drawback, since a combined representation of all genomic content of a species or population that captures the full information on similarity and variation between individual genomes is essential (Computational Pan-Genomics Consortium, 2018). Therefore, the more versatile concept of using multiple instead of a single reference genome for common analyses of NGS data is attracting more and more attention.

Initially defined to be the sum of core and dispensable genes of all strains of one bacterial organism (Tettelin et al., 2005), the term pan-genome is now more commonly used to describe any set of associated sequences aiming for a collective analysis. Gathered under a newly evolving field termed computational pan-genomics, several methods for the generation of data structures that can represent a set of multiple sequences have been developed. These data structures generally aim to fulfill the following requirements: (i) easy construction and maintenance, (ii) adding and retrieving of (biological) information, (iii) comparison to other sets of genomes or short or long sequences from individuals, (iv) easy visualization and (v) advanced data storage (Computational Pan-Genomics Consortium, 2018).

We assessed a collection of tools applied for the analysis of multiple sequences (Table 2.1). Many of these tools use graphs to represent the pan-genome and focus on efficiently building and storing that graph (Baier et al., 2016; Beller and Ohlebusch, 2016; Dawson, 2016; Minkin et al., 2016; Sirén, 2017). Some (Schneeberger et al., 2009; Huang et al., 2013; Sirén et al., 2014) focus on subsequent analyses such as mapping reads to the pan-genome, while others (Dilthey et al., 2015; Valenzuela et al., 2015) improve variant detection by using a set of reference sequences instead of a single one. The final category in our collection is made up by tools that introduce a complete data structure and provide methods for the construction, storage, processing and visualization of the pan-genome (Herbig et al., 2012; Rahn et al., 2014; Ernst and Rahmann, 2013; Dilthey et al., 2015; Garrison et al., 2018). Most of these tools depend on information on the (dis-)similarity of genomes from a multiple genome alignment or a reference sequence with an adjoining corresponding set of variants to create a pan-genome. This prerequisite cannot represent structural variants (e.g. large deletions or insertions or rearrangements of sequences) in most cases and has to be obtained via external tools.

While four of the analyzed tools - JST (Rahn et al., 2014), MHC-PRG (Dilthey et al., 2015), PanCake (Ernst and Rahmann, 2013), and vg (Garrison et al., 2018) - provide methods for adding or removing genomes from the pan-genome data structure, only GenomeRing (Herbig et al., 2012), JST (Rahn et al., 2014), panVC (Valenzuela et al., 2015) and vg (Garrison et al., 2018) offer the ability to annotate biological features. This is often caused by the representation of the pan-genome as graphs, for which there is no standard method providing a coordinate system, which severely complicates the use of existing annotation databases and formats. Proposed strategies for such coordinate systems (Rand et al., 2017) do not meet all preferential criteria (spatiality, readability, and backward compatibility) (Computational Pan-Genomics Consortium, 2018). Additionally, new methods for essential analyses such as comparing genetic information of individual samples with a graph of reference sequences have to be developed.

Name	Objective	Input	Visualization	Structural	Functionality		
			of pan-genome	Variants	Update		Possibility to
					add	remove	include annotation
svaha(Dawson, 2016)	Graph construction	reference sequence + variants	external	yes	no	no	no
cdbg(Baier et al., 2016)	Graph construction	multiple reference sequences	external	yes	no	no	no
cdbg_search(Beller and Ohlebusch, 2016)	Graph construction	multiple reference sequences	external	yes	no	no	no
SplitMEM(Marcus et al., 2014)	Graph construction	multiple reference sequences	external	yes	no	no	no
TwoPaCo(Minkin et al., 2016)	Graph construction	multiple reference sequences	external	yes	no	no	no
GCSA2(Sirén, 2017)	Graph indexing	variation graph	no	no	no	no	no
GCSA(Sirén et al., 2014)	Graph indexing Multiple sequence mapping	reference sequence + variants	no	no	no	no	no
BWBBLE(Huang et al., 2013)	Multiple sequence mapping	reference sequence + variants	no	no	no	no	no
GenomeMapper(Schneeberger et al., 2009)	Multiple sequence mapping	reference sequence + variants	no	no	no	no	no
panVC(Valenzuela et al., 2015)	Multiple sequence variant detection	whole genome alignment	external	yes	no	no	yes
MHC-PRG(Dilthey et al., 2015)	Multiple sequence variant detection Pan-genome data structure	multiple sequence alignment AND variants	no	no	yes	no	no
GenomeRing(Herbig et al., 2012)	Pan-genome data structure	whole genome alignment	yes	yes	no	no	yes
JST(Rahn et al., 2014)	Pan-genome data structure	reference sequence + variants	no	yes	yes	yes	yes
vg(Garrison et al., 2018)	Pan-genome data structure	reference sequence + variants OR multiple reference sequences	external	yes	yes*	yes*	yes
PanCake(Ernst and Rahmann, 2013)	Pan-genome data structure	multiple reference sequences AND pairwise alignment	external	yes	yes	no	no
seq-seq-pan	Pan-genome data structure	multiple reference sequences	external	yes	yes	yes	yes

Table 2.1: Comparison of pan-genome tools. We analyzed tools for pan-genome analysis that are available or currently under development. This table lists the corresponding publications or websites. We compared the intended use cases of the tools and the prerequisite data required in order to use them. We evaluated the availability of features needed to work with the pan-genome in subsequent analyses, e.g. updating the set of included genomes. Furthermore, we assessed whether the proposed data structures take into account structural variants and whether it is possible to visualize the resulting pan-genome. * Adding and removing of genomes in vg can be achieved using a combination of several steps.

Another (well-established) representation of sets of genomes is their alignment. Whole genome alignments (WGA) implicitly provide a coordinate system that allows the translation between pan-genome and strain genome position, enabling annotation of the alignment with biological features of the individual genomes. Due to the extensive research on whole genome alignment (Brudno et al., 2003; Blanchette et al., 2004; Kurtz et al., 2004; Darling et al., 2010; Nakato and Gotoh, 2010; Angiuoli and Salzberg, 2011; Di Tommaso et al., 2011; Paten et al., 2011; Kim and Ma, 2013; Sievers and Higgins, 2014), standard formats (eXtended Multi-FastA (Darling, 2015) (XMFA) and Multiple Alignment Format (UCSC Genome Bioinformatics Group, 2017) (MAF)), and methods for processing and visualizing WGA results are available (Shih et al., 2006; Hubisz et al., 2010; Herbig et al., 2012; Kearse et al., 2012; Dutheil et al., 2014; Poliakov et al., 2014). In the field of pan-genomics, whole genome aligners are used for the analysis of a set of closely related non-collinear genomes (e.g. several strains of a bacterial species). These genomes contain large insertions and deletions and also rearrangements and inversions of sequences that have to be detected and aligned properly (Angiuoli and Salzberg, 2011; Paten et al., 2011). Several methods (Mugsy (Angiuoli and Salzberg, 2011), progressiveCactus (Paten et al., 2011), progressiveMauve (Darling et al., 2010) and TBA (Blanchette et al., 2004)) have been developed to meet this challenge.

In summary, WGA structures presently fulfill most of the desirable properties of a pan-genome, but a severe drawback of existing methods is their final, non-updatable alignment result.

We here present seq-seq-pan, a framework that enables the usage of WGA as a pan-genome data structure. We provide methods for adding additional genomes or removing them from a set of aligned sequences and use them to sequentially align whole genome sequences. Throughout the sequential process we take measures to optimize the resulting whole genome alignment and provide a linear representation that can be used in place of a reference genome with established methods for subsequent analyses such as read mapping and variant detection.

Methods

seq-seq-pan Workflow

Overview

The key notion of seq-seq-pan is to use and optimize fast, well established whole genome alignment methods to construct a pan-genome from an *a priori* indefinite set of genomes. For this part, we use progressiveMauve (Darling et al., 2010), a fast whole genome aligner that accurately detects large genome rearrangements. The alignment

result is comprised of a set of blocks of aligned sequences that are internally free from genome rearrangements (referred to as locally collinear blocks (LCBs)). For each LCB we derive a consensus sequence using the concept of majority vote and combine all sequences with delimiter sequences of long stretches of the character 'N'. These delimiters are inserted to prevent alignment of sequences over block borders in the following step, because blocks are not consecutive in all genomes. After alignment of the consensus genome with the subsequent genome in the set, all LCBs stretching over block borders are separated. Unaligned sequences of each genome are analyzed again, to align sequences that are considered to be contextually unrelated (Darling et al., 2010). The resulting LCBs with sequences from one or both genomes are joined to the previously aligned blocks. The complete alignment of all genomes is reconstructed from the current and prevenient alignment. Optimizing measures are taken throughout the workflow to maintain the synteny of the original genomes and avoid accumulation of short, unrelated sequence blocks (Figure 2.1). Repetitive sequences within genomes are not aligned with each other but integrated into the alignment and its linear representation as they appear in the original genomes.
Below we describe all steps of the whole workflow in detail. The details of implementation and consecutive order of individual steps are depicted as well (see Implementation details below).

Consensus genome construction

LCBs are combined into a consensus genome by concatenating the consensus sequence of each block. At each position within the LCB, all aligned sequences are compared and the most abundant base is chosen for the consensus sequence. In case of ties, the base is drawn randomly from the available choices. To prevent alignments across block borders when the consensus genome is used for alignment, we integrate a sequence of 1000 'N' (undefined nucleic acid) between the consensus sequence blocks into the final sequence. In addition to the consensus genome, two accompanying index files are created. One contains the start positions of all delimiter sequences within the consensus genome and therefore enables the reconstruction of the alignment of all genomes from the alignment of the consensus genome with an additional sequence (referred to as "consensus index file"). The second index file contains the coordinates of all gaps of all sequences per block in the consensus genome, improving the performance of the reconstruction step and the mapping of coordinates between genomes. Furthermore, we make note of the sequence identifier and the description of all genomes and chromosomes, as this information is not contained in the final output of progressiveMauve.

Alignment step

When aligning two sequences with progressiveMauve (Darling et al., 2010), the alignment is partitioned in locally collinear blocks to allow for the representation of structural differences such as inversions or translocations. Each resulting block contains parts of either both or one of the genomes locally collinear. They form the basis for the subsequent workflow steps. We chose progressiveMauve for this step as we introduce artificial insertions and rearrangements by representing the genomes as a linear consensus genome, which are accurately resolved by this whole genome aligner.

Merging step

Sequences specific for one of the genomes are also reported as LCBs, but these only contain parts of this one genome. LCBs containing only a single unaligned sequence are typically moved to the end of the alignment file. They are created to ensure collinearity within blocks and can sometimes be of small length. When used in a sequential workflow, it can be advisable to avoid assembling short one-sequence-LCBs and attaching all of them to the end of the consensus genome. We prevent the accumulation of small blocks by merging short one-sequence-LCBs with their neighboring blocks within the genome and realigning consecutive gap stretches (see Realignment step). These short blocks can not only emerge in the alignment step but also result from the resolving step, when a LCB is split at block border positions (see Resolving step).

Realignment step

Alignment is improved by realigning genomes at sites where a gap ends in one sequence and starts in the other (referred to as "consecutive gaps"). We scan through the whole alignment and identify all positions with consecutive gaps. Then we extend the interval to the sequence on both sides of the gap sequences by the length of the longer sequence or up to block borders and align these sequences again.

Resolving step

Aligning a genome with a consensus genome can result in alignments that span the borders of the LCBs making up the consensus genome. We identify these blocks using the consensus index file. Then, we split them at the start and end of the delimiter sequence. If the alignment spans a complete delimiter sequence the separation results in three new blocks: the first and third one contain the aligned sequences of the two genomes. The second one includes only the sequence of the new genome that was aligned to the delimiter sequence. All gaps contained in this block are removed. In cases where the delimiter sequence is matched with a gap sequence only, we discard the complete block.

Alignment of initially unaligned sequences

We take the forward representation of all one-sequence-blocks per genome and sort them. We concatenate the sequences, again integrating stretches of 1000 'N', ending up with one sequence for each of the genomes. These sequences are then aligned using the same process as with the full genomes. Alignment with progressiveMauve, the optional Merging Step and the Realignment Step, are followed by a two-step Resolving and Reconstruction process using each of the initially concatenated sequences as "consensus sequence" (see Figure 2.1 and Figure 2.2). All blocks with newly aligned and unaligned sequences are joined with the initially aligned blocks for the final Reconstruction step.

Reconstruction step

All previous steps result in a set of LCBs that contain parts of the consensus genome, the aligned genome or both. For all LCBs that include consensus genome sequences, we reconstruct the alignment that formed this consensus sequence in the previous workflow iteration. For this we use the coordinate system of the consensus genome and the index file containing delimiter sequence positions. We translate the start and end positions of the consensus sequence in each LCB to their positions within the original genomes. With this information we can extract the bases, gaps and positional information of all sequences and report the complete alignment of all genomes for the current workflow iteration.

Removing a genome

After removing a genome from a pan-genome, gaps that were introduced only for the alignment of the removed genome are cut from the remaining genomes. Adjacent LCBs that are now composed of consecutive regions of the same set of genomes are joined to form one LCB.

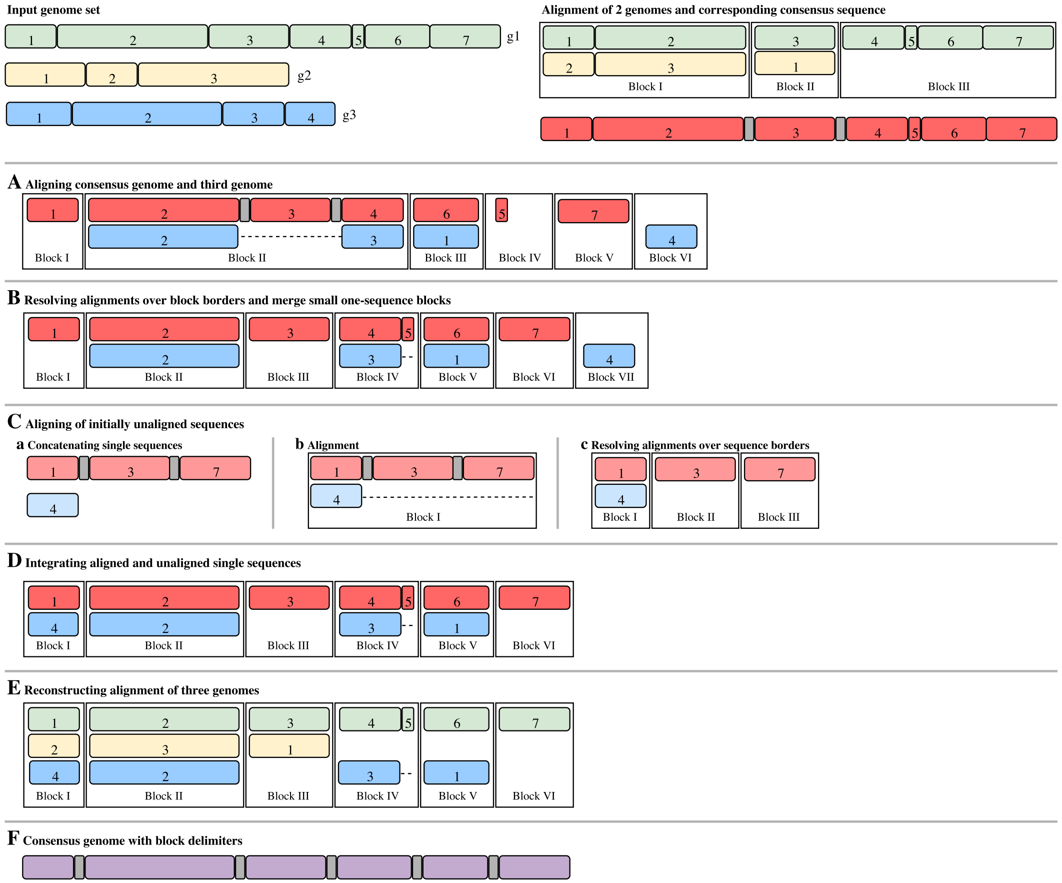
Input genome set
g1
g2
g3
Alignment of 2 genomes and corresponding consensus sequence
Block I
Block II
Block III
A Aligning consensus genome and third genome
Block I
Block II
Block III
Block IV
Block V
Block VI
B Resolving alignments over block borders and merge small one-sequence blocks
Block I
Block II
Block III
Block IV
Block V
Block VI
Block VII
C Aligning of initially unaligned sequences
a Concatenating single sequences
b Alignment
Block I
c Resolving alignments over sequence borders
Block I
Block II
Block III
D Integrating aligned and unaligned single sequences
Block I
Block II
Block III
Block IV
Block V
Block VI
E Reconstructing alignment of three genomes
Block I
Block II
Block III
Block IV
Block V
Block VI
F Consensus genome with block delimiters

Figure 2.1: **Visualization of the alignment workflow for an example with three genomes.** Input genomes (g1-3) are depicted as green, yellow and blue blocks. All sub-sequences are part of locally collinear blocks (LCBs) in the final result and are therefore marked within the whole genomes and numbered according to their appearance in the respective genome. The first two genomes are aligned and provided as separated blocks of aligned sub-sequences. Block I and II indicate a rearrangement of sub-sequence 3 of g1 when compared to g2 and parts of g1 are not present in g2. Consensus sequences are built individually for each LCB in the alignment and concatenated with stretches of 'N' as delimiters to form a consensus genome (depicted in red with delimiters in gray). It is used in the alignment with g3, which is presented in detail in steps A-E. (A) The consensus genome is aligned with the third genome (g3, blue), yielding six blocks. Block I and III represent a rearrangement of sub-sequence 6 of g1. Block II shows a large deletion in g3 compared to the consensus genome. Block IV-VI show single-sequence blocks. (B) Blocks resulting from alignment with the consensus genome are broken up into smaller blocks at delimiter positions (Block II in A is now Block II-VI in B). The small single-sequence block with sub-sequence 5 of the consensus genome (Block IV in A) is merged to its neighboring sub-sequence 4 of the consensus genome, introducing gaps into sub-sequence 3 of g3 (see Block IV in B). (C) Remaining single-sequence blocks of both genomes (depicted in lighter red and blue) are concatenated with stretches of 'N' as delimiters (C.a). Sequences are aligned (C.b) and resulting blocks are resolved at delimiter positions (C.c). Small single-sequences would also be merged to neighboring blocks (not shown). (D) Aligned and single-sequence blocks from step C are joined with initially aligned blocks and all blocks are sorted by their position in the consensus genome. (E) The full alignment is traced back using the newly formed blocks and the alignment of the first two genomes. (F) A consensus genome is built from the full alignment and alignment of additional genomes is achieved by consecutive repetition of steps A-F.

Implementation details

The order of all steps for each iteration of the sequential workflow is depicted in Figure 2.2. For the alignment of pairs of genomes we use progressiveMauve (snapshot 2015-02-13). All other parts of the sequential workflow are implemented in Python3.4. For performance reasons, the consensus construction step was additionally implemented in Java8. We use the following Biopython (version 1.68) modules (Cock et al., 2009) in our workflow: SeqIO, Seq and SeqRecord, for reading, writing and manipulation of single sequences and pairwise2 for aligning two sequences in the Realignment step. For performance reasons we use blat (Kent, 2002) for the alignment of large sequences in the Realignment step. For a straightforward construction of a pan-genome from a set of genomes, we combined all steps with the workflow management software Snakemake (Köster and Rahmann, 2012). The pipeline can be parametrized to include the optional merging steps and all necessary steps are determined automatically in each iteration. The output is composed of the final alignment of all input genomes and the corresponding consensus genome including index files necessary for following analyses and further updates of the pan-genome.

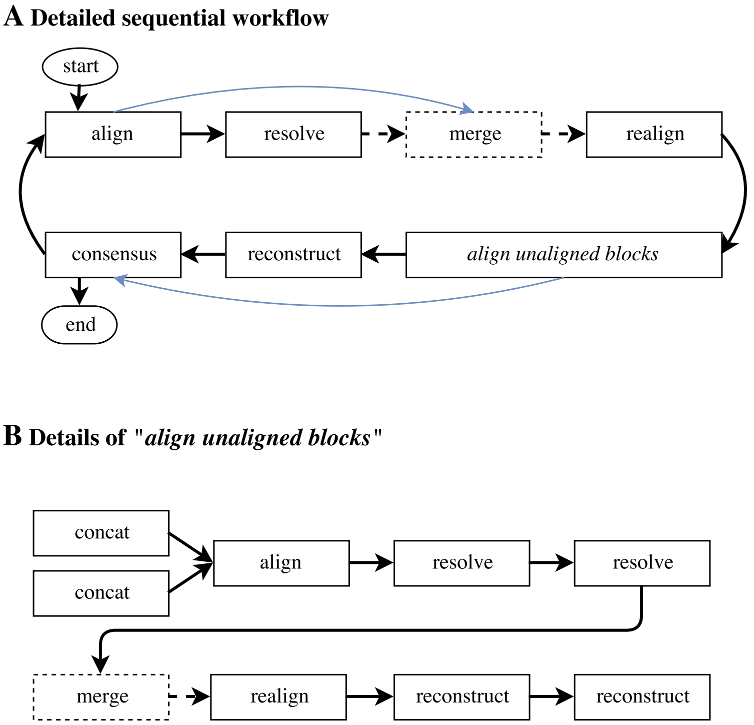

Figure 2.2: **Detailed sequential workflow.** Blocks represent steps in the workflow, dashed ones are optional. The first iteration is different and outlined with the blue lines. Subsequent iterations are represented with black arrows. (A) After aligning two genomes - two original ones in the first iteration and a consensus genome with another original one in all following - the result is optimized with an optional merge step and local realignment of sequences around consecutive gap stretches. Initially unaligned single-sequence blocks are aligned again and all resulting blocks are joined with the blocks resulting from the original alignment. Then the consensus genome is constructed using the optimized alignment. The consensus genome is aligned with the next original genome. Alignments over block borders in the consensus sequence are resolved and the alignment is optimized by merging and realignment. Again, initially unaligned blocks are aligned separately. After joining of all LCBs the full alignment of all genomes is reconstructed. (B) For aligning initially unaligned sequences of each genome, the same methods as for the main alignment are used. All unaligned sequence parts of each genome are sorted and concatenated with 'N' stretches as delimiters. The alignment of these two constructed sequences again results in several LCBs. Alignments stretching over borders of concatenated blocks are resolved successively for each of the genomes. The alignments are optionally optimized with the merging and realignment steps. After that the alignment of the original sequences of both genomes is sequentially reconstructed.

Setup for comparison experiments

Data

We use several sets of reference genomes available in the RefSeq database of the National Center for Biotechnology (NCBI) as of November 30th, 2016 for our experiments. The set of 43 *M. tuberculosis* genomes is used throughout all experiments. To demonstrate the ability of seq-seq-pan to align a large number of genomes, we use the set of all *Staphylococcus aureus* and *Escherichia coli* reference genomes. These sets contain 144 and 207 genomes, respectively. Accuracy of alignment was tested on a simulated dataset of twelve genomes with the genome of *E. coli* K12 as basis for the simulation of evolution. For evaluating the run-time when adding an additional genome to a pan-genome, we used another *M. tuberculosis* genome that became available on December 26^{th}, 2016 (for details on genomes see Appendix Tables 1.1, 1.2, 1.3).

Simulated Data

Accuracy of alignment was tested on a simulated dataset of 13 genomes. For the simulation of genomes with a known true alignment we used the EVOLVER software (Edgar et al., 2006) and the evolverSimControl suite (Earl et al., 2012) as described in the Alignathon project (Earl et al., 2014). The tool evolverSimControl enables the user to simulate several genomes along a phylogeny with EVOLVER. We used an *E. coli* K12 genome (NC_000913.3) as the origin of the evolution simulation. For the evolution parameters we adapted the example provided by the EVOLVER team. We fit the parameters to the smaller size of the *E. coli* genome by changing the probabilities of large insertion and deletion events and setting the maximum size of these events to 7000 – roughly the size of the longest *E. coli* gene. We simulated twelve genomes without using mutation acceptance constraints with the phylogeny depicted in Figure 2.3.

Comparison of alignments

We use an alignment comparison method to compare our results with the results of other whole-genome aligners: the tool mafComparator from the mafTools collection (Earl et al., 2014). To compare the alignments of the simulated dataset with the true alignment, we calculate recall, precision and F-score as described in the Alignathon project (Earl et al., 2014). To use the same method for comparing alignments of the *M. tuberculosis* we choose the alignment of the other aligners to act as the true alignment and our results as the prediction in each comparison. Here, we use the F-score to assess the similarity of the alignments. As in this case there is no ground truth to compare our results to we derive the accuracy of our alignment method from comparison to four other alignment tools.

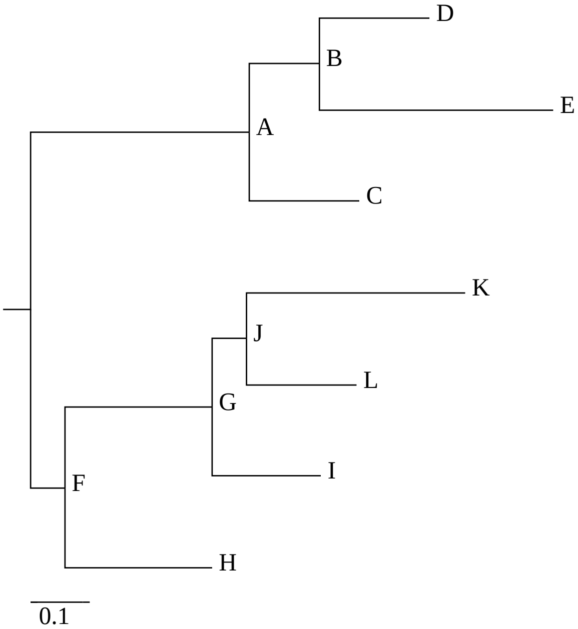

Figure 2.3: **Visualization of the phylogenetic tree used to simulate genomes with EVOLVER.** The corresponding NEWICK tree is (((D:0.015625, E:0.0333) B:0.01, C:0.015625) A:0.03125, (((K:0.03125, L:0.015625) J:0.005, I:0.015625) G:0.02083, H:0.02083) F:0.005);. (drawn with online version of Phylodendron (Gilbert, 1999))

Sorting and Merging

Due to the sequential nature of our workflow, the order in which genomes are added to the pan-genome might influence the resulting alignment. When additional genomes are added to existing alignments, they can not be set in relation to the genomes that are part of the alignment, but have to be added "on top" of them. Therefore, the alignment process should yield similar alignments irrespective of the order of genomes. To investigate the effect of sorting we arrange the genomes by similarity and by consecutive dissimilarity and compare the results. To sort the input genome sequences by similarity we apply the D2z score (Kantorovitz et al., 2007) on all pairs of sequences. The score reflects the sequence similarity, i.e. higher scores stand for more similar sequences. We calculate the upper quartile of all similarity scores and select the genome with the smallest distance from all others. The remaining sequences are ordered by their similarity to this genome.

As an alternative, to obtain a series of strongly differing genomes, we sort them as follows: we again start with the genome with the smallest distance from all others. Then we choose the one with the lowest similarity score as second and the genome most similar to the first for the third position and continue to alternate genomes in this manner throughout the complete set (so the sequence 1, 2, 3, 4, 5, 6 becomes 1, 6, 2, 5, 3, 4).

Additionally we constructed the alignment of the simulated dataset and the *M. tuberculosis* dataset a hundred times and the larger datasets (*S. aureus* and *E. coli*) 10 times with randomly sorted genomes and compare them to the alignment using genomes sorted by similarity. We also compare alignments that were created with and without using the merging steps (See order of genome sets in Appendix Tables 1.4, 1.5, 1.6, 1.7).

Whole genome alignment tools

We compared our sequential genome alignment approach with whole genome alignment tools to review the accuracy of the final alignment. For this, we chose progressiveMauve (Darling et al., 2010), Mugsy (Angiuoli and Salzberg, 2011), progressiveCactus (Paten et al., 2011) and TBA (Blanchette et al., 2004) as these are commonly used methods for WGA that allow aligning non-collinear genomes with large deletions and insertions, inversions and rearrangements. Each of these tools separates the final alignment into LCBs. We parametrized all tools to not report duplications and disabled filters on LCB sizes to fit the results to the methology of seq-seq-pan. In addition to comparing the final alignments, we analyze the time and memory needed to create these results. In cases where no ground truth for the alignment is available, we regard the concordance of the results of all tools.

Pan-genome tools

For comparison, we choose PanCake, which also accepts whole genomes as input and bases the construction of the pan-genome data structure on sequence alignment methods. PanCake represents all genomic sequences in the form of feature instances. Each feature contains part of a genome sequence and start and stop coordinates within the genome. By using the information of pairwise genome alignments, shared features can be extracted. These features contain a single version of the sequence and a list of edit operations and positional information describing all aligned sequences. Following the recommendations by the authors, nucmer (Kurtz et al., 2004) was used for pairwise alignments. We measure the time it takes to construct a pan-genome. Tasks that are part of many analyses of pan-genomes include adding an additional genome or removing a genome and extracting a genomic sequence from the pan-genome. Thus, we examine whether these steps are possible and which run-time they require. To account for differences in time needed to extract genomes based on their position within the pan-genome, we extracted each genome once and calculated the average time.

Results

As we sequentially construct a whole genome alignment, we compare our results with the alignments of progressiveMauve (Darling et al., 2010), Mugsy (Angiuoli and Salzberg, 2011), progressiveCactus (Paten et al., 2011) and TBA (Blanchette et al., 2004) for all datasets. We compare the run-time and memory requirements of pan-genome construction and the provided functional features between seq-seq-pan and PanCake (Ernst and Rahmann, 2013) using the *M. tuberculosis* dataset. We show that the order of genomes has minimal effects on the final alignment and that the merging step produces a less fragmented alignment. For all comparison analyses, we show the results for the whole genome alignment constructed with seq-seq-pan from genomes sorted by similarity using the merging step.

Sorting and Merging

We use 102 different orders for the simulated and the *M. tuberculosis* dataset and 12 different orders for the larger *S. aureus* and *E. coli* datasets and compare the results for all sort orders with the alignment using the genomes sorted by similarity. For these comparisons we use the mafComparator tool from the mafTools suite (Earl et al., 2014) and use the F-score for the assessment of the alignments similarity. As shown in Figure 2.4 the order of genomes has minimal effect on the resulting alignment.

For investigation of the effects of the merging steps we use the simulated and the *M. tuberculosis* dataset with genomes sorted by similarity. Using the sequential workflow without the merging step, results in the alignment of fewer genomes within each LCB and a high number of single-sequence blocks in both datasets (Table 2.2). This indicates that the alignment is more fragmented when small LCBs are not merged to their neighboring blocks. Nevertheless, the F-score comparing the results with and without merging with the truth in the simulated dataset indicates only small differences in the overall alignment (Table 2.2).

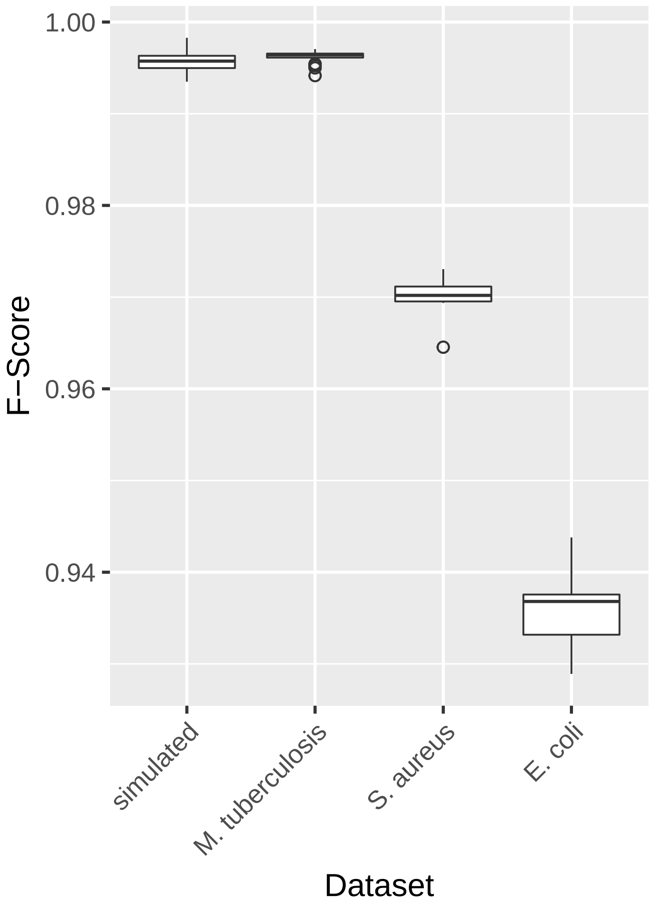

Figure 2.4: **F-scores for comparing alignments using different sort orders for genomes.** Genomes of each dataset were sorted by similarity and dissimilarity and randomly (100 times for the simulated and *M. tuberculosis* datasets and 10 times for the *S. aureus* and *E. coli* datasets and aligned using the sequential workflow. The F-score is used as measure of consistency for alignment when comparing alignments with the dissimilar and random sort orders to the alignment with genomes sorted by similarity. All F-scores were similar within datasets and greater than 0.93 for all comparisons.

	Total Alignment Length	Mean number of Sequences in LCB	Number of short LCBs	Number of short one-sequence LCBs	Precision	Recall	F-Score
Simulated dataset (13 genomes)							
With Merging Step	4809015	9.2	0	0	0.993	0.475	0.643
Without Merging Step	4789770	5.5	318	156	0.993	0.475	0.643
M. tuberculosis dataset (43 genomes)							
With Merging Step	4826979	16.1	0	0	-	-	-
Without Merging Step	4859842	7.5	154	109	-	-	-

Table 2.2: Effect of merging short one-sequence LCBs. We compare the results from sequentially aligning two genome datasets including and excluding the merging step in the workflow. For estimation of the fragmentation of the alignment we compare the total alignment length, the number of sequences per block and the number of small (<10 base pairs) LCBs and focus on the ones containing only sequences from one genome. By comparing the precision, recall and F-score of both alignments compared to the true alignment of the simulated dataset we show that the accuracy of the alignment is not affected by the merging step.

Comparison with whole genome alignment tools

The results of all whole genome aligners and our approach are competitive. In the simulated setting seq-seq-pan achieves similar precision and recall as progressiveMauve and Mugsy and better results than progressiveCactus (Table 2.3). We assessed whether the results of progressiveCactus and Mugsy improved when parametrized to detect duplications. This had almost no effect for Mugsy and improved the comparatively low precision for progressiveCactus, but reduced the recall. Our workflow achieves a precision as high as TBA, but all aligners show a significantly lower recall than TBA. However, comparably low recall values were also observed for simulated datasets used in the publication introducing the comparison method applied here (Earl et al., 2014). For the *M. tuberculosis* dataset, the result of seq-seq-pan is most similar to the alignment by Mugsy and closer to the one from TBA, the most accurate aligner for the simulated dataset, than all other tested aligners. ProgressiveCactus shows the least concordance with all aligners, but all F-scores for comparison between all aligners are greater than 0.9 (Table 2.4).

	Precision	Recall	F-score
TBA	0.993	0.999	0.997
progressiveMauve	0.992	0.477	0.644
seq-seq-pan	0.993	0.475	0.643
Mugsy	0.999	0.474	0.643
Mugsy with duplications	0.999	0.474	0.643
progressiveCactus	0.892	0.473	0.618
progressiveCactus with duplications	0.999	0.339	0.506

Table 2.3: Precision, Recall and F-Score for alignments of the simulated dataset. We compare the results of all alignment tools with the true alignment of the simulated genomes. Aligners are sorted first by F-score and then by Recall.

	seq-seq-pan	Mugsy*	progressive-Mauve	TBA	progressive-Cactus
seq-seq-pan	-	**0.996**	**0.991**	**0.975**	**0.934**
Mugsy*	**0.996**	-	0.990	0.974	**0.934**
progressiveMauve	**0.991**	0.990	-	0.972	0.928
TBA	**0.975**	0.974	0.972	-	0.914
progressiveCactus	**0.934**	**0.934**	0.928	0.914	-

Table 2.4: F-score for pairwise comparison of alignment results for the *M. tuberculosis* dataset. We estimate the similarity of alignments of progressiveMauve, Mugsy, progressiveCactus, seq-seq-pan and TBA, by calculating the pairwise F-score. The aligner with the most similar alignment is shown in bold for each aligner. * Aligning 43 *M. tuberculosis* genomes caused a segmentation fault in Mugsy. We were able to align 39 genomes and therefore compare the results only for this set of sequences.

Table 2.5 shows the high speed up seq-seq-pan achieves compared to the whole genome alignment tools. Seq-seq-pan aligns 13 simulated genomes within 30 minutes and 43 *M. tuberculosis* genomes within two hours - being at least five times faster than all other tools with the real data set. ProgressiveCactus required almost two days for the alignment of 43 genomes and we were unable to align the whole set with Mugsy. It took Mugsy almost 15 hours to align 39 (randomly chosen) genomes. TBA requires pairwise alignments for all genomes in the dataset and builds the alignment on top of these. Table 2.6 shows the cumulative run time for all steps in the alignment workflow. For alignment of the simulated dataset $\binom{13}{2} = 78$ pairwise alignments with a mean run time of 4 minutes 29 seconds were calculated, and for the *M. tuberculosis* dataset, $\binom{43}{2} = 903$ pairwise alignments with a mean run time of 10 hours 15 minutes were required. Of course, depending on the resources available, sets of pairwise analyses can be done in parallel.

	Elapsed wall clock time (hh:mm)	Maximum resident set size (GB)
Simulated dataset (13 genomes)		
seq-seq-pan	00:30	0.77
progressiveMauve	02:33	4.93
Mugsy	01:08	1.01
progressiveCactus	03:41	1.00
TBA	04:59	0.34
M. tuberculosis dataset (43 genomes)		
seq-seq-pan	02:06	1.20
progressiveMauve	09:03	2.79
Mugsy*	14:52	3.26
progressiveCactus	47:09	5.54
TBA	386 days	1.32
S. aureus dataset (144 genomes)		
seq-seq-pan	08:55	4.27
E. coli dataset (207 genomes)		
seq-seq-pan	68:19	8.5

Table 2.5: Run-time and memory usage. We compare seq-seq-pan to other whole genome aligners in terms of run-time and memory usage. Time and memory are indicated for single-threaded processes. Individual steps for TBA can be run in parallel. For the larger datasets (*S.aureus* and *E.coli*) only seq-seq-pan was used for the alignment due to run-time limitation of other tools. * Aligning 43 *M. tuberculosis* genomes caused a segmentation fault in Mugsy. This table lists data for aligning 39 genomes with Mugsy, but the whole set of 43 genomes for all other tools.

The memory requirements during the alignment construction are correlated with the elapsed time in most cases and are therefore lowest for seq-seq-pan, except for TBA. However, memory consumption of TBA will increase with the level of parallelization.

	seq-seq-pan	PanCake	nucmer
Time for construction (hh:mm:ss)	02:06:00	88:10:00	03:04:00
Maximum memory usage	1.20 GB	2.34 GB	0.10 GB
Pan-genome file size	198 MB	36 MB	-
Time to add genome	00:04:01	05:33:52	00:08:48
Mean time for extraction of sequence*	00:00:09	00:01:08	-
Mean time for removing genome**	00:00:19	not available	-
Time for consensus genome creation	00:00:47	not available	-

Table 2.6: Comparison of seq-seq-pan and PanCake. First we compare the run-time and memory usage of pan-genome creation for the set of 43 *M. tuberculosis* genomes. PanCake requires pairwise genome comparisons by nucmer. Run-time and memory requirements for nucmer are listed separately as these can be run in parallel. We also evaluate the file size of the resulting pan-genome. We clock all available features (adding a genome, extracting part of a genome or the whole genome, remove a genome and constructing a consensus genome). * Extraction times for whole genomes and parts of sequences are equal. We extracted the interval 500-1000 for all genomes. ** Each of the 43 genomes was removed from the whole set.

Comparison with pan-genome tools

In addition to the set of reference genomes, PanCake requires pairwise alignments of all genomes to construct a pan-genome. In the case of our experiments with 43 *M. tuberculosis* genomes, the construction of $\binom{43}{2} = 903$ pairwise alignments is required. For our comparison, we calculated these sequentially, but depending on the available hardware, this task can easily be parallelized. For this reason, we list the run-time and memory requirements of pairwise alignments with nucmer (Kurtz et al., 2004) separately (Table 2.6). Constructing the pan-genome with PanCake takes considerably longer than with seq-seq-pan. Also, the extraction of genomes or intervals of genomic sequences takes more time. The resulting pan-genome file from PanCake is smaller in size than the one created with seq-seq- pan. The reason for this difference in size and sequence extraction times is the strategy of PanCake of storing only the differences to a reference genome instead of the whole sequence for all genomes within a shared feature. Removing genomes and the generation of a consensus genome are features that are only provided by seq-seq-pan as listed in Table 2.1. PanCake detects and aligns sequence duplications within genomes and provides methods to compute core regions that are present in all aligned genomes. Arbitrary subsets of sequences can be extracted and singleton sequences that are only present in individual genomes can be identified (Ernst and Rahmann, 2013). Due to our choice to provide the results of seq-seq-pan in standard formats (XMFA, MAF) existing methods can be used for analysis and examination of alignment properties. For example, the maf_parse method of the Phast package (Hubisz et al., 2010) can be used to extract sub-alignments in specific regions or based on feature annotation files.

Discussion of Results

In this contribution, we introduced seq-seq-pan which enhances whole genome alignments by adding critical features for pan-genome data structures e.g. updating the set of genomes within the pan-genome. It provides a fast and simple construction process for whole genome alignments while optimizing the results for usage in subsequent analyses. The continuous merging of small unaligned blocks prevents the accumulation of sequences without context or position within the alignment and preserves the synteny of the original genomes, while the realignment of pairwise alignments avoids the introduction of additional repeats into the linear pan-genome representation. Both steps influence the composition of the linear consensus sequence and support its usage with mapping based methods such as read alignment.

The whole genome alignment format that we use as representation of a pan-genome in seq-seq-pan retains the full sequences and gaps for all aligned genomes in addition to meta-information about block borders. Therefore, it is not suitable to store the pan-genome efficiently. However, this format ensures loss-less and faster handling of the data. Further, it is thereby accessible by currently available downstream analysis tools without requiring subsequent novel tool implementations.

We demonstrate that the sort order of genomes does not substantially influence the result despite the sequential nature of our approach.

We compared seq-seq-pan with four whole genome aligners that offer alignment of non-collinear sequences. These tools use sophisticated methods for the identification of ortho- and even paralogs and conserved sequences. With these features, they identify similar but unrelated sequences within genomes, an aspect that is not considered in the field of pan-genomics. As we do not take such measures, we did not expect very high concordance between our results and the whole genome alignments. However, our comparison shows that our alignment differs as much from the results of progressiveMauve, progressiveCactus, Mugsy and TBA as their results differ among each other. Our approach is able to align a set of genomes much faster and with less memory usage than these whole genome alignment tools. Due to the focus on highly conserved sequences, some of these tools also provide a very fragmented alignment with many small blocks, which is prevented by the merge step in seq-seq-pan.

We compare our approach with currently available methods in terms of applicability and needed prerequisites (input data). For a detailed comparison, we chose PanCake as an approach by which a pan-genome can be constructed from a large set of genomes. We show that the construction of the pan-genome and using the structure for basic tasks requires substantially less time with seq-seq-pan than with PanCake. Some features, such as removing a genome from the pan-genome and the construction of a linear presentation of the pan-genome in the form of a consensus sequence, are not directly available in any other pan-genomics tool. For instance, the authors of

PanCake focused on the analysis of core and accessory gene sets and therefore provide different functionalities.

In the time between November 30th, 2016 and January 20th, 2017 eight new *M. tuberculosis* genomes became available in the NCBI Ref-Seq database. This already highlights the importance of having the ability to extend a pan-genome structure. Methods such as the investigated whole genome alignment tools that constrain the user to start the alignment afresh with the increased number of genomes are at risk of reaching computational limits (some indications could be observed for Mugsy in the experiments already) which is mitigated by our iterative approach which quickly adds new sequences without having to rebuild previously calculated results. Furthermore, publicly available sets of genomes, such as the collection of "Complete Genomes" in the NCBI RefSeq database, are subject to change due to altered quality standards or the redefinition of reference genomes, such as the commonly used *M. tuberculosis* H37Rv strain. Therefore, it is essential that pan-genome representations also provide the feature to easily remove genomes from the initial set without impacting the remaining genomes. Most of the evaluated tools do not provide methods for updating a constructed pan-genome. Particularly research like molecular surveillance, where new data is continuously analyzed and incorporated, depends on data structures that allow the integration of an up-to-date set of genomes.

In summary, we present a data structure for the representation of pan-genomes that provides a unique set of features needed for efficiently working with collections of related sequences and that can be integrated with existing methods for visualization and subsequent analyses.

Chapter 3

Computational Pan-genome Mapping and pairwise SNP-distance improve Detection of *Mycobacterium tuberculosis* Transmission Clusters

Background

Genotyping and sequencing methods have revolutionized infectious disease surveillance. So called molecular surveillance - molecular data in combination with classical epidemiological data - allows the investigation of the transmission of disease within the population and the sensitive detection of outbreaks. The employed methods shifted from fingerprinting, e.g. variable number of tandem repeats (VNTR) methods, and sequence-based genotyping assays, such as bacterial multilocus sequence typing (MLST) to next generation sequencing (NGS) based whole genome sequencing (WGS) in recent years (Struelens and Brisse, 2013). WGS gives access to all genetic information and enables studies on phylogeny, geographical spread of lineages, strain-specific differences, virulence and drug resistance (Wirth et al., 2008; Periwal et al., 2015). WGS allows for the comparison of pathogens on the level of single nucleotide polymorphisms (SNP) and thus is particularly useful for the transmission analysis of stable genomes with low mutation rates such as *Mycobacterium tuberculosis*.

Tuberculosis (TB) is one of the oldest communicable diseases in humankind and can be dated back to 8000 BCE (Frith et al., 2014). Although Robert Koch's characterization of the bacteria is known for more than hundred years, no effective way to eliminate TB has been found and *M. tuberculosis* is still one of the deadliest pathogens worldwide. Between 2000 and 2016 TB caused 53 million deaths. The World Health Organization (WHO) estimates more than 1.7 million deaths (including HIV coinfection) and 10.4 million new TB infections in 2016 (WHO, 2017).

Since available vaccination against TB is merely partly effective, recommended only for children and in high burden settings (WHO, 2018b), breaking transmission chains and successful treatment is needed to prevent new infections and decrease the spread to finally eliminate TB, which is the global goal to reach before 2050 (Uplekar et al., 2015). Here, multi-resistant (MDR-)TB is of special interest with about half a million new cases in 2017. MDR-TB alone is responsible for one third of all antimicrobial resistance deaths worldwide (Stop TB Partnership, 2017) and part of "WHO's Ten threats to global health in 2019" list in the context of antimicrobial resistance (WHO, 2019).

In recent outbreak investigations WGS has been an indispensable tool of outbreak detection and transmission analysis and is replacing genotyping (mycobacterial interspersed repetitive unit - variable number tandem repeat, MIRU-VNTR) as the method of choice in low incident countries such as much of West-Europe (Walker et al., 2013, 2018; Fiebig et al., 2017). WGS can validate participation of individuals to a common transmission event and thus associate TB cases that do not have a clear epidemiological link or vice versa (Ford et al., 2012; Gardy et al., 2011; Walker et al., 2018). With its high resolution WGS outperforms other molecular typing methods as MIRU-VNTR, Spoligotyping or RFLP (Meehan et al., 2018; Wyllie et al., 2018). Furthermore, NGS technology enabled, geographical distribution of different *M. tuberculosis* strains and the dynamics of *M. tuberculosis* evolution (Ford et al., 2012). Strain differentiation was of specific interest for the last decade and several studies showed how WGS outperforms other genotyping methods for detecting recent transmissions (Bryant et al., 2013; Nikolayevskyy et al., 2016) and clusters (Gurjav et al., 2016; Stucki et al., 2016). WGS enables base-by-base comparison between two samples with distinction of identical and non-identical regions. Reference genomes like the laboratory strain H37Rv (Lew et al., 2011) allow for comparison of multiple samples by comparing their sequences to the reference. Over the years a variety of additional reference genomes based on clinical strains with specific drug resistance patterns or specific to certain geographic appearance were created. Thereby, individual reference genomes help to match samples or sample groups like outbreak clusters (Wirth et al., 2008; Niemann et al., 2009; Merker et al., 2015).

Various methods for measuring the distance between samples using single nucleotide polymorphisms (SNPs) obtained with NGS have been proposed. Different strategies for making distance analysis less complex like core-genome MLST (Kohl et al., 2018a) or whole genome MLST (Maiden et al., 2013) have been described. However, these strategies use a gene-by-gene comparison approach.

As the molecular clock of *M. tuberculosis* is considered as very low with 0.3-0.5 mutations per genome per year (Ford et al., 2011; Roetzer et al., 2013), the MLST approaches can be insufficient to investigate transmission patterns in clusters or reconstruct direct transmission links. For this reason, we, as well as many outbreak analyses (Bryant et al., 2013; Pérez-Lago et al., 2013; Roetzer et al., 2013; Walker et al., 2013; Kohl et al., 2014; Guerra-Assunção et al., 2015; Gurjav et al., 2016; Hatherell et al., 2016; Fiebig et al., 2017; Walker et al., 2018), focus on SNP-counting methods as they retain higher resolution for patient-patient transmission. We assessed twelve studies on detection of *M. tuberculosis* transmission clusters (Gardy et al., 2011; Bryant et al., 2013; Kato-Maeda et al., 2013; Pérez-Lago et al., 2013; Roetzer et al., 2013; Walker et al., 2013; Mehaffy et al., 2014; Kohl et al., 2014; Guerra-Assunção et al., 2015; Witney et al., 2015; Gurjav et al., 2016; Fiebig et al., 2017). In nine of them the *M. tuberculosis* strain H37Rv (NC_000962.3)(Lew et al., 2011) is used as the reference genome to identify sample-specific variations. In order to achieve this, samples of patients' bacteria were sequenced and the resulting reads were mapped to the reference genome with commonly used short read aligners. Then, variant calling tools were used to detect single nucleotide polymorphisms in the sample data. For the majority of studies the main steps after variant calling were similar: SNPs were filtered using high quality standards such as a minimum number of reads mapped to the variant site (were e.g. 5-10 reads), with a high percentage of these reads supporting the variant (75%). In most of these studies additional regions of low confidence, such as repetitive regions, known resistance mutations, regions with more than one SNP, insertion or deletion within 12 bp of the SNP are identified and variant calls within these regions were excluded.

However, the described methods varied in how sites with insufficient coverage or low-quality variant calls were handled when comparing all samples in a dataset. There are different reasons for the occurrence of these regions: fluctuation in the coverage while sequencing, deletions within the sequence of the sample or large structural differences between the reference genome and the analyzed sample. Independent of the underlying cause, these regions of low coverage were either completely excluded from the analysis and all comparisons (e.g. (Walker et al., 2013; Fiebig et al., 2017)) by only considering positions with base calls of high quality in all analyzed samples. Or the low-quality base calls and regions with low coverage are ignored and substituted with the reference sequence. This is usually done by selecting SNPs of all samples and concatenating them, using the reference sequence for samples without SNPs at these positions to achieve equal sequence length (Gardy et al., 2011; Mehaffy et al., 2014; Gurjav et al., 2016). Some studies used pairwise comparisons for a small set of samples (Kato-Maeda et al., 2013; Witney et al., 2015). Several of these strategies are facilitated by the BugMat software (Mazariegos-Canellas et al., 2017).

Considering these studies, the questions that are posed for standardizing WGS-based molecular surveillance analyses include: Which genome should the reads be aligned to? How can regions with missing or low-quality information be handled? In alignment-based whole genome sequencing analyses, the choice of the reference genome predetermines the results of subsequent analysis (Lee and Behr, 2016). There is a need to avoid this bias and include multiple reference sequences into the analysis e.g. by using a pan-genome (Computational Pan-Genomics Consortium, 2018). In this context the term pan-genome describes a set of associated (whole genome) sequences rather than the core and accessory genes of all strains of an organism. The method of comparison of samples should consider all high-quality variants while excluding regions of low quality for each sample.

We present a new approach, PANPASCO (PAN-genome based PAirwise SNP COmparison), that combines an improved pairwise distance measure, that allows the comparison and clustering of a large number of diverse samples with the use of a computational pan-genome reference sequence. PANPASCO considers each detectable difference between pairs of samples, without sacrificing the ability to resolve intra-cluster patient-patient relationships.

Methods

PAN - Computational Pan-genome Mapping

In mapping based whole genome sequencing analyses, the choice of the reference genome can have significant impact on the results (Lee and Behr, 2016). For this reason we built a computational pan-genome from 146 *M. tuberculosis* genomes available in NCBI RefSeq by February 17th, 2018 with seq-seq-pan (Jandrasits et al., 2018). seq-seq-pan aligns all genomes in an iterative way, adding new genomic content step by step. This resulted in a computational pan-genome sequence with 5,205,216 bp (an increase of about 18% compared to 4,411,532 bp of the commonly used *M. tuberculosis* H37Rv strain) and contains all genomic regions shared by and specific to each included genome (see list of genomes in Appendix Table 2.1). We use this computational pan-genome sequence as reference sequence with a pipeline that includes various tools for quality control, mapping and variant calling and filtering, with bwa mem (Li, 2013) for read alignment and GATK for variant detection (DePristo et al.,

2011) (see below and Figure 3.2). Scripts for the whole analysis workflow are provided at `https://gitlab.com/rki_bioinformatics/panpasco`.

PASCO - Pairwise SNP Comparison

The first step of distance calculation is the identification of high-quality SNPs. For this we use several filters to identify regions with low coverage and low-quality and ambiguous sites for all samples. Additionally, SNPs in repetitive regions of the reference genome (Comas et al., 2010) are excluded from the analysis of real datasets (see below, in Figure 3.2 and Appendix Table 2.2). Then, we compare all samples *pairwise*, taking into account the set of all variant sites of high-quality (S) in the genomes of a pair of samples (x_1 and x_2) compared to a reference genome. This way we do not lose information about differing bases that are located in low-quality regions of other, unrelated samples. Opposed to the previously published exclusion and substitution methods, we end up with different numbers of sites for each comparison, due to differing number and length of low-quality regions in each sample. To account for this difference we normalize the SNP count by this number of compared sites. This score reflects the SNP difference per base.

We also determine common reference genome sites of the samples (G). For this we compare the low-quality regions of the samples with the whole genome alignment (WGA) that forms the computational pan-genome sequence. The common reference genome sites are composed of high-quality sites of the samples and the low-quality sites that do not overlap with gaps in the WGA of the computational pan-genome. Overlaps with gaps in the WGA indicate that the reason for lack of coverage are strain differences rather than low-quality sequencing (Figure 3.1).

To calculate the expected number of differences for the whole reference genome we multiply the SNP difference per base with the number of common reference genome sites (see Eq (3.1) and Eq (3.2)).

We define the distance between the genomes of a pair of samples x_1 and x_2 as

$$\frac{\sum_{i \epsilon S}^{i} d(x_{1,i}, x_{2,i})}{|S|} \times |G| \tag{3.1}$$

where

$$d(a,b) = \begin{cases} 1 & a \neq b \\ 0 & a = b \end{cases} \tag{3.2}$$

and $|S|$ and $|G|$ are the number of compared high-quality variant sites and the number of common reference genome sites, respectively.

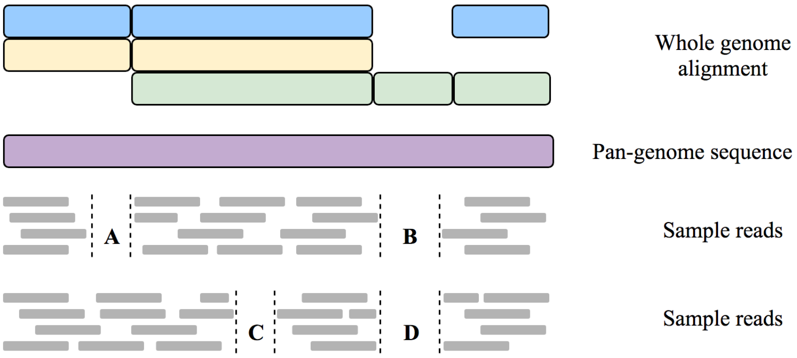

Figure 3.1: **Reads from two samples mapped to a computational pan-genome sequence with regions of zero coverage.** Regions with no coverage such as A and C are considered to contain as many difference between the samples as found in regions with sufficient coverage. To account for these regions with insufficient coverage the total expected difference between two samples is calculated using the SNP difference per base - derived from regions covered in both samples - and the set of common reference genome sites. This set is composed of all sites of the genome except regions such as B and D. These regions have low coverage in both samples and overlap with gaps in the whole genome alignment (blue, yellow and green) of the strains used to build the computational pan-genome (purple). This indicates that both samples are related to similar strains that both do not contain this specific genomic region, which should therefore not be considered when calculating the expected number of differences for the whole computational pan-genome.

Read mapping and variant detection workflow

We implemented a workflow for read mapping, variant discovery and detection of low-quality regions for paired-end next generation sequencing samples. It is depicted in Figure 3.2.

In this workflow we start with preparing the reads by removing sequence adapters from their ends with Trimmomatic (Bolger et al., 2014) and merging paired-end reads in case they overlap with at least 10 bp with Flash (Magoč and Salzberg, 2011). These steps are necessary in cases where the sequenced DNA fragments are shorter than twice the desired read length. This results in two sets of reads: non-overlapping paired-end reads and merged, longer single reads. Both sets are subjected to quality control with Trimmomatic (Bolger et al., 2014), where reads with bases of low quality are trimmed. In case reads are now shorter than 50 bp they are completely removed from the dataset. Remaining mates of excluded reads are added to the set of single reads.

Each of the sets is mapped to a reference genome with bwa mem (Li, 2013). This aligner performs a local alignment and can detect multiple hits for each reads. This feature is especially useful for aligning to bacterial genomes in the presence of rearrangements and aligning to the linear representation of the computational pan-genome where reference sequences might be interrupted due to the block-wise whole genome alignment that seq-seq-pan is based on(Jandrasits et al., 2018). The sets of mapped paired-end and single reads are joined for the following analysis with samtools (Li et al., 2009). Duplicated reads are marked and read groups added with picard tools (Broad Institute, Accessed: 2018-02-21) in preparation for variant detection. Reads with a mapping quality less than 10 are not considered in the following analysis steps.

We use the Genome Analysis Toolkit (GATK, (DePristo et al., 2011)) for variant detection. First we detect confidence scores for all sites, including reference sites with the HaplotypeCaller tool. After that we extract genotypes for each site using the GenotypeGVCFs tool. We call variants with a diploid model to be able to filter mixed base calls by allele frequency. We analyze all sites and separate them in variant sites, positions with uncalled genotypes and high-quality reference sites. Variant sites are then split into single nucleotide polymorphisms (SNPs), small deletions and insertions and structural variants with SelectVariants tool from GATK. We identify SNPs with an allele frequency of at least 75% and where 10 or more reads where used to call the SNP. We also use bedtools (Quinlan and Hall, 2010) to extract regions with less than 10 reads coverage.

By using these filters we separate the data into high-quality SNPs and five sets of low-quality regions with the following criteria:

- positions with less than 10 mapped reads
- positions at deletions
- positions with uncalled genotypes
- SNPs called from less than 10 reads
- SNPs with less than 75% allele frequency

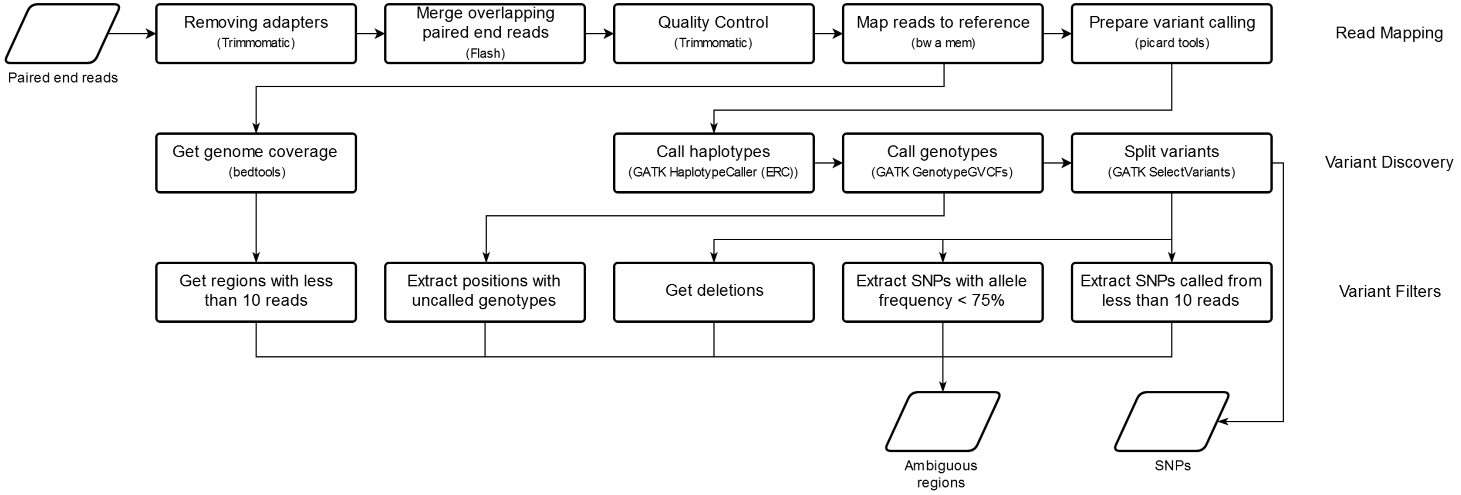
Paired end reads
Removing adapters (Trimmomatic)
Merge overlapping paired end reads (Flash)
Quality Control (Trimmomatic)
Map reads to reference (bw a mem)
Prepare variant calling (picard tools)
Read Mapping
Get genome coverage (bedtools)
Call haplotypes (GATK HaplotypeCaller (ERC))
Call genotypes (GATK GenotypeGVCFs)
Split variants (GATK SelectVariants)
Variant Discovery
Get regions with less than 10 reads
Extract positions with uncalled genotypes
Get deletions
Extract SNPs with allele frequency < 75%
Extract SNPs called from less than 10 reads
Variant Filters
Ambiguous regions
SNPs

Figure 3.2: **Workflow used to analyze the simulation and real datasets.** We divide the task in three parts: read mapping, variant calling and filtering of variants and detection of low-quality regions. We prepare the reads by removing sequence adapters and merging overlapping paired end reads. After that low-quality reads are filtered and reads with ends of low base quality are trimmed. High-quality reads are mapped to a reference and necessary steps for variant calling are taken. We use GATK (DePristo et al., 2011) for variant calling and bedtools (Quinlan and Hall, 2010) to calculate the coverage of the genome. The results are filtered to detect regions of low-quality and high-quality SNPs.

Definition of repetitive regions for the computational pan-genome

We use the list of repetitive regions that were provided in (Comas et al., 2010). Those regions were annotated on the *M. tuberculosis* H37Rv strain, so we mapped the start and end positions to the respective positions on the computational pan-genome with the map function of seq-seq-pan (Jandrasits et al., 2018). We switched the start and end positions in case a region was aligned as reverse complement in the whole genome alignment that the pan-genome is based on. We examined whether the positions were part of consecutively aligned blocks of the whole genome alignment. As this was the case - in fact most regions were mapped to a single block - we used those regions as is to exclude SNPs from the distance calculation. Excluded regions are provided in Appendix Table 2.2.

Simulation dataset

Following published real datasets (UKTB and RAGTB) we simulated 20 transmission clusters with 3 to 55 samples per cluster, resulting in a total number of 323 samples. The simulation dataset was set up to include several transmission clusters with varying numbers of samples based on different strain genomes, to reflect the properties of a real dataset. We chose four *M. tuberculosis* genomes and assigned five clusters to each of them (Table 3.1). For the genomes we chose the *M. tuberculosis* H37Rv strain, commonly used as reference genome, and three increasingly divergent genomes from the set of genomes that built up the computational pan-genome (NC_000962.3, NZ_CP023628.1, NZ_CP002871.1, NZ_CP017920.1). For the comparison of the genomes we used the whole genome alignment of the 146 *M.tuberculosis* genomes that make up our computational pan-genome.

We simulated the 20 transmission clusters with cluster sizes between 3 and 55. We used a beta distribution in R (v3.3, (R Core Team, 2014)) with non-negative parameters set to 2 and 9 for a right-skewed distribution of randomly chosen cluster sizes. For each cluster we generated intra- and inter-cluster SNPs and we assigned all inter-cluster and a selection of the intra-cluster SNPs to each sample. Number of inter-cluster SNPs were determined for each cluster by random uniform sampling between 6 and 40. A minimum of 6 was chosen so that the distance between two clusters is at least 12 SNPs - the cutoff we chose to separate transmission clusters. The number of intra-cluster SNPs is 1.5 times the cluster size and at least 11 (see cluster sizes and number of inter- and intra-cluster SNPs in Table 3.1). We sampled the positions for all SNPs on the respective genomes between 1 and the genome length without replacement. Then, we assigned inter-cluster SNPs to each sample within the respective cluster. We chose 1 to 11 intra-cluster SNPs for each sample from the set of intra-cluster SNPs assigned to each cluster.

Cluster	*M. tuberculosis* strain	number of samples	number of inter-cluster SNPs	number of intra-cluster SNPs
C1	H37Rv	19	36	29
C2	H37Rv	7	29	11
C3	H37Rv	8	32	12
C4	H37Rv	31	13	47
C5	H37Rv	9	6	14
C6	MDRMA2082	18	30	27
C7	MDRMA2082	30	31	45
C8	MDRMA2082	11	15	17
C9	MDRMA2082	11	20	17
C10	MDRMA2082	8	14	12
C11	HKBS1	4	40	11
C12	HKBS1	3	37	11
C13	HKBS1	16	27	24
C14	HKBS1	20	27	30
C15	HKBS1	22	40	33
C16	TB282	15	16	23
C17	TB282	4	35	11
C18	TB282	51	27	77
C19	TB282	11	31	17
C20	TB282	25	31	38

Table 3.1: **Description of clusters in the simulation dataset.**

We estimated the number and length of low coverage regions in the comprehensive dataset of (Walker et al., 2013). We created a list of all lengths of these regions, including duplicates. For each sample we chose the lengths for 70 to 550 such regions from this list and uniformly sampled their positions on the respective genomes. These regions were introduced into the dataset in the form of deletions.

This data was used to simulate short reads for each sample with NEAT (Stephens et al., 2016). For this purpose we created a variant calling file (VCF) with all assigned SNPs and deletions for each sample to simulate mismatches and regions with low coverage. As NEAT cannot handle ambiguous bases within the simulation VCF we replaced them with 'A' in all genomes. As they were only part of deletions this did not influence the simulated short reads. We randomly chose a replacement nucleotide from the set of A, C, T, G for each SNP position.

All read data for the simulated dataset can be downloaded at `https://doi.org/10.5281/zenodo.1346307`.

To create a fair simulated dataset we generated a whole genome alignment of all applied reference genomes with seq-seq-pan (Jandrasits et al., 2018) to calculate the true distance between the samples. To account for the differences between the genomes we used for the simulation, we generated a whole genome alignment of the four genomes with seq-seq-pan (Jandrasits et al., 2018) and counted the number of unequal bases in the alignment (Table 3.2). We combined these differences and all simulated SNPs into a distance matrix for all samples, which represents the true distance between all samples.

	H37Rv	MDRMA2082	HKBS1	TB282
H37Rv	0	1046	2321	2508
MDRMA2082	1279	0	2273	2461
HKBS1	104466	104367	0	408
TB282	132850	132766	41429	0

Table 3.2: **Description of genomes used for the simulation dataset.** Base differences were counted in a whole-genome alignment of the four genomes. Upper triangular part of table shows hamming distance of sequences in WGA ignoring gaps, while the lower part of the table lists all differences, including gaps.

We combined these differences and all simulated SNPs into a distance matrix for all samples, which represents the true distance between all samples. We assessed the number of simulated inter- and intra-cluster SNPs located in regions that are not part of the H37Rv strain by mapping their positions using the coordinate system of the pan-genome. Several SNPs simulated on the three genomes are not part of the H37Rv strain and therefore can not be detected when using this strain as reference genome (Table 3.3).

	inter-cluster SNPs	intra-cluster SNPs
MDRMA2082	0	1
HKBS1	4	3
TB282	8	9

Table 3.3: **Comparison of genomes used in the simulation dataset to the *M. tuberculosis* strain.** We count the simulated SNPs in all simulated samples located on genome specific regions that are not part of the H37Rv genome.

Scripts for generation of intra- and intercluster SNPs and deletions are provided at `https://gitlab.com/rki_bioinformatics/panpasco`.

Results

We developed PANPASCO, a novel method to determine the distance between samples based on SNP differences. We compare samples in a pairwise manner, considering all variant sites of high-quality for each pair. These high-quality sites are identified using an NGS variant calling workflow and a five step variant quality filter. To enable the comparison of pairs of samples with differing amount of missing data, the number of low-quality sites and regions with missing information is also incorporated into the distance measure in a normalization step. To minimize the information loss and avoid the problem of identifying each best fitting reference genome per sample, PANPASCO uses a computational pan-genome built from 146 *M. tuberculosis* genomes with seq-seq-pan (Jandrasits et al., 2018). This computational pan-genome is about 18% (<1 Mb) longer than a single *M. tuberculosis* genome, e.g. the H37Rv strain, and contains all genomic regions shared by and specific to each included genome. We use the computational pan-genome sequence in place of a lineage specific reference genome in our mapping and variant calling workflow. When comparing samples using their SNP difference, we can therefore also include SNPs that occur in regions that are not part of commonly used reference strains (for details see Methods).

Several studies have investigated the mutation rate for *M. tuberculosis* and evaluated the SNP-based difference cutoffs for detecting transmission between patients and within clusters ranging from 3 to 14 SNPs (Roetzer et al., 2013; Walker et al., 2013, 2015; Pérez-Lago et al., 2013; Guerra-Assunção et al., 2015). Here, we chose a conservative definition that assigns samples with a distance of fewer than 13 SNPs into transmission clusters and thereby distinguishes them from unrelated samples (Walker et al., 2013).

We compared PANPASCO to two other commonly used strategies for detecting SNP-based difference, where variant sites and regions with missing information in one of the samples are either completely excluded from the analysis (Walker et al., 2013; Fiebig et al., 2017) or are ignored and therefore substituted with the reference sequence when samples are compared (Gurjav et al., 2016) (referred to as exclusion method and substitution method). For these two methods we use the *M. tuberculosis* H37Rv strain as reference genome as described in the respective publications (Walker et al., 2013; Gurjav et al., 2016; Fiebig et al., 2017). Figure 3.3 shows how these methods work in detail and their individual characteristics: SNP differences in pairs of samples are missed if they are located in regions with missing information in unrelated samples with the exclusion method, while artificial differences are introduced by substituting missing data with the reference sequence when using the substitution method. Figure 3.3 also shows how all differences between pairs of samples are detected and incorporated in the distance measure with PANPASCO.

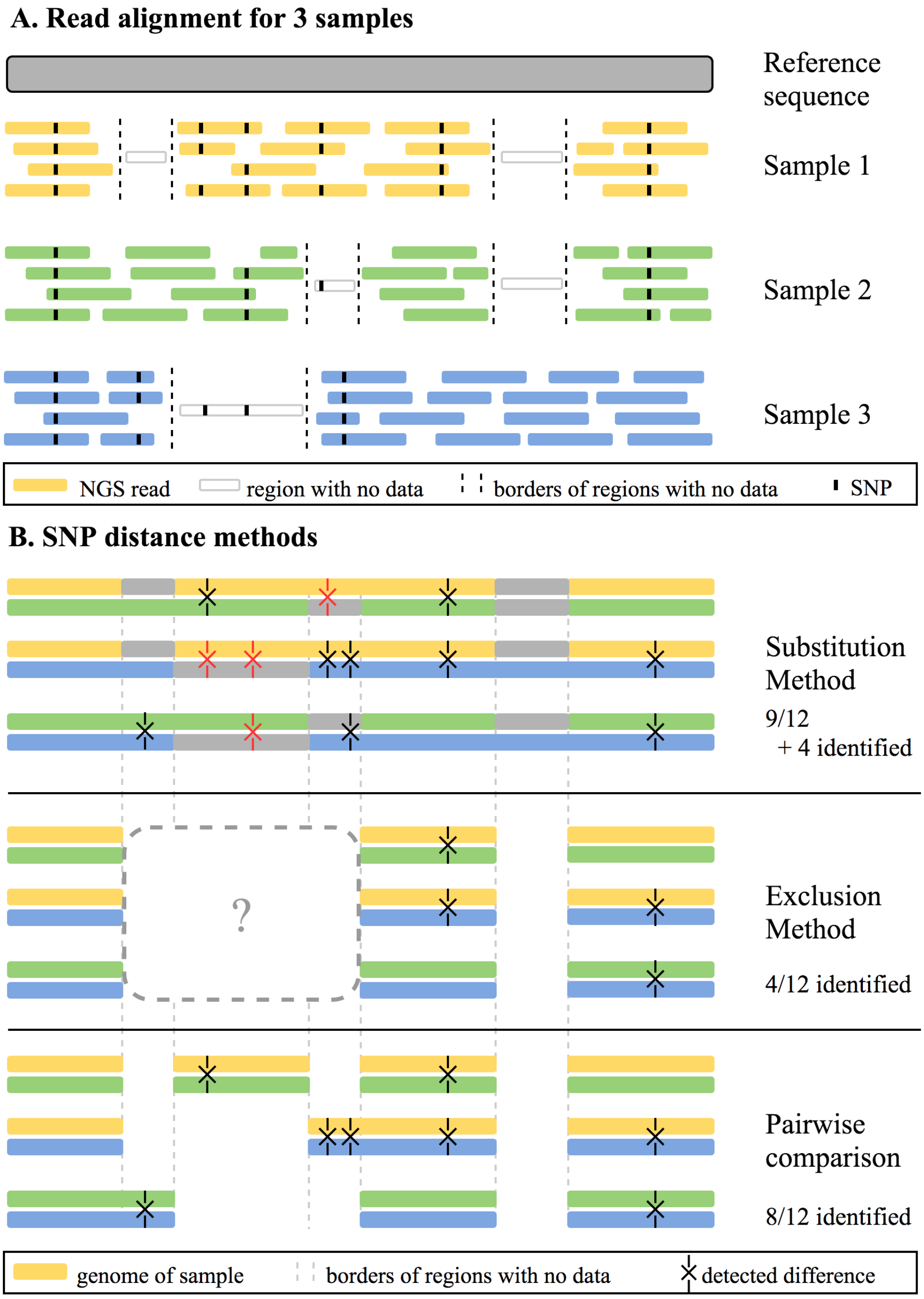
A. Read alignment for 3 samples
Reference sequence
Sample 1
Sample 2
Sample 3
NGS read
region with no data
borders of regions with no data
SNP
B. SNP distance methods
Substitution Method
9/12
+ 4 identified
Exclusion Method
4/12 identified
?
Pairwise comparison
8/12 identified
genome of sample
borders of regions with no data
detected difference

Figure 3.3: **Description and comparison of three SNP distance methods.** (A) Schematic representation of the next-generation sequencing (NGS) reads of three samples aligned to a reference sequence. Reference sequence is depicted in gray, while samples are colored yellow, green and blue. Dashed lines represent borders of regions of the samples that are not covered by any NGS reads. Transparent reads represent the true genome of the samples in regions with no coverage. Reads reveal single nucleotide polymorphisms (SNPs) when comparing the samples to the reference sequence (depicted as black points on the reads). SNPs also occur in sequences of samples where no read data is available, these can not be detected using any method that is based on these aligned reads. In sum, the samples have 12 differences, while only 8 can be identified with the available reads. (B) Representation of compared sequence parts using three different distance measuring methods. Each method compares pairs of samples. Samples are depicted as whole genomes in yellow, blue, and green. Differences between samples are marked with X, true differences are colored black and incorrect ones are colored red. The three methods differ in how regions with missing information in one sample are treated when comparing other samples of the dataset. With the substitution method, missing parts in the samples are replaced with the corresponding parts of the reference sequence (in gray). This leads to incorrectly identified differences, where the true sequence of the samples differs from the reference sequence but has no read coverage. Using the exclusion method, all regions with no coverage of all samples are excluded in all comparisons, e.g. low coverage in Sample 1 influences the comparison of Samples 2 and 3. This leads to overlooking differences between samples, in this example only 4 of 12 detectable differences were identified. The pairwise SNP comparison method used in PANPASCO determines all comparable regions for each pair of samples. Regions with low coverage in each pair are excluded for pairwise comparison. This way all detectable differences are identified, without introducing additional, incorrect ones.

Experimental setup

To evaluate the different SNP-counting strategies, we compare the number of transmission cluster links identified by each method in four different datasets.

We created a simulation dataset with the properties (e.g. number of mutations and coverage distribution) of real datasets (Brudey et al., 2006; Walker et al., 2013; Yang et al., 2017) . It includes 20 transmission clusters with 3 to 55 samples per cluster. Samples were simulated from four different genomes including the commonly used *M. tuberculosis* reference strain H37Rv, with five clusters for each genome (for details see Methods). With this simulation we compare the accuracy of the classification of sample links of the different methods.

We also evaluated PANPASCO's performance for the analysis of three published datasets, to demonstrate the relevance of our improved distance measure. We chose a study with a large national dataset of 217 samples, with one transmission cluster described in detail (Walker et al., 2013) and a dataset with a small number of patients (Fiebig et al., 2017) (referred to as UKTB (United-Kingdom-TB) and RAGTB (Romania-Austria-Germany-TB), respectively). We also included a study including isolates from a high-burden setting in China (Yang et al., 2017) (referred to as CTB).

For the comparison and evaluation, we classify all links between samples into transmission cluster links (SNP difference of fewer than 13) and unrelated links following previously published results (Walker et al., 2013) and group samples into transmission clusters. To be assigned to a transmission cluster the distance of one sample to only one of the other samples has to be classified as transmission cluster link, therefore unrelated samples can belong to the same transmission cluster when they are connected by other samples.

Simulation dataset results

We applied the exclusion method to the simulation dataset and compared the detected transmission links to the simulated ones. Using the exclusion method results in a high number of predicted transmission cluster links. The sensitivity is very high (1.000), as there are no false negative classifications, however the specificity (0.782) and F1-Score (0.326) are low. In detail, the strategy of excluding low-quality variant sites found in any sample from the transmission analysis resulted in a high number of false positive classifications - differences are overlooked and a transmission is assumed where there is no relation between the samples (Table 3.4).

The substitution method classifies a high number of links between samples as unrelated. Comparison of this classification with the simulated links showed that there were many true negative classifications (specificity = 1.000), but also in many false negative ones (sensitivity = 0.179). Substituting regions of low quality with the sequence of the reference genome introduces artificial differences between the samples and therefore this method cannot be used to identify transmission links, as the relation between samples is obscured (Table 3.4).

Transmission cluster links can be accurately identified using PANPASCO resulting in the highest accuracy and F-Score for the simulation dataset (>0.99 and >0.94, respectively, Table 3.4).

To analyze the influence of the reference genome on the results of the different

	TP	TN	FP	FN	Sensitivity	Specificity	Accuracy	F-Score
exclusion	**2604**	38654	10745	**0**	**1.000**	0.782	0.793	0.326
substitution	465	**49394**	**5**	2139	0.179	**1.000**	0.959	0.303
PANPASCO	2525	49174	225	79	0.970	0.995	**0.994**	**0.943**

Table 3.4: **Comparison of SNP-counting methods in all clusters of the simulation dataset.** Samples in this dataset were simulated from four different *M. tuberculosis* genomes including the H37Rv strain. This dataset includes 2604 tranmission cluster links and 49399 unrelated links. We compare three SNP-counting approaches: exclusion = low-quality variable sites are excluded for all samples in the analysis, substitution = SNPs at low-quality sites are substituted with the reference base, PANPASCO = high-quality sites are identified and differences counted for each pair separately and a computational pan-genome sequence built from 146 *M. tuberculosis* genomes was used as reference genome. Samples with a SNP-difference of fewer than 13 SNPs belong to the same transmission cluster. The data shows that using PANPASCO results in the highest accuracy and F-score. TP = true positives, TN = true negatives, FP = false positives, FN = false negatives

methods, we group all samples by the genome used for simulation and calculated the accuracy of transmission link classification for each of the four sets individually. The 74 samples of the first set of clusters (C1-C5) were simulated from the *M. tuberculosis* H37Rv strain, which is commonly used for SNP distance analyses (Walker et al., 2013; Gurjav et al., 2016; Fiebig et al., 2017). The samples in the rest of the clusters are simulated from three other genomes that are increasingly different from the H37Rv strain (see Table 3.2) and therefore increase the diversity between analyzed samples.

Due to the characteristics of the exclusion method, results change with the number of samples and the sequence diversity among the samples that are included in the analysis. For this reason we used the exclusion method in two ways, once including only the clusters simulated from the respective genome and once including all samples but show the results for the respective samples only. The differing results for these strategies are clearly evident in Table 3.5. Using the exclusion method for analyzing groups of very similar samples works well. In contrast, when more samples, originating from different genomes are included, more genomic regions are excluded from the analysis and therefore fewer differences are detected. This results in a strong increase of false-positive classifications of sample links and therefore a large decrease in specificity and accuracy.

The choice of reference genome strongly affects the results obtained with the substitution method. Substituting regions of low quality with the reference genome works well only if the analyzed samples are mapped to the best fitting reference genome: almost all transmission cluster links between samples simulated from genomes other than the H37Rv strain were misclassified as unrelated (Table 3.5).

Using PANPASCO, the classification results do not differ between the four groups of samples. We achieved the highest accuracy and F-Score (>0.96 and >0.92, respectively) for all links between samples compared to the other two methods (Table 3.5). This detailed analysis shows the advantage of our approach: links between all samples in a large, diverse dataset are classified equally well, independent of the number of included samples or strains, as all SNPs, even those in strain specific genomic regions are taken into account for measuring the distance between pairs of samples.

Detailed inspection of sample links for each simulated cluster (Appendix Table 2.3) and comparing the number of SNPs detected for each link within each simulated cluster underlines the properties of the other methods the number of SNPs detected for each link within each simulated cluster underlines the properties of the other methods (see Figure 3.4): Due to the diversity of the genomes used for the simulation and the resulting exclusion of genomic regions, most samples within clusters have a reported difference of 0 or 1 with the exclusion method. The opposite is true for the substitution method as most transmission cluster links were reported as being unrelated links (> 12 SNPs).

Genome	Method	TP	TN	FP	FN	Sensitivity	Specificity	Accuracy	F-Score
H37Rv	exclusion, cluster samples*	499	1915	287	0	**1.000**	0.870	0.894	0.777
	exclusion*	499	0	2202	0	**1.000**	0.000	0.185	0.312
	substitution	463	2197	5	36	0.928	**0.998**	0.985	0.958
	PANPASCO	489	2187	15	10	0.980	0.993	**0.991**	**0.975**
MDRMA208	exclusion, cluster samples*	464	2279	260	0	**1.000**	0.898	0.913	0.781
	exclusion*	464	0	2539	0	**1.000**	0.000	0.155	0.268
	substitution	2	2539	0	462	0.004	**1.000**	0.846	0.009
	PANPASCO	425	2520	19	39	0.916	0.993	**0.981**	**0.936**
HKBS1	exclusion, cluster samples*	440	1531	109	0	**1.000**	0.934	0.948	0.890
	exclusion*	440	0	1640	0	**1.000**	0.000	0.212	0.349
	substitution	0	1640	0	440	0.000	**1.000**	0.788	0.000
	PANPASCO	425	1616	24	15	0.966	0.985	**0.981**	**0.956**
TB282	exclusion, cluster samples*	1201	3679	685	0	**1.000**	0.843	0.877	0.778
	exclusion*	1201	0	4364	0	**1.000**	0.000	0.216	0.355
	substitution	0	4364	0	1201	0.000	**1.000**	0.784	0.000
	PANPASCO	1186	4197	167	15	0.988	0.962	**0.967**	**0.929**

Table 3.5: **Comparison of SNP-counting methods for simulated samples grouped by the genome used for simulation.** Comparison of sample link classification for all samples of the simulated dataset grouped by the genome used for simulation shows the impact of reference genome for the different methods. For all samples the same reference genome was used in the analysis: the *M. tuberculosis* H37Rv strain for the exclusion and substitution method and the computational pan-genome reference sequence for PANPASCO. The results show a great reduction in accuracy and F-Score for the groups of samples that were simulated from genomes other than the H37Rv strain for the substitution method and the exclusion method when using all samples in the distance calculation. PANPASCO achieves very good classifications with no difference between the groups of samples. * Results change with the number and diversity of samples when using the exclusion method. We show the results for each group, including only the respective cluster samples or all simulated samples in the analysis. TP = true positives, TN = true negatives, FP = false positives, FN = false negatives

For a complete assessment of the different methods we also we also analyzed the simulation dataset with our computational pan-genome as the reference genome with the exclusion and substitution methods and PANPASCO with the*M. tuberculosis* H37Rv as the reference genome (see Appendix Table 2.4). This comparison shows that the classification results do not improve for the exclusion method. The combination of the substitution method and the pan-genome achieves very poor results as not a single transmission cluster link was detected. PANPASCO works best with the computational pan-genome reference sequence, but also achieves better results than the other two methods using the standard reference genome.

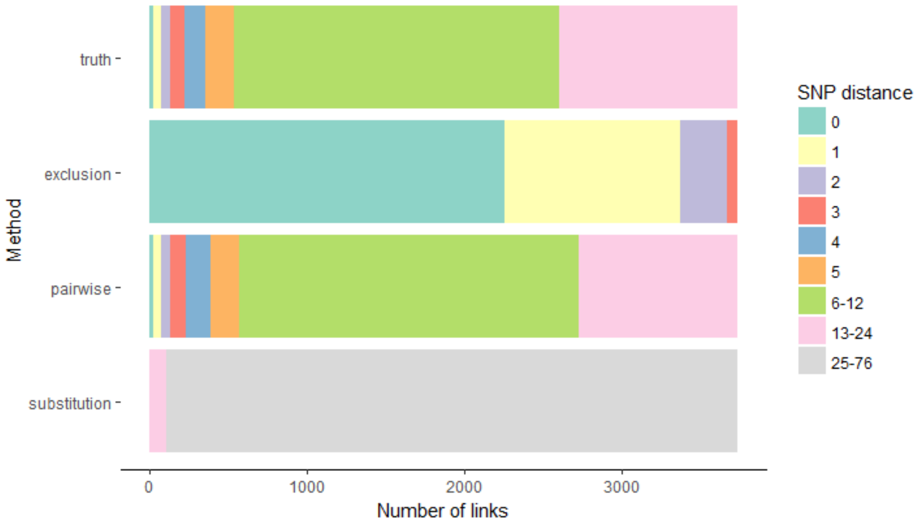

Figure 3.4: **Counting difference between samples within clusters.** We select each cluster individually and count the differences between samples within the clusters. Each cluster can contain closely related samples (0 SNPs), transmission cluster links (fewer than 13 SNPs) and unrelated samples. Samples are assigned to a cluster if the difference is fewer than 13 SNPs compared to at least one sample within the cluster. We evaluate the frequency of each distance value for the three methods and compare them to the true distance. The exclusion method reports distances of 0-3 SNPs for all samples and with the substitution method all distances are greater than 12. The distribution of distances using PANPASCO closely resemble the true distances in the dataset.

Real datasets results

Both previously published datasets were analyzed using the computational pan-genome as reference sequence and the mapping and variant calling workflow described in Methods.

For the UKTB dataset we focused our comparison on the cluster for which an epidemiological network and nucleotide variants were provided (cluster seven, UKTB7) (Walker et al., 2013). This cluster was initially defined by the shared MIRU-VNTR profile of the samples and includes 17 sequenced isolates of ten patients with one central, treatment non-compliant individual. When calculating the SNP distance using PANPASCO, we identified one pair of samples with 4 differing SNPs in addition to the ones described in the original manuscript: P066 and P175. These patients were reported with a distance of 0, indicating a close transmission event or even direct transmission. The minimum spanning tree we calculated with Cytoscape App Spanning Tree (Shaik et al., 2015) based on our SNP differences better matches the described epidemiological network (Figure 3.5). In contrast to the findings in (Walker et al., 2013) it is more likely that P076 infected the other two cases, rather than one of these the other one. We investigated the variant sites that we identified and compared them to the published ones. We identified 36 variant sites including all 20 sites listed in (Walker et al., 2013) (depicted in Appendix Table 2.5). In the original manuscript the exclusion method was used and the additional 24 sites we identified were excluded because they are not covered in samples assigned to MIRU-VNTR cluster six (data not shown). This lack of coverage is likely explained by the differing phylotypes of cluster six and seven (see Table in (Walker et al., 2013)) and in combination with the exclusion method is the reason why the 4 differing SNPs between P066 and P175 were not reported in the original publication.

The second dataset, RAGTB (Fiebig et al., 2017) consists of 13 patients and two replicates for one of the patients. We again calculated a minimum spanning tree using distances calculated with PANPASCO (see Figure 3.6). We identified the same transmission clusters as described in the original manuscript with one exception: We identified more than 12 differing SNPs for patients VI and III. The reason for that is, that the patients were previously analyzed using the exclusion method, with additional filtering of SNPs associated with drug resistance or located in repetitive regions of the genome (Fiebig et al., 2017). As this is a small dataset, there is not a large difference between the results of our approach and the exclusion method.

In the original study of the CTB dataset three clusters of patients (A, B, C) were analyzed (Yang et al., 2017). The clusters contained 19, 9 and 4 patients from two regions in China. We used PANPASCO to calculate the distances between all 32 isolates and annotated all identified transmission links in the Median-Joining networks from the original publication (Figure 3.7). Using the cutoff of fewer than 13 SNPs we identified the same transmission clusters, which shows that PANPASCO can also be used in this high-burden setting.

Discussion of Results

We present a new approach for measuring genomic distance of *M. tuberculosis* isolates, integrating commonly used mapping and variant calling methods. Reads of isolates are mapped to a computational pan-genome built from over a hundred *M. tuberculosis* genomes to include strain specific genomic regions in the analysis. SNP differences are evaluated pairwise so all high-quality regions and incorporating regions of low-quality or missing information are considered.

Our method PANPASCO accurately determines transmission clusters within large sets of samples. Previously published and commonly used methods struggle with large, diverse datasets, showing either low specificity or sensitivity when identifying transmissions. Any sort of error is problematic in a disease outbreak investigation. Methods lacking sensitivity result in overlooking transmission and failing outbreak de-

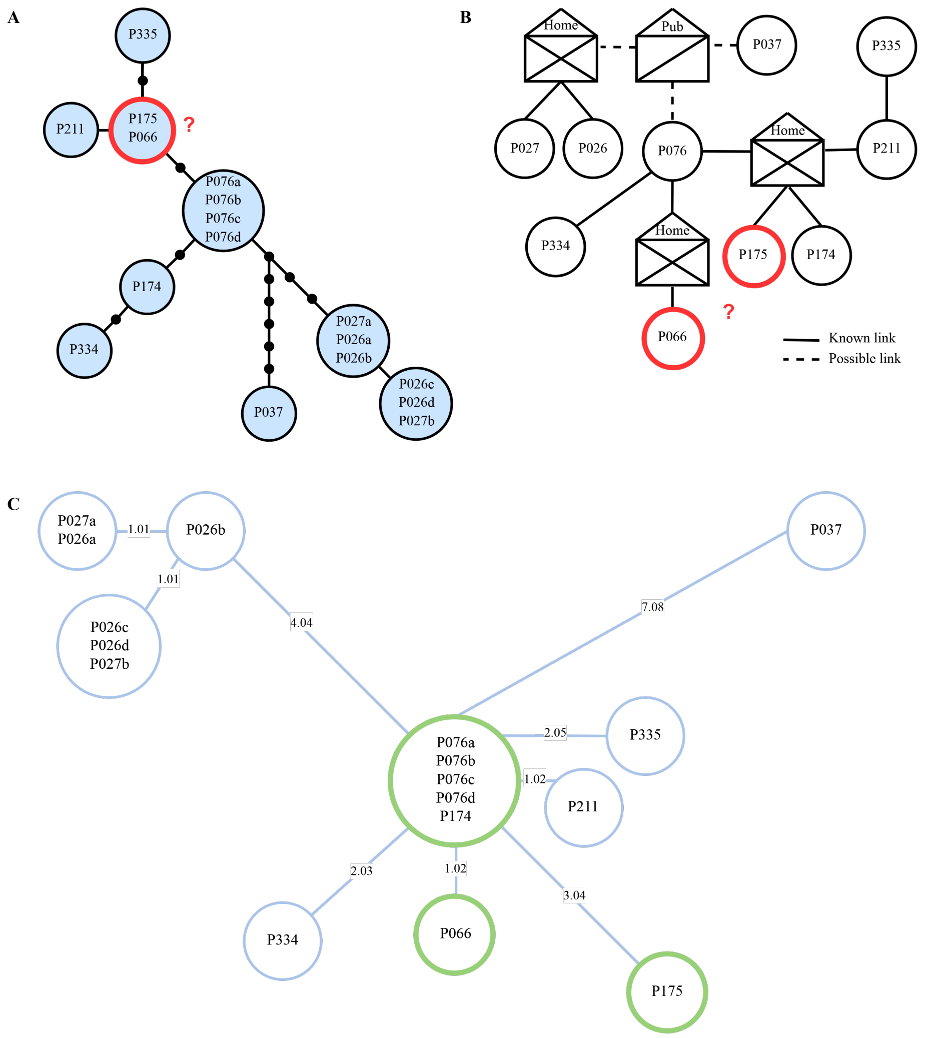

Figure 3.5: **Minimum spanning tree computed with PANPASCO between samples of the UKTB7 dataset.** (A) Genetic distances estimated by Walker et al. SNP distances are represented by dots on edges, isolates within blue circles are separated by 0 SNPs. (B) Epidemiological network as published by Walker et al. (C) Minimum spanning tree from distances calculated with PANPASCO. Adjoining isolates are separated by 0 SNPs, edges are labelled with pairwise distance. Transmission links between P066 and P175 are marked in all three parts - more SNPs were detected using PANPASCO and the resulting tree better fits the epidemiological network. ((A) and (B) taken from (Walker et al., 2013) and edited.)

tection. However, also methods lacking specificity produce unsatisfying results. Any health care system can only follow-up on a limited number of possible transmission events and must prioritize its actions and concentrate capacities and efforts. Having large numbers of spurious and incorrect transmission events detected, will automatically impact the availability of resources to focus on the true underlying transmission

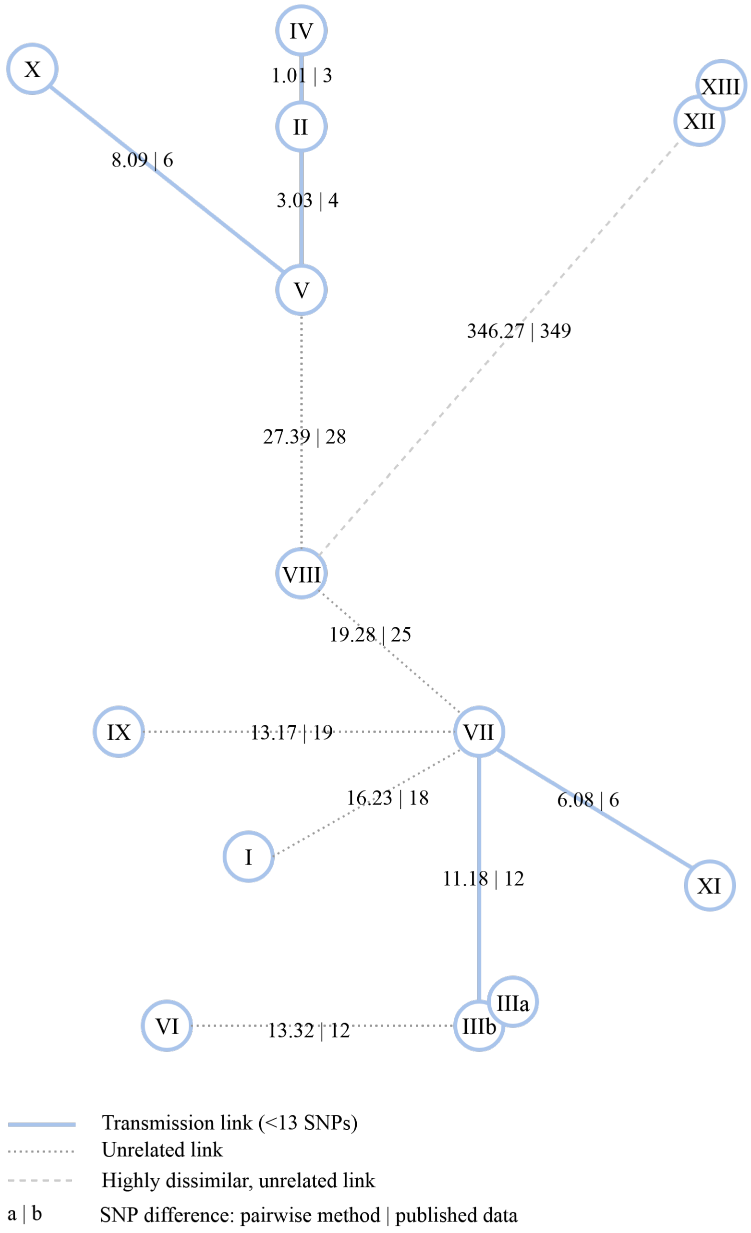

Figure 3.6: **Minimum spanning tree computed from pairwise distances between samples of the RAGTB dataset.** Edges are labeled with SNP distances (format: PANPASCO | published data). The exclusion method and the *M. tuberculosis* H37Rv strain were used in the original publication (Fiebig et al., 2017). Blue lines mark distances of fewer than 13 SNPs detected with PANPASCO, clustering patients into three transmission clusters and four independent samples. Samples XII and XIII show no difference with both analyses. The comparison of these two distances shows that there is no large difference in the results as this is a small dataset with very similar samples.

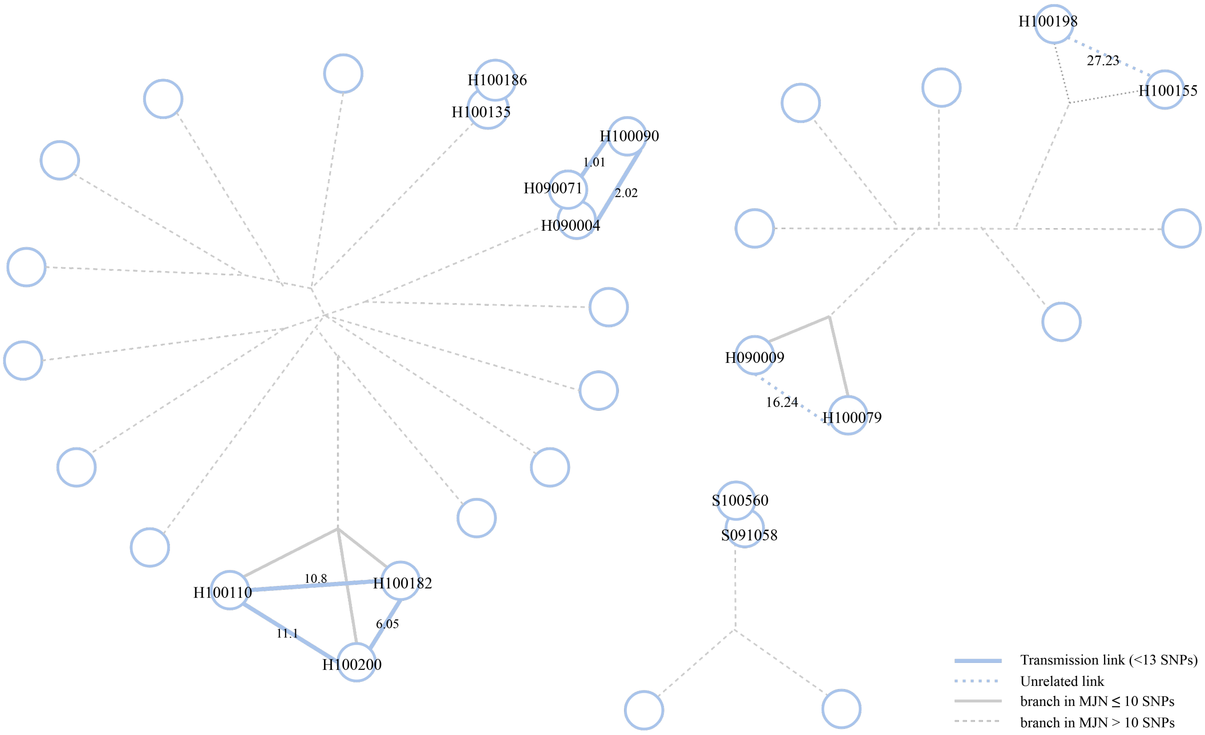

Figure 3.7: **Median-Joining networks for samples of the CTB dataset annotated with pairwise distances.** Edges between isolates with genetic distance of fewer than 10 SNPs identified in the original publication are labeled with SNP distances calculated with PANPASCO. Blue lines mark distances detected with PANPASCO, while gray ones mark original results. The comparison of these two analyses showed very similar results.

events within short time.

The major drawback of the exclusion method is its limited resolution, which is driven by the sample with the lowest coverage. Another disadvantage of this method for usage in disease surveillance is that each time a new sample is added to the dataset, the results for previously analyzed isolates change as more regions of the genome have to be excluded to be able to compare the new sample to the dataset. Taking this further, with a rising number of samples at some point, due to random sequencing errors and varying coverage distribution, there will be fewer and fewer sites of the genome with high-quality information available for comparison. This problem is of special interest and high importance regarding long time infectious disease surveillance rather than outbreak investigations and recent transmission analysis.

This problem can be avoided when low-quality or missing regions are substituted with the reference sequence. With the substitution method all detectable difference between samples are considered in the distance calculation. However, many artificial differences are introduced as well. When an *M. tuberculosis* isolate is compared to a reference sequence, usually several hundred SNPs are detected. Within a transmission cluster all but a few of these SNPs are the same. When a subset of these sites cannot be detected due to missing or low-quality reads and are substituted with the reference base, they are counted as differences between the samples and transmissions cannot be detected. The problem increases the more the isolates diverge from the reference sequence.

Strains of *M. tuberculosis* can be assigned to different sub-lineages that differ in many loci and blocks of deletions (Brudey et al., 2006). For analysis of small transmission clusters a lineage specific genome can be used as the reference genome for optimal results. In larger studies, samples of different strains are often analyzed together, choosing one reference genome that naturally represents only one of the lineages adequately (Bryant et al., 2013; Walker et al., 2013; Guerra-Assunção et al., 2015; Gurjav et al., 2016). This means, that for a part of the set of samples a suboptimal reference is used for mapping and variant calling. Nevertheless, identifying the best fitting reference sequence for each sample is no optimal solution. Clusters within a dataset and datasets from different studies will not be comparable with each other. The computational pan-genome approach solves this problem as it allows the integration of genomic information of several *M. tuberculosis* strains. The computational pan-genome enables the detection of SNPs in the core genome and in strain specific genomic regions at the same time.

We showed the specific weaknesses of the exclusion and substitution approach by applying them to a simulation dataset that was constructed to resemble a real dataset. We demonstrate the superior classification result of combining the usage of the computational pan-genome reference and the pairwise comparison. Using PANPASCO to analyze previously published datasets provided additional insight for the epidemiological investigation of these transmission clusters.

Having differing numbers of comparable sites for each pair in the analysis, complicates distance calculation. To account for the fact that fewer differences can be detected with fewer comparable sites we project the number of SNPs found for each pair to the full length of the genome, e.g. detecting 2 differences within 2 Mb is different from detecting 2 differences in 4 Mb using a normalized score. For identifying comparable sites and incorporating missing regions into the distance measure we keep track of all, instead of only variable high-quality sites of all samples. In this normalization step we work with the assumption that SNPs in samples are equally distributed over the length of the computational pan-genome. Mutation hot spots and the frequency of variants in the genomes used for the computational pan-genome can be taken into account for a more accurate normalized score.

For further improvement of transmission cluster detection, several steps can be taken. The pan-genome used in this study was built from all available reference sequences for *M. tuberculosis* in NCBI Refseq (Jandrasits et al., 2018). However, some of the seven major lineages (Coll et al., 2014) are underrepresented in this set of genomes (see information about lineages in Appendix Table 2.1). Reference genomes representing these lineages and animal strains such as *Mycobacterium bovis* could be included additionally for the analysis of more diverse datasets.

To increase differentiation between samples, information about additional genetic variation such as mixed base calls for SNPs, insertions, deletions or tandem repeats and microsatellites could be incorporated into the distance measure. There is a need for standard methods for evaluating quality and validity to include these additional measures of genetic variation. This can be achieved by using graph representation of reference genomes and sample sequences such as Cortex (Iqbal et al., 2012, 2013), panVC (Valenzuela et al., 2015) or vg (Garrison et al., 2018). With these approaches all sequence information of the samples can be used in the analysis to detect differences among samples or compared to one or more reference sequences. However, methods for annotation of identified differences are lacking and interpretation of results can be challenging.

Small parts of missing information in the analyzed samples can also be imputed from known haplotypes or using a phylogenetic tree of the analyzed samples (Jobin et al., 2018). This method has the potential to improve the results of all methods described in this study. The success of these methods vary with the availability of information on haplotypes and the variation among related sequences for constructing informative phylogenetic trees.

Several studies investigated appropriate SNP distance cutoffs for transmission cluster definition (Hatherell et al., 2016). These cutoffs have to be defined specifically for each species and the environment of transmission. While the cutoffs described for *M. tuberculosis* seem to be stable between different settings (Yang et al., 2017), there are also alternative approaches that can be used for transmission cluster definition (Stimson et al., 2019).

Today, WGS-based molecular surveillance of TB is established in a number of low-incident countries, e.g. USA, UK, Netherlands. Such systems allow for event specific adaption of public health action, patient care, medication and treatment based on pathogen specifics like resistance, virulence and actual spread (outbreak size). Early detection of antibiotic resistance and prevention of further transmission are one of the main tasks on the path to TB elimination. This implies a large number of samples that has to be analyzed - e.g. in Germany there were more than 4000 cases (4099 of 5915 reported cases; corresponding to 83.4%) laboratory confirmed by culture of tuberculosis in 2016 (European Centre for Disease Prevention and Control/WHO Regional Office for Europe, 2018). Nowadays, higher mobility and worldwide migration cause a larger geospatial spread of specific pathogens and increase the diversity of samples.

Several studies analyzed and assessed the integration of WGS into routine tuberculosis diagnosis and investigation (Walker et al., 2013; Casali et al., 2014; Witney et al., 2015; Pankhurst et al., 2016). The authors show the added benefit of using SNP-based analysis in transmission cluster and drug resistance detection in large groups of patients. However, while they discuss the importance of common standards for sequencing techniques and quality, the implications of integrating a large number of samples across different lineages in the same analysis are not addressed.

Balancing sensitivity and specificity is key for the analysis of large and diverse groups of samples during outbreak investigations and in TB surveillance, when it is of importance to find each case and expensive to investigate large numbers of false positives. PANPASCO can contribute to achieving these goals by usage of pan-genomic references and improved pairwise SNP-distances.

Chapter 4

Improving tuberculosis surveillance by detecting international transmission using publicly available whole genome sequencing data

Background

Improving the surveillance of tuberculosis (TB) is one of the eight core activities identified by the World Health Organization (WHO) and the European Respiratory Society to achieve TB elimination, defined as less than one incident case per million (Matteelli et al., 2018). Monitoring transmission is especially important for multidrug-resistant (MDR)-TB isolates – defined as being resistant to rifampicin and isoniazid – and for extensively drug-resistant (XDR)-TB isolates – defined as MDR-TB isolates with additional resistant to at least one of the fluoroquinolones and to at least one of the second-line injectable drugs. In 2017, the WHO estimated that worldwide more than 450,000 people fell ill with MDR-TB and among these, more than 38,000 fell ill with XDR-TB (WHO, 2018a). The rapid advance in molecular typing technology – especially the availability of whole genome sequencing (WGS) to identify and characterize pathogens – gives us the chance of integrating this information into disease surveillance. For TB surveillance it is possible to combine the results of molecular typing of *M. tuberculosis* complex isolates with traditional epidemiological information to infer or to exclude TB transmission (Roetzer et al., 2013; Hatherell et al., 2016). This is of particular relevance if transmission occurs among multiple countries, where epidemiological data such as social contacts are more difficult to get and where data exchange is more difficult to organize. The European Centre for Disease Prevention and Control (ECDC) identified 44 events of international transmission (international clusters) of MDR-TB isolates collected in different European countries between 2012 and 2015 (ECDC, 2017). In this example, the authors inferred TB transmission using the mycobacterial interspersed repetitive units variable number of tandem repeats (MIRU-VNTR) typing method. However, this method has limitations such as low correlation with epidemiological information in outbreak settings and low discriminatory power (Roetzer et al., 2013; Wyllie et al., 2018). In comparison, WGS analysis offers a much higher discriminatory power and allows for inferring (or rejecting) TB transmission at a higher resolution (Hatherell et al., 2016). In a recent systematic review, van der Werf and co-authors identified three studies that used WGS to investigate the international transmission of TB (van der Werf and Ködmön, 2019). In recent years, the amount of WGS data available has increased, especially due to the reduction of sequencing costs (Muir et al., 2016). In addition, more and more authors deposit the raw data of their projects in open access public repositories such as the Sequence Read Archive (SRA) of the National Center for Biotechnology (NCBI) (Leinonen et al., 2010). These raw WGS data of thousands of isolates – together with their public availability – enable the re-use and the additional analysis at a larger and global scale from different perspectives (Ohta et al., 2017). However, standards in bioinformatics analysis and interpretation of these WGS data for surveillance purposes are not yet fully established (Meehan et al., 2019). In addition, it is still unclear if and how far we can use this high amount of publicly available data to improve TB surveillance. Our aim was to investigate to what extent we could use raw WGS data of global MDR/XDR-TB isolates available from public repositories for TB surveillance. Specifically, we wanted to identify potential international events of TB transmission and to compare the international isolates with a collection of *M. tuberculosis* isolates collected in Germany in 2012-2013.

Methods

Data collection: public dataset

The SRA database is a public repository provided by the NCBI (U.S. National Library of Medicine, Bethesda, USA) which stores raw sequencing data derived from high-throughput sequencing platforms (Leinonen et al., 2010). We queried the repository for the pathogen "Mycobacterium tuberculosis" and restricted the results to "genomic", "WGS" data from the "Illumina" sequencing technology using the appropriate query keywords. After excluding single-end sequenced and missing raw data, 8,716 isolates remained, which were further filtered for sequence characteristics. We excluded samples with reads shorter than 100 bp, as well as samples with a relatively low (< 20x) or high (> 500x) average coverage depth of the reference genome (see below) to obtain a more homogenous dataset. In addition, we excluded samples with less than 90% reads aligned to the reference genome to prevent having contaminated or incorrectly annotated samples in the set. Samples for which over 50% of all single-nucleotide variant calls were inconclusive were also excluded. For this, we extracted all single nucleotide base calls in high quality regions called from at least 5 reads. We determined the number of SNPs with an allele frequency between 25% and 75% (mixed base calls). Samples where 50% or more SNPs were mixed base calls were excluded from our analyzed set. To identify duplicates (e.g. the same file uploaded more than once in different projects) within the public dataset, we compared numbers of reads and detected variants at every step of the analysis. We excluded samples that were identical in all those numbers and their corresponding epidemiological data. After all filtering steps, 7,620 isolates remained and we will refer to these isolates as the "public dataset" throughout the manuscript. In addition to the raw reads, we also collected metadata available in the SRA repository (Leinonen et al., 2010) (for relevant metadata for all samples see Appendix Table 3.1).

Data collection: German dataset In addition to the international public dataset, we analyzed isolates from Germany, which will be referred to as "German dataset" throughout the manuscript. The German dataset includes all *M. tuberculosis* isolates processed at the National Reference Center for Mycobacteria (Forschungszentrum Borstel, Germany) and classified as MDR-TB or XDR-TB in 2012-2013 by drug susceptibility tests (DSTs) according to the German TB surveillance system. We extracted the epidemiological data available for the *M. tuberculosis* isolates using the laboratory ID of the National Reference Center for Mycobacteria. Then, we identified the respective isolate in the German TB surveillance system and thus matched molecular with epidemiological data. We collected information on year of isolation, federal state of isolation, DST results, and patient-related information such as age, gender, citizenship, and country of birth. The epidemiological data of the German dataset was obtained through the national surveillance system at the Robert Koch Institute, the German public health institute. Ethical approval was not required for this study, as the data extraction were performed on anonymized notification data.

NGS analysis workflow

Our NGS workflow includes quality control, read mapping, variant discovery and detection of low-quality regions for paired-end next generation sequencing samples. First, adapter sequences were removed from the ends of the reads with Trimmomatic

(Bolger et al., 2014), then they were merged with Flash (Magoč and Salzberg, 2011) in case they overlap with at least 5 base pairs. This resulted in two sets of reads: non-overlapping paired-end reads and merged, longer single reads. We used Trimmomatic (Bolger et al., 2014) to trim bases of low base quality in both sets. Reads with a length of less than 35 were removed from the dataset and remaining mates of excluded reads added to the set of single reads. We mapped all reads to two reference genomes with bwa mem (Bolger et al., 2014) and the sets of mapped paired-end and single reads were joined for the following analysis with samtools (Li et al., 2009). We used the Mycobacterium tuberculosis H37Rv strain (Accession number NC_000962.3) for resistance mutation detection and the linear pan-genome consensus sequence built from 146 *M. tuberculosis* genomes (Jandrasits et al. (2018); Jandrasits et al. (in revision)) for SNP distance calculation. In preparation for variant detection, duplicated reads were marked and read groups added with picard tools (Broad Institute, Accessed: 2018-02-21). Reads with a mapping quality of less than 10 were excluded from the set of mapped reads. For variant detection we used the Genome Analysis Toolkit (GATK (DePristo et al., 2011)) for both genomes. As described in the workflow for PANPASCO (Jandrasits et al. (in revision)), we first detected confidence scores for all sites, including reference sites with the HaplotypeCaller tool. After that we extracted genotypes for each site using the GenotypeGVCFs tool. We called variants with a diploid model to be able to filter mixed base calls by allele frequency. We analyzed all sites and separated them in variant sites, positions with uncalled genotypes and high-quality reference sites. Variant sites were then split into SNPs, small deletions and insertions and structural variants with the SelectVariants tool from GATK. We identified SNPs with an allele frequency of at least 75% and where 5 or more reads were used to call the SNP. We also used bedtools (Quinlan and Hall, 2010) to extract regions with less than 5 reads coverage. By using these filters we separated the data into high-quality SNPs and five sets of low-quality regions with the following criteria: - positions with less than 5 mapped reads - positions at deletions - positions with uncalled genotypes - SNPs called from less than 5 reads - SNPs with less than 75% allele frequency

Only high quality SNPs are used in subsequent analyses.

Drug-resistance prediction

We used Phyresse (Feuerriegel et al., 2015) and TBDreamDB (Sandgren et al., 2009) to identify drug-resistance mutations in our datasets (last access October 18th, 2018). We filtered both lists to include only single nucleotide substitutions. For TBDreamDB we mapped the provided locations within resistance genes to positions on the *M. tuberculosis* H37Rv genome where necessary. We excluded mutations not associated with drug-resistance according to the WHO (Miotto et al., 2017) (data not shown).

We intersected this list of mutations with the variants detected from reads mapped to the *M. tuberculosis* H37Rv genome from each sample to identify resistance-associated mutations within samples. We also identified uncovered or low-quality regions that overlap with locations of resistance mutations. For the classification of isolates into resistance classes (MDR-TB and XDR-TB), we used the definitions of the WHO (WHO, 2018a).

Molecular clustering

We used PANPASCO (Jandrasits et al.) to calculate relative pairwise SNP distance between all isolates classified as MDR-TB or XDR-TB in the public and German dataset. This method builds on two parts to enable distance calculation for large, diverse datasets: mapping all reads to a computational pan-genome including 146 *M. tuberculosis* genomes and distance calculation for each individual pair of samples. For this we identify all positions with high quality for each pair of samples and calculate the SNP distance based on this set of positions. SNPs in repeat-rich genes were not used for distance calculations as studies have shown that variants found in these regions are often false positives (Roetzer et al., 2011, 2013). The list of genes provided by Comas et al. (2009) was used for filtering. We applied single linkage agglomerative clustering for defining transmission clusters and used a threshold of fewer than 13 SNPs, based on a previous study (Walker et al., 2013). PANPASCO calculates distances based on data available for each pair separately. For this reason, an individual sample can potentially have small distances to samples that have a much greater distance in a direct comparison, due to a higher number of compared high quality sites. In this study, we aimed to discover clusters of closely related samples. Therefore, the implemented agglomerative clustering approach evaluated the distance from the sample, that should be added to two instead of one sample of an existing cluster -we did not only compare pairs of samples but two sets of trios. The sample was added to the cluster only if the maximum distance in the trio is below twice the SNP threshold. Samples that violated this condition were iteratively removed from the clustering and marked for potential follow-up analyses. Clusters were visualized with Cytoscape 3.7 (Shannon et al., 2003). We classified all clustered samples into TB lineages using lineage specific SNPs provided in (Coll et al., 2014) and (Merker et al., 2015) (see Appendix Table 3.1). We compared and validated clustering results of a subset of isolates using the pipeline MTBSeq (Kohl et al., 2018b) (see Appendix Table 3.5).

Data availability

The raw whole genome sequencing data used in this study are available in the NCBI SRA repository. The accession numbers for all samples of the public dataset are available in Appendix Table 3.1. The German dataset will be available as ENA Bioproject. Software for creating a pan-genome sequence (seq-seq-pan) is accessible at `https://gitlab.com/rki_bioinformatics/seq-seq-pan` and scripts for the NGS workflow and the SNP-distance method (PANPASCO) are available at `https://gitlab.com/rki_bioinformatics/panpasco`. The code for the clustering method is available at `https://gitlab.com/rki_bioinformatics/snp_distance_clustering`.

Results

Final dataset

After the filtering steps, 7,620 of initially 8,716 downloaded isolates remained in the Public dataset and 131 isolates from the German dataset (Figure 4.1). We focused our study on MDR/XDR-TB, and therefore the final dataset contained overall 1,339 isolates after filtering using resistant associated SNPs. Appendix Table 3.1 shows the cluster assignment, molecular drug-resistance prediction and extracted metadata of these 1,339 isolates.

Metadata availability and drug-resistance prediction: public dataset (N=1,208)

The majority of metadata collected from the public dataset consisted of country of isolation (1,052/1,208, 87.09%), year of isolation (924/1,208, 74.49%) and the source of the isolate (1,000/1,208, 82.78%) (Table 4.1). For other metadata we could collect less information, for example in the case of patient age (178/1,208, 14.73 %), patient gender (175/1,208, 14.49 %), or patient HIV status (161/1,208, 13.33 %). For 915 isolates, we had information on both country and year of isolation. Initially, we identified 336 isolates with missing data for the country of isolation. After examining the Bioproject information (SRA, Leinonen et al. (2010)) of these 336 isolates, we could further identify the country of isolation of 177 isolates. We identified 923/1,208 MDR (76.41%) and 285/1,208 XDR (23.59%) isolates.

Metadata availability and drug-resistance prediction: German dataset (N=131)

We identified all isolates (N=131) in the German TB surveillance system and could retrieve demographic, epidemiological information and DST results for 129/131 (98.47%) of the isolates. Table 4.2 and Appendix Table 3.2 show the collected metadata. The 131 German isolates came from 15/16 (93.75%) of the German federal states. The most frequent countries of birth of the patients were Russia (27/131, 20.61%), Germany (19/131, 14.50%) and Romania (10/131, 7.63%) (Table 4.2). We identified discrepancies in the identification of rifampicin resistance between the results of the phenotypic DST and the detection of drug-resistance mutations in 14 isolates (Appendix Table 3.2). Four isolates were classified as MDR in the TB surveillance system (isolates 4556-12, 9165-12, 72-13 and 14102-13) while they were classified as non-MDR according to the molecular analysis, due to the absence of any drug-resistance mutations against rifampicin. However, in one of these four isolates (isolate 72-13), we found insufficient sequencing coverage in some of the genomic regions with known resistance mutations for rifampicin; while in another isolate (isolate 14102-13), we found an insertion of 3 nucleotides near a region with known resistance mutations for rifampicin. In addition, ten isolates classified as MDR in the TB surveillance system (isolates 11355-13, 12016-13, 2955-12, 3007-13, 4245-13, 5096-13, 5190-13, 7712-13, 8291-13 and 8565-12), were classified as XDR according to the analysis of the drug-resistance mutations. The reason for such discrepancy was that a drug-resistance mutation against amikacin, kanamycin or capreomycin was identified in these ten isolates, but no DST results were available for these antibiotics.

Molecular clustering and comparison between the public and the German dataset Among all the isolates of our study, we identified 133 molecular clusters – with at least 2 isolates – and 595 singletons. The 133 clusters included 744 isolates (Appendix Table 3.3). Appendix Table 3.4 shows a summary of distances between all isolates for all molecular clusters. In 16 clusters, the isolates were from at least two different countries of isolation, suggesting international transmission of TB (Appendix Table 3.3). For example, cluster2 included 56 MDR/XDR-TB isolates from three countries – Moldova, Georgia and Germany. A total of 51/56 isolates in this cluster were part of a previous study (Bioproject PRJNA318002, Rosenthal et al. (2017)). Figure 2 shows the country of isolation and the year of isolation of the isolates belonging to cluster2. Cluster1 is the largest cluster identified in our study. According to the metadata (such as host subject, isolate name, year of isolation, patient age, and patient gender), it included 79 autopsy samples from different anatomic sites (such as lung or liver) of the same patient, marked as "P21". Similarly, cluster3, cluster14, cluster16, cluster18, and cluster28 contained multiple isolates from single patients from South Africa, which were part of a study including 2,693 autopsy samples of 44 subjects (Lieberman et al., 2016). In line with previous findings (Lieberman et al., 2016), our analysis showed very low variability within these clusters (Appendix Table 3.4). In addition, analysis of the respective metadata revealed that cluster26, cluster32 and cluster33 included multiple isolates from single patients. These isolates were part of a study investigating the evolution of drug-resistant TB in patients during long-term treatment (Xu et al., 2018). When we compared the German dataset with the public dataset, we observed that in 11 clusters there was at least one isolate from Germany and at least one isolate from another country. Table 4.3 shows the relation between the German isolates and the international isolates from the public dataset. The epidemiological information collected from the German isolates correlates well with molecular clusters in 7/11 cases. For example, in cluster9 there were 16 isolates from Georgia and two isolates from Germany; the country of birth recorded for one of these two isolates from Germany was Georgia. Moreover, cluster24, cluster35, and cluster103 included isolates from Georgia and Germany, and the country of birth recorded for the isolates from Germany was Georgia. Three further examples of agreement between molecular and epidemiological data were cluster13, which included isolates from Germany and Kazakhstan, and cluster53, which included isolates from Germany and from Romania, and cluster58, which included isolates from Germany and from India (Table 4.3). By comparing the molecular data of the German and of the public dataset, we could connect previously epidemiologically unlinked cases. For example, in cluster2 (Figure 4.2), two isolates from Germany (in orange) were connected through several isolates from Georgia and Moldova (in dark and light blue), and the distance between the two German isolates was >13 SNPs. Similarly, in cluster53, two isolates from Romania were connected through a German isolate, and the distance between the two isolates from Romania was > 13 SNPs (data not shown).

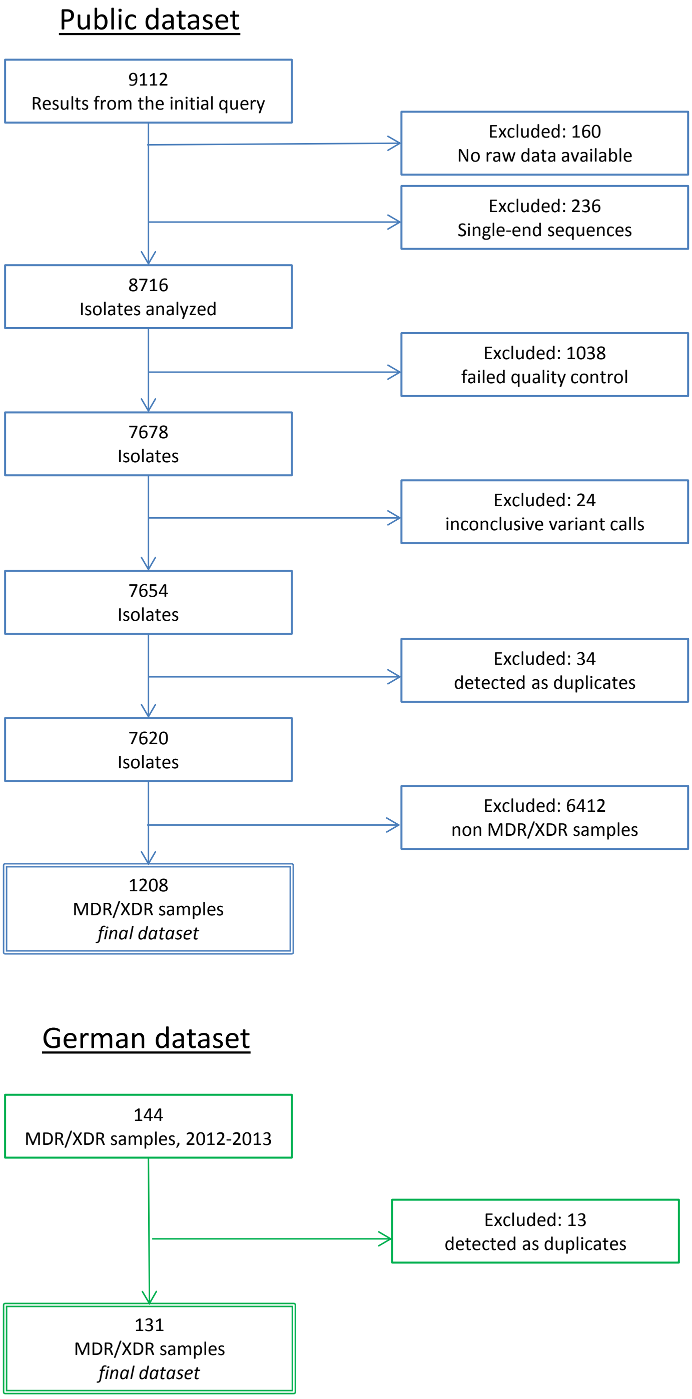

Figure 4.1: **F**lowchart of the inclusion and exclusion of isolates in our study from the public and the German dataset. The final dataset included 1,339 isolates: 1,208 from the public and 131 from the German dataset.

Characteristic		n	%
Country of isolation	South Africa	295	24.42
	Georgia	160	13.24
	Moldova	135	11.17
	Vietnam	68	5.63
	Azerbaijan	58	4.80
	Bangladesh	46	3.81
	Romania	38	3.15
	Djibouti	31	2.57
	Ivory Coast	29	2.40
	India	28	2.32
	Nigeria	28	2.32
	Thailand	24	1.99
	Peru	23	1.90
	China	23	1.90
	Tanzania	17	1.41
	Other	49	4.06
	NA	156	12.91
Year of isolation	2016	53	4.39
	2015	254	21.03
	2014	106	8.78
	2013	147	12.17
	2012	86	7.12
	2011	60	4.97
	2010	87	7.20
	2009	65	5.38
	2008	27	2.23
	2007	11	0.91
	2006	6	0.50
	2005	14	1.16
	2004	6	0.50
	2003	1	0.08
	1996	1	0.08
	NA	284	23.51
Source of the isolate	Sputum	836	69.20
	Morgue	167	13.82
	Other	6	0.50
	NA	199	16.47

Table 4.1: Characteristics of the 1,208 multi- and extensively drug resistant *M. tuberculosis* isolates from the public dataset analyzed in this study. NA: not available

Characteristic		n	%
Molecular drug resistance prediction	MDR	110	83.97
	XDR	17	12.98
	Non MDR non XDR	4	3.05
Phenotypic drug Resistance prediction	MDR	122	93.13
	XDR	7	5.34
	NA	2	1.53
Year of isolation	2013	80	61.07
	2012	50	38.17
	2014	1	0.76
Federal state of isolation	North Rhine-Westphalia	32	24.43
	Bavaria	13	9.92
	Baden-Württemberg	15	11.45
	Saxony	10	7.63
	Lower Saxony	10	7.63
	Berlin	10	7.63
	Hamburg	8	6.11
	Hesse	8	6.11
	Schleswig-Holstein	5	3.82
	Saxony-Anhalt	5	3.82
	Other	11	8.40
	NA	4	3.05
Patient age	Median	34 (2-83)	
	Mean	35.73	
Patient gender	Male	79	60.31
	Female	50	38.17
	NA	2	1.53
Patient citizenship	Germany	30	22.90
	Russia	25	19.08
	India	8	6.11
	Georgia	7	5.34
	Romania	7	5.34
	Kazakhstan	6	4.58
	Ukraine	5	3.82
	other	39	29.78
	NA	4	3.05
Patient country of birth	Russia	27	20.61
	Germany	19	14.50
	Romania	10	7.63
	Ukraine	8	6.11
	India	8	6.11
	Kazakhstan	8	6.11
	Georgia	7	5.34
	Other	41	31.30
	NA	3	2.29

Table 4.2: **Characteristics of the 131 multi- and extensively drug resistant Mycobacterium tuberculosis isolates from Germany analyzed in this study.** Demographic information, epidemiological information and drug susceptibility test- results were available in the German TB surveillance system for 129/131 isolates. MDR: multidrug-resistant; XDR: extensively drug-resistant; NA: not available

Cluster name	No. of isolates in the cluster	No. of MDR	Country of isolation of MDR (n)	No. of XDR	Country of isolation of XDR (n)	Characteristics of the German isolates within the clusters	
						Patient country of birth (n)	Patient nationality (n)
2	56	54	Moldova (47) Germany (2) Georgia (1) NA (3)	2	Moldova (2)	Romania (1) Germany (1)	Romania (1) Germany (1)
5	30	5	South Africa (4) Germany (1)	25	South Africa (25)	Abroad (1)	Abroad (1)
9	18	18	Georgia (16) Germany (2)	0	0	Georgia (1) Romania (1)	Georgia (1) Germany (1)
13	10	1	Germany (1)	9	Kazakhstan (9)	Kazakhstan (1)	Germany (1)
21	6	6	Georgia (5) Germany (1)	0	0	Syria (1)	Syria (1)
24	5	5	Georgia (3) Germany (2)	0	0	Georgia (2)	Georgia (2)
35	4	0	0	4	Georgia (3) Germany (1)	Georgia (1)	Georgia (1)
53	3	2	Romania (1) Germany (1)	1	Romania (1)	Romania (1)	Romania (1)
58	3	3	India (2) Germany (1)	0	0	India (1)	India (1)
59	3	3	Georgia (1) Germany (2)	0	0	Georgia (1) Ukraine(1)	Georgia (1) Ukraine(1)
103	2	0	0	2	Georgia (1) Germany (1)	Georgia (1)	Georgia (1)

Table 4.3: Characteristics of the 11 molecular clusters identified in this study which contain at least one isolate from Germany and at least one isolate from another country. In bold the isolates from Germany. Within each cluster, information about the country of birth, the nationality and the federal state of isolation of the German isolates is provided. MDR: multidrug-resistant; XDR: extensively drug-resistant; NA: not available

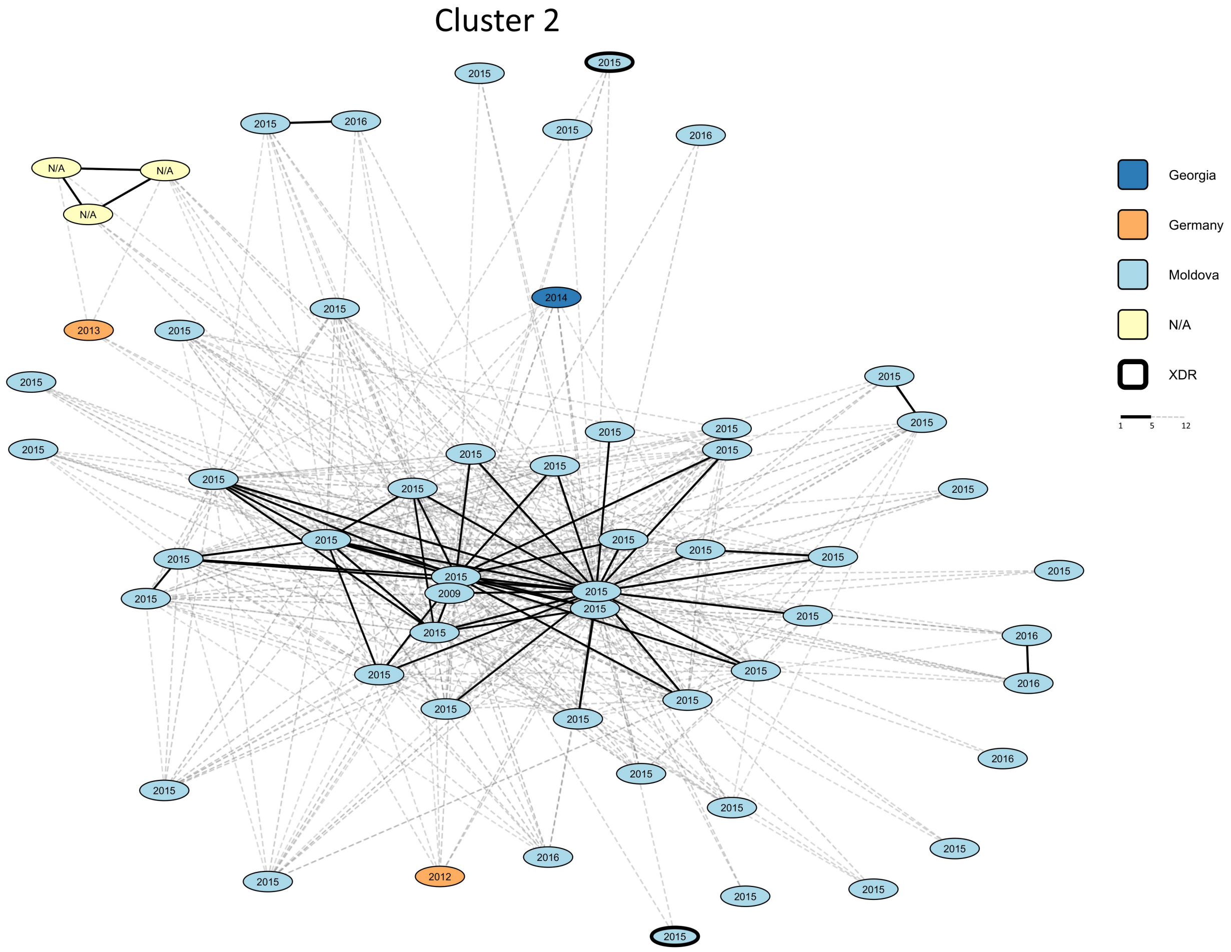

Figure 4.2: **Visualization of transmission cluster2 (N=56) identified among the 1,339 *M. tuberculosis* isolates analyzed in our study.** The country of isolation, multi- and extensive drug resistance classification and year of isolation are represented in the clusters. SNP distances where calculated for each pair of isolates individually. Links with fewer than 6 SNPs are marked black, those with fewer than 13 SNPs are marked in grey. Connections with 13 SNPs or more than are not shown in the network.

Discussion of results

In this study, we investigated to what extent WGS data of MDR/XDR-TB isolates available from public sequence repositories can be used for improving TB surveillance. We identified several molecular clusters which contained isolates from multiple countries suggesting international transmission of TB. We expected to find international TB-transmission events, also considering previous studies reporting cross-border molecular clusters (ECDC, 2017; van der Werf and Ködmön, 2019). Looking at the collected metadata, we identified several clusters with multiple isolates from the same patient or multiple autopsy samples collected from the same patient (Lieberman et al., 2016; Xu et al., 2018). This shows the importance of providing complete metadata together with the publicly available molecular data; based on the metadata we could, distinguish between clusters of isolates taken from different patients – the "real" transmission clusters – and clusters of isolates taken from a single patient. The real transmission clusters are crucial for the routine TB surveillance, while the clusters of isolates taken from the same patient are useful to study the intra-host variability of isolates. We observed agreement between molecular and epidemiological data by comparing the public and the German datasets. This is clear for example in the clusters containing isolates from both the German dataset and the public dataset originating from Georgia. It is therefore likely that migrants from Georgia acquired the TB infections in their country – or during visits there – and were diagnosed later when they moved or returned to Germany, as already described (Odone et al., 2015). This shows that we could identify events of potential international transmission (between Germany and Georgia), that we could have missed by looking only at the German molecular clusters. We observed discrepancies in the identification of rifampicin resistance between the results of the phenotypic DST and the detection of drug-resistance mutations. Specifically, four isolates were phenotypically resistant to rifampicin but they did not contain any known drug-resistance mutation against rifampicin or the genetic regions containing the known mutation had lower sequencing quality. This means that in our study the drug resistance mutations correctly predicted the resistance to rifampicin in 125/129 of the isolates, resulting in a sensitivity of 96.90%. This sensitivity is in accordance with a study by the CRyPTIC Consortium, where the authors reported a sensitivity of 97.50% (CRyPTIC Consortium and the 100,000 Genomes Project, 2018). The incorrect identification of rifampicin resistance misclassifies four isolates which were MDR by phenotype, but non-MDR by genotype. This might have had consequences for patient therapy if we would have replaced the phenotypic DST with the molecular detection of drug resistance mutations. Therefore, we suggest being careful in the transition from phenotypic to genotypic drug resistance determination as suggested by the CRyPTIC Consortium (CRyPTIC Consortium and the 100,000 Genomes Project, 2018). Specifically, laboratories and national reference laboratories should still perform the phenotypic DST, for example on a representative set of isolates or on isolates with low sequencing quality and coverage. Our study has one major implication: we demonstrated that by considering the international context (the public dataset), while analysing the national molecular data (the German dataset), we could identify previously unknown transmissions between patients, and thus we could detect larger and international events of TB transmission. To improve the WGS-based TB surveillance we, therefore, suggest to regularly compare the national molecular clusters with the international molecular clusters available in the public sequence repositories. Our study has two major limitations: first, the raw WGS data uploaded in the SRA repository (Leinonen et al., 2010) were either from single studies or from outbreaks, and therefore they were not representative of the TB situation in the different countries. This sampling bias is, however, a well-known bias in molecular epidemiology studies (Murray and Alland, 2002). Second, the metadata collected were incomplete, especially regarding patient information. Both limitations can be overcome by genotyping all TB isolates, by including the genotyping results in the TB surveillance systems and by making genotyping datapublicly available. In conclusion, we demonstrated that using WGS data from public repositories improved the surveillance of TB. The comparison between the German and the international molecular clusters was indeed useful to identify potential international events of transmission. Kohl and co-authors suggested a similar approach and used the core genome multilocus sequence typing to detect clusters (Kohl et al., 2018a). Lastly, supranational institutions such as the WHO, the ECDC or international TB networks could perform such analysis at a global scale, improving the global surveillance of TB.

Chapter 5

Inferring transmission chains of tuberculosis from genetic and epidemiological data

Background

Knowing 'who infected whom' is one of the most important parts of infectious disease surveillance and outbreak investigations. The increased availability of pathogen sequencing results has opened the path to improve surveillance with genomic data (Jombart et al., 2014). This data can be a valuable complement to contact tracing as it is independent from social contact reporting of patients, that can be obscured by recall bias, mobility of patients and reluctance to report contacts (Andrés et al., 2017). Analysis of outbreaks have to incorporate several unobserved processes: transmission, case observation, within-host microevolution and mutation. A range of heuristic methods modelling those processes and integrating genomic and epidemiologcial data have been developed over the last decade (Jombart et al., 2014). Firestone et al. (2019) recently reviewed and compared many of them on simulations of foot-and-mouth disease outbreaks (Firestone et al., 2019). This evaluation is not applicable to human pathogens as the control over infected cases and their location and infectious periods could hardly be more different. Epidemiological data and increasingly also genomic data is available for a high number of TB cases as part of national surveillance systems (Andrés et al., 2019). However, transmission inferring methods have rarely been used for *M. tuberculosis* outbreak investigation so far (Hatherell et al., 2016). In this study we compare seven transmission chain detection methods for the use with *M. tuberculosis* isolates. Two of those methods (SeqTrack (Jombart et al.,

2011) and Transphylo (Didelot et al., 2017)) were used to analyze *M. tuberculosis* outbreaks before (Guerra-Assunção et al., 2015; Didelot et al., 2017; Ayabina et al., 2018). However, the reasons for the choice of method were not apparent. We aim to identify the most appropriate method by evaluating all of them on four thoroughly analyzed datasets and establish the minimal amount of epidemiological information necessary for accurate transmission chain inference.

Methods

In a literature review we assessed several studies investigating transmission chains for bacterial and viral outbreaks. Among them we identified seven implemented algorithms: BadTrIP (Maio et al., 2018), outbreaker (Jombart et al., 2014), outbreaker2 (Campbell et al., 2018), phybreak (Klinkenberg et al., 2017), SCOTTI (Maio et al., 2016), SeqTrack (Jombart et al., 2011) and TransPhylo (Didelot et al., 2017) that were designed for transmission analysis, use aligned genomic sequences and that allow direct interpretation of the results. All methods depend on genomic data and sampling times of isolates. Additional epidemiological data required by the analyzed methods is summarized in 5.1. SeqTrack depends on genomic data and sampling times only, while outbreaker, outbreaker2, phybreak, TransPhylo require an estimation on generation time distribution. BadTrIP and SCOTTI allow the definition of infection time intervals. SCOTTI and BadTrIP are extensions of BEAST2 (Bouckaert et al., 2014) and depend on timed phylogenetic trees calculated on the genomic data of the outbreak samples. TransPhylo is also based on a such a tree. A timed phylogenetic tree can be used with phybreak as one option of transmission chain initialisation. Some methods explicitly model within-host evolution and allow for non-observed cases when inferring transmission (Table 5.1).

Method	Generation Time Distribution	Infection Time Intervals	Multiple Samples Per Host	Based on phylogenetics	Allows non-observed hosts	Models within-host evolution
BadTrIP (Maio et al., 2018)	−	+	+	+	−	+
outbreaker (Jombart et al., 2014)	+	−	−	−	+	−
outbreaker2 (Campbell et al., 2018)	+	−	−	−	+	−
phybreak (Klinkenberg et al., 2017)	+	−	−	±*	−	+
SCOTTI (Maio et al., 2016)	−	+	+	+	+	+
SeqTrack (Jombart et al., 2011)	−	−	−	−	−	−
TransPhylo (Didelot et al., 2017)	+	−	−	+	+	+

Table 5.1: Features of methods for transmission chain detection from genomic and epidemiologcial data. We compare required input data and modelling features. All methods are depending on genomic data in the form of aligned sequences and isolate sampling times. * Phybreak can use a timed phylogenetic tree for chain initialisation.

We used these seven methods on four real datasets with epidemiological data. The selected studies provide genomic data for investigated samples either in the form of already aligned sequences or complete raw whole genome sequencing (WGS) data. In the latter case, we identify high-quality single nucleotide polymorphisms (SNPs) of all samples using the analysis workflow described before (Jandrasits et al.). We used the *M. tuberculosis* strain H37Rv as reference genomes to be able to compare results between studies all datasets. Characteristics of the four datasets are shown in Table 5.2. For the NL and US datasets, where dates of sampling times were not explicitly listed we made a best guess based on provided timelines. We set the date to January 1st for all isolates of the DE and GB datasets, as only years were provided for sampling times.

For the DE dataset, we only included those samples in the analysis where epidemiological data was available (Kohl et al., 2014). We included all sequenced isolates for the remaining three datasets, also in cases where several isolates were provided for a single host: In the NL dataset, three isolates of one patient with two relapses were sequenced (Schürch et al., 2010). In the GB dataset, four isolates of one patient were sampled in 2002, 2004 (2 times) and 2005 and three isolates of two patients were sequenced in three consecutive years starting in 2006 (Walker et al., 2013).

Dataset	Location	Number of isolates	Number of hosts	Sampling Interval	WGS data
DE (Kohl et al., 2014)	Germany, Hamburg	33	15	2001-2010	+
GB (Walker et al., 2013)	United Kingdom, Midlands	17	10	2002-2011	+
NL (Schürch et al., 2010)	Netherlands, Harlingen	18	16	1994-2006	−
US (Kato-Maeda et al., 2013)	United States, San Francisco	9	9	2000-2001	−

Table 5.2: Characteristics of datasets used for comparison of transmission chain detection methods.

Generation Time and Infection Intervals

The time span during that a patient is infectious with TB depends on the state of treatment and whether the patient was infected with a drug-resistant *M. tuberculosis* strain. Susceptible individuals are considered noninfectious two weeks after treatment initiation (Schwartzman and Menzies, 2000). Starting time and state of treatment were not reported for most cases used in this analysis. Also, latent infection has to be considered when choosing an infection time interval or distribution for modelling. Individuals will fall ill with TB after the infectious period of their infectors in case they developed latent infection first. Provided time intervals between cases in this analysis mostly spanned more than two weeks. Therefore, we limited the infection time intervals for BadTrIP and SCOTTI to the maximum time interval between sample times. For the methods that required an infection time distribution (see Table 5.1) we chose a gamma distribution with shape and scale dependent on the mean time interval between sampling times with higher probabilities for earlier infections.

Results

We evaluated the transmission event predictions of the seven described methods with all four datasets. Results of predictions for the DE, GB, NL, and US datasets are shown in Table 5.3. In cases with an unclear order of infection in the epidemiological data (e.g. patient1 infected patient2 and patient3, or patient2 infected patient3 after being infected by patient1, or patient3 infected patient2 after being infected by patient1) we grouped all possible transmission events. In case one of the possible links within a group was predicted by a method the group was considered as predicted correctly.

All methods described provide a confidence score or weight for each predicted transmission. We filtered all results using only links with scores higher than the mean score of all predictions for each method. For SeqTrack we excluded predictions with a weight of "0". We examined whether the results improve when only predictions with high scores are considered in the analysis Table 5.3.

SeqTrack and outbreaker2 achieved a high number of correct predictions in the GB, NL, and US datasets, while also reporting a low number of false predictions. Outbreaker2 provided a higher number of true predictions, while SeqTrack demonstrated fewer false ones. The benefit of filtering predictions varied among the datasets and methods.

SeqTrack outperformed most of the others while only using genomic data and sampling times for modelling. Outbreaker, Phybreak, TransPhylo have mediocre prediction results, while it is not clear if the estimated infection time played a major role as Outbreaker2 was used with the same distribution. BadTrip and SCOTTI both provided a relatively high number of false predictions for two datasets (NL, US). A more limited infection time interval may improve results with these methods. However, reanalysis with shorter intervals for all datasets caused problems in finding any solution (data not shown). Some transmission events reported in the original studies were not predicted by any method. These isolates are probably genetically unrelated and patients erroneously linked in contact tracing. Another explanation could be mixed infection of the patient were one strain was sampled and sequenced while the other was transmitted.

BadTrIP, outbreaker2, outbreaker, phybreak, SCOTTI, SeqTrack provided prediction of an index patient. Most of the times all methods agree on the index case. BadTrIP predicts a different patient for the NL and US dataset (Table 5.3). The

		all predictions				predictions with high score			
		DE (4)	GB (7)	NL (10)	US (6)	DE (4)	GB (7)	NL (10)	US (6)
correct predictions	BadTrIP	**3**	**7**	**6**	5	**2**	**6**	**6**	5
	Outbreaker	1	4	4	**6**	1	2	2	5
	Outbreaker2	**3**	6	**6**	**6**	1	4	4	**6**
	Phybreak	2	5	2	4	**2**	2	1	1
	SCOTTI	1	5	**6**	5	1	3	**6**	4
	SeqTrack	2	6	5	**6**	1	5	1	5
	TransPhylo	**3**	2	3	2	**2**	2	2	2
false predictions	BadTrIP	21	25	81	25	2	9	24	6
	Outbreaker	11	9	25	3	4	4	10	**0**
	Outbreaker2	5	7	30	2	1	**1**	14	**0**
	Phybreak	**3**	4	13	4	3	3	7	2
	SCOTTI	8	16	87	18	1	**0**	19	5
	SeqTrack	5	**2**	**7**	**0**	**0**	2	**6**	**0**
	TransPhylo	14	14	27	8	6	7	13	6
not detected by any method		0	0	3	0	1	0	4	0
index case	BadTrIP	4155-03	P076	H7	A2				
	Outbreaker2	*	P076	H7	A1				
	Phybreak	7679-03	P076	H7	A1				
	SCOTTI	7679-03	P076	H7	A1				
	SeqTrack	*	P076	H7	A1				

Table 5.3: Comparison of results for all seven methods and four datasets. We count the number of correct predictions based on grouped epidemiological links (correct predictions). Grouped links could not be resolved into clear transmission events with contact tracing data, therefore groups were counted as correctly predicted in case at least one was predicted. Number of correct groups are provided in the header. We counted every single link predicted that was not reported in the epidemiological data of the datasets (false predictions). We also examined results when filtering predicted links by the reported scores or weights. We reported the number of groups that were not detected by any method and predicted index cases. Best results for each dataset are marked in bold.* More than one index case was identified: Outbreaker2: 4155-03, 8073-03, 11114-03, 164-04, 7679-03, 9956-03, 1608-04, 7684-04, 8506-04, 8370-04; SeqTrack: 4155-03, 8073-03, 11114-03, 7679-03, 9956-03

most frequently predicted index cases are 7679-06 for the DE dataset, P076 for the GB dataset, H7 for the NL dataset and A1 for the US dataset and this is in agreement with the epidemiological data. SeqTrack and Outbreaker2 reported several patients as potential infection source for the DE dataset. For this group of patients, only years were provided for onset of TB, resulting in several isolates with the exact same sampling time. The same is true for the GB dataset, however here the earliest sampling time was assigned to a single isolate. Outbreaker2 and SeqTrack seem to be sensitive to this lack of exact time point.

Discussion of Results

We investigated seven transmission chain inference methods on four *M. tuberculosis* outbreak datasets. Among those, outbreaker2 and SeqTrack provided results that fit most closely with contact tracing data. We discovered that sampling times at least at detail level of the month and high quality genetic data are necessary for accurate index case identification and transmission inference. Estimates of generation time and infection intervals could potentially improve results, but the benefit for the resolution of the analyzed datasets was limited.

Epidemiological links will not always reflect the true transmission chain, even if the greatest care is taken in contact tracing (Andrés et al., 2017). Evaluation of transmission detection methods on datasets simulated from real *M. tuberculosis* outbreaks could further strengthen the propositions in this study. Furthermore, only one estimation on generation time and infection time intervals has been investigated for now. For a complete benchmark, also the run time and memory usage of all methods should be recorded.

All discussed methods rely heavily on highly informative, high quality genetic sequences, which might not be available in many cases. Campbell et al. (2019) recently introduced a framework with a novel approach of integrating a full contact model using information on dates of symptom onset, incubation period and generation time alongside WGS pathogen sequences. This model accounts for partially sampled cases and therefore aids outbreak reconstruction in the case of missing links. While the necessary contact model can be inferred from national surveillance systems, availability of full contact data for international surveillance and research projects will be limited due to privacy protection.

In conclusion, we were able to compare the performance of transmission chain inference methods on available genetic and sampling time data. We identified two methods, outbreaker2 and SeqTrack that showed promising results in their applicability to *M. tuberculosis* surveillance and outbreak investigation.

Chapter 6

Summary and Outlook

Summary

Tuberculosis has plagued human populations for millennia and prevention and treatment still pose a major challenge on modern medicine. It is therefore essential to fully investigate and understand *M. tuberculosis*, the bacteria causing this disease, and its spread among humans. Molecular surveillance has enabled great progress in *M. tuberculosis* research and methods are under constant development. Several methods for using molecular typing methods as a complementary approach to contact tracing in disease surveillance have been proposed. WGS based methods currently offer the highest resolution and enable many relevant analyses in parallel. Additionally to measuring SNP distances for transmission cluster detection, sequencing results can be used for classifying lineages and predicting drug resistance (Meehan et al., 2019).

This thesis presents methods to improve next generation sequencing (NGS) based surveillance allowing the collective analysis of more diverse datasets and incorporating information of more than one *M. tuberculosis* reference genome. Chapter 2 introduces seq-seq-pan, that can be used to build a computational pan-genome from several genomes from different lineages and strains. The pan-genome is constructed from sequential whole genome alignment (WGA). To do this, in each step two genomes are aligned at each iterative step with the whole genome aligner progressiveMauve: in the first step two whole genomes are aligned, while in every consecutive step an additional genome is aligned to the consensus sequence of the previous alignment. In the latter case, the WGA of all genomes is reconstructed before the new consensus sequence is determined. ProgressiveMauve provides results separated in locally collinear blocks (LCBs) to resolve genomic rearrangements and large structural variants. Seq-seq-pan postprocesses those LCBs to improve whole genome representation. Consensus sequences of all blocks are concatenated to form the consensus sequence for the whole sequence. The result is a WGA with an accompanying linear consensus sequence. This method is faster than other whole genome alignments and can work with a higher number of sequences. Using the whole genome alignment (WGA), differences between genomes and features of individual sequences can be analyzed. The linear representation of the pan-genome allows its integration in common WGS alignment workflows. Existing information on positional features such as gene annotations or drug resistance associated mutations can be used in subsequent analyses as positions within all included genomes can be translated unto the pan-genome.

A computational pan-genome built of 146 complete genomes from the NCBI RefSeq database with seq-seq-pan forms the basis of single nucleotide polymorphism (SNP) distance based transmission cluster detection with PANPASCO. The combination of using the pan-genome as a reference genome with an alternative distance measure is described in Chapter 3. Previously described methods are using only one reference genome and therefore introduce bias into the analysis. A single reference approach should only be used for well defined groups of samples that are closely related and the best fitting reference genome should be determined for each analysis. However, a group wise approach would limit transferability of results and comparability of studies from different sources. Furthermore, with one type of previously defined distance methods, regions of the genome are excluded in case they are of low quality for a small number of samples. Other approaches attempt to infer those regions using the reference genome. The distance measure used in PANPASCO takes into account all high quality regions for each pair of samples when determining SNP distances, thereby eliminating the need for assigning samples into closely related groups for analysis. For this calculation high quality regions with reference or variant calls are determined with filters on coverage depth, absence of deletions, mapping quality and allele frequency.

This integrative approach is essential for the analysis of large, diverse sets of samples such as the dataset in Chapter 4. To detect potential international transmission we combined WGS data available from NCBI's SRA database with a well defined national set of drug-resistant *M. tuberculosis* isolates from Germany. We filtered the Public dataset based on sequencing depth and quality of mapping. We predicted drug resistance for the Public dataset using resistance associated mutations and combined MDR-TB and XDR-TB isolates with the German dataset for joint analysis. The final dataset included 1339 samples, for which we calculated SNP distances with PANPASCO. As PANPASCO determines SNP distances from pairwise available data, individual samples have the potential to connect two samples that are not closely related due to fluctuation of coverage or a high number of mixed base calls. Therefore, we extended a standard single-linkage agglomerative clustering approach to compare more than two samples before clusters are joined. With this approach we determined

133 clusters made up from 744 isolates where the close relation of isolates within clusters is indicative of transmission. We used available metadata (country of isolation, date of isolation and sample name) to investigate the identified clusters. We discovered that nine clusters were groups of isolates from single patients that were resequenced for investigating emergence of drug resistance or heterogeneity among samples from different sites of the body. These clusters could be distinguished from transmission clusters only with the provided metadata. Among the remaining clusters, we identified links between isolates sampled in different countries in 16 clusters.

To investigate 'who infected whom', we compared seven transmission inference methods on data of four *M. tuberculosis* outbreaks. In Chapter 5, predictions of all methods were compared to available epidemiological data. We identified two methods, SeqTrack and outbreaker2 that outperformed the other methods, providing a higher number of correct predictions and fewer false ones. We established that sampling time points should be reported at least at detail level of the month to ensure optimal prediction results with these methods.

Approachess integrating as much information as possible are necessary for adequate analysis of *M. tuberculosis*. Due to its low mutation rate every detected difference between individual samples greatly improves investigation and reconstruction of transmission routes. To take into account human mobility and accompanying spread of *M. tuberculosis* strains, surveillance programs should opt for WGS based typing methods that incorporate genomic information of all *M. tuberculosis* lineages.

Outlook

Computational Pan-genomics

With the reduction of cost of NGS, the amount of available genetic data increased considerably. This enabled access to a vast amount of data on variation within populations. Therefore, general analysis methods based on single reference genomes will not suffice for the comprehensive analysis of this new information. Any discipline of genetic research, including human genetics, oncology, breeding, microbiology and virology, will benefit from joint and complete analyses of collections of sequences: Pan-genomes have many applications including the analysis of structural variation, population genetics, investigation of highly variable regions of a genome and comparative genomics (Computational Pan-Genomics Consortium, 2018). seq-seq-pan offers one possibility to combine several genomes to be used as reference genome in WGS analyses. This approach is useful in bacteriology where a high number of relatively small and stable complete genomes are available. Alternative methods are needed for highly variable sequences such as viral genomes or much larger sequences such as human genomes. As assemblies of large genomes are rare, WGA based methods cannot be used sufficiently for pan-genome analyses. Several methods for integrating variants detected from sequencing reads into subsequent analyses have been proposed (Valenzuela et al., 2015; Computational Pan-Genomics Consortium, 2018; Garrison et al., 2018) and can be used to improve human population studies. Decades of research have been spent on detecting, evaluating and recording functions of genes, disease related or drug resistance associated mutations, insertion, deletions, structural variants and copy number variations (Brent, 2005). These analyses have been carried out with the single reference paradigm and have to be translated to the field of computational pan-genomics with accessible coordinate or reference systems to allow the community to reach the next milestones in genome analysis.

Molecular Typing

Integrating pan-genomes and enhancing distance measurements by considering all detectable high quality SNP as proposed in Chapter 3, are only the first steps for improving differential analysis of *M. tuberculosis*. Within-host heterogeneity plays a major role in drug resistance treatment and transmission detection (Cohen et al., 2019; Hatherell et al., 2016), but in current methods mixed base calls are often regarded to be of "bad quality" and therefore excluded from analyses (Gardy et al. (2011); Bryant et al. (2013); Kato-Maeda et al. (2013); Pérez-Lago et al. (2013); Roetzer et al. (2013); Walker et al. (2013); Mehaffy et al. (2014); Kohl et al. (2014); Guerra-Assunção et al. (2015); Witney et al. (2015); Gurjav et al. (2016); Fiebig et al. (2017); Kohl et al. (2018b); Jandrasits et al. (in revision)). Detection of samples with mixed infections and establishment of haplotypes from detected variants to differentiate co-infecting strains could improve *M. tuberculosis* surveillance significantly. However, great care must be taken to distinguish mixed base calls caused by heterogeneity from sequencing errors and errors resulting from ambiguous read mapping. Detection and utilisation of within-host microevolution has the potential to resolve transmission chains by differentiating isolates where no difference could be detected before. Within-host heterogeneity should already be considered in the analysis of *M. tuberculosis* before sequencing. Current drug resistance detection as well as NGS methods rely on bacterial culture. However, culture can delay time to drug resistance test results for weeks (Doyle et al., 2018). Furthermore, within-host diversity is reduced as not all strains and clones are equally suited to grow in culture (Nimmo et al., 2019). For this reason, methods for sequencing directly from sputum have been proposed (Doyle et al., 2018; Nimmo et al., 2019). However, implications of drug resistance mutations with low allele frequency have yet to be explored and metagenomics approaches have to be integrated to distinguish *M. tuberculosis* sequences from other sequences in the sputum samples (e.g. from bacteria composing the oral flora).

Traditional molecular typing methods are mostly based on repetitive regions in the genome of *M. tuberculosis*. Integration of information about repeats could potentially improve differential analysis for transmission cluster detection, however common analyses of *second generation sequencing* are not suitable for the detection of repetitive regions. On the contrary, repeats are often masked from reference genomes in common alignment workflows as they increase the ambiguity in NGS mapping results (Sims et al., 2014). *Third generation sequencing* methods that provide long reads with a length of up to several kb could improve repeat analysis and might prove as a valuable complementary approach for SNP based typing methods.

Surveillance

For effective control and surveillance of disease, every single case has to be recorded and integrated into surveillance systems. Even with the cost of WGS constantly decreasing, this approach is feasible only in high-income countries and low-burden settings. However, using genomic relatedness additionally to contact tracing would be much more beneficial in high-burden settings to resolve complex transmission chains and epidemiological links. International collaboration and joint investment is necessary to include cases from low-income, high-burden and also intermediate income and burden countries into global surveillance systems. As there is evidence that drug resistance often emerged in low-income countries and then migrated to low-burden settings, global surveillance could be beneficial also for TB control in high-income countries (Wheeler, 2019). WGS surveillance systems have to be extended or developed to handle challenges arising with high amounts of genetic data. Due to the high incidence of latent infection all data has to be kept for efficient surveillance, as individuals can progress to active disease years after initial infection through reactivation. The risk to develop active tuberculosis is low for individuals with long-established latent infection. However, reactivated cases can make up the majority of cases in settings with high prevalence of latent TB infection or low-burden settings were transmission rates are declining (Ahmad, 2010). In a recent survey among 26 members of the European Union (EU)/European Economic Assocation (EEA), Andrés et al. (2019) identified 20 countries that use molecular typing for TB control, with nine countries among them using WGS based methods. The majority of countries participating in the survey considered the establishment of systems for *M. tuberculosis* typing or to extend existing systems to integrate WGS. The survey investigated barriers for the establishment for molecular surveillance systems. The most named barrier was "financial constraints", followed by "human resources" and "data management and analysis". Also, the majority of survey particpants rated standardization of analysis and sharing of data as highly important (Andrés et al., 2019).

Transmission inference methods based on epidemiological and genetic data could further improve reconstruction of acute and ongoing TB outbreaks. Data available in national surveillance systems can be used to improve models of TB transmission and also take into account reactivated cases more accurately.

Many approaches and pipelines have been established for the analysis of WGS of *M. tuberculosis*, including the one described in this thesis. However, for successful global surveillance storage and analysis methods should be combined and standardized. The recent adaption of ReSeqTB (Starks et al., 2015) by the WHO can be considered as a step in the right direction given that it will be constantly improved and refined with new data and developments.

Appendix

Appendix 1 - Appendix for Chapter 2

Appendix Table 1.1: Information on genomes used in the *Mycobacterium tuberculosis* dataset in Chapter 2.

Accession number	Description	Assembly Accession	ASM Name	Set
NZ_CP016888.1	Mycobacterium tuberculosis strain SCAID 252.0, complete genome	GCF_001708265.1	ASM170826v1	set of 43 genomes
NZ_CP010340.1	Mycobacterium tuberculosis strain 26105, complete genome	GCF_001545055.1	ASM154505v1	set of 43 genomes
NC_020089.1	Mycobacterium tuberculosis 7199-99 complete genome	GCF_000331445.1	ASM33144v1	set of 43 genomes
NC_022350.1	Mycobacterium tuberculosis str. Haarlem, complete genome	GCF_000153685.2	ASM15368v2	set of 43 genomes
NC_009565.1	Mycobacterium tuberculosis F11, complete genome	GCF_000016925.1	ASM1692v1	set of 43 genomes
NZ_CP010338.1	Mycobacterium tuberculosis strain 37004, complete genome	GCF_001544985.1	ASM154498v1	set of 43 genomes
NZ_HG813240.1	Mycobacterium tuberculosis 49-02 complete genome	GCF_000786505.1	MT49-02	set of 43 genomes
NC_021251.1	Mycobacterium tuberculosis CCDC5079, complete genome	GCF_000400615.1	ASM40061v1	set of 43 genomes
NZ_CP002882.1	Mycobacterium tuberculosis BT2, complete genome	GCF_000572155.1	ASM57215v1	set of 43 genomes
NZ_CP002885.1	Mycobacterium tuberculosis CCDC5180, complete genome	GCF_000572195.1	ASM57219v1	set of 43 genomes
NZ_CP010330.1	Mycobacterium tuberculosis strain F28, complete genome	GCF_001544705.1	ASM154470v1	set of 43 genomes
NZ_CP010339.1	Mycobacterium tuberculosis strain 22103, complete genome	GCF_001545015.1	ASM154501v1	set of 43 genomes
NC_018143.2	Mycobacterium tuberculosis H37Rv, complete genome	GCF_000277735.2	ASM27773v2	set of 43 genomes
NC_009525.1	Mycobacterium tuberculosis H37Ra, complete genome	GCF_000016145.1	ASM1614v1	set of 43 genomes
NC_000962.3	Mycobacterium tuberculosis H37Rv, complete genome	GCF_000195955.2	ASM19595v2	set of 43 genomes
NZ_CP009101.1	Mycobacterium tuberculosis strain ZMC13-88, complete genome	GCF_000738475.1	ASM73847v1	set of 43 genomes
NZ_CP009100.1	Mycobacterium tuberculosis strain ZMC13-264, complete genome	GCF_000738445.1	ASM73844v1	set of 43 genomes
NZ_CP007027.1	Mycobacterium tuberculosis H37RvSiena, complete genome	GCF_000827085.1	ASM82708v1	set of 43 genomes
NZ_CP002871.1	Mycobacterium tuberculosis HKBS1, complete genome	GCF_000572125.1	ASM57212v1	set of 43 genomes
NZ_CP007809.1	Mycobacterium tuberculosis strain KIT87190, complete genome	GCF_000706665.1	ASM70666v1	set of 43 genomes
NZ_CP013475.1	Mycobacterium tuberculosis strain 1458, complete genome	GCF_001855255.1	ASM185525v1	set of 43 genomes
NC_002755.2	Mycobacterium tuberculosis CDC1551, complete genome	GCF_000008585.1	ASM858v1	set of 43 genomes
NC_021740.1	Mycobacterium tuberculosis EAI5, complete genome	GCF_000422125.1	ASM42212v1	set of 43 genomes
NZ_CP012506.2	Mycobacterium tuberculosis strain SCAID 187.0, complete genome	GCF_001275565.2	ASM127556v2	set of 43 genomes

NZ_CP012090.1	Mycobacterium tuberculosis W-148, complete genome	GCF_000193185.2	ASM19318v2	set of 43 genomes
NC_017522.1	Mycobacterium tuberculosis CCDC5180, complete genome	GCF_000270365.1	ASM27036v1	set of 43 genomes
NZ_CP009427.1	Mycobacterium tuberculosis strain 96121, complete genome	GCF_000756545.1	ASM75654v1	set of 43 genomes
NZ_AP014573.1	Mycobacterium tuberculosis str. Kurono DNA, complete genome	GCF_000828995.1	ASM82899v1	set of 43 genomes
NC_017524.1	Mycobacterium tuberculosis CTRI-2, complete genome	GCF_000224435.1	ASM22443v1	set of 43 genomes
NC_021194.1	Mycobacterium tuberculosis EAI5/NITR206, complete genome	GCF_000389945.1	ASM38994v1	set of 43 genomes
NC_021054.1	Mycobacterium tuberculosis str. Beijing/NITR203, complete genome	GCF_000364825.1	ASM36482v1	set of 43 genomes
NZ_CP016794.1	Mycobacterium tuberculosis strain SCAID 320.0, complete genome	GCF_001702435.1	ASM170243v1	set of 43 genomes
NC_012943.1	Mycobacterium tuberculosis KZN 1435, complete genome	GCF_000023625.1	ASM2362v1	set of 43 genomes
NC_016768.1	Mycobacterium tuberculosis KZN 4207, complete genome	GCF_000154585.2	ASM15458v2	set of 43 genomes
NC_018078.1	Mycobacterium tuberculosis KZN 605, complete genome	GCF_000154605.2	ASM15460v2	set of 43 genomes
NZ_CP002883.1	Mycobacterium tuberculosis BT1, complete genome	GCF_000572175.1	ASM57217v1	set of 43 genomes
NZ_CP007803.1	Mycobacterium tuberculosis K, complete genome	GCF_000698475.1	ASM69847v1	set of 43 genomes
NZ_CP009426.1	Mycobacterium tuberculosis strain 96075, complete genome	GCF_000756525.1	ASM75652v1	set of 43 genomes
NZ_CP010337.1	Mycobacterium tuberculosis strain 22115, complete genome	GCF_001544955.1	ASM154495v1	set of 43 genomes
NC_020559.1	Mycobacterium tuberculosis str. Erdman = ATCC 35801 DNA, complete genome	GCF_000350205.1	ASM35020v1	set of 43 genomes
NZ_CP009480.1	Mycobacterium tuberculosis H37Rv, complete genome	GCF_000831245.1	ASM83124v1	set of 43 genomes
NZ_CP011510.1	Mycobacterium tuberculosis strain Beijing, complete genome	GCF_001750865.1	ASM175086v1	set of 43 genomes
NZ_CP017920.1	Mycobacterium tuberculosis strain TB282, complete genome	GCF_001870145.1	ASM187014v1	set of 43 genomes
NZ_CP018778.1	Mycobacterium tuberculosis strain DK9897, complete genome	GCF_001922485.1	ASM192248v1	additional genome

Appendix Table 1.2: Information on genomes used in the *Staphylococcus aureus* dataset in Chapter 2.

Accession number	Description	Assembly Accession	ASM Name
NZ_CP007539.1	Staphylococcus aureus strain NRS 100, complete genome	GCF_000626615.1	ASM62661v1
NZ_CP012979.1	Staphylococcus aureus strain ST20130940, complete genome	GCF_001611325.1	ASM161132v1
NC_002951.2	Staphylococcus aureus subsp. aureus COL, complete genome	GCF_000012045.1	ASM1204v1
NZ_CP007672.1	Staphylococcus aureus strain CA12, complete genome	GCF_001045795.2	ASM104579v2
NZ_CP015646.1	Staphylococcus aureus strain 08-02300, complete genome	GCF_001656075.1	ASM165607v1
NZ_CP012978.1	Staphylococcus aureus strain ST20130941, complete genome	GCF_001611345.1	ASM161134v1
NZ_LN831036.1	Staphylococcus aureus genome assembly NCTC13435, chromosome : 1	GCF_001457495.1	NCTC13435
NC_003923.1	Staphylococcus aureus subsp. aureus MW2 DNA, complete genome	GCF_000011265.1	ASM1126v1
NZ_CP013231.1	Staphylococcus aureus strain UTSW MRSA 55, complete sequence	GCF_001580515.1	ASM158051v1
NC_002953.3	Staphylococcus aureus strain MSSA476, complete genome	GCF_000011525.1	ASM1152v1
NZ_CP012972.1	Staphylococcus aureus strain ST20130938, complete genome	GCF_001611405.1	ASM161140v1
NZ_CP012970.1	Staphylococcus aureus strain ST20130939, complete genome	GCF_001611425.1	ASM161142v1
NZ_CP020020.1	Staphylococcus aureus subsp. aureus strain ATCC 6538, complete genome	GCF_002025145.1	ASM202514v1
NZ_AP014942.1	Staphylococcus aureus DNA, complete genome, strain: FDA209P	GCF_001548295.1	ASM154829v1
NZ_CP013132.1	Staphylococcus aureus strain FORC_026, complete genome	GCF_001879545.1	ASM187954v1
NZ_CP007670.1	Staphylococcus aureus strain M121, complete genome	GCF_001021875.1	ASM102187v1
NZ_CP007676.1	Staphylococcus aureus strain HUV05, complete genome	GCF_001045995.2	ASM104599v2
NZ_CP013621.1	Staphylococcus aureus strain RIVM3897, complete genome	GCF_001465755.1	ASM146575v1
NZ_CP011526.1	Staphylococcus aureus subsp. aureus DSM 20231, complete genome	GCF_001027105.1	ASM102710v1
NZ_CP007674.1	Staphylococcus aureus strain CA15, complete genome	GCF_001021895.1	ASM102189v1
NZ_CP007657.1	Staphylococcus aureus strain V2200, complete genome	GCF_001046095.2	ASM104609v2
NZ_CP012974.1	Staphylococcus aureus strain ST20130943, complete genome	GCF_001611385.1	ASM161138v1
NZ_CP012976.1	Staphylococcus aureus strain ST20130942, complete genome	GCF_001611365.1	ASM161136v1
NZ_LT671859.1	Staphylococcus aureus subsp. aureus isolate Clinical isolate genome assembly, chromosome: I	GCF_900129335.1	PP_HGAG_ QV30_2SC_T4
NZ_CP010998.1	Staphylococcus aureus strain FORC_012, complete genome	GCF_001580495.1	ASM158049v1
NC_017763.1	Staphylococcus aureus subsp. aureus HO 5096 0412 complete genome	GCF_000284535.1	ASM28453v1

NZ_CP007176.1	Staphylococcus aureus USA300-ISMMS1, complete genome	GCF_000568455.1	ASM56845v1
NZ_CP007499.1	Staphylococcus aureus strain 2395 USA500, complete genome	GCF_000746505.1	ASM74650v1
NC_010079.1	Staphylococcus aureus subsp. aureus USA300_TCH1516, complete genome	GCF_000017085.1	ASM1708v1
NZ_CP007690.1	Staphylococcus aureus strain UA-S391_USA300, complete genome	GCF_000695875.1	ASM69587v1
NC_007793.1	Staphylococcus aureus subsp. aureus USA300_FPR3757, complete genome	GCF_000013465.1	ASM1346v1
NZ_CP014407.1	Staphylococcus aureus strain USA300-SUR12, complete genome	GCF_002000625.1	ASM200062v1
NZ_CP014432.1	Staphylococcus aureus strain USA300-SUR20, complete genome	GCF_002000785.1	ASM200078v1
NZ_CP014438.1	Staphylococcus aureus strain USA300-SUR22, complete genome	GCF_002000825.1	ASM200082v1
NZ_CP014441.1	Staphylococcus aureus strain USA300-SUR23, complete genome	GCF_002000845.1	ASM200084v1
NZ_CP010298.1	Staphylococcus aureus strain 26b_MRSA, complete genome	GCF_000815165.1	ASM81516v1
NZ_CP014415.1	Staphylococcus aureus strain USA300-SUR15, complete genome	GCF_002000685.1	ASM200068v1
NZ_CP014444.1	Staphylococcus aureus strain USA300-SUR24, complete genome	GCF_002000865.1	ASM200086v1
NZ_LT615218.1	Staphylococcus aureus strain AUS0325 genome assembly, chromosome: 1	GCF_900096745.1	AUS0325
NZ_CP010295.1	Staphylococcus aureus strain 29b_MRSA, complete genome	GCF_000815045.1	ASM81504v1
NZ_CP010296.1	Staphylococcus aureus strain 31b_MRSA, complete genome	GCF_000815085.1	ASM81508v1
NZ_CP010297.1	Staphylococcus aureus strain 33b, complete genome	GCF_000815125.1	ASM81512v1
NZ_CP010299.1	Staphylococcus aureus strain 25b_MRSA, complete genome	GCF_000815205.1	ASM81520v1
NZ_CP010300.1	Staphylococcus aureus strain 27b_MRSA, complete genome	GCF_000815245.1	ASM81524v1
NZ_CP010402.1	Staphylococcus aureus subsp. aureus strain GR2, complete genome	GCF_001296985.1	ASM129698v1
NZ_CP014420.1	Staphylococcus aureus strain USA300-SUR16, complete genome	GCF_002000705.1	ASM200070v1
NZ_CP014423.1	Staphylococcus aureus strain USA300-SUR17, complete genome	GCF_002000725.1	ASM200072v1
NZ_CP014426.1	Staphylococcus aureus strain USA300-SUR18	GCF_002000745.1	ASM200074v1
NZ_CP014429.1	Staphylococcus aureus strain USA300-SUR19, complete genome	GCF_002000765.1	ASM200076v1
NC_017333.1	Staphylococcus aureus subsp. aureus ST398 complete genome	GCF_000009585.1	ASM958v1
NC_013450.1	Staphylococcus aureus subsp. aureus ED98, complete genome	GCF_000024585.1	ASM2458v1
NZ_CP014412.1	Staphylococcus aureus strain USA300-SUR14, complete genome	GCF_002000665.1	ASM200066v1
NC_016912.1	Staphylococcus aureus subsp. aureus VC40, complete genome	GCF_000245495.1	ASM24549v1
NC_017349.1	Staphylococcus aureus subsp. aureus LGA251 complete genome sequence	GCF_000237265.1	ASM23726v1
NZ_CP014392.1	Staphylococcus aureus strain USA300-SUR9, complete genome	GCF_002000565.1	ASM200056v1
NZ_CP014397.1	Staphylococcus aureus strain USA300-SUR10, complete genome	GCF_002000585.1	ASM200058v1
NZ_CP014402.1	Staphylococcus aureus strain USA300-SUR11, complete genome	GCF_002000605.1	ASM200060v1
NZ_CP014409.1	Staphylococcus aureus strain USA300-SUR13, complete genome	GCF_002000645.1	ASM200064v1
NZ_CP014435.1	Staphylococcus aureus strain USA300-SUR21, complete genome	GCF_002000805.1	ASM200080v1
NC_009641.1	Staphylococcus aureus subsp. aureus str. Newman DNA, complete genome	GCF_000010465.1	ASM1046v1
NZ_LT598688.1	Staphylococcus aureus isolate Sa_Newman_UoM genome assembly, chromosome: 1	GCF_900092595.1	Sa_Newman_UoM
NZ_CP007659.1	Staphylococcus aureus subsp. aureus strain H-EMRSA-15, complete genome	GCF_000695215.1	ASM69521v1
NZ_CP013619.1	Staphylococcus aureus strain RIVM1607, complete genome	GCF_001465675.1	ASM146567v1
NZ_CP012756.1	Staphylococcus aureus subsp. aureus strain JS395, complete genome	GCF_001307235.1	ASM130723v1
NZ_CP009361.1	Staphylococcus aureus subsp. aureus strain ATCC 25923, complete genome	GCF_000756205.1	ASM75620v1
NZ_CP013616.1	Staphylococcus aureus strain RIVM1295, complete genome	GCF_001465635.1	ASM146563v1
NC_018608.1	Staphylococcus aureus 08BA02176, complete genome	GCF_000296595.1	ASM29659v1
NC_007795.1	Staphylococcus aureus subsp. aureus NCTC 8325 chromosome, complete genome	GCF_000013425.1	ASM1342v1
NZ_CP020019.1	Staphylococcus aureus strain 08S00974, complete genome	GCF_002025125.1	ASM202512v1
NZ_CP010890.1	Staphylococcus aureus strain SA564, complete genome	GCF_001281145.1	ASM128114v1
NZ_CP013218.1	Staphylococcus aureus subsp. aureus strain LA-MRSA ST398 isolate E154, complete genome	GCF_001887075.1	ASM188707v1
NC_017351.1	Staphylococcus aureus subsp. aureus 11819-97, complete genome	GCF_000239235.1	ASM23923v1
NC_021554.1	Staphylococcus aureus CA-347, complete genome	GCF_000412775.1	ASM41277v1
NC_009487.1	Staphylococcus aureus subsp. aureus JH9, complete genome	GCF_000016805.1	ASM1680v1
NC_009632.1	Staphylococcus aureus subsp. aureus JH1, complete genome	GCF_000017125.1	ASM1712v1
NZ_CP014064.1	Staphylococcus aureus strain FDAARGOS_159, complete genome	GCF_001558795.1	ASM155879v1
NZ_CP018768.1	Staphylococcus aureus subsp. aureus strain UCI 28, complete genome	GCF_001975045.1	ASM197504v1
NZ_CP019117.1	Staphylococcus aureus strain SJTUF_J27, complete genome	GCF_001956755.1	ASM195675v1
NZ_AP014652.1	Staphylococcus aureus subsp. aureus DNA, complete genome, strain: TMUS2126	GCF_001549655.1	ASM154965v1
NZ_AP014653.1	Staphylococcus aureus subsp. aureus DNA, complete genome, strain: TMUS2134	GCF_001549675.1	ASM154967v1
NZ_CP009423.1	Staphylococcus aureus subsp. aureus strain USA300_SUR1, complete genome	GCF_001956815.1	ASM195681v1
NZ_CP012593.1	Staphylococcus aureus strain HOU1444-VR, complete genome	GCF_001278745.1	ASM127874v1
NZ_CP013182.1	Staphylococcus aureus strain SA40TW, complete genome	GCF_001880265.1	ASM188026v1
NZ_CP018205.1	Staphylococcus aureus subsp. aureus strain HG001, complete genome	GCF_001900185.1	ASM190018v1
NZ_CP018766.1	Staphylococcus aureus strain UCI62, complete genome	GCF_001975005.1	ASM197500v1
NZ_CP014791.1	Staphylococcus aureus strain MCRF184, complete genome	GCF_001594205.1	ASM159420v1
NC_017340.1	Staphylococcus aureus 04-02981, complete genome	GCF_000025145.1	ASM2514v1
NZ_CP010526.1	Staphylococcus aureus subsp. aureus ST772-MRSA-V strain DAR4145, complete genome	GCF_000828035.1	ASM82803v1
NZ_CP012015.1	Staphylococcus aureus subsp. aureus strain Gv51, complete genome	GCF_001515665.1	ASM151566v1
NC_022226.1	Staphylococcus aureus subsp. aureus CN1, complete genome	GCF_000463055.1	ASM46305v1

NC_017341.1	Staphylococcus aureus subsp. aureus str. JKD6008, complete genome	GCF_000145595.1	ASM14559v1
NC_017343.1	Staphylococcus aureus subsp. aureus ECT-R 2 complete genome	GCF_000253135.1	ASM25313v1
NZ_CP011528.1	Staphylococcus aureus strain RKI4, complete genome	GCF_001027045.1	ASM102704v1
NZ_CP012013.1	Staphylococcus aureus subsp. aureus strain Be62, complete genome	GCF_001515685.1	ASM151568v1
NZ_CP012018.1	Staphylococcus aureus subsp. aureus strain Gv88, complete genome	GCF_001515705.1	ASM151570v1
NZ_CP012011.1	Staphylococcus aureus subsp. aureus strain HC1340, complete genome	GCF_001515745.1	ASM151574v1
NC_016928.1	Staphylococcus aureus subsp. aureus M013, complete genome	GCF_000237125.1	ASM23712v1
NC_022442.1	Staphylococcus aureus subsp. aureus SA957, complete genome	GCF_000470845.1	ASM47084v1
NC_022443.1	Staphylococcus aureus subsp. aureus SA40, complete genome	GCF_000470865.1	ASM47086v1
NC_017338.1	Staphylococcus aureus subsp. aureus JKD6159, complete genome	GCF_000144955.1	ASM14495v1
NZ_CP012409.1	Staphylococcus aureus subsp. aureus Tager 104, complete genome	GCF_000452385.2	ASM45238v2
NZ_CP012120.1	Staphylococcus aureus subsp. aureus strain USA300_2014.C02, complete genome	GCF_001183725.1	ASM118372v1
NZ_AP017320.1	Staphylococcus aureus DNA, complete genome, strain: MI	GCF_001548415.1	ASM154841v1
NZ_CP015645.1	Staphylococcus aureus strain 08-02119, complete genome	GCF_001656045.1	ASM165604v1
NZ_LT009690.1	Staphylococcus aureus strain NZAK3 genome assembly, chromosome: 1	GCF_900017775.1	NZAK3
NC_002952.2	Staphylococcus aureus subsp. aureus strain MRSA252, complete genome	GCF_000011505.1	ASM1150v1
NZ_CP009554.1	Staphylococcus aureus subsp. aureus strain FORC_001, complete genome	GCF_000772025.1	ASM77202v1
NZ_CP019563.1	Staphylococcus aureus strain SR434, complete genome	GCF_001986135.1	ASM198613v1
NZ_CP007454.1	Staphylococcus aureus strain 502A, complete genome	GCF_000597965.1	ASM59796v1
NZ_CP006630.1	Staphylococcus aureus subsp. aureus SA268, complete genome	GCF_000737615.1	ASM73761v1
NZ_CP012119.1	Staphylococcus aureus subsp. aureus strain USA300_2014.C01, complete genome	GCF_001183705.1	ASM118370v1
NZ_CP012012.1	Staphylococcus aureus subsp. aureus strain HC1335, complete genome	GCF_001515765.1	ASM151576v1
NC_017342.1	Staphylococcus aureus subsp. aureus TCH60, complete genome	GCF_000159535.2	ASM15953v2
NC_017337.1	Staphylococcus aureus subsp. aureus ED133, complete genome	GCF_000210315.1	ASM21031v1
NC_020533.1	Staphylococcus aureus subsp. aureus ST228 complete genome, isolate 16035	GCF_000382965.1	ASM38296v1
NC_020566.1	Staphylococcus aureus subsp. aureus ST228 complete genome, isolate 16125	GCF_000382985.1	ASM38298v1
NC_020568.1	Staphylococcus aureus subsp. aureus ST228 complete genome, isolate 18583	GCF_000383005.1	ASM38300v1
NC_020529.1	Staphylococcus aureus subsp. aureus ST228 complete genome, isolate 10388	GCF_000967325.1	ASM96732v1
NC_020564.1	Staphylococcus aureus subsp. aureus ST228 complete genome, isolate 10497	GCF_000967345.1	ASM96734v1
NC_020532.1	Staphylococcus aureus subsp. aureus ST228 complete genome, isolate 15532	GCF_000967365.1	ASM96736v1
NC_020536.1	Staphylococcus aureus subsp. aureus ST228 complete genome, isolate 18341	GCF_000967385.1	ASM96738v1
NC_020537.1	Staphylococcus aureus subsp. aureus ST228 complete genome, isolate 18412	GCF_000967405.1	ASM96740v1
NC_007622.1	Staphylococcus aureus RF122 complete genome	GCF_000009005.1	ASM900v1
NC_002745.2	Staphylococcus aureus subsp. aureus N315 DNA, complete genome	GCF_000009645.1	ASM964v1
NC_002758.2	Staphylococcus aureus subsp. aureus Mu50 DNA, complete genome	GCF_000009665.1	ASM966v1
NC_009782.1	Staphylococcus aureus subsp. aureus Mu3 DNA, complete genome	GCF_000010445.1	ASM1044v1
NC_017331.1	Staphylococcus aureus subsp. aureus TW20, complete genome	GCF_000027045.1	ASM2704v1
NC_022113.1	Staphylococcus aureus subsp. aureus 55/2053, complete genome	GCF_000160335.2	ASM16033v2
NC_017347.1	Staphylococcus aureus subsp. aureus T0131, complete genome	GCF_000204665.1	ASM20466v1
NC_021670.1	Staphylococcus aureus Bmb9393, complete genome	GCF_000418345.1	ASM41834v1
NC_022222.1	Staphylococcus aureus subsp. aureus 6850, complete genome	GCF_000462955.1	ASM46295v1
NC_022604.1	Staphylococcus aureus subsp. aureus Z172, complete genome	GCF_000485885.1	ASM48588v1
NZ_CP009681.1	Staphylococcus aureus subsp. aureus strain Gv69, complete genome	GCF_000769575.1	ASM76957v1
NZ_LN626917.1	Staphylococcus aureus genome assembly Staphylococcus_aureus ILRI_Eymole1/1, chromosome : I	GCF_000953255.1	Staphylococcus_aureus_Sa_ILRI_217
NZ_CP011147.1	Staphylococcus aureus strain FCFHV36, complete genome	GCF_000969225.1	ASM96922v1
NZ_CP013137.1	Staphylococcus aureus strain XQ, complete genome	GCF_001444345.1	ASM144434v1
NZ_CP009828.1	Staphylococcus aureus strain MS4, complete genome	GCF_001456215.1	ASM145621v1
NZ_CP015173.1	Staphylococcus aureus strain RIVM6519, complete genome	GCF_001618305.1	ASM161830v1
NZ_CP011685.1	Staphylococcus aureus strain ZJ5499, complete genome	GCF_001640885.1	ASM164088v1
NZ_CP013953.1	Staphylococcus aureus strain NCCP14558, complete genome	GCF_001640905.1	ASM164090v1
NZ_CP013955.1	Staphylococcus aureus strain NCCP14562, complete genome	GCF_001640925.1	ASM164092v1
NZ_CP013957.1	Staphylococcus aureus strain V521, complete genome	GCF_001641025.1	ASM164102v1
NZ_CP012692.1	Staphylococcus aureus strain FORC_027, complete genome	GCF_001725965.1	ASM172596v1
NZ_LN854556.1	Staphylococcus aureus strain BB155 genome assembly, chromosome: 1	GCF_900004855.1	BB155

Appendix Table 1.3: Information on genomes used in the *Escherichia coli* dataset in Chapter 2.

Accession number	Description	Assembly Accession	ASM Name
NZ_CP009859.1	Escherichia coli strain ECONIH1, complete genome	GCF_000784925.1	ASM78492v1
NZ_CP015241.1	Escherichia coli strain 2013C-4465, complete genome	GCF_001644745.1	ASM164474v1
NZ_CP008957.1	Escherichia coli O157:H7 str. EDL933, complete genome	GCF_000732965.1	ASM73296v1
NZ_CP016358.1	Escherichia coli strain K-15KW01, complete genome	GCF_001683435.1	ASM168343v1
NC_017656.1	Escherichia coli O55:H7 str. RM12579, complete genome	GCF_000245515.1	ASM24551v1
NZ_CP015831.1	Escherichia coli O157 strain 644-PT8, complete genome	GCF_001650295.1	ASM165029v1
NZ_CP008805.1	Escherichia coli O157:H7 str. SS17, complete genome	GCF_000730345.1	ASM73034v1
NC_013008.1	Escherichia coli O157:H7 str. TW14359, complete genome	GCF_000022225.1	ASM2222v1
NC_011353.1	Escherichia coli O157:H7 str. EC4115, complete genome	GCF_000021125.1	ASM2112v1
NZ_CP010304.1	Escherichia coli O157:H7 str. SS52, complete genome	GCF_000803705.1	ASM80370v1
NZ_CP017249.1	Escherichia coli strain NADC 5570/86-24/6565 isolate mutant, complete genome	GCF_001806265.1	ASM180626v1
NZ_CP017251.1	Escherichia coli strain NADC 5570/86-24/6564, complete sequence	GCF_001806285.1	ASM180628v1
NC_017906.1	Escherichia coli Xuzhou21, complete genome	GCF_000262125.1	ASM26212v1
NZ_CP014314.1	Escherichia coli O157:H7 strain JEONG-1266, complete genome	GCF_001558995.2	ASM155899v2
NZ_CP015846.1	Escherichia coli O157:H7 strain FRIK2069, complete genome	GCF_001651925.1	ASM165192v1
NC_013941.1	Escherichia coli O55:H7 str. CB9615, complete genome	GCF_000025165.1	ASM2516v1
NC_002695.1	Escherichia coli O157:H7 str. Sakai, complete genome	GCF_000008865.1	ASM886v1
NZ_CP015842.1	Escherichia coli O157:H7 strain FRIK2533, complete genome	GCF_001651945.1	ASM165194v1
NZ_CP014667.1	Escherichia coli strain ECONIH2, complete genome	GCF_001675145.1	ASM167514v1
NZ_CP015843.1	Escherichia coli O157:H7 strain FRIK2455, complete genome	GCF_001651965.1	ASM165196v1
NC_004431.1	Escherichia coli CFT073, complete genome	GCF_000007445.1	ASM744v1
NZ_CP014670.1	Escherichia coli strain CFSAN004177, complete genome	GCF_001721205.1	ASM172120v1
NC_010498.1	Escherichia coli SMS-3-5, complete genome	GCF_000019645.1	ASM1964v1
NZ_CP015834.1	Escherichia coli strain MS6198, complete genome	GCF_001721525.1	ASM172152v1
NZ_CP015023.1	Escherichia coli strain SRCC 1675, complete genome	GCF_001612495.1	ASM161249v1
NZ_CP017434.1	Escherichia coli O157:H7 strain 1130, complete genome	GCF_001753445.1	ASM175344v1
NZ_CP017444.1	Escherichia coli O157:H7 strain 8368, complete genome	GCF_001753485.1	ASM175348v1
NZ_CP017438.1	Escherichia coli O157:H7 strain 2159, complete genome	GCF_001753505.1	ASM175350v1
NZ_CP017446.1	Escherichia coli O157:H7 strain 9234, complete genome	GCF_001753525.1	ASM175352v1
NZ_CP017436.1	Escherichia coli O157:H7 strain 2149, complete genome	GCF_001753465.1	ASM175346v1
NZ_CP017440.1	Escherichia coli O157:H7 strain 3384, complete genome	GCF_001753545.1	ASM175354v1
NZ_CP017442.1	Escherichia coli O157:H7 strain 4276, complete genome	GCF_001753565.1	ASM175356v1
NZ_CP016625.1	Escherichia coli O157:H7 strain FRIK944, complete genome	GCF_001695515.1	ASM169551v1
NZ_CP014583.1	Escherichia coli strain CFSAN004176, complete genome	GCF_001721225.1	ASM172122v1
NC_017646.1	Escherichia coli O7:K1 str. CE10, complete genome	GCF_000227625.1	ASM22762v1
NZ_CP012802.1	Escherichia coli O157:H7 strain WS4202, complete genome	GCF_001307215.1	ASM130721v1
NZ_CP017669.1	Escherichia coli strain PA20, complete genome	GCF_001865295.1	ASM186529v1
NC_017626.1	Escherichia coli 042 complete genome	GCF_000027125.1	ASM2712v1
NZ_CP007799.1	Escherichia coli Nissle 1917, complete genome	GCF_000714595.1	ASM71459v1
NZ_CP015020.1	Escherichia coli strain 28RC1, complete genome	GCF_001612475.1	ASM161247v1
NZ_CP015229.1	Escherichia coli strain 06-00048, complete genome	GCF_001677475.1	ASM167747v1
NC_008563.1	Escherichia coli APEC O1, complete genome	GCF_000014845.1	ASM1484v1
NZ_CP015832.1	Escherichia coli O157 strain 180-PT54, complete genome	GCF_001650275.1	ASM165027v1
NZ_CP016497.1	Escherichia coli strain UPEC 26-1, complete genome	GCF_001693315.1	ASM169331v1
NZ_CP008697.1	Escherichia coli strain ST648, complete genome	GCF_001485455.1	ASM148545v1
NC_008253.1	Escherichia coli 536, complete genome	GCF_000013305.1	ASM1330v1
NZ_CP009072.1	Escherichia coli ATCC 25922, complete genome	GCF_000743255.1	ASM74325v1
NC_013364.1	Escherichia coli O111:H- str. 11128 DNA, complete genome	GCF_000010765.1	ASM1076v1
NC_017651.1	Escherichia coli str. ’clone D i2’, complete genome	GCF_000233875.1	ASM23387v1
NC_017652.1	Escherichia coli str. ’clone D i14’, complete genome	GCF_000233895.1	ASM23389v1
NZ_CP007592.1	Escherichia coli O157:H16 strain Santai, complete genome	GCF_000827105.1	ASM82710v1
NC_017631.1	Escherichia coli ABU 83972, complete genome	GCF_000148365.1	ASM14836v1
NC_011601.1	Escherichia coli 0127:H6 E2348/69 complete genome, strain E2348/69	GCF_000026545.1	ASM2654v1
NZ_CP013112.1	Escherichia coli strain YD786, complete genome	GCF_001442495.1	ASM144249v1
NZ_CP012693.1	Escherichia coli strain FORC_028, complete genome	GCF_001721125.1	ASM172112v1
NC_007946.1	Escherichia coli UTI89, complete genome	GCF_000013265.1	ASM1326v1
NC_013361.1	Escherichia coli O26:H11 str. 11368 DNA, complete genome	GCF_000091005.1	ASM9100v1
NC_013353.1	Escherichia coli O103:H2 str. 12009 DNA, complete genome	GCF_000010745.1	ASM1074v1
NC_009801.1	Escherichia coli E24377A, complete genome	GCF_000017745.1	ASM1774v1
NZ_CP014495.1	Escherichia coli strain SaT040, complete geome	GCF_001566615.1	ASM156661v1
NZ_CP006027.1	Escherichia coli O145:H28 str. RM13514, complete genome	GCF_000520035.1	ASM52003v1
NZ_CP007136.1	Escherichia coli O145:H28 str. RM12581, complete genome	GCF_000671295.1	ASM67129v1
NZ_CP007392.1	Escherichia coli strain ST2747, complete genome	GCF_000599665.1	ASM59966v1
NZ_CP007149.1	Escherichia coli RS218, complete genome	GCF_000800845.1	ASM80084v2
NC_011750.1	Escherichia coli IAI39 chromosome, complete genome	GCF_000026345.1	ASM2634v1
NZ_CP006262.1	Escherichia coli O145:H28 str. RM13516, complete genome	GCF_000520055.1	ASM52005v1
NZ_CP007133.1	Escherichia coli O145:H28 str. RM12761, complete genome	GCF_000662395.1	ASM66239v1
NZ_CP007393.1	Escherichia coli strain ST2747, complete genome	GCF_000599685.1	ASM59968v1
NC_011748.1	Escherichia coli 55989 chromosome, complete genome	GCF_000026245.1	ASM2624v1
NZ_CP012625.1	Escherichia coli strain SF-468, complete genome	GCF_001280345.1	ASM128034v1

NZ_CP015228.1	Escherichia coli strain 09-00049, complete genome	GCF_001677495.1	ASM167749v1
NZ_CP016007.1	Escherichia coli strain NGF1, complete genome	GCF_001660585.1	ASM166058v1
NZ_CP005930.1	Escherichia coli APEC IMT5155, complete genome	GCF_000813165.1	ASM81316v1
NZ_CP012631.1	Escherichia coli strain SF-173, complete genome	GCF_001280405.1	ASM128040v1
NZ_CP012635.1	Escherichia coli strain SF-088, complete genome	GCF_001280325.1	ASM128032v1
NZ_CP012633.1	Escherichia coli strain SF-166, complete genome	GCF_001280385.1	ASM128038v1
NC_017632.1	Escherichia coli UM146, complete genome	GCF_000148605.1	ASM14860v1
NZ_CP016546.1	Escherichia coli strain O177:H21, complete genome	GCF_001693635.1	ASM169363v1
NZ_CP013029.1	Escherichia coli strain 2012C-4227, complete genome	GCF_001420935.1	ASM142093v1
NC_017628.1	Escherichia coli IHE3034, complete genome	GCF_000025745.1	ASM2574v1
NC_017634.1	Escherichia coli O83:H1 str. NRG 857C chromosome, complete genome	GCF_000183345.1	ASM18334v1
NZ_CP014488.1	Escherichia coli strain G749, complete genome	GCF_001566635.1	ASM156663v1
NZ_LT601384.1	Escherichia coli isolate NCTC86EC genome assembly, chromosome: I	GCF_900092615.1	PRJEB14041
NC_011993.1	Escherichia coli LF82 chromosome, complete sequence	GCF_000284495.1	ASM28449v1
NC_020163.1	Escherichia coli APEC O78, complete genome	GCF_000332755.1	ASM33275v1
NZ_CP007442.1	Escherichia coli ACN001, complete genome	GCF_001051135.1	ASM105113v1
NZ_CP009106.2	Escherichia coli strain 94-3024, complete genome	GCF_000801185.2	ASM80118v2
NZ_CP007491.1	Escherichia coli strain ACN002, complete genome	GCF_001515725.1	ASM151572v1
NZ_CP014522.1	Escherichia coli strain ZH063, complete genome	GCF_001577325.1	ASM157732v1
NZ_CP006632.1	Escherichia coli PCN033, complete genome	GCF_000219515.2	ASM21951v3
NZ_HF572917.1	Escherichia coli HUSEC2011 complete genome	GCF_000967155.2	HUSEC2011 CHR1
NC_017641.1	Escherichia coli UMNK88, complete genome	GCF_000212715.2	ASM21271v2
NC_018661.1	Escherichia coli O104:H4 str. 2009EL-2071, complete genome	GCF_000299475.1	ASM29947v1
NZ_CP011331.1	Escherichia coli O104:H4 str. C227-11, complete genome	GCF_000986765.1	ASM98676v1
NC_018658.1	Escherichia coli O104:H4 str. 2011C-3493 chromosome, complete genome	GCF_000299455.1	ASM29945v1
NZ_CP015069.1	Escherichia coli strain Ecol_743, complete genome	GCF_001618325.1	ASM161832v1
NZ_CP009166.1	Escherichia coli 1303, complete genome	GCF_000829985.1	ASM82998v1
NZ_CP015076.1	Escherichia coli strain Ecol_448, complete genome	GCF_001618365.1	ASM161836v1
NZ_CP013663.1	Escherichia coli strain GB089, complete genome	GCF_001678965.1	ASM167896v1
NZ_CP015159.1	Escherichia coli strain Eco889, complete genome	GCF_001663475.1	ASM166347v1
NC_017633.1	Escherichia coli ETEC H10407, complete genome	GCF_000210475.1	ASM21047v1
NZ_CP007594.1	Escherichia coli strain SEC470, complete genome	GCF_000987875.1	ASM98787v1
NZ_HG941718.1	Escherichia coli ST131 strain EC958 chromosome, complete genome	GCF_000285655.3	EC958.v1
NC_018650.1	Escherichia coli O104:H4 str. 2009EL-2050, complete genome	GCF_000299255.1	ASM29925v1
NZ_CP011061.1	Escherichia coli str. Sanji, complete genome	GCF_001610755.1	ASM161075v1
NZ_CP013662.1	Escherichia coli strain 08-00022, complete genome	GCF_001677515.1	ASM167751v1
NZ_CP015995.1	Escherichia coli strain S51, complete genome	GCF_001660565.1	ASM166056v1
NC_013654.1	Escherichia coli SE15 DNA, complete genome	GCF_000010485.1	ASM1048v1
NZ_CP010876.1	Escherichia coli strain MNCRE44, complete genome	GCF_000931565.1	ASM93156v1
NZ_CP013025.1	Escherichia coli strain 2009C-3133, complete genome	GCF_001420955.1	ASM142095v1
NZ_CP013835.1	Escherichia coli strain JJ2434, complete genome	GCF_001513635.1	ASM151363v1
NZ_CP013658.1	Escherichia coli strain uk_P46212, complete sequence	GCF_001469815.1	ASM146981v1
NZ_CP011416.1	Escherichia coli strain CFSAN029787, complete genome	GCF_001007915.1	ASM100791v1
NZ_CP015138.1	Escherichia coli strain Ecol_732, complete genome	GCF_001617565.1	ASM161756v1
NZ_CP009104.1	Escherichia coli strain RM9387, complete genome	GCF_000801165.1	ASM80116v1
NZ_CP014316.1	Escherichia coli JJ1887, complete genome	GCF_001593565.1	ASM159356v1
NZ_CP015074.2	Escherichia coli strain Ecol_745, complete genome	GCF_001618345.2	ASM161834v2
NZ_CP013190.1	Escherichia coli strain FORC_031, complete genome	GCF_001750845.1	ASM175084v1
NZ_CP016628.1	Escherichia coli strain FORC_041, complete genome	GCF_001886935.1	ASM188693v1
NZ_CP014497.1	Escherichia coli strain ZH193, complete genome	GCF_001566675.1	ASM156667v1
NZ_CP007394.1	Escherichia coli strain ST2747, complete genome	GCF_000599705.1	ASM59970v1
NC_022648.1	Escherichia coli JJ1886, complete genome	GCF_000493755.1	ASM49375v1
NZ_CP008801.1	Escherichia coli KLY, complete genome	GCF_000725305.1	ASM72530v1
NZ_CP010315.1	Escherichia coli strain 789, complete genome	GCF_000819645.1	ASM81964v1
NZ_CP011495.1	Escherichia coli strain NCM3722, complete genome	GCF_001043215.1	ASM104321v1
NZ_CP014492.1	Escherichia coli strain MVAST0167, complete genome	GCF_001566655.1	ASM156665v1
NZ_CP013031.1	Escherichia coli strain H1827/12, complete genome	GCF_001678925.1	ASM167892v1
NZ_CP010344.1	Escherichia coli ECC-1470, complete genome	GCF_000831565.1	ASM83156v1
NZ_CP011018.1	Escherichia coli strain CI5, complete genome	GCF_000971615.1	ASM97161v1
NZ_CP014348.1	Escherichia coli str. K-12 substr. MG1655 strain JW5437-1, complete genome	GCF_001566335.1	ASM156633v1
NZ_CP015912.1	Escherichia coli strain 210205630, complete genome	GCF_001679985.1	ASM167998v1
NZ_CP014270.1	Escherichia coli K-12 strain DHB4, complete genome	GCF_001559655.1	ASM155965v1
NC_011415.1	Escherichia coli SE11 DNA, complete genome	GCF_000010385.1	ASM1038v1
NC_012967.1	Escherichia coli B str. REL606, complete genome	GCF_000017985.1	ASM1798v1
NZ_CP010445.1	Escherichia coli K-12 strain ER3435, complete genome	GCF_000974885.1	ASM97488v1
NZ_CP011324.1	Escherichia coli strain SQ2203, complete genome	GCF_000988465.1	ASM98846v1
NZ_CP011321.1	Escherichia coli strain SQ88, complete genome	GCF_000988385.1	ASM98838v1
NZ_CP013831.1	Escherichia coli strain CD306, complete genome	GCF_001513615.1	ASM151361v1
NZ_CP014225.1	Escherichia coli str. K-12 substr. MG1655, complete genome	GCF_001544635.1	ASM154463v1
NC_010473.1	Escherichia coli str. K12 substr. DH10B, complete genome	GCF_000019425.1	ASM1942v1
NZ_CP009273.1	Escherichia coli BW25113, complete genome	GCF_000750555.1	ASM75055v1
NZ_CP011134.1	Escherichia coli VR50, complete genome	GCF_000968515.1	ASM96851v1
NZ_CP010442.1	Escherichia coli K-12 strain ER3466, complete genome	GCF_000974575.1	ASM97457v1
NC_011741.1	Escherichia coli IAI1 chromosome, complete genome	GCF_000026265.1	ASM2626v1

NZ_CP011320.1	Escherichia coli strain SQ37, complete genome	GCF_000988355.1	ASM98835v1
NZ_CP017100.1	Escherichia coli strain K-12 NEB 5-alpha, complete genome	GCF_001723505.1	ASM172350v1
NC_000913.3	Escherichia coli str. K-12 substr. MG1655, complete genome	GCF_000005845.2	ASM584v2
NC_017635.1	Escherichia coli W, complete genome	GCF_000184185.1	ASM18418v1
NZ_CP010438.1	Escherichia coli K-12 strain ER3454, complete genome	GCF_000974405.1	ASM97440v1
NZ_CP010440.1	Escherichia coli K-12 strain ER3476, complete genome	GCF_000974505.1	ASM97450v1
NZ_CP013253.1	Escherichia coli strain CQSW20, complete genome	GCF_001455385.1	ASM145538v1
NZ_CP014272.1	Escherichia coli K-12 strain C3026, complete genome	GCF_001559675.1	ASM155967v1
NC_012759.1	Escherichia coli BW2952, complete genome	GCF_000022345.1	ASM2234v1
NC_016902.1	Escherichia coli KO11, complete genome	GCF_000147855.2	ASM14785v3
NC_017664.1	Escherichia coli W, complete genome	GCF_000258145.1	ASM25814v1
NZ_CP009685.1	Escherichia coli str. K-12 substr. MG1655, complete genome	GCF_000801205.1	ASM80120v1
NZ_LM995446.1	Escherichia coli genome assembly EcRV308Chr, chromosome : 1	GCF_000952955.1	EcRV308Chr
NZ_CP012868.1	Escherichia coli str. K-12 substr. MG1655, complete genome	GCF_001308065.1	ASM130806v1
NZ_CP012869.1	Escherichia coli strain K-12 substrain MG1655_TMP32XR1, complete genome	GCF_001308125.1	ASM130812v1
NZ_CP012870.1	Escherichia coli strain K-12 substrain MG1655_TMP32XR2, complete genome	GCF_001308165.1	ASM130816v1
NC_007779.1	Escherichia coli str. K-12 substr. W3110 DNA, complete genome	GCF_000010245.2	ASM1024v1
NZ_CP010816.1	Escherichia coli strain BL21 (TaKaRa), complete genome	GCF_000833145.1	ASM83314v1
NZ_CP010439.1	Escherichia coli K-12 strain ER3440, complete genome	GCF_000974465.1	ASM97446v1
NZ_CP010441.1	Escherichia coli K-12 strain ER3445, complete genome	GCF_000974535.1	ASM97453v1
NZ_CP010443.1	Escherichia coli K-12 strain ER3446, complete genome	GCF_000974825.1	ASM97482v1
NZ_CP010444.1	Escherichia coli K-12 strain ER3475, complete genome	GCF_000974865.1	ASM97486v1
NZ_CP011342.2	Escherichia coli K-12 GM4792 Lac+, complete genome	GCF_001020945.2	ASM102094v2
NZ_CP011343.2	Escherichia coli K-12 GM4792 Lac-, complete genome	GCF_001021005.2	ASM102100v2
NZ_CP010371.1	Escherichia coli strain 6409, complete genome	GCF_000814145.2	ASM81414v2
NC_017625.1	Escherichia coli DH1, complete genome	GCF_000023365.1	ASM2336v1
NC_022364.1	Escherichia coli LY180, complete genome	GCF_000468515.1	ASM46851v1
NZ_CP009644.1	Escherichia coli ER2796, complete genome	GCF_000800215.1	ASM80021v1
NZ_CP009789.1	Escherichia coli K-12 strain ER3413, complete genome	GCF_000800765.1	ASM80076v1
NZ_LN832404.1	Escherichia coli K-12 genome assembly EcoliK12AG100, chromosome : I	GCF_000981485.1	EcoliK12AG100
NZ_CP011113.2	Escherichia coli strain RR1, complete genome	GCF_001276585.2	ASM127658v2
NZ_CP013483.1	Escherichia coli strain Y5, complete genome	GCF_001860505.1	ASM186050v1
NZ_CP018115.1	Escherichia coli strain MRSN346638, complete genome	GCF_001886555.1	ASM188655v1
NC_012892.2	Escherichia coli BL21(DE3), complete genome	GCF_000009565.1	ASM956v1
NC_009800.1	Escherichia coli HS, complete genome	GCF_000017765.1	ASM1776v1
NC_012971.2	Escherichia coli BL21(DE3), complete genome	GCF_000022665.1	ASM2266v1
NC_017638.1	Escherichia coli DH1 (ME8569) DNA, complete genome	GCF_000270105.1	ASM27010v1
NZ_CP010585.1	Escherichia coli strain C41(DE3), complete genome	GCF_000830035.1	ASM83003v1
NZ_LM993812.1	Escherichia coli genome assembly EcHMS174Chr, chromosome : 1	GCF_000953515.1	EcHMS174Chr
NZ_CP006636.1	Escherichia coli PCN061, complete genome	GCF_001029125.1	ASM102912v1
NZ_CP011938.1	Escherichia coli strain C43(DE3), complete genome	GCF_001039415.1	ASM103941v1
NZ_CP012125.1	Escherichia coli strain DH1Ec095, complete genome	GCF_001183645.1	ASM118364v1
NZ_CP012126.1	Escherichia coli strain DH1Ec104, complete genome	GCF_001183665.1	ASM118366v1
NZ_CP012127.1	Escherichia coli strain DH1Ec169, complete genome	GCF_001183685.1	ASM118368v1
NZ_CP014268.2	Escherichia coli B strain C2566, complete genome	GCF_001559615.2	ASM155961v2
NZ_CP014269.1	Escherichia coli B strain C3029, complete genome	GCF_001559635.1	ASM155963v1
NZ_CP016182.1	Escherichia coli strain EC590, complete genome	GCF_001682305.1	ASM168230v1
NZ_CP018103.1	Escherichia coli strain MRSN352231, complete genome	GCF_001886535.1	ASM188653v1
NZ_CP018121.1	Escherichia coli strain MRSN346355, complete genome	GCF_001886575.1	ASM188657v1
NZ_CP018109.1	Escherichia coli strain MRSN346595, complete genome	GCF_001886755.1	ASM188675v1
NC_010468.1	Escherichia coli ATCC 8739, complete genome	GCF_000019385.1	ASM1938v1
NC_012947.1	Escherichia coli 'BL21-Gold(DE3)pLysS AG', complete genome	GCF_000023665.1	ASM2366v1
NC_017663.1	Escherichia coli P12b, complete genome	GCF_000257275.1	ASM25727v1
NC_017660.1	Escherichia coli KO11FL, complete genome	GCF_000258025.1	ASM25802v1
NC_020518.1	Escherichia coli str. K-12 substr. MDS42 DNA, complete genome	GCF_000350185.1	ASM35018v1
NZ_HG738867.1	Escherichia coli str. K-12 substr. MC4100 complete genome	GCF_000499485.1	MYMC4100
NZ_CP007265.1	Escherichia coli strain ST540, complete genome	GCF_000597845.1	ASM59784v1
NZ_CP007390.1	Escherichia coli strain ST540, complete genome	GCF_000599625.1	ASM59962v1
NZ_CP007391.1	Escherichia coli strain ST540, complete genome	GCF_000599645.1	ASM59964v1
NZ_CP014197.1	Escherichia coli strain MRE600, complete genome	GCF_001542675.2	ASM154267v2
NZ_CP015240.1	Escherichia coli strain 2011C-3911, complete genome	GCF_001644725.1	ASM164472v1
NZ_CP016018.1	Escherichia coli strain ER1821R, complete genome	GCF_001663075.1	ASM166307v1
NZ_CP016404.1	Escherichia coli strain 210221272, complete genome	GCF_001735705.1	ASM173570v1

Appendix Table 1.4: Sort order of simulated dataset in Chapter 2.

Genome	Similarity	Dissimilarity	Random sort orders																									
	order	order	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
F	1	1	6	3	9	9	1	1	6	5	8	3	3	11	2	9	13	10	4	10	10	2	7	13	1	13	13	13
G	2	3	12	13	3	5	6	5	12	2	2	12	4	10	1	4	4	4	7	9	12	8	9	1	7	10	1	8
H	3	5	7	7	5	7	13	9	4	3	12	4	9	3	6	6	10	1	11	8	4	10	13	3	5	6	7	5
A	4	7	10	4	12	1	7	2	7	7	13	1	1	12	5	10	9	11	1	13	9	6	8	4	13	11	2	7
J	5	9	3	5	10	10	4	8	1	12	11	10	11	6	12	5	7	6	12	6	11	9	2	5	9	3	6	12
original	6	11	13	11	13	8	2	11	11	13	5	2	8	7	13	7	12	2	5	7	8	4	1	7	12	7	11	4
B	7	13	9	8	1	11	3	12	5	8	1	5	7	2	11	12	1	8	2	3	5	7	11	12	11	9	12	11
I	8	12	5	6	7	4	5	4	2	9	3	6	2	5	4	3	11	12	10	4	7	13	3	9	8	8	4	3
C	9	10	8	2	11	2	11	7	10	1	7	11	5	1	3	11	2	5	9	2	6	12	4	8	2	4	10	1
D	10	8	2	9	4	3	8	13	8	10	10	8	6	8	7	8	5	13	6	5	3	1	10	11	6	1	8	9
E	11	6	4	1	6	6	12	6	9	6	4	9	13	9	9	1	3	7	3	12	13	3	5	6	10	5	3	2
K	12	4	11	10	2	13	9	10	13	4	6	7	12	4	10	13	6	3	8	11	1	5	12	2	3	12	5	10
L	13	2	1	12	8	12	10	3	3	11	9	13	10	13	8	2	8	9	13	1	2	11	6	10	4	2	9	6

Random sort orders *cont.*																																				
27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63
12	1	2	5	9	5	2	11	13	6	10	5	10	8	10	2	7	11	5	1	10	2	4	6	7	5	3	12	8	11	7	11	9	10	9	10	11
9	2	4	10	4	11	5	4	5	7	9	6	8	13	1	11	13	12	10	13	4	1	2	10	13	1	13	2	2	2	11	5	4	5	8	11	12
10	3	7	8	2	10	10	12	7	4	6	1	1	6	3	8	4	13	3	12	5	7	6	11	5	2	8	5	12	3	1	3	8	13	3	1	3
4	6	8	9	11	1	4	9	6	10	5	8	9	3	2	10	1	2	7	5	1	9	3	13	12	3	10	1	10	8	8	12	5	7	7	2	7
6	8	5	7	8	12	6	10	8	5	13	2	2	12	11	5	6	5	6	9	12	12	10	7	2	9	7	7	3	4	13	2	13	4	4	5	10
2	9	1	3	5	13	9	5	12	8	11	11	5	10	9	6	11	10	11	4	9	3	8	3	10	13	5	13	9	12	10	13	2	2	11	3	4
13	11	11	6	12	8	7	1	9	12	8	4	13	1	7	9	12	9	9	3	3	4	1	2	3	12	1	3	13	13	12	7	10	3	12	8	2
11	7	6	13	13	3	11	13	10	2	1	10	11	5	12	13	5	1	4	8	13	13	7	1	11	6	2	4	7	10	6	1	11	1	13	13	9
3	13	12	11	3	4	3	6	3	13	7	12	7	7	5	3	3	3	8	11	8	6	13	5	9	10	11	10	1	6	4	6	7	6	10	9	6
5	5	13	1	1	7	1	3	4	11	3	9	6	4	6	4	10	6	1	2	7	11	12	4	1	4	6	11	4	9	5	4	1	8	5	6	1
1	4	3	4	6	9	12	2	1	1	4	7	3	2	8	12	8	8	13	10	2	8	9	12	6	8	9	8	11	7	9	9	12	12	6	7	13
8	12	10	12	10	6	13	7	11	9	2	13	4	11	4	1	2	4	12	7	11	10	11	8	4	7	4	9	6	1	3	8	3	11	1	4	5
7	10	9	2	7	2	8	8	2	3	12	3	12	9	13	7	9	7	2	6	6	5	5	9	8	11	12	6	5	5	2	10	6	9	2	12	8

Random sort orders *cont.*																																				
64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100
12	9	13	7	2	8	7	11	3	10	6	8	9	2	11	6	5	2	4	5	6	5	3	13	10	12	5	7	1	7	2	4	3	10	9	3	8
8	2	10	13	6	13	9	5	5	7	2	10	11	9	3	11	1	13	1	1	7	1	10	2	7	9	4	11	4	8	8	12	12	5	10	1	3
1	3	5	12	4	7	2	4	8	6	12	7	5	1	8	7	13	6	8	7	5	13	6	12	4	11	10	1	13	6	9	1	5	13	2	7	2
6	7	8	1	5	11	3	8	11	4	1	11	13	11	12	1	7	11	3	2	2	4	8	3	1	10	2	5	8	4	1	5	6	4	8	2	9
7	4	11	10	13	12	5	1	12	2	10	13	1	7	7	10	6	12	2	6	3	2	7	1	8	2	9	2	3	5	5	13	11	3	1	13	11
2	8	9	6	8	9	11	3	7	5	5	3	12	12	4	3	8	4	6	4	13	7	4	7	11	1	7	13	12	3	6	10	8	1	11	8	13
11	12	3	3	11	2	10	9	13	13	13	2	3	6	2	12	12	10	9	8	12	6	9	4	3	13	3	10	10	2	11	3	10	2	12	6	1
13	11	6	5	9	1	13	7	10	1	9	1	4	13	6	9	10	3	11	13	11	10	5	9	9	4	6	8	9	1	4	2	7	12	3	9	6
9	5	1	9	12	10	12	12	1	11	8	9	7	5	5	8	2	1	13	12	1	11	1	11	13	3	11	9	5	11	3	7	13	9	7	10	10
3	13	4	11	1	5	1	10	6	8	3	5	8	8	10	5	11	9	7	10	4	8	2	5	12	5	13	6	6	9	12	11	2	6	13	11	7
10	10	12	2	7	6	6	6	4	3	11	12	10	3	9	2	4	8	10	11	9	3	11	10	5	7	1	3	11	12	13	9	4	7	4	12	5
5	6	2	8	10	3	8	2	9	9	7	6	2	4	1	4	3	7	5	9	8	12	12	8	2	6	8	12	2	13	10	6	1	11	6	4	12
4	1	7	4	3	4	4	13	2	12	4	4	6	10	13	13	9	5	12	3	10	9	13	6	6	8	12	4	7	10	7	8	9	8	5	5	4

Appendix Table 1.5: Sort order of *Mycobacterium tuberculosis* dataset in Chapter 2.

Accession	Similarity	Dissimilarity	Random sort orders																						
number	order	order	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
NZ_CP016888.1	1	1	40	30	2	12	24	10	29	30	24	18	3	27	27	6	28	15	16	33	18	17	41	25	15	30
NZ_CP010340.1	2	3	33	4	24	42	15	36	14	42	32	40	43	37	4	11	31	25	3	42	28	1	13	35	19	4
NC_020089.1	3	5	35	28	14	30	10	4	5	36	9	39	34	8	14	23	37	2	13	8	3	11	12	41	29	3
NC_022350.1	4	7	34	38	8	10	22	9	4	34	33	21	19	33	2	3	1	11	20	17	39	2	9	43	10	39
NC_009565.1	5	9	30	7	29	17	34	5	8	14	29	43	18	28	39	37	20	6	21	25	29	13	5	12	18	5
NZ_CP010338.1	6	11	4	41	5	18	20	27	3	11	1	15	11	26	40	22	34	21	27	37	5	14	10	34	17	43
NZ_HG813240.1	7	13	7	6	16	4	35	2	37	18	28	12	8	13	32	14	22	28	10	4	24	39	16	19	12	42
NC_021251.1	8	15	37	8	13	43	18	23	40	41	14	17	24	32	43	32	26	40	32	9	17	20	24	23	8	13
NZ_CP002882.1	9	17	36	20	18	39	30	21	10	39	2	14	4	31	24	9	19	7	34	22	14	41	18	17	38	8
NZ_CP002885.1	10	19	5	13	30	6	7	38	24	31	42	8	33	2	26	2	18	38	14	31	33	22	36	29	11	33
NZ_CP010330.1	11	21	26	32	32	20	3	18	41	32	18	11	12	24	15	39	15	33	39	41	27	36	14	21	5	31
NZ_CP010339.1	12	23	43	42	3	15	1	35	28	40	37	13	37	21	41	24	2	17	43	15	32	12	34	40	25	14
NC_018143.2	13	25	38	1	20	40	4	30	36	37	30	30	27	14	6	5	9	12	31	35	15	28	17	39	7	22
NC_009525.1	14	27	11	5	34	33	14	8	31	29	5	26	25	39	16	4	21	10	35	2	8	42	39	2	14	41
NC_000962.3	15	29	3	2	43	37	31	39	22	3	20	41	14	36	1	31	13	18	40	13	13	34	20	18	27	25
NZ_CP009101.1	16	31	14	26	15	38	39	41	11	8	43	4	13	3	10	20	40	39	29	7	7	8	22	32	42	36
NZ_CP009100.1	17	33	28	43	4	3	27	3	30	27	27	34	40	22	28	15	43	27	23	29	36	27	28	22	31	27
NZ_CP007027.1	18	35	17	37	41	24	43	17	34	15	36	29	39	16	42	35	32	16	25	18	9	24	21	4	6	20
NZ_CP002871.1	19	37	39	10	35	36	37	34	19	1	8	22	38	41	9	28	30	31	12	32	19	23	38	27	2	16
NZ_CP007809.1	20	39	31	23	33	41	21	19	35	2	7	2	41	18	8	38	29	32	8	5	16	4	42	14	26	38
NZ_CP013475.1	21	41	22	18	40	23	23	24	33	38	6	10	10	15	36	19	16	37	36	20	4	10	27	16	35	26
NC_002755.2	22	43	12	33	42	9	29	43	42	13	35	9	31	11	37	16	23	23	42	6	35	30	1	3	9	7
NC_021740.1	23	42	18	16	19	29	28	6	39	21	4	32	42	6	34	33	25	26	1	3	38	6	43	7	32	19
NZ_CP012506.2	24	40	42	15	25	5	42	40	1	17	26	28	35	25	5	18	12	13	28	16	1	32	2	36	21	2
NZ_CP012090.1	25	38	20	14	10	2	32	31	13	5	16	33	36	38	25	12	5	36	37	12	30	9	35	15	24	34
NC_017522.1	26	36	29	17	7	14	6	16	21	10	12	35	22	9	30	10	11	5	11	21	41	15	6	5	28	9
NZ_CP009427.1	27	34	41	40	26	31	13	37	12	4	17	24	1	20	23	7	24	42	41	19	34	26	3	38	34	11
NZ_AP014573.1	28	32	8	31	1	7	19	42	43	43	22	7	32	40	19	13	27	4	17	38	20	35	31	8	20	29
NC_017524.1	29	30	2	3	12	25	25	12	27	12	11	42	16	29	18	42	7	1	30	30	23	3	19	30	43	32
NC_021194.1	30	28	1	12	31	27	17	25	16	20	40	6	28	42	3	17	36	24	22	10	22	37	11	10	41	28
NC_021054.1	31	26	19	21	17	16	33	26	25	16	31	20	26	34	33	40	35	20	24	39	12	40	23	11	13	37
NZ_CP016794.1	32	24	15	11	9	13	2	1	38	22	41	5	7	35	17	34	6	43	33	36	40	33	26	6	22	1
NC_012943.1	33	22	27	25	28	35	26	22	20	28	3	38	29	43	20	43	39	3	9	1	6	18	4	37	33	10
NC_016768.1	34	20	23	29	38	26	38	14	2	26	25	16	30	4	35	30	41	35	26	43	10	29	29	24	30	17
NC_018078.1	35	18	10	27	6	19	8	11	6	7	21	3	23	10	12	25	3	22	7	14	11	25	40	9	36	21
NZ_CP002883.1	36	16	25	35	11	11	5	7	15	9	23	19	21	12	29	8	8	29	15	26	2	19	30	33	16	15
NZ_CP007803.1	37	14	16	39	37	8	36	20	26	6	13	31	5	7	22	36	10	34	2	24	43	31	33	26	23	12
NZ_CP009426.1	38	12	21	36	36	21	11	28	9	19	15	36	20	30	31	27	38	30	18	11	25	43	25	20	37	18
NZ_CP010337.1	39	10	9	22	23	28	9	32	17	35	39	37	17	17	21	1	4	41	6	27	37	7	7	1	1	23
NC_020559.1	40	8	32	24	22	32	41	29	32	25	34	1	6	1	7	41	42	8	38	28	31	16	37	13	4	40
NZ_CP009480.1	41	6	13	19	27	1	40	13	23	33	19	27	9	23	11	29	17	14	4	23	21	21	8	28	3	6
NZ_CP011510.1	42	4	24	9	21	22	12	15	18	24	38	23	2	19	38	26	33	9	5	40	42	5	15	31	39	24
NZ_CP017920.1	43	2	6	34	39	34	16	33	7	23	10	25	15	5	13	21	14	19	19	34	26	38	32	42	40	35

Random sort orders *cont.*																																					
25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62
43	37	33	26	9	20	4	33	33	8	35	20	36	34	2	40	25	35	13	21	15	5	27	29	32	3	39	11	6	34	12	23	30	1	1	26	22	1
9	22	41	3	4	2	6	37	31	32	10	26	8	21	22	13	29	42	30	27	24	21	43	25	19	18	21	21	31	10	33	26	39	30	36	14	10	36
30	25	20	29	11	34	1	27	2	3	32	32	33	42	3	19	37	14	32	43	16	4	18	2	35	6	34	1	11	13	17	19	21	12	42	42	25	37
42	36	26	40	19	30	28	16	8	2	29	17	39	30	9	30	8	4	40	17	37	37	22	37	37	35	23	38	23	30	35	15	29	5	25	29	16	21
10	10	32	30	10	7	22	30	16	17	21	1	43	14	1	26	1	26	28	6	5	1	41	28	12	21	40	35	1	42	15	1	27	31	37	18	31	3
4	43	4	10	21	15	3	28	25	9	1	15	22	10	35	24	28	36	26	15	42	26	12	12	5	32	28	14	3	16	28	13	33	8	2	3	12	15
7	12	5	39	8	14	9	43	13	38	30	16	7	40	27	14	43	27	2	2	26	19	39	32	13	24	29	32	34	25	37	17	24	40	30	38	29	28
27	33	23	15	18	35	32	32	36	31	6	14	21	20	43	20	7	28	36	19	4	6	28	43	11	20	32	4	2	8	23	22	40	21	19	12	20	19
14	31	30	9	24	21	31	26	40	24	17	29	2	26	11	42	20	11	18	31	13	42	40	38	16	28	43	7	18	36	38	34	18	37	13	31	36	7
20	17	12	7	27	1	34	6	43	7	20	27	9	25	20	16	22	40	41	40	25	32	7	23	31	9	33	27	19	31	41	31	17	4	17	27	19	42
37	32	42	35	35	26	38	5	27	39	41	43	29	38	4	2	4	19	4	39	10	27	14	26	36	26	10	40	8	32	2	6	13	26	5	15	37	9
8	38	2	41	1	11	42	18	38	19	22	25	12	3	39	28	31	9	27	13	28	11	26	30	10	36	15	26	14	23	4	27	20	32	18	16	32	11
40	2	37	28	13	28	13	36	19	13	26	7	37	18	5	31	15	24	1	23	3	8	1	24	4	34	31	30	33	29	7	4	5	15	40	32	35	16
12	34	1	14	25	24	29	1	39	15	36	40	11	19	38	43	30	12	10	28	21	12	38	7	21	8	38	34	15	27	27	12	14	14	15	25	2	32
41	7	43	36	39	43	19	14	35	5	42	31	18	36	12	32	41	39	7	7	39	22	36	20	40	7	22	18	21	9	30	29	35	7	39	20	39	40
31	16	6	12	5	4	41	39	11	25	15	22	32	7	19	3	12	18	12	10	1	33	19	8	30	40	14	42	28	3	34	42	4	17	22	23	11	13
11	24	19	37	33	27	24	13	23	16	34	24	20	1	25	25	6	33	21	4	38	38	16	14	28	11	18	6	16	2	9	32	6	33	12	17	41	24
34	26	17	43	16	17	21	21	29	22	43	18	23	11	24	23	11	31	24	24	30	28	11	15	20	12	19	31	12	37	16	38	11	18	10	28	40	33
29	9	36	27	30	37	25	2	42	14	24	19	3	15	30	38	39	30	23	33	43	17	4	10	9	17	30	3	4	40	32	41	36	23	28	34	18	10
6	18	40	34	3	8	43	24	34	6	4	6	35	8	37	9	5	5	17	3	6	18	34	4	14	10	17	24	40	12	25	7	9	2	7	11	27	5
13	1	3	22	15	23	23	7	22	41	31	21	17	6	34	27	18	16	5	14	9	36	5	17	18	15	41	15	32	35	39	33	19	38	35	41	28	14
25	13	8	1	12	29	36	40	20	11	28	42	28	24	17	5	32	2	29	8	29	9	13	33	15	33	25	41	39	17	21	3	32	3	38	2	23	31
3	5	16	25	34	39	10	29	32	34	3	34	10	12	13	8	2	34	20	16	8	3	20	27	26	14	37	39	22	38	1	16	1	13	4	39	5	26
17	11	22	23	32	32	33	10	17	40	7	41	15	39	10	41	42	17	19	22	11	14	15	34	1	19	5	2	43	5	18	24	25	43	34	9	4	35
18	20	15	38	2	9	39	22	14	23	40	8	34	33	36	33	36	7	3	32	41	13	8	42	34	13	3	37	10	24	19	39	28	27	43	7	34	20
39	39	14	42	40	3	30	3	21	20	11	4	25	32	8	10	13	21	38	35	40	43	37	16	27	42	24	9	17	26	10	14	10	22	6	43	7	41
21	21	35	31	17	12	11	11	6	1	33	10	26	9	16	21	27	38	14	11	7	15	31	39	23	22	35	25	20	14	42	2	34	34	27	35	6	29
33	27	38	5	36	41	2	4	7	29	13	5	42	31	14	15	38	8	34	18	32	10	21	9	38	38	2	23	27	39	5	8	42	19	3	8	13	43
2	15	10	16	22	18	26	31	5	28	19	13	1	29	28	22	14	32	15	34	33	31	30	35	42	23	36	5	9	21	22	5	43	24	33	24	26	12
38	35	28	20	23	22	20	19	37	4	37	30	27	28	23	17	34	6	31	41	18	40	3	22	17	39	8	28	7	20	31	30	16	36	23	19	8	4
19	41	24	8	26	10	35	15	26	21	38	37	31	22	18	7	40	10	35	29	34	35	25	21	33	41	20	17	35	4	8	10	15	29	29	40	3	8
23	3	13	4	43	36	8	41	18	10	9	11	13	23	32	4	23	37	8	5	2	16	17	1	8	2	12	10	30	22	43	43	22	20	16	1	1	22
26	29	18	2	7	25	15	8	9	12	25	33	4	27	6	18	16	13	9	26	23	41	33	6	2	4	4	33	29	28	40	37	23	6	11	4	42	6
22	19	39	21	14	42	17	25	10	18	18	2	19	41	40	37	10	25	43	25	36	24	2	13	24	31	27	29	38	1	13	28	38	39	26	5	21	34
16	14	11	13	20	33	14	17	41	35	16	9	6	2	41	35	24	43	6	9	17	39	32	40	6	43	9	13	24	18	24	11	41	42	31	13	38	23
15	8	29	17	37	6	37	23	12	36	5	38	24	35	31	34	35	29	22	38	35	23	24	11	29	29	6	8	37	6	11	36	7	16	14	37	43	27
1	23	9	11	28	16	12	34	30	26	12	23	41	16	7	36	9	41	25	1	31	20	35	18	39	1	1	12	13	7	6	25	26	10	41	21	14	2
36	28	34	32	6	19	40	42	3	37	8	36	40	5	21	12	26	23	33	12	22	25	42	31	41	27	26	20	25	43	20	20	31	25	20	30	30	38
32	40	31	18	42	13	27	35	4	43	23	3	5	43	29	6	3	3	42	30	27	30	9	36	3	16	16	43	36	33	36	18	3	11	32	33	9	17
5	30	25	33	41	38	18	9	15	42	14	28	16	13	33	11	21	1	16	20	20	2	23	41	7	37	42	36	5	41	29	9	8	9	21	6	24	25
24	6	7	6	38	40	5	20	28	30	39	12	30	17	15	39	17	20	39	37	14	7	29	5	43	25	11	22	26	15	26	40	12	28	9	36	15	30
28	42	21	19	29	31	16	38	24	33	27	39	38	37	42	29	19	15	37	42	19	29	6	19	25	30	7	16	41	19	14	21	37	35	8	22	33	18
35	4	27	24	31	5	7	12	1	27	2	35	14	4	26	1	33	22	11	36	12	34	10	3	22	5	13	19	42	11	3	35	2	41	24	10	17	39

Random sort orders *cont.*

63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100
36	29	27	31	10	23	14	15	36	18	43	3	19	27	5	5	19	22	3	29	15	43	15	4	41	15	20	10	36	27	27	27	22	4	26	28	22	41
5	31	39	43	15	11	4	27	22	42	3	8	25	40	21	18	1	4	14	40	7	23	9	38	24	39	2	21	26	18	5	16	14	32	10	11	8	30
10	23	3	41	34	10	23	24	12	32	36	34	29	10	37	16	28	42	38	2	31	11	42	41	38	7	24	5	42	37	41	12	19	16	43	17	42	12
38	42	43	20	35	25	25	34	43	41	20	14	42	43	38	21	37	36	22	16	25	34	13	21	22	3	3	20	16	7	11	24	38	15	21	19	35	4
16	40	13	19	37	2	29	19	13	35	14	32	39	21	15	22	29	37	39	8	35	36	21	22	33	8	13	6	40	1	34	23	33	34	37	7	29	8
12	21	11	23	6	9	24	11	9	7	40	20	23	34	2	3	39	14	28	38	21	31	10	28	20	9	36	28	43	3	18	3	20	40	2	18	3	40
3	2	32	22	29	39	16	23	42	11	7	36	34	16	14	14	22	12	32	7	12	19	7	14	14	29	9	4	8	2	15	22	31	19	12	12	39	38
20	12	6	10	7	42	21	26	28	15	31	27	2	38	4	31	2	1	23	41	6	32	40	31	31	32	8	38	3	24	26	28	41	26	16	23	36	31
28	41	21	11	39	22	36	7	27	10	33	39	11	20	40	29	16	9	36	42	24	6	29	24	11	41	10	23	38	32	29	11	34	38	8	39	32	20
40	33	25	12	21	17	1	40	4	21	16	5	30	37	20	38	32	20	5	39	42	35	36	36	6	38	14	36	11	15	32	20	8	5	32	15	31	24
8	34	36	18	17	37	38	4	20	9	34	9	3	29	7	8	33	24	2	11	13	21	32	19	27	34	39	42	27	22	9	21	39	30	30	16	11	23
30	5	5	37	23	41	6	28	19	20	19	18	14	25	12	7	8	21	26	18	40	5	18	7	5	33	25	43	9	38	7	4	37	25	33	4	21	43
11	19	29	40	19	15	3	6	3	3	18	23	4	3	34	15	23	16	30	25	38	24	43	9	36	25	31	18	1	42	30	34	18	12	29	34	5	37
41	17	24	15	32	3	26	33	5	38	32	35	26	35	10	40	27	5	19	14	32	27	25	11	13	23	43	16	15	11	21	6	32	43	36	14	14	28
9	38	9	7	41	19	9	25	30	34	1	12	7	31	32	4	38	33	10	43	16	42	5	27	39	42	32	40	31	28	14	26	25	35	1	42	38	18
37	39	14	9	42	1	40	9	10	5	12	10	31	24	18	13	4	19	12	10	19	4	8	13	21	27	40	12	29	14	13	5	27	23	15	29	9	27
17	13	40	28	30	7	35	10	6	19	17	17	33	26	27	26	15	32	15	31	20	17	14	2	4	10	12	25	7	20	39	30	5	29	11	10	13	25
7	15	8	27	20	24	15	38	8	40	37	31	41	17	23	10	34	6	25	32	11	10	26	40	30	30	5	2	4	41	1	14	43	20	27	13	7	9
34	18	20	26	24	16	2	8	37	2	8	7	35	32	39	42	5	39	16	36	2	28	20	16	3	24	18	27	18	39	40	36	29	22	41	33	28	36
24	28	19	32	31	20	19	16	14	26	9	38	8	1	26	36	12	7	7	26	27	30	12	23	43	40	1	39	41	33	17	17	11	27	28	9	18	15
14	30	16	24	33	4	30	29	39	27	26	29	36	41	31	33	11	17	6	23	22	2	38	37	26	1	38	32	22	10	12	39	24	7	22	20	26	10
13	36	34	13	1	28	31	2	31	8	24	4	6	8	13	28	41	40	9	34	18	18	35	1	17	28	28	35	5	4	31	13	6	8	42	32	12	22
26	7	1	17	18	29	10	5	16	36	39	30	16	28	17	1	40	11	37	1	41	40	3	42	23	14	26	9	14	26	4	7	28	13	35	36	43	16
43	10	37	38	28	5	22	41	25	13	30	19	24	36	25	39	24	18	43	15	26	20	31	17	25	5	11	41	2	8	3	18	26	41	18	24	4	13
18	3	2	25	25	38	12	42	23	30	15	28	1	2	42	35	26	28	34	35	4	39	24	39	1	16	23	33	39	25	37	33	30	17	17	5	24	35
1	20	17	14	5	35	11	12	40	17	13	25	38	22	35	17	3	34	33	19	36	37	17	6	28	31	17	24	34	6	42	43	7	9	13	31	20	29
42	14	23	42	3	27	5	17	26	39	28	13	18	42	43	11	13	2	4	6	28	3	33	43	18	18	41	11	17	21	24	41	12	39	40	8	37	39
25	26	31	1	43	34	37	31	21	43	25	37	37	7	8	9	17	30	20	24	29	12	34	8	10	4	16	22	21	31	6	40	16	2	38	43	34	2
27	6	4	8	9	32	13	1	38	12	38	40	13	33	33	20	31	3	13	33	37	14	4	15	32	17	29	17	10	5	43	9	10	14	31	35	25	42
21	35	30	16	16	8	42	37	17	24	2	26	21	12	6	41	42	25	18	27	8	15	28	26	42	2	7	14	35	35	8	38	21	1	34	6	2	11
29	8	38	33	12	13	18	39	29	14	21	1	43	5	16	37	20	29	31	21	33	26	19	3	29	22	33	3	12	40	36	8	9	31	9	27	27	17
23	22	12	3	40	40	7	20	2	28	29	33	10	39	9	43	21	10	24	13	30	25	27	25	2	37	30	7	30	36	28	10	23	18	6	38	10	34
2	4	41	35	13	31	39	36	32	1	35	16	20	11	30	25	35	41	27	5	23	41	30	29	15	20	6	37	33	13	22	19	17	3	23	41	33	14
32	16	7	36	38	33	8	21	34	22	42	22	32	19	11	2	43	35	1	20	9	13	39	18	34	19	37	15	20	17	25	32	36	33	14	25	16	1
31	27	18	4	27	36	32	32	18	25	11	15	17	6	3	30	36	26	21	22	39	38	11	5	40	43	35	30	13	9	38	1	4	10	5	30	15	26
33	9	10	30	2	14	17	22	15	23	6	42	5	9	22	19	7	27	35	37	1	8	6	12	37	13	19	1	19	30	33	25	13	37	39	22	19	21
15	11	15	34	26	21	27	30	35	31	23	2	9	18	24	32	10	8	11	30	34	16	22	30	12	6	22	34	23	23	23	42	35	36	24	40	30	32
35	37	42	39	14	12	34	14	33	16	10	24	28	15	28	23	18	43	29	4	43	33	1	20	7	35	34	26	32	16	35	31	1	28	7	21	1	19
6	43	33	21	11	43	28	18	41	6	22	11	40	4	29	24	6	31	42	17	3	29	23	34	9	21	4	29	6	29	19	37	42	21	20	37	6	33
4	1	28	6	22	18	43	43	24	4	41	6	22	14	19	12	30	38	40	3	14	1	16	33	16	12	27	19	25	43	20	2	2	42	25	1	41	3
22	32	22	5	4	6	41	35	7	37	4	21	27	13	1	6	9	23	8	28	10	7	2	35	8	36	21	31	28	34	2	15	3	6	3	26	17	5
19	24	35	2	8	30	33	13	1	33	27	41	15	30	41	27	14	15	17	12	5	22	41	32	19	26	15	13	37	12	16	29	40	24	19	2	23	6
39	25	26	29	36	26	20	3	11	29	5	43	12	23	36	34	25	13	41	9	17	9	37	10	35	11	42	8	24	19	10	35	15	11	4	3	40	7

Appendix Table 1.6: Sort order of *Staphylococcus aureus* dataset in Chapter 2.

Accession	Similarity	Dissimilarity	Random sort orders									
number	order	order	1	2	3	4	5	6	7	8	9	10
NZ_CP007539.1	1	1	10	76	67	7	55	115	144	42	35	123
NZ_CP012979.1	2	3	90	101	64	31	123	61	101	50	141	91
NC_002951.2	3	5	97	134	48	40	21	26	139	87	78	54
NZ_CP007672.1	4	7	128	27	78	1	70	139	99	104	32	104
NZ_CP015646.1	5	9	57	17	47	140	98	99	142	71	101	102
NZ_CP012978.1	6	11	31	141	136	51	119	87	79	15	16	79
NZ_LN831036.1	7	13	49	121	84	124	102	122	1	70	102	28
NC_003923.1	8	15	46	111	38	39	42	55	41	52	94	96
NZ_CP013231.1	9	17	101	109	112	82	134	104	31	23	44	101
NC_002953.3	10	19	82	136	103	92	49	64	21	115	20	51
NZ_CP012972.1	11	21	45	130	42	121	27	39	19	64	96	74
NZ_CP012970.1	12	23	42	48	102	52	25	76	63	136	137	116
NZ_CP020020.1	13	25	109	106	72	73	107	78	69	16	79	142
NZ_AP014942.1	14	27	18	103	54	28	120	75	56	142	139	13
NZ_CP013132.1	15	29	1	72	135	59	108	98	55	110	89	113
NZ_CP007670.1	16	31	141	57	77	80	127	41	133	12	63	106
NZ_CP007676.1	17	33	130	104	5	65	81	35	123	131	143	86
NZ_CP013621.1	18	35	39	21	115	143	128	88	108	5	82	30
NZ_CP011526.1	19	37	81	45	40	44	65	77	50	144	62	66
NZ_CP007674.1	20	39	116	56	96	90	43	111	3	112	133	81
NZ_CP007657.1	21	41	71	124	142	96	37	69	2	35	72	60
NZ_CP012974.1	22	43	27	108	31	74	30	56	6	11	67	58
NZ_CP012976.1	23	45	62	4	60	130	9	85	35	22	113	18
NZ_LT671859.1	24	47	99	39	49	50	93	109	16	109	121	17
NZ_CP010998.1	25	49	55	85	7	37	41	30	136	102	38	137
NC_017763.1	26	51	142	140	3	94	13	102	4	88	100	73
NZ_CP007176.1	27	53	15	25	25	107	79	136	91	57	30	120
NZ_CP007499.1	28	55	67	47	82	3	23	49	98	128	69	108
NC_010079.1	29	57	11	6	111	99	135	7	134	106	6	103
NZ_CP007690.1	30	59	26	114	34	72	111	116	34	138	129	52
NC_007793.1	31	61	78	38	116	19	60	20	80	3	132	138
NZ_CP014407.1	32	63	60	31	110	125	118	47	8	43	98	88
NZ_CP014432.1	33	65	94	137	29	17	140	53	44	126	136	37
NZ_CP014438.1	34	67	75	26	113	16	2	81	95	122	95	117
NZ_CP014441.1	35	69	16	79	80	60	129	57	30	130	112	80
NZ_CP010298.1	36	71	8	7	69	71	97	128	52	40	127	56
NZ_CP014415.1	37	73	134	18	66	75	26	129	39	56	114	134
NZ_CP014444.1	38	75	61	122	12	78	85	130	119	117	135	43
NZ_LT615218.1	39	77	79	55	105	93	104	67	75	129	144	133
NZ_CP010295.1	40	79	13	80	22	102	72	86	105	47	59	99
NZ_CP010296.1	41	81	38	128	140	64	29	65	129	20	21	45
NZ_CP010297.1	42	83	6	115	121	110	82	3	13	38	142	125
NZ_CP010299.1	43	85	123	13	57	112	7	133	23	141	138	119
NZ_CP010300.1	44	87	37	53	100	38	56	27	24	89	26	110
NZ_CP010402.1	45	89	85	1	109	35	91	17	38	30	134	124
NZ_CP014420.1	46	91	131	135	134	27	109	118	89	116	73	21
NZ_CP014423.1	47	93	138	78	131	26	143	144	54	135	57	107
NZ_CP014426.1	48	95	110	33	128	13	19	97	102	124	140	49
NZ_CP014429.1	49	97	80	36	68	41	88	63	93	69	4	139
NC_017333.1	50	99	43	95	74	54	47	33	132	96	103	50
NC_013450.1	51	101	33	52	99	5	34	74	22	6	48	64
NZ_CP014412.1	52	103	14	96	125	129	3	14	10	140	71	6
NC_016912.1	53	105	4	118	63	70	80	100	88	133	93	126
NC_017349.1	54	107	44	42	45	115	83	51	51	1	116	143
NZ_CP014392.1	55	109	112	16	43	36	69	8	58	36	86	20
NZ_CP014397.1	56	111	24	98	123	142	15	43	68	120	115	59
NZ_CP014402.1	57	113	32	59	65	86	20	110	14	54	66	63
NZ_CP014409.1	58	115	140	116	33	85	45	5	29	77	31	92
NZ_CP014435.1	59	117	20	93	83	61	40	15	97	91	28	31
NC_009641.1	60	119	76	8	4	128	122	60	60	25	110	121
NZ_LT598688.1	61	121	135	49	15	48	142	59	49	28	104	62
NZ_CP007659.1	62	123	68	107	39	32	110	142	76	99	128	132
NZ_CP013619.1	63	125	22	60	88	49	68	132	111	45	53	70
NZ_CP012756.1	64	127	19	67	21	123	95	37	37	86	46	10
NZ_CP009361.1	65	129	92	24	18	57	35	143	70	63	8	5
NZ_CP013616.1	66	131	21	142	87	63	90	68	121	37	39	144
NC_018608.1	67	133	118	46	35	24	10	94	43	53	74	3
NC_007795.1	68	135	104	22	124	134	16	22	90	55	70	76
NZ_CP020019.1	69	137	48	37	28	126	53	31	66	100	14	136
NZ_CP010890.1	70	139	137	28	97	106	46	40	62	62	40	36
NZ_CP013218.1	71	141	143	11	92	98	71	140	124	80	97	7
NC_017351.1	72	143	86	126	141	20	86	70	94	48	61	114
NC_021554.1	73	144	77	133	117	79	112	112	36	10	11	135
NC_009487.1	74	142	83	73	14	118	84	21	122	118	118	75

NC_009632.1	75	140	9	23	19	30	57	45	32	137	117	129
NZ_CP014064.1	76	138	66	99	119	81	124	96	127	44	25	118
NZ_CP018768.1	77	136	117	129	73	46	96	105	103	7	77	89
NZ_CP019117.1	78	134	34	100	71	58	139	28	72	81	85	47
NZ_AP014652.1	79	132	23	19	37	55	99	117	27	27	24	97
NZ_AP014653.1	80	130	51	82	137	138	31	46	137	85	99	87
NZ_CP009423.1	81	128	89	81	44	137	137	66	120	73	84	85
NZ_CP012593.1	82	126	50	32	51	100	116	95	115	33	27	78
NZ_CP013182.1	83	124	74	113	89	9	117	90	78	19	41	53
NZ_CP018205.1	84	122	113	110	85	6	138	93	15	119	5	55
NZ_CP018766.1	85	120	56	132	46	22	5	23	5	143	49	40
NZ_CP014791.1	86	118	64	51	76	77	144	119	46	75	131	38
NC_017340.1	87	116	3	3	126	116	4	134	138	32	68	61
NZ_CP010526.1	88	114	125	139	27	11	48	19	71	79	50	1
NZ_CP012015.1	89	112	106	9	23	14	12	2	67	58	13	33
NC_022226.1	90	110	95	127	2	105	141	25	47	66	60	26
NC_017341.1	91	108	105	14	55	89	131	127	110	113	81	90
NC_017343.1	92	106	102	90	30	62	89	38	33	78	119	94
NZ_CP011528.1	93	104	25	40	52	104	114	124	92	18	52	4
NZ_CP012013.1	94	102	5	75	138	34	74	36	25	107	120	84
NZ_CP012018.1	95	100	47	50	95	132	101	12	18	41	90	100
NZ_CP012011.1	96	98	121	94	79	127	24	29	7	8	109	57
NC_016928.1	97	96	139	83	143	122	106	24	11	90	34	105
NC_022442.1	98	94	88	5	56	88	126	34	59	98	54	35
NC_022443.1	99	92	103	63	104	120	61	141	45	46	55	68
NC_017338.1	100	90	126	143	9	76	8	18	109	51	64	44
NZ_CP012409.1	101	88	28	97	94	135	18	10	140	103	51	93
NZ_CP012120.1	102	86	115	125	93	8	125	9	53	134	15	23
NZ_AP017320.1	103	84	36	102	90	66	28	11	26	24	23	98
NZ_CP015645.1	104	82	73	91	144	21	62	32	28	59	88	2
NZ_LT009690.1	105	80	124	70	114	139	14	137	61	74	10	115
NC_002952.2	106	78	70	105	13	45	92	6	82	83	42	111
NZ_CP009554.1	107	76	41	61	53	68	17	138	104	65	65	34
NZ_CP019563.1	108	74	93	86	129	114	87	125	125	82	22	22
NZ_CP007454.1	109	72	122	68	107	47	67	13	126	72	105	130
NZ_CP006630.1	110	70	30	62	8	2	58	80	130	125	37	131
NZ_CP012119.1	111	68	54	10	58	136	38	42	42	60	19	83
NZ_CP012012.1	112	66	107	77	24	25	78	91	84	76	56	11
NC_017342.1	113	64	63	54	101	23	63	79	112	34	43	48
NC_017337.1	114	62	12	92	41	56	50	50	141	95	76	39
NC_020533.1	115	60	52	71	6	97	136	58	20	123	111	65
NC_020566.1	116	58	58	131	62	29	73	108	86	9	58	67
NC_020568.1	117	56	40	2	59	95	6	62	96	121	7	82
NC_020529.1	118	54	2	15	139	42	39	123	17	97	108	72
NC_020564.1	119	52	129	74	132	67	115	92	48	14	122	9
NC_020532.1	120	50	111	88	118	108	130	72	57	139	91	122
NC_020536.1	121	48	35	119	86	33	100	121	65	68	126	29
NC_020537.1	122	46	87	43	81	113	36	83	73	105	17	41
NC_007622.1	123	44	69	58	26	83	59	107	143	4	36	77
NC_002745.2	124	42	17	138	133	144	103	73	107	67	18	12
NC_002758.2	125	40	65	64	108	15	54	113	114	108	12	95
NC_009782.1	126	38	84	117	50	101	44	114	81	61	75	128
NC_017331.1	127	36	91	41	75	111	75	120	106	39	87	15
NC_022113.1	128	34	29	87	70	87	94	1	40	92	106	14
NC_017347.1	129	32	114	120	98	10	52	82	77	84	92	140
NC_021670.1	130	30	96	144	32	103	51	4	85	21	130	141
NC_022222.1	131	28	100	29	91	69	77	52	9	2	124	69
NC_022604.1	132	26	136	69	120	91	133	131	117	132	107	8
NZ_CP009681.1	133	24	72	20	106	141	132	89	83	94	83	32
NZ_LN626917.1	134	22	108	89	127	43	113	106	87	29	45	24
NZ_CP011147.1	135	20	127	112	20	109	105	84	131	111	47	127
NZ_CP013137.1	136	18	53	66	61	119	11	48	116	13	29	71
NZ_CP009828.1	137	16	133	44	16	84	64	103	113	49	1	109
NZ_CP015173.1	138	14	98	34	10	133	121	135	100	93	125	25
NZ_CP011685.1	139	12	119	35	1	53	33	44	64	17	33	112
NZ_CP013953.1	140	10	120	12	11	117	1	16	128	101	123	42
NZ_CP013955.1	141	8	132	65	17	12	76	126	74	26	2	16
NZ_CP013957.1	142	6	59	30	36	18	32	54	118	31	9	46
NZ_CP012692.1	143	4	144	123	130	131	66	101	12	114	3	27
NZ_LN854556.1	144	2	7	84	122	4	22	71	135	127	80	19

Appendix Table 1.7: Sort order of *Escherichia coli* dataset in Chapter 2.

Accession	Similarity	Dissimilarity	Random sort orders									
number	order	order	1	2	3	4	5	6	7	8	9	10
NZ_CP009859.1	1	1	207	111	45	179	166	34	78	58	15	88
NZ_CP015241.1	2	3	154	182	54	79	118	158	72	3	176	30
NZ_CP008957.1	3	5	158	27	152	97	134	89	110	158	161	118
NZ_CP016358.1	4	7	73	38	42	123	27	85	32	138	139	42
NC_017656.1	5	9	72	76	187	100	40	7	107	74	174	23
NZ_CP015831.1	6	11	98	95	7	27	173	83	73	15	19	15
NZ_CP008805.1	7	13	145	44	103	66	13	44	178	10	110	20
NC_013008.1	8	15	133	163	170	177	90	183	131	190	23	184
NC_011353.1	9	17	153	29	36	96	111	25	207	67	48	201
NZ_CP010304.1	10	19	6	122	56	14	44	194	74	101	49	18
NZ_CP017249.1	11	21	119	49	104	85	204	91	202	116	111	196
NZ_CP017251.1	12	23	77	168	113	139	1	47	151	191	24	181
NC_017906.1	13	25	2	148	37	28	38	187	103	203	51	135
NZ_CP014314.1	14	27	87	12	137	167	107	45	144	7	186	62
NZ_CP015846.1	15	29	130	101	191	15	128	190	192	54	132	81
NC_013941.1	16	31	110	196	21	92	76	180	148	24	172	205
NC_002695.1	17	33	161	53	101	16	124	35	84	25	107	19
NZ_CP015842.1	18	35	202	79	100	20	58	197	170	62	10	32
NZ_CP014667.1	19	37	24	107	164	175	26	206	179	6	152	96
NZ_CP015843.1	20	39	4	19	153	81	116	169	1	50	130	26
NC_004431.1	21	41	15	65	188	3	135	139	46	95	33	58
NZ_CP014670.1	22	43	142	3	141	55	171	199	106	38	85	195
NC_010498.1	23	45	47	167	201	163	29	167	197	89	82	115
NZ_CP015834.1	24	47	199	97	47	117	139	135	49	12	165	47
NZ_CP015023.1	25	49	10	47	78	40	4	141	2	90	169	55
NZ_CP017434.1	26	51	113	130	194	18	181	104	93	85	148	71
NZ_CP017444.1	27	53	162	161	28	10	110	161	141	118	101	35
NZ_CP017438.1	28	55	206	21	49	9	188	33	30	110	37	178
NZ_CP017446.1	29	57	39	86	50	194	97	156	109	21	96	90
NZ_CP017436.1	30	59	195	190	98	44	22	130	130	161	65	67
NZ_CP017440.1	31	61	175	25	155	161	122	192	14	162	153	180
NZ_CP017442.1	32	63	122	46	108	127	41	164	70	120	144	193
NZ_CP016625.1	33	65	97	120	25	143	70	26	121	176	193	167
NZ_CP014583.1	34	67	21	119	74	109	16	57	15	144	177	43
NC_017646.1	35	69	66	88	175	7	137	14	5	14	71	48
NZ_CP012802.1	36	71	186	121	182	36	154	92	26	102	200	56
NZ_CP017669.1	37	73	181	106	140	46	62	115	37	97	78	148
NC_017626.1	38	75	26	133	117	95	35	134	113	107	117	105
NZ_CP007799.1	39	77	134	180	111	145	96	184	81	93	17	192
NZ_CP015020.1	40	79	171	174	120	190	93	150	186	151	108	166
NZ_CP015229.1	41	81	37	139	20	146	17	27	187	112	86	119
NC_008563.1	42	83	129	166	148	110	202	95	194	189	135	49
NZ_CP015832.1	43	85	157	113	86	207	174	122	123	51	204	188
NZ_CP016497.1	44	87	49	84	161	148	64	43	180	180	197	204
NZ_CP008697.1	45	89	182	186	73	126	113	38	62	77	30	45
NC_008253.1	46	91	117	45	2	138	60	88	90	174	184	168
NZ_CP009072.1	47	93	41	100	99	58	19	163	16	166	8	2
NC_013364.1	48	95	88	177	48	201	75	170	42	43	2	136
NC_017651.1	49	97	193	132	163	6	21	195	166	35	66	69
NC_017652.1	50	99	42	83	61	59	73	5	168	56	18	84
NZ_CP007592.1	51	101	173	162	67	77	197	29	51	80	191	41
NC_017631.1	52	103	92	136	189	22	39	159	63	139	158	16
NC_011601.1	53	105	80	169	9	204	141	94	54	69	75	46
NZ_CP013112.1	54	107	5	184	173	52	50	72	127	136	93	137
NZ_CP012693.1	55	109	185	66	91	23	184	181	85	48	25	111
NC_007946.1	56	111	143	89	18	124	82	36	69	167	180	5
NC_013361.1	57	113	27	104	97	41	120	56	189	165	32	74
NC_013353.1	58	115	189	143	46	48	101	31	64	131	41	126
NC_009801.1	59	117	198	203	3	21	164	116	139	64	136	73
NZ_CP014495.1	60	119	59	205	138	61	176	51	126	47	179	112
NZ_CP006027.1	61	121	174	125	34	135	143	12	114	70	54	140
NZ_CP007136.1	62	123	118	128	195	45	37	60	96	141	3	132
NZ_CP007392.1	63	125	54	37	27	56	148	82	152	135	38	97
NZ_CP007149.1	64	127	126	193	196	26	2	182	201	49	72	191
NC_011750.1	65	129	7	35	190	107	99	143	3	68	189	139
NZ_CP006262.1	66	131	114	28	53	12	177	140	82	45	155	176
NZ_CP007133.1	67	133	205	151	23	51	23	58	169	177	27	154
NZ_CP007393.1	68	135	155	98	90	129	192	78	61	140	31	59
NC_011748.1	69	137	163	200	124	67	67	178	161	2	87	158
NZ_CP012625.1	70	139	166	197	178	72	72	39	108	91	104	116
NZ_CP015228.1	71	141	38	63	59	5	178	111	12	98	175	202
NZ_CP016007.1	72	143	30	150	122	115	170	136	80	202	141	76
NZ_CP005930.1	73	145	85	152	75	112	53	77	116	30	14	72
NZ_CP012631.1	74	147	197	67	10	37	10	114	171	20	203	200

NZ_CP012635.1	75	149	51	41	133	189	89	173	98	207	98	64
NZ_CP012633.1	76	151	70	18	60	13	112	137	150	125	91	110
NC_017632.1	77	153	128	5	70	137	61	46	75	128	11	147
NZ_CP016546.1	78	155	16	1	135	160	95	198	143	63	59	87
NZ_CP013029.1	79	157	102	2	126	114	33	63	60	149	9	8
NC_017628.1	80	159	25	54	149	193	131	133	45	111	145	53
NC_017634.1	81	161	204	52	177	172	24	204	21	65	57	127
NZ_CP014488.1	82	163	100	99	38	118	18	6	149	196	102	153
NZ_LT601384.1	83	165	96	34	114	128	157	107	4	32	100	186
NC_011993.1	84	167	192	82	106	103	127	129	91	133	60	38
NC_020163.1	85	169	160	144	87	31	169	131	29	81	121	85
NZ_CP007442.1	86	171	106	135	144	181	6	117	35	82	115	57
NZ_CP009106.2	87	173	89	36	79	142	105	40	43	137	88	131
NZ_CP007491.1	88	175	167	13	77	199	115	152	92	27	138	130
NZ_CP014522.1	89	177	19	138	181	196	121	21	7	42	167	65
NZ_CP006632.1	90	179	187	195	32	75	149	138	165	34	147	4
NZ_HF572917.1	91	181	180	11	69	120	51	123	34	44	105	161
NC_017641.1	92	183	60	183	66	63	66	165	79	185	58	37
NC_018661.1	93	185	50	85	44	159	46	18	181	109	160	173
NZ_CP011331.1	94	187	28	140	33	106	20	20	48	115	12	151
NC_018658.1	95	189	14	145	89	198	100	119	157	92	123	101
NZ_CP015069.1	96	191	64	78	52	24	68	13	174	28	89	7
NZ_CP009166.1	97	193	200	43	68	153	91	80	125	84	92	175
NZ_CP015076.1	98	195	33	61	51	49	163	70	134	123	119	109
NZ_CP013663.1	99	197	12	126	65	69	207	37	57	23	143	17
NZ_CP015159.1	100	199	43	179	116	144	87	145	104	178	109	10
NC_017633.1	101	201	169	93	166	54	106	118	175	153	84	78
NZ_CP007594.1	102	203	120	10	143	132	28	19	162	194	113	25
NZ_HG941718.1	103	205	68	187	174	188	185	61	160	179	68	155
NC_018650.1	104	207	125	7	85	8	150	201	97	206	77	114
NZ_CP011061.1	105	206	9	48	200	39	151	109	11	104	190	144
NZ_CP013662.1	106	204	67	202	41	169	172	67	142	37	79	100
NZ_CP015995.1	107	202	140	26	159	149	133	160	146	152	34	21
NC_013654.1	108	200	144	90	118	4	144	106	124	5	40	197
NZ_CP010876.1	109	198	132	185	11	173	86	151	58	205	83	107
NZ_CP013025.1	110	196	164	81	107	86	132	142	25	57	195	89
NZ_CP013835.1	111	194	56	4	205	166	9	101	17	156	206	31
NZ_CP013658.1	112	192	52	171	6	192	55	202	100	121	52	150
NZ_CP011416.1	113	190	99	102	160	182	160	155	76	29	6	121
NZ_CP015138.1	114	188	86	30	13	98	205	121	65	46	50	75
NZ_CP009104.1	115	186	11	108	17	2	12	102	36	19	95	6
NZ_CP014316.1	116	184	146	188	185	122	5	207	55	127	55	120
NZ_CP015074.2	117	182	17	70	96	34	36	17	129	88	39	156
NZ_CP013190.1	118	180	150	8	146	158	182	53	205	99	20	123
NZ_CP016628.1	119	178	141	115	4	180	7	147	176	181	26	164
NZ_CP014497.1	120	176	34	124	193	29	117	16	190	4	162	183
NZ_CP007394.1	121	174	62	159	186	184	54	100	77	53	126	99
NC_022648.1	122	172	109	158	63	62	88	148	31	17	4	141
NZ_CP008801.1	123	170	22	6	92	30	187	120	28	172	134	86
NZ_CP010315.1	124	168	63	129	94	57	175	125	40	39	120	203
NZ_CP011495.1	125	166	115	64	109	191	159	66	132	188	21	51
NZ_CP014492.1	126	164	147	198	81	151	183	28	158	31	70	138
NZ_CP013031.1	127	162	44	173	204	162	138	2	59	36	163	124
NZ_CP010344.1	128	160	76	172	184	152	198	98	44	117	129	44
NZ_CP011018.1	129	158	81	123	102	121	179	149	156	145	196	1
NZ_CP014348.1	130	156	71	74	202	156	130	188	183	200	94	142
NZ_CP015912.1	131	154	116	73	19	116	161	162	199	119	168	194
NZ_CP014270.1	132	152	93	77	156	70	81	79	83	155	43	13
NC_011415.1	133	150	127	31	83	171	59	50	18	40	47	91
NC_012967.1	134	148	136	160	197	140	140	84	164	147	36	174
NZ_CP010445.1	135	146	31	80	151	131	196	189	13	143	146	82
NZ_CP011324.1	136	144	29	59	147	84	126	171	102	41	63	66
NZ_CP011321.1	137	142	83	42	142	202	199	75	182	163	137	198
NZ_CP013831.1	138	140	172	62	199	78	195	179	119	18	149	34
NZ_CP014225.1	139	138	123	192	71	35	189	112	128	16	164	61
NC_010473.1	140	136	165	175	171	80	125	64	86	129	142	106
NZ_CP009273.1	141	134	23	118	105	17	147	127	137	75	118	146
NZ_CP011134.1	142	132	82	58	123	130	108	157	188	22	35	171
NZ_CP010442.1	143	130	137	69	132	203	119	3	56	106	205	9
NC_011741.1	144	128	36	71	12	60	52	154	68	159	192	79
NZ_CP011320.1	145	126	95	14	198	111	42	11	117	175	124	95
NZ_CP017100.1	146	124	108	50	158	65	94	186	135	52	188	113
NC_000913.3	147	122	40	60	39	74	78	68	38	9	181	50
NC_017635.1	148	120	201	194	35	157	193	86	115	11	150	179
NZ_CP010438.1	149	118	121	22	16	113	156	41	67	195	178	108
NZ_CP010440.1	150	116	184	204	64	76	109	132	95	60	67	172
NZ_CP013253.1	151	114	45	170	129	53	56	32	204	182	112	170
NZ_CP014272.1	152	112	107	155	1	89	145	108	154	105	46	94

NC_012759.1	153	110	1	147	130	105	69	42	118	83	166	93
NC_016902.1	154	108	57	39	82	88	152	73	138	26	42	54
NC_017664.1	155	106	90	206	157	102	186	69	20	199	28	145
NZ_CP009685.1	156	104	156	103	72	11	84	174	19	164	64	199
NZ_LM995446.1	157	102	8	114	88	83	153	71	105	1	103	133
NZ_CP012868.1	158	100	35	117	176	25	83	193	24	78	171	33
NZ_CP012869.1	159	98	13	149	14	71	162	185	196	169	80	83
NZ_CP012870.1	160	96	74	201	93	141	165	4	133	173	140	14
NC_007779.1	161	94	53	40	145	200	74	128	22	154	1	129
NZ_CP010816.1	162	92	94	87	127	108	206	22	89	103	194	39
NZ_CP010439.1	163	90	101	109	43	134	8	203	23	71	198	98
NZ_CP010441.1	164	88	196	72	136	82	71	124	8	204	173	125
NZ_CP010443.1	165	86	159	92	15	33	85	8	10	122	7	152
NZ_CP010444.1	166	84	112	116	168	93	47	205	122	168	131	80
NZ_CP011342.2	167	82	131	55	76	133	203	23	71	8	97	27
NZ_CP011343.2	168	80	191	191	115	154	146	191	47	72	185	207
NZ_CP010371.1	169	78	46	91	112	174	103	15	112	55	202	60
NC_017625.1	170	76	194	189	5	195	11	99	173	13	187	103
NC_022364.1	171	74	32	96	95	170	63	176	203	96	122	77
NZ_CP009644.1	172	72	152	131	84	32	180	168	177	187	106	92
NZ_CP009789.1	173	70	177	157	134	68	79	24	120	171	170	128
NZ_LN832404.1	174	68	104	20	167	206	45	30	147	73	90	63
NZ_CP011113.2	175	66	20	105	24	164	129	90	52	33	128	157
NZ_CP013483.1	176	64	138	15	162	155	48	52	87	94	114	190
NZ_CP018115.1	177	62	149	32	203	87	34	93	140	148	183	206
NC_012892.2	178	60	3	137	192	185	194	126	206	184	5	169
NC_009800.1	179	58	84	134	139	205	57	97	66	124	125	29
NC_012971.2	180	56	151	178	80	168	167	54	184	113	29	12
NC_017638.1	181	54	170	17	22	42	200	87	27	126	76	163
NZ_CP010585.1	182	52	111	24	154	125	3	9	39	142	44	22
NZ_LM993812.1	183	50	124	23	29	94	191	74	155	193	151	40
NZ_CP006636.1	184	48	61	56	125	178	190	49	159	59	127	185
NZ_CP011938.1	185	46	139	154	165	176	201	200	6	130	156	104
NZ_CP012125.1	186	44	55	142	110	104	65	48	136	76	74	182
NZ_CP012126.1	187	42	69	51	26	186	31	146	193	160	116	122
NZ_CP012127.1	188	40	148	207	179	90	14	65	185	134	81	149
NZ_CP014268.2	189	38	188	127	207	101	123	81	167	201	61	3
NZ_CP014269.1	190	36	91	165	30	64	32	62	41	100	157	177
NZ_CP016182.1	191	34	179	57	57	73	15	103	145	170	99	143
NZ_CP018103.1	192	32	18	110	31	38	155	166	101	146	45	102
NZ_CP018121.1	193	30	135	164	169	187	114	113	153	186	207	70
NZ_CP018109.1	194	28	178	33	180	1	43	10	163	150	53	117
NC_010468.1	195	26	75	112	131	119	168	55	191	183	154	159
NC_012947.1	196	24	190	153	119	197	104	105	99	192	73	11
NC_017663.1	197	22	78	156	183	165	158	175	111	108	201	189
NC_017660.1	198	20	65	9	128	136	92	59	50	66	22	165
NC_020518.1	199	18	176	141	121	43	49	144	172	132	13	162
NZ_HG738867.1	200	16	183	199	172	183	30	172	198	197	133	160
NZ_CP007265.1	201	14	48	146	40	91	80	1	9	86	199	36
NZ_CP007390.1	202	12	103	16	150	50	25	196	53	114	182	28
NZ_CP007391.1	203	10	58	94	8	147	77	110	88	87	16	187
NZ_CP014197.1	204	8	105	176	55	99	98	153	33	198	69	68
NZ_CP015240.1	205	6	203	181	206	150	136	177	195	61	56	24
NZ_CP016018.1	206	4	79	68	58	47	142	76	94	79	159	134
NZ_CP016404.1	207	2	168	75	62	19	102	96	200	157	62	52

Appendix 2 - Appendix for Chapter 3

Appendix Table 2.1: Details for genomes used for building the computational *M. tuberculosis* pan-genome in Chapter 3 including accession numbers, sort order and lineage.

Assembly Accession	Accession number	Description	Position in pan-genome	Length of genome	Lineage	Used for simulation
GCF_001708265.1	NZ_CP016888.1	Mycobacterium tuberculosis strain SCAID 252.0 chromosome, complete genome	1	4439387	Beijing	
GCF_000331445.1	NC_020089.1	Mycobacterium tuberculosis 7199-99 complete genome	2	4421197	Haarlem	
GCF_002116775.1	NZ_CP017594.1	Mycobacterium tuberculosis strain Beijing-like/36918 chromosome, complete genome	3	4441591	Beijing	
GCF_000153685.2	NC_022350.1	Mycobacterium tuberculosis str. Haarlem, complete genome	4	4408224	Haarlem	
GCF_002072775.2	NZ_CP020381.2	Mycobacterium tuberculosis strain MTB1, complete genome	5	4433542	New-1	
GCF_002116755.1	NZ_CP017593.1	Mycobacterium tuberculosis strain Beijing-like/35049 chromosome, complete genome	6	4427062	Beijing	
GCF_000016925.1	NC_009565.1	Mycobacterium tuberculosis F11, complete genome	7	4424435	LAM	
GCF_002116835.1	NZ_CP017597.1	Mycobacterium tuberculosis strain Beijing-like/50148 chromosome, complete genome	8	4444417	Beijing	
GCF_002208235.1	NZ_CP022014.1	Mycobacterium tuberculosis strain MTB2 chromosome, complete genome	9	4417716	Beijing	
GCF_002116795.1	NZ_CP017595.1	Mycobacterium tuberculosis strain Beijing-like/38774 chromosome, complete genome	10	4431885	Beijing	
GCF_002357955.1	NZ_AP018035.1	Mycobacterium tuberculosis DNA, complete genome, strain: HN-321	11	4421540	Beijing	
GCF_000786505.1	NZ_HG813240.1	Mycobacterium tuberculosis 49-02 complete genome	12	4412379	Beijing	
GCF_001938725.1	NZ_CP016972.1	Mycobacterium tuberculosis H37Ra chromosome, complete genome	13	4426109	Euro-American (4.9)	
GCF_001545015.1	NZ_CP010339.1	Mycobacterium tuberculosis strain 22103, complete genome	14	4399422	Euro-American (4.2)	
GCF_002357935.1	NZ_AP018034.1	Mycobacterium tuberculosis DNA, complete genome, strain: HN-205	15	4411033	Beijing	
GCF_001544705.1	NZ_CP010330.1	Mycobacterium tuberculosis strain F28, complete genome	16	4421903	Euro-American (4.9)	
GCF_000400615.1	NC_021251.1	Mycobacterium tuberculosis CCDC5079, complete genome	17	4414325	Beijing	
GCF_002356255.1	NZ_AP018033.1	Mycobacterium tuberculosis DNA, complete genome, strain: HN-024	18	4399916	EAI	
GCF_002886865.1	NZ_CP025607.1	Mycobacterium tuberculosis strain GG-186-10 chromosome, complete genome	19	4411478	Haarlem	
GCF_000572195.1	NZ_CP002885.1	Mycobacterium tuberculosis CCDC5180, complete genome	20	4414346	Beijing	
GCF_000572155.1	NZ_CP002882.1	Mycobacterium tuberculosis BT2, complete genome	21	4401899	Beijing	
GCF_002887145.1	NZ_CP025599.1	Mycobacterium tuberculosis strain GG-45-11 chromosome, complete genome	22	4411469	Euro-American (4.8)	
GCF_002886945.1	NZ_CP025596.1	Mycobacterium tuberculosis strain GG-27-11 chromosome, complete genome	23	4411443	X-type	
GCF_002886165.1	NZ_CP025594.1	Mycobacterium tuberculosis strain GG-5-10 chromosome, complete genome	24	4411442	Euro-American (4.8)	
GCF_000277735.2	NC_018143.2	Mycobacterium tuberculosis H37Rv, complete genome	25	4411709	Euro-American (4.9)	
GCF_002886405.1	NZ_CP025604.1	Mycobacterium tuberculosis strain GG-129-11 chromosome, complete genome	26	4411413	Euro-American (4.1.1.3)	
GCF_002886585.1	NZ_CP025608.1	Mycobacterium tuberculosis strain GG-229-10 chromosome, complete genome	27	4411519	LAM	
GCF_001922485.1	NZ_CP018778.1	Mycobacterium tuberculosis strain DK9897, complete genome	28	4411511	LAM	
GCF_002886195.1	NZ_CP025595.1	Mycobacterium tuberculosis strain GG-20-11 chromosome, complete genome	29	4411504	LAM	
GCF_002886145.1	NZ_CP025593.1	Mycobacterium tuberculosis strain GG-111-10 chromosome, complete genome	30	4411563	Haarlem	
GCF_002886335.1	NZ_CP025601.1	Mycobacterium tuberculosis strain GG-90-10 chromosome, complete genome	31	4411602	LAM	
GCF_000016145.1	NC_009525.1	Mycobacterium tuberculosis H37Ra, complete genome	32	4419977	Euro-American (4.9)	
GCF_002357975.1	NZ_AP018036.1	Mycobacterium tuberculosis DNA, complete genome, strain: HN-506	33	4413362	Beijing	
GCF_000738475.1	NZ_CP009101.1	Mycobacterium tuberculosis strain ZMC13-88, complete genome	34	4411515	Beijing	

GCF_002887065.1	NZ_CP025597.1	Mycobacterium tuberculosis strain 36-11 chromosome, complete genome	35	4411469	Euro-American (4.8)	
GCF_000195955.2	NC_000962.3	Mycobacterium tuberculosis H37Rv, complete genome	36	4411532	Euro-American (4.9)	yes - H37Rv
GCF_002886225.1	NZ_CP025600.1	Mycobacterium tuberculosis strain 77-11 chromosome, complete genome	37	4411508	LAM	
GCF_002887255.1	NZ_CP025602.1	Mycobacterium tuberculosis strain GG-109-10 chromosome, complete genome	38	4411463	LAM	
GCF_000738445.1	NZ_CP009100.1	Mycobacterium tuberculosis strain ZMC13-264, complete genome	39	4411507	Beijing	
GCF_000827085.1	NZ_CP007027.1	Mycobacterium tuberculosis H37RvSiena, complete genome	40	4410911	Euro-American (4.9)	
GCF_002886685.1	NZ_CP025598.1	Mycobacterium tuberculosis strain GG-37-11 chromosome, complete genome	41	4411526	LAM	
GCF_002886775.1	NZ_CP025603.1	Mycobacterium tuberculosis strain GG-121-10 chromosome, complete genome	42	4411510	LAM	
GCF_002887335.1	NZ_CP025606.1	Mycobacterium tuberculosis strain GG-137-10 chromosome, complete genome	43	4411446	LAM	
GCF_000572125.1	NZ_CP002871.1	Mycobacterium tuberculosis HKBS1, complete genome	44	4407929	Beijing	yes - HKBS1
GCF_002447575.1	NZ_CP023608.1	Mycobacterium tuberculosis strain LE410 chromosome, complete genome	45	4411365	Euro-American (4.8)	
GCF_002447735.1	NZ_CP023616.1	Mycobacterium tuberculosis strain LN3668 chromosome, complete genome	46	4411494	Euro-American (4.8)	
GCF_002886505.1	NZ_CP025605.1	Mycobacterium tuberculosis strain GG-134-11 chromosome, complete genome	47	4411399	LAM	
GCF_002448095.1	NZ_CP023634.1	Mycobacterium tuberculosis strain TBDM2487 chromosome, complete genome	48	4411314	Euro-American (4.8)	
GCF_002447475.1	NZ_CP023603.1	Mycobacterium tuberculosis strain LE63 chromosome, complete genome	49	4411415	Euro-American (4.9)	
GCF_002447695.1	NZ_CP023614.1	Mycobacterium tuberculosis strain LN3588 chromosome, complete genome	50	4411379	LAM	
GCF_002447515.1	NZ_CP023605.1	Mycobacterium tuberculosis strain LE79 chromosome, complete genome	51	4411475	Haarlem	
GCF_002447975.1	NZ_CP023628.1	Mycobacterium tuberculosis strain MDRMA2082 chromosome, complete genome	52	4411479	Haarlem	yes - MDRMA2082
GCF_002448155.1	NZ_CP023637.1	Mycobacterium tuberculosis strain TBDM2717 chromosome, complete genome	53	4411442	Haarlem	
GCF_002447755.1	NZ_CP023617.1	Mycobacterium tuberculosis strain LN3672 chromosome, complete genome	54	4411335	LAM	
GCF_002447995.1	NZ_CP023629.1	Mycobacterium tuberculosis strain MDRMA2260 chromosome, complete genome	55	4411352	LAM	
GCF_002446995.1	NZ_CP023579.1	Mycobacterium tuberculosis strain LE492 chromosome, complete genome	56	4411208	LAM	
GCF_002448135.1	NZ_CP023636.1	Mycobacterium tuberculosis strain TBDM2699 chromosome, complete genome	57	4411457	Haarlem	
GCF_002447715.1	NZ_CP023615.1	Mycobacterium tuberculosis strain LN3589 chromosome, complete genome	58	4411392	Haarlem	
GCF_002448015.1	NZ_CP023630.1	Mycobacterium tuberculosis strain MDRMA2441 chromosome, complete genome	59	4411454	Euro-American (4.7)	
GCF_002447195.1	NZ_CP023589.1	Mycobacterium tuberculosis strain TBV5000 chromosome, complete genome	60	4411318	LAM	
GCF_002448175.1	NZ_CP023638.1	Mycobacterium tuberculosis strain TBV4766 chromosome, complete genome	61	4411310	LAM	
GCF_002447775.1	NZ_CP023618.1	Mycobacterium tuberculosis strain LN3695 chromosome, complete genome	62	4411449	Haarlem	
GCF_002447415.1	NZ_CP023600.1	Mycobacterium tuberculosis strain CSV3611 chromosome, complete genome	63	4411439	Euro-American (4.7)	
GCF_002447655.1	NZ_CP023612.1	Mycobacterium tuberculosis strain LN2978 chromosome, complete genome	64	4411331	Euro-American (4.9)	
GCF_002448115.1	NZ_CP023635.1	Mycobacterium tuberculosis strain TBDM2489 chromosome, complete genome	65	4411369	LAM	
GCF_002446875.1	NZ_CP023573.1	Mycobacterium tuberculosis strain CSV4519 chromosome, complete genome	66	4411288	LAM	
GCF_002447215.1	NZ_CP023590.1	Mycobacterium tuberculosis strain TBV5362 chromosome, complete genome	67	4411331	LAM	
GCF_002447895.1	NZ_CP023624.1	Mycobacterium tuberculosis strain MDRMA701 chromosome, complete genome	68	4411338	LAM	
GCF_002447955.1	NZ_CP023627.1	Mycobacterium tuberculosis strain MDRMA2019 chromosome, complete genome	69	4411229	LAM	
GCF_002448075.1	NZ_CP023633.1	Mycobacterium tuberculosis strain TBDM2444 chromosome, complete genome	70	4411284	LAM	

GCF_002447455.1	NZ_CP023602.1	Mycobacterium tuberculosis strain LE13 chromosome, complete genome	71	4411412	Haarlem	
GCF_002448035.1	NZ_CP023631.1	Mycobacterium tuberculosis strain TBDM1506 chromosome, complete genome	72	4411157	Euro-American (4.1.1)	
GCF_002448055.1	NZ_CP023632.1	Mycobacterium tuberculosis strain TBDM2189 chromosome, complete genome	73	4411316	Euro-American (4.9)	
GCF_000706665.1	NZ_CP007809.1	Mycobacterium tuberculosis strain KIT87190, complete genome	74	4410788	Beijing	
GCF_002447015.1	NZ_CP023580.1	Mycobacterium tuberculosis strain LN180 chromosome, complete genome	75	4411436	Haarlem	
GCF_002447855.1	NZ_CP023622.1	Mycobacterium tuberculosis strain MDRDM827 chromosome, complete genome	76	4411315	LAM	
GCF_002446895.1	NZ_CP023574.1	Mycobacterium tuberculosis strain CSV4644 chromosome, complete genome	77	4411271	LAM	
GCF_002446935.1	NZ_CP023576.1	Mycobacterium tuberculosis strain CSV10399 chromosome, complete genome	78	4411180	Euro-American (4.1.1.3)	
GCF_002446975.1	NZ_CP023578.1	Mycobacterium tuberculosis strain LE486 chromosome, complete genome	79	4411180	LAM	
GCF_002447155.1	NZ_CP023587.1	Mycobacterium tuberculosis strain ME1473 chromosome, complete genome	80	4411217	LAM	
GCF_002447255.1	NZ_CP023592.1	Mycobacterium tuberculosis strain SLM036 chromosome, complete genome	81	4411342	LAM	
GCF_002447295.1	NZ_CP023594.1	Mycobacterium tuberculosis strain SLM056 chromosome, complete genome	82	4411306	LAM	
GCF_002447635.1	NZ_CP023611.1	Mycobacterium tuberculosis strain LN763 chromosome, complete genome	83	4411321	LAM	
GCF_002447795.1	NZ_CP023619.1	Mycobacterium tuberculosis strain LN1100 chromosome, complete genome	84	4411392	Haarlem	
GCF_002448215.1	NZ_CP023640.1	Mycobacterium tuberculosis strain TBV4952 chromosome, complete genome	85	4411414	Haarlem	
GCF_002356015.1	NZ_AP017901.1	Mycobacterium tuberculosis DNA, complete genome, strain: NCGM946K2	86	4380602	LAM	
GCF_002447035.1	NZ_CP023581.1	Mycobacterium tuberculosis strain LN2358 chromosome, complete genome	87	4411353	LAM	
GCF_002447115.1	NZ_CP023585.1	Mycobacterium tuberculosis strain MDRDM1098 chromosome, complete genome	88	4411148	LAM	
GCF_002447175.1	NZ_CP023588.1	Mycobacterium tuberculosis strain TBDM425 chromosome, complete genome	89	4411143	LAM	
GCF_002447495.1	NZ_CP023604.1	Mycobacterium tuberculosis strain LE76 chromosome, complete genome	90	4411211	LAM	
GCF_002447835.1	NZ_CP023621.1	Mycobacterium tuberculosis strain LN2900 chromosome, complete genome	91	4411327	LAM	
GCF_002446915.1	NZ_CP023575.1	Mycobacterium tuberculosis strain CSV5769 chromosome, complete genome	92	4411312	Euro-American (4.9)	
GCF_002447375.1	NZ_CP023598.1	Mycobacterium tuberculosis strain SLM100 chromosome, complete genome	93	4411394	Haarlem	
GCF_002447675.1	NZ_CP023613.1	Mycobacterium tuberculosis strain LN3584 chromosome, complete genome	94	4411326	LAM	
GCF_002447815.1	NZ_CP023620.1	Mycobacterium tuberculosis strain LN1856 chromosome, complete genome	95	4411299	LAM	
GCF_002447875.1	NZ_CP023623.1	Mycobacterium tuberculosis strain MDRMA203 chromosome, complete genome	96	4411290	Haarlem	
GCF_002447915.1	NZ_CP023625.1	Mycobacterium tuberculosis strain MDRMA863 chromosome, complete genome	97	4411432	Haarlem	
GCF_000008585.1	NC_002755.2	Mycobacterium tuberculosis CDC1551, complete genome	98	4403837	Euro-American (4.1.1.3)	
GCF_000422125.1	NC_021740.1	Mycobacterium tuberculosis EAI5, complete genome	99	4391174	EAI	
GCF_001855255.1	NZ_CP013475.1	Mycobacterium tuberculosis strain 1458, complete genome	100	4402033	Beijing	
GCF_001895845.1	NZ_CP018300.1	Mycobacterium tuberculosis strain I0002353-6, complete genome	101	4385578	LAM	
GCF_002447275.1	NZ_CP023593.1	Mycobacterium tuberculosis strain SLM040 chromosome, complete genome	102	4411408	Haarlem	
GCF_002447535.1	NZ_CP023606.1	Mycobacterium tuberculosis strain LE103 chromosome, complete genome	103	4411243	LAM	
GCF_001275565.2	NZ_CP012506.2	Mycobacterium tuberculosis strain SCAID 187.0 chromosome, complete genome	104	4411829	Beijing	
GCF_002446955.1	NZ_CP023577.1	Mycobacterium tuberculosis strain CSV11678 chromosome, complete genome	105	4411382	Haarlem	
GCF_002447055.1	NZ_CP023582.1	Mycobacterium tuberculosis strain LN3756 chromosome, complete genome	106	4411315	Euro-American (4.9)	

GCF_002447075.1	NZ_CP023583.1	Mycobacterium tuberculosis strain MDRDM260 chromosome, complete genome	107	4411280	Euro-American (4.9)	
GCF_002447135.1	NZ_CP023586.1	Mycobacterium tuberculosis strain MDRMA2491 chromosome, complete genome	108	4411121	Beijing	
GCF_002447315.1	NZ_CP023595.1	Mycobacterium tuberculosis strain SLM060 chromosome, complete genome	109	4411134	Beijing	
GCF_002447335.1	NZ_CP023596.1	Mycobacterium tuberculosis strain SLM063 chromosome, complete genome	110	4411337	Haarlem	
GCF_002447355.1	NZ_CP023597.1	Mycobacterium tuberculosis strain SLM088 chromosome, complete genome	111	4411385	Haarlem	
GCF_002447395.1	NZ_CP023599.1	Mycobacterium tuberculosis strain CSV383 chromosome, complete genome	112	4411115	Beijing	
GCF_002447435.1	NZ_CP023601.1	Mycobacterium tuberculosis strain CSV9577 chromosome, complete genome	113	4411230	LAM	
GCF_002447555.1	NZ_CP023607.1	Mycobacterium tuberculosis strain LE371 chromosome, complete genome	114	4411137	LAM	
GCF_002447595.1	NZ_CP023609.1	Mycobacterium tuberculosis strain LN55 chromosome, complete genome	115	4411186	Beijing	
GCF_002447935.1	NZ_CP023626.1	Mycobacterium tuberculosis strain MDRMA1565 chromosome, complete genome	116	4411159	Beijing	
GCF_002448195.1	NZ_CP023639.1	Mycobacterium tuberculosis strain TBV4768 chromosome, complete genome	117	4411173	Beijing	
GCF_000193185.2	NZ_CP012090.1	Mycobacterium tuberculosis W-148, complete genome	118	4418548	Beijing	
GCF_000270365.1	NC_017522.1	Mycobacterium tuberculosis CCDC5180, complete genome	119	4405981	Beijing	
GCF_000828995.1	NZ_AP014573.1	Mycobacterium tuberculosis str. Kurono DNA, complete genome	120	4415078	Euro-American (4.9)	
GCF_000224435.1	NC_017524.1	Mycobacterium tuberculosis CTRI-2, complete genome	121	4398525	LAM	
GCF_000756545.1	NZ_CP009427.1	Mycobacterium tuberculosis strain 96121, complete genome	122	4410945	EAI Manila	
GCF_002116815.1	NZ_CP017596.1	Mycobacterium tuberculosis strain Beijing/391 chromosome, complete genome	123	4406925	Beijing	
GCF_000364825.1	NC_021054.1	Mycobacterium tuberculosis str. Beijing/NITR203, complete genome	124	4411128	Beijing	
GCF_000389945.1	NC_021194.1	Mycobacterium tuberculosis EAI5/NITR206, complete genome	125	4390306	EAI	
GCF_001895865.1	NZ_CP018304.1	Mycobacterium tuberculosis strain M0002959-6, complete genome	126	4386447	LAM	
GCF_000572175.1	NZ_CP002883.1	Mycobacterium tuberculosis BT1, complete genome	127	4399405	Beijing	
GCF_001702435.1	NZ_CP016794.1	Mycobacterium tuberculosis strain SCAID 320.0 chromosome, complete genome	128	4406628	Beijing	
GCF_001895765.1	NZ_CP018303.1	Mycobacterium tuberculosis strain I0004241-1, complete genome	129	4386132	LAM	
GCF_002447095.1	NZ_CP023584.1	Mycobacterium tuberculosis strain MDRDM627 chromosome, complete genome	130	4411215	LAM	
GCF_000023625.1	NC_012943.1	Mycobacterium tuberculosis KZN 1435, complete genome	131	4398250	LAM	
GCF_000154585.2	NC_016768.1	Mycobacterium tuberculosis KZN 4207, complete genome	132	4394985	LAM	
GCF_000154605.2	NC_018078.1	Mycobacterium tuberculosis KZN 605, complete genome	133	4399120	LAM	
GCF_000698475.1	NZ_CP007803.1	Mycobacterium tuberculosis K, complete genome	134	4385518	Beijing	
GCF_000756525.1	NZ_CP009426.1	Mycobacterium tuberculosis strain 96075, complete genome	135	4379376	Beijing	
GCF_001544955.1	NZ_CP010337.1	Mycobacterium tuberculosis strain 22115, complete genome	136	4401829	New-1	
GCF_001895785.1	NZ_CP018305.1	Mycobacterium tuberculosis strain M0018684-2, complete genome	137	4359825	LAM	
GCF_001895805.1	NZ_CP018302.1	Mycobacterium tuberculosis strain I0004000-1, complete genome	138	4365724	LAM	
GCF_002447235.1	NZ_CP023591.1	Mycobacterium tuberculosis strain TBV5365 chromosome, complete genome	139	4411398	Euro-American (4.9)	
GCF_002447615.1	NZ_CP023610.1	Mycobacterium tuberculosis strain LN317 chromosome, complete genome	140	4411340	Euro-American (4.9)	
GCF_000350205.1	NC_020559.1	Mycobacterium tuberculosis str. Erdman = ATCC 35801 DNA, complete genome	141	4392353	Haarlem	
GCF_000831245.1	NZ_CP009480.1	Mycobacterium tuberculosis H37Rv, complete genome	142	4396119	Euro-American (4.9)	
GCF_001750865.1	NZ_CP011510.1	Mycobacterium tuberculosis strain Beijing, complete genome	143	4378588	Beijing	

GCF_001870145.1	NZ_CP017920.1	Mycobacterium tuberculosis strain TB282 chromosome, complete genome	144	4425860	Beijing	yes - TB282
GCF_001895825.1	NZ_CP018301.1	Mycobacterium tuberculosis strain I0002801-4, complete genome	145	4376067	Euro-American (4.1.1.3)	
GCF_002116855.1	NZ_CP017598.1	Mycobacterium tuberculosis strain Beijing-like/1104 chromosome, complete genome	146	4380156	Beijing	

Appendix Table 2.2: Regions filtered in real datasets in Chapter 3 with coordinates for the *M. tuberculosis* H37Rv strain genome and coordinates mapped to the computational pan-genome.

GeneID	Symbol	Aliases	Description	H37Rv		pan-genome		
				Start	End	Segment ID	Start	End
887049	Rv0031	Rv0031	Possible remnant of a transposase	33582	33794	1	36422	36634
886938	PPE1	Rv0096	PPE family protein PPE1	105324	106715	2	108638	110029
886912	PE_PGRS1	Rv0109	PE-PGRS family protein PE_PGRS1	131382	132872	3	134701	136207
886883	PE_PGRS2	Rv0124	PE-PGRS family protein PE_PGRS2	149533	150996	4	152871	154623
886857	PE1	Rv0151c	PE family protein PE1	177543	179309	5	181279	183045
886838	PE2	Rv0152c	PE family protein PE2	179319	180896	6	183055	184632
886826	PE3	Rv0159c	PE family protein PE3	187433	188839	7	191170	192576
886825	PE4	Rv0160c	PE family protein PE4	188931	190439	8	192668	194176
886684	PPE2	Rv0256c	PPE family protein PPE2	307877	309547	9	344783	346455
886623	Rv0278c	Rv0278c	PE-PGRS family protein PE_PGRS3	333437	336310	10	370354	380011
886621	PE_PGRS4	Rv0279c	PE-PGRS family protein PE_PGRS4	336560	339073	11	380312	383320
886619	PPE3	Rv0280	PPE family protein PPE3	339364	340974	12	383615	385266
886608	PE5	Rv0285	PE family protein PE5	349624	349932	13	393920	394228
886607	PPE4	Rv0286	PPE family protein PPE4	349935	351476	14	394231	395773
885981	PE_PGRS5	Rv0297	PE-PGRS family protein PE_PGRS5	361334	363109	15	405632	407456
886592	PPE5	Rv0304c	PPE family protein PPE5	366150	372764	16	410498	417114
885978	PPE6	Rv0305c	PPE family protein PPE6	372820	375711	17	417170	420092
886527	PE6	Rv0335c	PE family protein PE6	399535	400050	18	443920	444435
886498	PPE7	Rv0354c	PPE family protein PPE7	424269	424694	19	468763	469189
886491	PPE8	Rv0355c	PPE family protein PPE8	424777	434679	20	469272	479222
886436	Rv0387c	Rv0387c	pseudo	466672	468001	21	512589	513920
886439	PPE9	Rv0388c	PPE family protein PPE9	467459	468001	21	512589	513920
886340	PPE10	Rv0442c	PPE family protein PPE10	530751	532214	22	578361	579839
886317	PPE11	Rv0453	PPE family protein PPE11	543174	544730	23	590803	592359
887391	PE_PGRS6	Rv0532	PE-PGRS family protein PE_PGRS6	622793	624577	24	672718	674895
887725	PE_PGRS7	Rv0578c	PE-PGRS family protein PE_PGRS7	671996	675916	25	722329	727115
888644	Rv0741	Rv0741	Probable transposase (fragment)	832534	832848	26	884616	884930
888645	Rv0742	Rv0742	hypothetical protein	832981	833508	27	885063	885591
888664	PE_PGRS9	Rv0746	PE-PGRS family protein PE_PGRS9	835701	838052	28	887784	890552
888662	PE_PGRS10	Rv0747	PE-PGRS family protein PE_PGRS10	838451	840856	29	890951	904025
888695	PE_PGRS11	Rv0754	PE-PGRS family protein PE_PGRS11	846159	847913	30	909340	911094
888708	PPE12	Rv0755c	PPE family protein PPE12	848103	850040	31	911284	913252
3205072	Rv0755A	Rv0755A	Putative transposase (fragment)	850342	850527	32	915450	915635
885454	Rv0795	Rv0795	insertion sequence elementIS6110 transposase (fragment)	889072	889398	33	959541	960802
885099	Rv0796	Rv0796	insertion sequence element IS986/IS6110 transposase	889347	890333	33	959541	960802
885476	Rv0797	Rv0797	insertion sequence element IS1547 transposase	890388	891482	34	960884	961980
885236	PE_PGRS12	Rv0832	PE-PGRS family protein PE_PGRS12	924951	925364	35	995916	999047
885391	PE_PGRS13	Rv0833	PE-PGRS family protein PE_PGRS13	925361	927610	35	995916	999047
885369	PE_PGRS14	Rv0834c	PE-PGRS family protein PE_PGRS14	927837	930485	36	999274	1002402
885054	Rv0850	Rv0850	Putative transposase (fragment)	947312	947644	37	1024532	1024864
885742	PE_PGRS15	Rv0872c	PE-PGRS family protein PE_PGRS15	968424	970244	38	1045956	1047797
885617	PPE13	Rv0878c	PPE family protein PPE13	976872	978203	39	1054425	1055770
885069	PPE14	Rv0915c, MTB41	PPE family protein PPE14	1020058	1021329	40	1097640	1098911
885167	PE7	Rv0916c, MTB10	PE family protein PE7	1021344	1021643	41	1098926	1099225
885549	Rv0920c	Rv0920c	transposase	1025497	1026816	42	1103079	1104399
885564	Rv0922	Rv0922	transposase	1027685	1029337	43	1105268	1106921
885264	PE_PGRS16	Rv0977	PE-PGRS family protein PE_PGRS16	1090373	1093144	44	1172555	1175357
885077	PE_PGRS17	Rv0978c	PE-PGRS family protein PE_PGRS17	1093361	1094356	45	1175574	1176875
885327	PE_PGRS18	Rv0980c	PE-PGRS family protein PE_PGRS18	1095078	1096451	46	1177597	1179456
886010	Rv1034c	Rv1034c	Probable transposase (fragment)	1158918	1159307	47	1241948	1242337
888206	Rv1035c	Rv1035c	Probable transposase (fragment)	1159375	1160061	48	1242405	1243092

888227	Rv1036c	Rv1036c	Probable IS1560 transposase (fragment)	1160095	1160433	49	1243126	1243464
888477	PPE15	Rv1039c	PPE family protein PPE15	1161297	1162472	50	1244328	1245505
888533	PE8	Rv1040c	PE family protein PE8	1162549	1163376	51	1245582	1247767
888546	Rv1041c	Rv1041c	IS2-like transposase	1164572	1165435	52	1248963	1249890
888607	Rv1042c	Rv1042c	IS2-like transposase	1165092	1165499	52	1248963	1249890
886060	Rv1047	Rv1047	transposase	1169423	1170670	53	1254003	1255250
887139	Rv1054	Rv1054	Probable integrase (fragment)	1176928	1177242	54	1262871	1263185
887122	PE_PGRS19	Rv1067c	PE-PGRS family protein PE_PGRS19	1188421	1190424	55	1274366	1277792
887123	PE_PGRS20	Rv1068c	PE-PGRS family protein PE_PGRS20	1190757	1192148	56	1278125	1280512
887094	PE_PGRS21	Rv1087	PE-PGRS family protein PE_PGRS21	1211560	1213863	57	1299926	1302791
887096	PE9	Rv1088	PE family protein PE9	1214513	1214947	58	1303442	1304060
887090	PE10	Rv1089	PE family protein PE10	1214769	1215131	58	1303442	1304060
885258	PE_PGRS22	Rv1091	PE-PGRS family protein PE_PGRS22	1216469	1219030	59	1305422	1308861
885131	PPE16	Rv1135c	PPE family protein PPE16	1262272	1264128	60	1352154	1355377
885164	Rv1149	Rv1149	transposase	1277893	1278300	61	1369147	1369554
885990	PPE17	Rv1168c	PPE family protein PPE17	1298764	1299804	62	1390024	1391064
885930	lipX	Rv1169c, PE11	lipase LipX	1299822	1300124	63	1391082	1391384
885988	PE12	Rv1172c	PE family protein PE12	1301755	1302681	64	1393017	1393943
886044	PE13	Rv1195	PE family protein PE13	1339003	1339302	65	1430339	1430638
886073	PPE18	Rv1196, mtb39a	PPE family protein PPE18	1339349	1340524	66	1430685	1431866
886092	Rv1199c	Rv1199c	insertion sequence element IS1081 transposase	1341358	1342605	67	1432749	1433996
888362	PE14	Rv1214c	PE family protein PE14	1357293	1357625	68	1448689	1449021
887109	PE_PGRS23	Rv1243c	PE-PGRS family protein PE_PGRS23	1384989	1386677	69	1476396	1478131
886922	Rv1313c	Rv1313c	insertion sequence element IS1557 transposase	1468171	1469505	70	1564143	1565477
886899	PE_PGRS24	Rv1325c	PE-PGRS family protein PE_PGRS24	1488154	1489965	71	1588017	1590841
886819	PPE19	Rv1361c, mtb39b	PPE family protein PPE19	1532443	1533633	72	1637397	1638592
886789	Rv1369c	Rv1369c	insertion sequence element IS986/IS6110 transposase	1541994	1542980	73	1646962	1648223
886791	Rv1370c	Rv1370c	insertion sequence element IS6110 transposase(fragment)	1542929	1543255	73	1646962	1648223
886757	PE15	Rv1386	PE family protein PE15	1561464	1561772	74	1669168	1671092
886784	PPE20	Rv1387	PPE family protein PPE20	1561769	1563388	74	1669168	1671092
886745	PE_PGRS25	Rv1396c	PE-PGRS family protein PE_PGRS25	1572127	1573857	75	1679831	1681571
886652	PE16	Rv1430	PE family protein PE16	1606386	1607972	76	1714109	1715695
886626	PE_PGRS26	Rv1441c	PE-PGRS family protein PE_PGRS26	1618209	1619684	77	1725986	1727528
886605	PE_PGRS27	Rv1450c	PE-PGRS family protein PE_PGRS27	1630638	1634627	78	1738490	1744196
886595	PE_PGRS28	Rv1452c	PE-PGRS family protein PE_PGRS28	1636004	1638229	79	1745573	1748163
886556	PE_PGRS29	Rv1468c	PE-PGRS family protein PE_PGRS29	1655609	1656721	80	1765596	1769384
886384	PPE21	Rv1548c	PPE family protein PPE21	1751297	1753333	81	1895264	1897300
886337	Rv1573	Rv1573	phage protein	1779314	1779724	82	1926323	1926789
886331	Rv1574	Rv1574	phage protein	1779930	1780241	83	1926995	1929130
886335	Rv1575	Rv1575	phage protein	1780199	1780699	83	1926995	1929130
886327	Rv1576c	Rv1576c	phage capsid protein	1780643	1782064	83	1926995	1929130
886329	Rv1577c	Rv1577c	phage prohead protease	1782072	1782584	84	1929138	1929650
886322	Rv1578c	Rv1578c	phage protein	1782758	1783228	85	1929824	1930294
886369	Rv1579c	Rv1579c	phage protein	1783309	1783623	86	1930375	1930958
886313	Rv1580c	Rv1580c	phage protein	1783620	1783892	86	1930375	1930958
886318	Rv1581c	Rv1581c	phage protein	1783906	1784301	87	1930972	1931367
886311	Rv1582c	Rv1582c	phage protein	1784497	1785912	88	1931563	1933594
886315	Rv1583c	Rv1583c	phage protein	1785912	1786310	88	1931563	1933594
886307	Rv1584c	Rv1584c	phage protein	1786307	1786528	88	1931563	1933594
886309	Rv1585c	Rv1585c	phage protein	1786584	1787099	89	1933650	1935571
886305	Rv1586c	Rv1586c	phage integrase	1787096	1788505	89	1933650	1935571
885486	PE17	Rv1646	PE family protein PE17	1855764	1856696	90	2116522	2117454
885174	PE_PGRS30	Rv1651c	PE-PGRS family protein PE_PGRS30	1862347	1865382	91	2123105	2126165
885068	PPE22	Rv1705c	PPE family protein PPE22	1931497	1932654	92	2197800	2198957
885070	PPE23	Rv1706c	PPE family protein PPE23	1932694	1933878	93	2198997	2200181
885544	PPE24	Rv1753c	PPE family protein PPE24	1981614	1984775	94	2251185	2256637
885541	Rv1756c	Rv1756c	Putative transposase	1987745	1988731	95	2270615	2271876
885558	Rv1757c	Rv1757c	Putative transposase for insertion sequence element IS6110 (fragment)	1988680	1989006	95	2270615	2271876
885372	Rv1763	Rv1763	Putative transposase for insertion sequence element IS6110 (fragment)	1996152	1996478	96	2287069	2290574
885238	Rv1764	Rv1764	Putative transposase	1996427	1997413	96	2287069	2290574
3205098	Rv1765A	Rv1765A	Putative transposase (fragment)	1999142	1999357	97	2293963	2294178
885429	PE_PGRS31	Rv1768	PE-PGRS family protein PE_PGRS31	2000614	2002470	98	2295435	2297319
885827	PPE25	Rv1787	PPE family protein PPE25	2025301	2026398	99	2321511	2322610
885895	PE18	Rv1788	PE family protein PE18	2026477	2026776	100	2322689	2324346
885333	PPE26	Rv1789	PPE family protein PPE26	2026790	2027971	101	2324360	2325541
885859	PPE27	Rv1790	PPE family protein PPE27	2028425	2029477	102	2325995	2327049
885445	PE19	Rv1791	PE family protein PE19	2029904	2030203	103	2327476	2327775
885465	PPE28	Rv1800	PPE family protein PPE28	2039453	2041420	104	2339774	2343892
885491	PPE29	Rv1801	PPE family protein PPE29	2042001	2043272	105	2345831	2347104
885542	PPE30	Rv1802	PPE family protein PPE30	2043384	2044775	106	2347216	2349965
885730	PE_PGRS32	Rv1803c	PE-PGRS family protein PE_PGRS32	2044923	2046842	107	2350113	2352043
885537	PE20	Rv1806	PE family protein PE20	2048072	2048371	108	2354631	2354930
885072	PPE31	Rv1807	PPE family protein PPE31	2048398	2049597	109	2354957	2356157

885590	PPE32	Rv1808	PPE family protein PPE32	2049921	2051150	110	2356483	2357712
885555	PPE33	Rv1809	PPE family protein PPE33	2051282	2052688	111	2357844	2359250
885551	PE_PGRS33	Rv1818c	PE-PGRS family protein PE_PGRS33	2061178	2062674	112	2367757	2369313
885753	PE_PGRS34	Rv1840c	PE-PGRS family protein PE_PGRS34	2087971	2089518	113	2396374	2397927
885362	PPE34	Rv1917c	PPE family protein PPE34	2162932	2167311	114	2471535	2484365
885506	PPE35	Rv1918c	PPE family protein PPE35	2167649	2170612	115	2484733	2487701
885921	PE_PGRS35	Rv1983	PE-PGRS family protein PE_PGRS35	2226244	2227920	116	2555194	2556880
887546	Rv2013	Rv2013	Transposase	2260665	2261144	117	2589659	2590682
887547	Rv2014	Rv2014	Transposase	2261098	2261688	117	2589659	2590682
888395	Rv2105	Rv2105	Putative transposase for insertion sequence element IS6110 (fragment)	2365465	2365791	118	2714913	2716174
888398	Rv2106	Rv2106	Probable transposase	2365740	2366726	118	2714913	2716174
887811	PE22	Rv2107	PE family protein PE22	2367359	2367655	119	2716819	2717115
887814	PPE36	Rv2108	PPE family protein PPE36	2367711	2368442	120	2717171	2723171
888710	PPE37	Rv2123, irg2	PPE family protein PPE37	2381071	2382492	121	2736032	2737453
887791	PE_PGRS37	Rv2126c	PE-PGRS family protein PE_PGRS37	2387202	2387972	122	2742163	2742934
887300	PE_PGRS38	Rv2162c	PE-PGRS family protein PE_PGRS38	2423240	2424838	123	2778334	2779963
888197	Rv2167c	Rv2167c	insertion sequence element IS986/IS6110 transposase	2430159	2431145	124	2785285	2786546
888459	Rv2168c	Rv2168c	Putative transposase for insertion sequence element IS6110 (fragment)	2431094	2431420	124	2785285	2786546
888326	Rv2177c	Rv2177c	transposase	2439282	2439947	125	2794425	2795090
888602	Rv2278	Rv2278	insertion sequence element IS6110 transposase(fragment)	2550065	2550391	126	2909783	2911044
887746	Rv2279	Rv2279	insertion sequence element IS986/IS6110 transposase	2550340	2551326	126	2909783	2911044
888111	PE23	Rv2328	PE family protein PE23	2600731	2601879	127	2967174	2968322
888961	PE_PGRS39	Rv2340c	PE-PGRS family protein PE_PGRS39	2617667	2618908	128	2990997	2992238
888959	PPE38	Rv2352c	PPE family protein PPE38	2632923	2634098	129	3018365	3019544
886003	PPE39	Rv2353c	PPE family protein PPE39	2634528	2635592	130	3019976	3024138
888963	Rv2354	Rv2354	insertion sequence element IS6110 transposase(fragment)	2635628	2635954	131	3024186	3025447
888957	Rv2355	Rv2355	insertion sequence element IS986/IS6110 transposase	2635903	2636889	131	3024186	3025447
888950	PPE40	Rv2356c	PPE family protein PPE40	2637688	2639535	132	3026276	3030216
885141	PE_PGRS40	Rv2371	PE-PGRS family protein PE_PGRS40	2651753	2651938	133	3042435	3042620
885517	PE_PGRS41	Rv2396, aprC	acid and phagosome regulated protein AprC	2692799	2693884	134	3084909	3086014
885511	PE24	Rv2408	PE family protein PE24	2705762	2706736	135	3097917	3098891
885699	Rv2424c	Rv2424c	transposase	2720776	2721777	136	3112937	3113938
885945	PPE41	Rv2430c	PPE family protein PPE41	2727336	2727920	137	3119497	3120081
885703	PE25	Rv2431c	PE family protein PE25	2727967	2728266	138	3120128	3120427
887201	Rv2479c	Rv2479c	insertion sequence element IS986/IS6110 transposase	2784657	2785643	139	3176839	3178100
887328	Rv2480c	Rv2480c	insertion sequence element IS6110 transposase(fragment)	2785592	2785918	139	3176839	3178100
887909	PE_PGRS42	Rv2487c	PE-PGRS family protein PE_PGRS42	2795301	2797385	140	3187495	3189585
887941	PE_PGRS43	Rv2490c	PE-PGRS family protein PE_PGRS43	2801254	2806236	141	3193454	3199797
888515	Rv2512c	Rv2512c	insertion sequence element IS1081 transposase	2828556	2829803	142	3236142	3237389
888172	PE26	Rv2519	PE family protein PE26	2835785	2837263	143	3243376	3244854
887992	PE_PGRS44	Rv2591	PE-PGRS family protein PE_PGRS44	2921551	2923182	144	3329839	3331480
888204	PPE42	Rv2608	PPE family protein PPE42	2935046	2936788	145	3343347	3345089
888215	PE_PGRS45	Rv2615c	PE-PGRS family protein PE_PGRS45	2943600	2944985	146	3351902	3362479
888573	PE_PGRS46	Rv2634c	PE-PGRS family protein PE_PGRS46	2960105	2962441	147	3377604	3379950
887706	Rv2646	Rv2646	integrase	2970551	2971549	148	3388062	3389060
887828	Rv2648	Rv2648	insertion sequence element IS6110 transposase(fragment)	2972160	2972486	149	3389758	3391020
888553	Rv2649	Rv2649	insertion sequence element IS986/IS6110 transposase	2972435	2973421	149	3389758	3391020
887478	Rv2650c	Rv2650c	prophage protein	2973795	2975234	150	3391397	3392839
887837	Rv2651c	Rv2651c	prophage protease	2975242	2975775	151	3392847	3393380
888577	Rv2652c	Rv2652c	prophage protein	2975928	2976554	152	3393541	3394167
887367	Rv2653c	Rv2653c	toxin	2976586	2976909	153	3394200	3394523
888154	Rv2654c	Rv2654c	antitoxin	2976989	2977234	154	3394603	3396272
887388	Rv2655c	Rv2655c	prophage protein	2977231	2978658	154	3394603	3396272
888179	Rv2656c	Rv2656c	prophage protein	2978660	2979052	155	3396274	3396923
887399	Rv2657c	Rv2657c	prophage protein	2979049	2979309	155	3396274	3396923
885098	Rv2659c	Rv2659c	prophage integrase	2979691	2980818	156	3397305	3398432
888904	Rv2666	Rv2666	Probable transposase for insertion sequence element IS1081 (fragment)	2983071	2983874	157	3402044	3402847
888339	PE_PGRS47	Rv2741	PE-PGRS family protein PE_PGRS47	3053914	3055491	158	3486888	3488471
887765	PPE43	Rv2768c	PPE family protein PPE43	3076894	3078078	159	3511236	3512420
888461	PE27	Rv2769c	PE family protein PE27	3078158	3078985	160	3512500	3513328
888456	PPE44	Rv2770c	PPE family protein PPE44	3079309	3080457	161	3513652	3514800
888281	Rv2791c	Rv2791c	transposase	3100202	3101581	162	3537266	3538645
887784	Rv2810c	Rv2810c	Probable transposase	3115741	3116142	163	3555527	3555928
888942	Rv2812	Rv2812	transposase	3116818	3118227	164	3556604	3558013
887839	Rv2814c	Rv2814c	insertion sequence element IS986/IS6110 transposase	3120566	3121552	165	3564995	3566256
888511	Rv2815c	Rv2815c	insertion sequence element IS6110 transposase(fragment)	3121501	3121827	165	3564995	3566256
888171	PE_PGRS48	Rv2853	PE-PGRS family protein PE_PGRS48	3162268	3164115	166	3612510	3614399
887173	Rv2885c	Rv2885c	transposase	3194166	3195548	167	3644669	3646055
887824	PPE45	Rv2892c	PPE family protein PPE45	3200794	3202020	168	3651303	3652530
887834	Rv2943	Rv2943	insertion sequence element IS1533 transposase	3288464	3289705	169	3739862	3741904
3205061	Rv2943A	Rv2943A	transposase	3289705	3290235	169	3739862	3741904
887636	Rv2944	Rv2944	insertion sequence element IS1533 transposase	3289790	3290506	169	3739862	3741904
887316	Rv2961	Rv2961	transposase	3313283	3313672	170	3766227	3766616

887390	Rv2978c	Rv2978c	transposase	3333785	3335164	171	3786735	3788114
888940	PPE46	Rv3018c	PPE family protein PPE46	3376939	3378243	172	3830927	3832245
3205087	PE27A	Rv3018A	PE family protein PE27A	3378329	3378415	173	3832331	3832423
888924	PPE47	Rv3021c	pseudo	3379376	3380452	174	3843448	3850569
888512	PPE48	Rv3022c	pseudo	3380440	3380682	174	3843448	3850569
3205088	PE29	Rv3022A	PE family protein PE29	3380679	3380993	174	3843448	3850569
888525	Rv3023c	Rv3023c	transposase	3381375	3382622	175	3851607	3853379
888790	Rv3115	Rv3115	transposase	3481451	3482698	176	3954127	3955374
888892	PPE49	Rv3125c	PPE family protein PPE49	3490476	3491651	177	3964511	3968404
888153	PPE50	Rv3135	PPE family protein PPE50	3501334	3501732	178	3980869	3982608
888835	PPE51	Rv3136	PPE family protein PPE51	3501794	3502936	179	3982670	3983812
887930	PPE52	Rv3144c	PPE family protein PPE52	3510088	3511317	180	3990965	3992194
888794	PPE53	Rv3159c	PPE family protein PPE53	3527391	3529163	181	4008272	4012224
888796	Rv3184	Rv3184	insertion sequence element IS6110 transposase(fragment)	3551281	3551607	182	4041274	4042535
887441	Rv3185	Rv3185	insertion sequence element IS986/IS6110 transposase	3551556	3552542	182	4041274	4042535
888024	Rv3186	Rv3186	insertion sequence element IS6110 transposase(fragment)	3552764	3553090	183	4042764	4044025
887604	Rv3187	Rv3187	insertion sequence element IS986/IS6110 transposase	3553039	3554025	183	4042764	4044025
887628	Rv3191c	Rv3191c	transposase	3557311	3558345	184	4048748	4049782
887314	Rv3325	Rv3325	insertion sequence element IS6110 transposase(fragment)	3710433	3710759	185	4213364	4214625
887563	Rv3326	Rv3326	insertion sequence element IS986/IS6110 transposase	3710708	3711694	185	4213364	4214625
887965	Rv3327	Rv3327	transposase fusion protein	3711749	3713461	186	4214953	4222316
888033	PPE54	Rv3343c	PPE family protein PPE54	3729364	3736935	187	4239582	4253069
888115	PE_PGRS49	Rv3344c	PE-PGRS family protein PE_PGRS49	3736984	3738000	188	4253118	4254308
888114	PE_PGRS50	Rv3345c	PE-PGRS family protein PE_PGRS50	3738158	3742774	189	4254699	4263661
888120	PPE55	Rv3347c	PPE family protein PPE55	3743711	3753184	190	4264600	4275439
888110	Rv3348	Rv3348	transposase	3753765	3754256	191	4276020	4276511
888126	Rv3349c	Rv3349c	transposase	3754293	3755237	192	4276548	4277492
888113	PPE56	Rv3350c	PPE family protein PPE56	3755952	3767102	193	4278207	4289375
887404	PE_PGRS51	Rv3367	PE-PGRS family protein PE_PGRS51	3778568	3780334	194	4300842	4302628
887411	Rv3380c	Rv3380c	insertion sequence element IS986/IS6110 transposase	3795100	3796086	195	4317399	4318660
887646	Rv3381c	Rv3381c	insertion sequence element IS6110 transposase(fragment)	3796035	3796361	195	4317399	4318660
888044	Rv3386	Rv3386	transposase	3800092	3800796	196	4325289	4326660
887820	Rv3387	Rv3387	transposase	3800786	3801463	196	4325289	4326660
888151	PE_PGRS52	Rv3388	PE-PGRS family protein PE_PGRS52	3801653	3803848	197	4326850	4329423
887635	PPE57	Rv3425	PPE family protein PPE57	3842239	3842769	198	4368876	4369406
887622	PPE58	Rv3426	PPE family protein PPE58	3843036	3843734	199	4369673	4370383
887631	Rv3427c	Rv3427c	transposase	3843885	3844640	200	4370534	4371289
887621	Rv3428c	Rv3428c	transposase	3844738	3845970	201	4372756	4373990
887630	PPE59	Rv3429	PPE family protein PPE59	3847165	3847701	202	4397378	4399069
887615	Rv3430c	Rv3430c	transposase	3847642	3848805	202	4397378	4399069
888097	Rv3474	Rv3474	insertion sequence element IS6110 transposase(fragment)	3890830	3891156	203	4460421	4461683
888055	Rv3475	Rv3475	insertion sequence element IS986/IS6110 transposase	3891105	3892091	203	4460421	4461683
888474	PE31	Rv3477	PE family protein PE31	3894093	3894389	204	4472297	4472593
888047	PPE60	Rv3478, mtb39c	PE family protein PPE60	3894426	3895607	205	4472630	4473816
888256	PE_PGRS53	Rv3507	PE-PGRS family protein PE_PGRS53	3926569	3930714	206	4504786	4611956
888270	PE_PGRS54	Rv3508	PE-PGRS family protein PE_PGRS54	3931005	3936710	207	4612247	4625244
888273	PE_PGRS55	Rv3511	PE-PGRS family protein PE_PGRS55	3939617	3941761	208	4628153	4631030
888306	PE_PGRS56	Rv3512	PE-PGRS family protein PE_PGRS56	3943812	3944963	209	4633925	4635114
888294	PE_PGRS57	Rv3514	PE-PGRS family protein PE_PGRS57	3945794	3950263	210	4635945	4641776
888370	PPE61	Rv3532	PPE family protein PPE61	3969343	3970563	211	4660862	4662082
888385	PPE62	Rv3533c	PPE family protein PPE62	3970705	3972453	212	4662224	4663978
888438	PPE63	Rv3539	PPE family protein PPE63	3978059	3979498	213	4669599	4671039
887822	PPE64	Rv3558	PPE family protein PPE64	3997980	3999638	214	4689528	4691186
887874	PE_PGRS58	Rv3590c	PE-PGRS family protein PE_PGRS58	4031404	4033158	215	4722955	4724996
885464	PE_PGRS59	Rv3595c	PE-PGRS family protein PE_PGRS59	4036731	4038050	216	4728569	4729925
885097	PPE65	Rv3621c	PPE family protein PPE65	4060648	4061889	217	4753121	4754363
885712	PE32	Rv3622c	PE family protein PE32	4061899	4062198	218	4754373	4754672
885274	Rv3636	Rv3636	pseudo	4075752	4076984	219	4768239	4770217
885496	Rv3637	Rv3637	Possible transposase	4076484	4076984	219	4768239	4770217
885803	Rv3638	Rv3638	transposase	4076984	4077730	219	4768239	4770217
885324	Rv3640c	Rv3640c	transposase	4078520	4079749	220	4772555	4773784
885832	PE33	Rv3650	PE family protein PE33	4091233	4091517	221	4785333	4785617
886260	PE_PGRS60	Rv3652	PE-PGRS family-related protein PE_PGRS60	4093632	4093946	222	4787733	4788755
886259	PE_PGRS61	Rv3653	PE-PGRS family-related protein PE_PGRS61	4093940	4094527	222	4787733	4788755
886262	PPE66	Rv3738c	PPE family protein PPE66	4189285	4190232	223	4885358	4886305
886257	PPE67	Rv3739c	PPE family protein PPE67	4190284	4190517	224	4886357	4886590
885764	PE34	Rv3746c	PE family protein PE34	4196171	4196506	225	4892244	4892579
885857	Rv3751	Rv3751	Probable integrase (fragment)	4198874	4199089	226	4894952	4895167
886268	Rv3798	Rv3798	insertion sequence element IS1557 transposase	4252993	4254327	227	4949119	4950454
886143	PE_PGRS62	Rv3812	PE-PGRS family protein PE_PGRS62	4276571	4278085	228	4972700	4974214
886151	Rv3827c	Rv3827c	transposase	4301563	4302789	229	4997694	4998920
886180	Rv3844	Rv3844	transposase	4318775	4319266	230	5014923	5015414
886191	PE35	Rv3872	PE family protein PE35	4350745	4351044	231	5047064	5047363
886201	PPE68	Rv3873	PPE family protein PPE68	4351075	4352181	232	5047394	5048501
886227	PPE69	Rv3892c	PPE family protein PPE69	4374484	4375683	233	5070915	5072114
886213	PE36	Rv3893c	PE family protein PE36	4375762	4375995	234	5072193	5072427

Appendix Table 2.3: Details of transmission detection in simulated clusters in Chapter 3. Counts for transmission cluster links for the exclusion and substitution method (mapped to the M. tuberculosis H37Rv strain) and PANPASCO for each cluster separately.

Cluster	Truth		Exclusion				Substitution				PANPASCO			
	transmission links intra-cluster	no relation intra-cluster	TP	TN	FP	FN	TP	TN	FP	FN	TP	TN	FP	FN
C1	127	44	127	0	44	0	106	44	0	21	125	35	9	2
C2	21	0	21	0	0	0	21	0	0	0	21	0	0	0
C3	28	0	28	0	0	0	28	0	0	0	28	0	0	0
C4	287	178	287	0	178	0	272	173	5	15	279	172	6	8
C5	36	0	36	0	0	0	36	0	0	0	36	0	0	0
C6	125	28	125	0	28	0	2	28	0	123	125	15	13	0
C7	205	230	205	0	230	0	0	230	0	205	170	225	5	35
C8	54	1	54	0	1	0	0	1	0	54	54	0	1	0
C9	52	3	52	0	3	0	0	3	0	52	48	3	0	4
C10	28	0	28	0	0	0	0	0	0	28	28	0	0	0
C11	6	0	6	0	0	0	0	0	0	6	6	0	0	0
C12	3	0	3	0	0	0	0	0	0	3	3	0	0	0
C13	101	19	101	0	19	0	0	19	0	101	99	6	13	2
C14	158	32	158	0	32	0	0	32	0	158	149	31	1	9
C15	172	59	172	0	59	0	0	59	0	172	168	49	10	4
C16	95	10	95	0	10	0	0	10	0	95	95	6	4	0
C17	6	0	6	0	0	0	0	0	0	6	6	0	0	0
C18	837	438	837	0	438	0	0	438	0	837	829	300	138	8
C19	53	2	53	0	2	0	0	2	0	53	53	2	0	0
C20	210	90	210	0	90	0	0	90	0	210	203	65	25	7

Appendix Table 2.4: Comparison of SNP-counting methods using different reference genomes in Chapter 3. We used the all three methods (exclusion, substitution, PANPASCO) with the commonly used M. tuberculosis H37Rv and the computational pan-genome reference genomes for classification of links between samples in the simulated dataset. The pairwise method used in PANPASCO gives the best results for both genomes and using the computational pan-genome results in higher sensitivity, accuracy and F-Score.

Method	Reference genome	Sensitivity	Specificity	Accuracy	F-Score
H37Rv	exclusion	**1.000**	0.782	0.793	0.326
	substitution	0.179	**1.000**	0.959	0.303
	PANPASCO	0.896	0.998	**0.993**	**0.930**
pan-genome	exclusion	**1.000**	0.782	0.793	0.326
	substitution	0.000	**1.000**	0.950	0.000
	PANPASCO	0.970	0.995	**0.994**	**0.943**

Appendix Table 2.5: Comparison of differential SNPs detected in the UKTB7 dataset in Chapter 3.

Published order	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	Locus on pan-genome	Locus on H37Rv strain	In Walker et al. 2013
Year of sample isolation	2008	2007	2008	2006	2006	2007	2002	2004	2004	2005	2011	2004	2011	2011	2008	2007	2005			
Published name	P027b	P026c	P026d	P027a	P026a	P026b	P076a	P076b	P076c	P076d	P334	P174	P335	P175	P066	P211	P037			
	G	X	G	X	X	X	X	X	G	X	C	X	C	X	G	X	X	42168	39017	
	G	G	G	T	T	T	T	T	T	T	T	T	T	T	T	T	T	170358	166624	y
	C	C	C	C	C	C	C	C	C	C	C	C	G	C	C	C	C	514289	468370	y
	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	T	682920	632594	y
	X	C	C	X	C	X	C	X	X	C	C	G	C	C	X	C	C	939618	873144	
	X	C	C	X	C	X	C	X	X	C	C	G	C	C	X	C	C	939619	873145	
	G	G	G	G	G	G	X	G	X	X	G	G	A	A	A	A	G	946431	879955	y
	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	T	969914	899416	y
	T	T	T	T	T	T	C	C	C	C	C	C	C	C	C	C	C	1023003	945783	y
	A	A	A	A	A	A	G	G	G	G	G	G	G	G	G	G	G	1167980	1085807	y
	G	G	G	G	G	G	G	G	G	G	G	G	X	C	G	G	G	1179475	1096470	
	C	C	C	C	C	C	C	C	C	C	C	C	C	G	C	C	C	1179513	1096508	
	T	T	T	T	T	T	T	T	T	T	T	T	T	C	T	T	T	1179515	1096510	
	X	X	X	X	G	X	X	X	T	X	X	X	X	X	X	X	X	1260326	-	
	X	G	G	X	C	G	X	G	G	G	G	X	G	X	X	X	X	1375328	1284071	
	G	G	G	G	G	G	X	X	X	G	C	C	G	G	G	G	G	1466454	1375047	y
	G	G	G	G	G	G	X	X	X	G	C	C	G	G	G	G	G	1466455	1375048	y
	X	X	C	X	C	X	G	C	X	X	X	X	X	X	X	C	X	1651259	-	
	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	T	1691909	1584192	y
	A	A	A	A	A	A	G	G	G	G	G	G	G	G	G	G	G	2160699	1897180	y
	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	G	2404646	2096225	y
	C	C	C	C	C	C	C	C	C	C	T	C	C	C	C	C	C	2649456	2303882	y
	G	G	G	G	G	G	A	A	A	A	A	A	A	A	A	A	G	2738716	2383755	y
	X	G	X	C	X	G	C	X	X	X	G	X	G	X	X	X	X	2768703	2413615	
	C	C	C	C	C	C	C	C	C	C	C	C	G	C	C	C	C	2897425	2537713	y
	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	A	3261991	2854386	y
	X	X	X	X	X	X	X	X	X	X	T	X	X	X	X	X	C	3362566	2945072	
	C	C	C	X	G	C	X	C	C	X	C	X	X	C	C	C	X	3896906	3426077	
	C	X	C	X	X	X	X	X	X	X	G	X	X	C	X	X	X	4386001	-	
	C	C	C	C	C	C	C	C	C	C	T	C	C	C	C	C	C	4814383	4120151	y
	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	A	4852391	4156479	y
	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	C	G	4946995	4250896	y
	G	G	G	G	G	G	G	G	G	G	G	G	G	G	T	G	G	4993297	4297167	
	X	C	X	C	C	X	C	X	X	X	C	C	X	C	X	G	C	5082697	-	
	G	G	G	G	G	G	X	G	X	X	G	G	C	C	C	C	G	5096147	4398748	y
	X	X	X	X	G	X	X	X	X	X	C	X	X	X	X	X	X	5141599	-	

Appendix 3 - Appendix for Chapter 4

Appendix Table 3.1: **Cluster assignment, molecular drug resistance prediction, lineage classicifation and parts of extracted metadata of the 1339 multi- and extensive drug resistant *M. tuberculosis* isolates analyzed in Chapter 4.** Four isolates were classified as not multi-resistent according to the molecualr drug resistance prediction. SRA: sequence read archive; MDR: multidrug resistant; XDR: extensively drug-resistant; NA: not available

Isolate id	Cluster name	Cluster size	Dataset	Drug resistance prediction	Year of isolation	Sample Name	Country of isolation	Lineage Classification
SRR4035489	1	79	Public dataset	MDR	2013	P21-Lung1-C1	South Africa	3 Delhi-CAS
SRR4035490	1	79	Public dataset	MDR	2013	P21-Lung1-C2	South Africa	3 Delhi-CAS
SRR4035491	1	79	Public dataset	MDR	2013	P21-Lung1-C3	South Africa	3 Delhi-CAS
SRR4035492	1	79	Public dataset	MDR	2013	P21-Lung1-C4	South Africa	3 Delhi-CAS
SRR4035493	1	79	Public dataset	MDR	2013	P21-Lung1-C5	South Africa	3 Delhi-CAS
SRR4035494	1	79	Public dataset	MDR	2013	P21-Lung2-A1	South Africa	3 Delhi-CAS
SRR4035495	1	79	Public dataset	MDR	2013	P21-Lung2-A2	South Africa	3 Delhi-CAS
SRR4035496	1	79	Public dataset	MDR	2013	P21-Lung2-A3	South Africa	3 Delhi-CAS
SRR4035497	1	79	Public dataset	MDR	2013	P21-Lung2-A4	South Africa	3 Delhi-CAS
SRR4035498	1	79	Public dataset	MDR	2013	P21-Lung2-A5	South Africa	3 Delhi-CAS
SRR4035500	1	79	Public dataset	MDR	2013	P21-Lung2-B1	South Africa	3 Delhi-CAS
SRR4035501	1	79	Public dataset	MDR	2013	P21-Lung2-B2	South Africa	3 Delhi-CAS
SRR4035502	1	79	Public dataset	MDR	2013	P21-Lung2-B3	South Africa	3 Delhi-CAS
SRR4035503	1	79	Public dataset	MDR	2013	P21-Lung2-B5	South Africa	3 Delhi-CAS
SRR4035504	1	79	Public dataset	MDR	2013	P21-Lung2-C1	South Africa	3 Delhi-CAS
SRR4035505	1	79	Public dataset	MDR	2013	P21-Lung2-C2	South Africa	3 Delhi-CAS
SRR4035506	1	79	Public dataset	MDR	2013	P21-Lung2-C3	South Africa	3 Delhi-CAS
SRR4035507	1	79	Public dataset	MDR	2013	P21-Lung2-C4	South Africa	3 Delhi-CAS
SRR4035508	1	79	Public dataset	MDR	2013	P21-Lung2-C5	South Africa	3 Delhi-CAS
SRR4035509	1	79	Public dataset	MDR	2013	P21-Lung3-A1	South Africa	3 Delhi-CAS
SRR4035511	1	79	Public dataset	MDR	2013	P21-Lung3-B1	South Africa	3 Delhi-CAS
SRR4035512	1	79	Public dataset	MDR	2013	P21-Lung3-B2	South Africa	3 Delhi-CAS
SRR4035513	1	79	Public dataset	MDR	2013	P21-Lung3-B3	South Africa	3 Delhi-CAS
SRR4035514	1	79	Public dataset	MDR	2013	P21-Lung3-B4	South Africa	3 Delhi-CAS
SRR4035515	1	79	Public dataset	MDR	2013	P21-Lung3-B5	South Africa	3 Delhi-CAS
SRR4035516	1	79	Public dataset	MDR	2013	P21-Lung3-C1	South Africa	3 Delhi-CAS
SRR4035517	1	79	Public dataset	MDR	2013	P21-Lung3-C2	South Africa	3 Delhi-CAS
SRR4035518	1	79	Public dataset	MDR	2013	P21-Lung3-C3	South Africa	3 Delhi-CAS
SRR4035519	1	79	Public dataset	MDR	2013	P21-Lung3-C4	South Africa	3 Delhi-CAS
SRR4035520	1	79	Public dataset	MDR	2013	P21-Lung3-C5	South Africa	3 Delhi-CAS
SRR4035522	1	79	Public dataset	MDR	2013	P21-Lung4-A3	South Africa	3 Delhi-CAS
SRR4035523	1	79	Public dataset	MDR	2013	P21-Lung4-B1	South Africa	3 Delhi-CAS
SRR4035524	1	79	Public dataset	MDR	2013	P21-Lung4-B2	South Africa	3 Delhi-CAS
SRR4035525	1	79	Public dataset	MDR	2013	P21-Lung4-B3	South Africa	3 Delhi-CAS
SRR4035526	1	79	Public dataset	MDR	2013	P21-Lung4-B4	South Africa	3 Delhi-CAS
SRR4035527	1	79	Public dataset	MDR	2013	P21-Lung4-B5	South Africa	3 Delhi-CAS
SRR4035528	1	79	Public dataset	MDR	2013	P21-Lung4-C1	South Africa	3 Delhi-CAS
SRR4035529	1	79	Public dataset	MDR	2013	P21-Lung4-C2	South Africa	3 Delhi-CAS
SRR4035530	1	79	Public dataset	MDR	2013	P21-Lung4-C3	South Africa	3 Delhi-CAS
SRR4035531	1	79	Public dataset	MDR	2013	P21-Lung4-C4	South Africa	3 Delhi-CAS
SRR4035535	1	79	Public dataset	MDR	2013	P21-Lung5-A2	South Africa	3 Delhi-CAS
SRR4035536	1	79	Public dataset	MDR	2013	P21-Lung5-A3	South Africa	3 Delhi-CAS
SRR4035537	1	79	Public dataset	MDR	2013	P21-Lung5-A4	South Africa	3 Delhi-CAS
SRR4035538	1	79	Public dataset	MDR	2013	P21-Lung5-A5	South Africa	3 Delhi-CAS
SRR4035539	1	79	Public dataset	MDR	2013	P21-Lung5-B1	South Africa	3 Delhi-CAS
SRR4035540	1	79	Public dataset	MDR	2013	P21-Lung5-B2	South Africa	3 Delhi-CAS
SRR4035541	1	79	Public dataset	MDR	2013	P21-Lung5-B4	South Africa	3 Delhi-CAS
SRR4035544	1	79	Public dataset	MDR	2013	P21-Lung5-C1	South Africa	3 Delhi-CAS
SRR4035545	1	79	Public dataset	MDR	2013	P21-Lung5-C2	South Africa	3 Delhi-CAS
SRR4035546	1	79	Public dataset	MDR	2013	P21-Lung5-C3	South Africa	3 Delhi-CAS
SRR4035547	1	79	Public dataset	MDR	2013	P21-Lung5-C4	South Africa	3 Delhi-CAS
SRR4035548	1	79	Public dataset	MDR	2013	P21-Lung6-A2	South Africa	3 Delhi-CAS
SRR4035550	1	79	Public dataset	MDR	2013	P21-Lung6-A5	South Africa	3 Delhi-CAS
SRR4035551	1	79	Public dataset	MDR	2013	P21-Lung6-B1	South Africa	3 Delhi-CAS
SRR4035552	1	79	Public dataset	MDR	2013	P21-Lung6-B2	South Africa	3 Delhi-CAS
SRR4035553	1	79	Public dataset	MDR	2013	P21-Lung6-B3	South Africa	3 Delhi-CAS
SRR4035555	1	79	Public dataset	MDR	2013	P21-Lung6-B5	South Africa	3 Delhi-CAS
SRR4035556	1	79	Public dataset	MDR	2013	P21-Lung6-C2	South Africa	3 Delhi-CAS
SRR4035558	1	79	Public dataset	MDR	2013	P21-Lung6-C4	South Africa	3 Delhi-CAS
SRR4036006	1	79	Public dataset	MDR	2013	P21-Eta-A1	South Africa	3 Delhi-CAS
SRR4036008	1	79	Public dataset	MDR	2013	P21-Eta-A3	South Africa	3 Delhi-CAS
SRR4036009	1	79	Public dataset	MDR	2013	P21-Eta-A4	South Africa	3 Delhi-CAS

SRR4036010	1	79	Public dataset	MDR	2013	P21-Liver-A1	South Africa	3 Delhi-CAS
SRR4036011	1	79	Public dataset	MDR	2013	P21-Liver-A3	South Africa	3 Delhi-CAS
SRR4036012	1	79	Public dataset	MDR	2013	P21-Liver-B1	South Africa	3 Delhi-CAS
SRR4036013	1	79	Public dataset	MDR	2013	P21-Liver-B2	South Africa	3 Delhi-CAS
SRR4036014	1	79	Public dataset	MDR	2013	P21-Liver-B3	South Africa	3 Delhi-CAS
SRR4036015	1	79	Public dataset	MDR	2013	P21-Liver-B4	South Africa	3 Delhi-CAS
SRR4036016	1	79	Public dataset	MDR	2013	P21-Liver-B5	South Africa	3 Delhi-CAS
SRR4036017	1	79	Public dataset	MDR	2013	P21-Liver-C2	South Africa	3 Delhi-CAS
SRR4036019	1	79	Public dataset	MDR	2013	P21-Liver-C4	South Africa	3 Delhi-CAS
SRR4036020	1	79	Public dataset	MDR	2013	P21-Liver-C5	South Africa	3 Delhi-CAS
SRR4036021	1	79	Public dataset	MDR	2013	P21-Lung1-A1	South Africa	3 Delhi-CAS
SRR4036022	1	79	Public dataset	MDR	2013	P21-Lung1-A2	South Africa	3 Delhi-CAS
SRR4036023	1	79	Public dataset	MDR	2013	P21-Lung1-A3	South Africa	3 Delhi-CAS
SRR4036024	1	79	Public dataset	MDR	2013	P21-Lung1-A4	South Africa	3 Delhi-CAS
SRR4036026	1	79	Public dataset	MDR	2013	P21-Lung1-B1	South Africa	3 Delhi-CAS
SRR4036027	1	79	Public dataset	MDR	2013	P21-Lung1-B2	South Africa	3 Delhi-CAS
SRR4036028	1	79	Public dataset	MDR	2013	P21-Lung1-B5	South Africa	3 Delhi-CAS
10962-13	2	56	German dataset	MDR	2013		Germany	4.2.1 Ural
3593-12	2	56	German dataset	MDR	2012		Germany	4.2.1 Ural
ERR1664643	2	56	Public dataset	MDR	NA	SRMTB25		4.2.1 Ural
ERR1664653	2	56	Public dataset	MDR	NA	SRMTB35		4.2.1 Ural
ERR1664663	2	56	Public dataset	MDR	NA	SRMTB45		4.2.1 Ural
SRR3743369	2	56	Public dataset	MDR	2015	Mycobacterium tuberculosis Moldova 14660	Moldova	4.2.1 Ural
SRR3743378	2	56	Public dataset	MDR	2015	Mycobacterium tuberculosis Moldova 1059	Moldova	4.2.1 Ural
SRR3743381	2	56	Public dataset	MDR	2015	Mycobacterium tuberculosis Moldova 22253	Moldova	4.2.1 Ural
SRR3743382	2	56	Public dataset	MDR	2015	Mycobacterium tuberculosis Moldova 10197	Moldova	4.2.1 Ural
SRR3743383	2	56	Public dataset	MDR	2015	Mycobacterium tuberculosis Moldova 19353	Moldova	4.2.1 Ural
SRR3743385	2	56	Public dataset	MDR	2015	Mycobacterium tuberculosis Moldova 11314	Moldova	4.2.1 Ural
SRR3743389	2	56	Public dataset	MDR	2015	Mycobacterium tuberculosis Moldova 14381	Moldova	4.2.1 Ural
SRR3743394	2	56	Public dataset	XDR	2015	Mycobacterium tuberculosis Moldova 7725	Moldova	4.2.1 Ural
SRR3743398	2	56	Public dataset	MDR	2015	Mycobacterium tuberculosis Moldova 5230	Moldova	4.2.1 Ural
SRR3743402	2	56	Public dataset	MDR	2015	Mycobacterium tuberculosis Moldova 2582	Moldova	4.2.1 Ural
SRR3743403	2	56	Public dataset	MDR	2015	Mycobacterium tuberculosis Moldova 16266	Moldova	4.2.1 Ural
SRR3743405	2	56	Public dataset	MDR	2015	Mycobacterium tuberculosis Moldova 21445	Moldova	4.2.1 Ural
SRR3743408	2	56	Public dataset	MDR	2015	Mycobacterium tuberculosis Moldova 16757	Moldova	4.2.1 Ural
SRR3743412	2	56	Public dataset	MDR	2015	Mycobacterium tuberculosis Moldova 24564	Moldova	4.2.1 Ural
SRR3743416	2	56	Public dataset	MDR	2015	Mycobacterium tuberculosis Moldova 5079	Moldova	4.2.1 Ural
SRR3743449	2	56	Public dataset	MDR	2015	Mycobacterium tuberculosis Moldova 5900	Moldova	4.2.1 Ural
SRR3743459	2	56	Public dataset	MDR	2015	Mycobacterium tuberculosis Moldova 10775	Moldova	4.2.1 Ural
SRR3743460	2	56	Public dataset	MDR	2015	Mycobacterium tuberculosis Moldova 21950	Moldova	4.2.1 Ural
SRR3743461	2	56	Public dataset	MDR	2015	Mycobacterium tuberculosis Moldova 14713	Moldova	4.2.1 Ural
SRR3743462	2	56	Public dataset	MDR	2015	Mycobacterium tuberculosis Moldova 13152	Moldova	4.2.1 Ural
SRR3743472	2	56	Public dataset	MDR	2016	Mycobacterium tuberculosis Moldova 157	Moldova	4.2.1 Ural
SRR3743479	2	56	Public dataset	MDR	2015	Mycobacterium tuberculosis Moldova 11678	Moldova	4.2.1 Ural
SRR3743480	2	56	Public dataset	XDR	2015	Mycobacterium tuberculosis Moldova 19029	Moldova	4.2.1 Ural
SRR3743484	2	56	Public dataset	MDR	2015	Mycobacterium tuberculosis Moldova 11473	Moldova	4.2.1 Ural
SRR3743486	2	56	Public dataset	MDR	2015	Mycobacterium tuberculosis Moldova 17146	Moldova	4.2.1 Ural
SRR3743491	2	56	Public dataset	MDR	2015	Mycobacterium tuberculosis Moldova 22473	Moldova	4.2.1 Ural
SRR3743498	2	56	Public dataset	MDR	2015	Mycobacterium tuberculosis Moldova 2063	Moldova	4.2.1 Ural
SRR5153333	2	56	Public dataset	MDR	2014	Mycobacterium tuberculosis G-081	Georgia	4.2.1 Ural

SRR5153821	2	56	Public dataset	MDR	2015	Mycobacterium tuberculosis complex 15408	Moldova	4.2.1 Ural
SRR5153824	2	56	Public dataset	MDR	2015	Mycobacterium tuberculosis complex 15287	Moldova	4.2.1 Ural
SRR5153827	2	56	Public dataset	MDR	2015	Mycobacterium tuberculosis complex 18122	Moldova	4.2.1 Ural
SRR5153828	2	56	Public dataset	MDR	2015	Mycobacterium tuberculosis complex 12637	Moldova	4.2.1 Ural
SRR5153832	2	56	Public dataset	MDR	2015	Mycobacterium tuberculosis complex 20030	Moldova	4.2.1 Ural
SRR5153846	2	56	Public dataset	MDR	2009	Mycobacterium tuberculosis complex 435	Moldova	4.2.1 Ural
SRR5153849	2	56	Public dataset	MDR	2015	Mycobacterium tuberculosis complex 16237	Moldova	4.2.1 Ural
SRR5153854	2	56	Public dataset	MDR	2016	Mycobacterium tuberculosis complex 13166	Moldova	4.2.1 Ural
SRR5153860	2	56	Public dataset	MDR	2015	Mycobacterium tuberculosis complex 22812	Moldova	4.2.1 Ural
SRR5153865	2	56	Public dataset	MDR	2015	Mycobacterium tuberculosis complex 18922	Moldova	4.2.1 Ural
SRR5153877	2	56	Public dataset	MDR	2015	Mycobacterium tuberculosis complex 19217	Moldova	4.2.1 Ural
SRR5153878	2	56	Public dataset	MDR	2015	Mycobacterium tuberculosis complex 16467	Moldova	4.2.1 Ural
SRR5153882	2	56	Public dataset	MDR	2015	Mycobacterium tuberculosis complex 22083	Moldova	4.2.1 Ural
SRR5153885	2	56	Public dataset	MDR	2015	Mycobacterium tuberculosis complex 13252	Moldova	4.2.1 Ural
SRR5153887	2	56	Public dataset	MDR	2015	Mycobacterium tuberculosis complex 19266	Moldova	4.2.1 Ural
SRR5153902	2	56	Public dataset	MDR	2015	Mycobacterium tuberculosis complex 14040	Moldova	4.2.1 Ural
SRR5153906	2	56	Public dataset	MDR	2015	Mycobacterium tuberculosis complex 12299	Moldova	4.2.1 Ural
SRR5153910	2	56	Public dataset	MDR	2016	Mycobacterium tuberculosis complex 10981	Moldova	4.2.1 Ural
SRR5153915	2	56	Public dataset	MDR	2015	Mycobacterium tuberculosis complex 19495	Moldova	4.2.1 Ural
SRR5153917	2	56	Public dataset	MDR	2016	Mycobacterium tuberculosis complex 13346	Moldova	4.2.1 Ural
SRR5153921	2	56	Public dataset	MDR	2016	Mycobacterium tuberculosis complex 13597	Moldova	4.2.1 Ural
SRR5153929	2	56	Public dataset	MDR	2015	Mycobacterium tuberculosis complex 15201	Moldova	4.2.1 Ural
SRR5163781	2	56	Public dataset	MDR	2016	Mycobacterium tuberculosis complex 11224	Moldova	4.2.1 Ural
SRR4033152	3	54	Public dataset	MDR	2013	P16-Eta-A1	South Africa	2.2.1 Beijing Asian/Africa 2
SRR4033153	3	54	Public dataset	MDR	2013	P16-Eta-A2	South Africa	2.2.1 Beijing Asian/Africa 2
SRR4033154	3	54	Public dataset	MDR	2013	P16-Eta-A3	South Africa	2.2.1 Beijing Asian/Africa 2
SRR4033155	3	54	Public dataset	MDR	2013	P16-Eta-A5	South Africa	2.2.1 Beijing Asian/Africa 2
SRR4033157	3	54	Public dataset	MDR	2013	P16-Eta-B2	South Africa	2.2.1 Beijing Asian/Africa 2
SRR4033158	3	54	Public dataset	MDR	2013	P16-Eta-C1	South Africa	2.2.1 Beijing Asian/Africa 2
SRR4033159	3	54	Public dataset	MDR	2013	P16-Eta-C3	South Africa	2.2.1 Beijing Asian/Africa 2
SRR4033160	3	54	Public dataset	MDR	2013	P16-Eta-C4	South Africa	2.2.1 Beijing Asian/Africa 2
SRR4033161	3	54	Public dataset	MDR	2013	P16-Liver-A1	South Africa	2.2.1 Beijing Asian/Africa 2
SRR4033163	3	54	Public dataset	MDR	2013	P16-Liver-A2	South Africa	2.2.1 Beijing Asian/Africa 2
SRR4033164	3	54	Public dataset	MDR	2013	P16-Liver-A3	South Africa	2.2.1 Beijing Asian/Africa 2
SRR4033165	3	54	Public dataset	MDR	2013	P16-Liver-B1	South Africa	2.2.1 Beijing Asian/Africa 2
SRR4033166	3	54	Public dataset	MDR	2013	P16-Liver-B2	South Africa	2.2.1 Beijing Asian/Africa 2
SRR4033167	3	54	Public dataset	MDR	2013	P16-Liver-B3	South Africa	2.2.1 Beijing Asian/Africa 2
SRR4033169	3	54	Public dataset	MDR	2013	P16-Lung1-A1	South Africa	2.2.1 Beijing Asian/Africa 2
SRR4033170	3	54	Public dataset	MDR	2013	P16-Lung1-A2	South Africa	2.2.1 Beijing Asian/Africa 2
SRR4033171	3	54	Public dataset	MDR	2013	P16-Lung1-A3	South Africa	2.2.1 Beijing Asian/Africa 2
SRR4033172	3	54	Public dataset	MDR	2013	P16-Lung1-A5	South Africa	2.2.1 Beijing Asian/Africa 2
SRR4033173	3	54	Public dataset	MDR	2013	P16-Lung1-B1	South Africa	2.2.1 Beijing Asian/Africa 2
SRR4033174	3	54	Public dataset	MDR	2013	P16-Lung1-B2	South Africa	2.2.1 Beijing Asian/Africa 2
SRR4033175	3	54	Public dataset	MDR	2013	P16-Lung1-B4	South Africa	2.2.1 Beijing Asian/Africa 2
SRR4033176	3	54	Public dataset	MDR	2013	P16-Lung1-B5	South Africa	2.2.1 Beijing Asian/Africa 2
SRR4033177	3	54	Public dataset	MDR	2013	P16-Lung1-C1	South Africa	2.2.1 Beijing Asian/Africa 2
SRR4033178	3	54	Public dataset	MDR	2013	P16-Lung1-C3	South Africa	2.2.1 Beijing Asian/Africa 2
SRR4033180	3	54	Public dataset	MDR	2013	P16-Lung1-C4	South Africa	2.2.1 Beijing Asian/Africa 2
SRR4033181	3	54	Public dataset	MDR	2013	P16-Lung1-C5	South Africa	2.2.1 Beijing Asian/Africa 2
SRR4033182	3	54	Public dataset	MDR	2013	P16-Lung2-A1	South Africa	2.2.1 Beijing Asian/Africa 2
SRR4033183	3	54	Public dataset	MDR	2013	P16-Lung2-A2	South Africa	2.2.1 Beijing Asian/Africa 2
SRR4033184	3	54	Public dataset	MDR	2013	P16-Lung2-A3	South Africa	2.2.1 Beijing Asian/Africa 2
SRR4033185	3	54	Public dataset	MDR	2013	P16-Lung2-A4	South Africa	2.2.1 Beijing Asian/Africa 2
SRR4033186	3	54	Public dataset	MDR	2013	P16-Lung2-A5	South Africa	2.2.1 Beijing Asian/Africa 2
SRR4033187	3	54	Public dataset	MDR	2013	P16-Lung2-B2	South Africa	2.2.1 Beijing Asian/Africa 2

SRR4033188	3	54	Public dataset	MDR	2013	P16-Lung2-B3	South Africa	2.2.1 Beijing Asian/Africa 2
SRR4033189	3	54	Public dataset	MDR	2013	P16-Lung4-A1	South Africa	2.2.1 Beijing Asian/Africa 2
SRR4033191	3	54	Public dataset	MDR	2013	P16-Lung4-A2	South Africa	2.2.1 Beijing Asian/Africa 2
SRR4033192	3	54	Public dataset	MDR	2013	P16-Lung4-B1	South Africa	2.2.1 Beijing Asian/Africa 2
SRR4033193	3	54	Public dataset	MDR	2013	P16-Lung4-B2	South Africa	2.2.1 Beijing Asian/Africa 2
SRR4033194	3	54	Public dataset	MDR	2013	P16-Lung4-B3	South Africa	2.2.1 Beijing Asian/Africa 2
SRR4033195	3	54	Public dataset	MDR	2013	P16-Lung4-B4	South Africa	2.2.1 Beijing Asian/Africa 2
SRR4033196	3	54	Public dataset	MDR	2013	P16-Lung4-C1	South Africa	2.2.1 Beijing Asian/Africa 2
SRR4033197	3	54	Public dataset	MDR	2013	P16-Lung5-A1	South Africa	2.2.1 Beijing Asian/Africa 2
SRR4033198	3	54	Public dataset	MDR	2013	P16-Lung5-A2	South Africa	2.2.1 Beijing Asian/Africa 2
SRR4033199	3	54	Public dataset	MDR	2013	P16-Lung5-A3	South Africa	2.2.1 Beijing Asian/Africa 2
SRR4033201	3	54	Public dataset	MDR	2013	P16-Lung5-A4	South Africa	2.2.1 Beijing Asian/Africa 2
SRR4033204	3	54	Public dataset	MDR	2013	P16-Lung5-A5	South Africa	2.2.1 Beijing Asian/Africa 2
SRR4033205	3	54	Public dataset	MDR	2013	P16-Lung5-B1	South Africa	2.2.1 Beijing Asian/Africa 2
SRR4033206	3	54	Public dataset	MDR	2013	P16-Lung5-B2	South Africa	2.2.1 Beijing Asian/Africa 2
SRR4033207	3	54	Public dataset	MDR	2013	P16-Lung5-B3	South Africa	2.2.1 Beijing Asian/Africa 2
SRR4033208	3	54	Public dataset	MDR	2013	P16-Lung5-B4	South Africa	2.2.1 Beijing Asian/Africa 2
SRR4033209	3	54	Public dataset	MDR	2013	P16-Lung5-C1	South Africa	2.2.1 Beijing Asian/Africa 2
SRR4033210	3	54	Public dataset	MDR	2013	P16-Lung5-C2	South Africa	2.2.1 Beijing Asian/Africa 2
SRR4033211	3	54	Public dataset	MDR	2013	P16-Lung5-C3	South Africa	2.2.1 Beijing Asian/Africa 2
SRR4033212	3	54	Public dataset	MDR	2013	P16-Lung5-C5	South Africa	2.2.1 Beijing Asian/Africa 2
SRR4033213	3	54	Public dataset	MDR	2013	P16-Spleen-A1	South Africa	2.2.1 Beijing Asian/Africa 2
ERR1633796	4	33	Public dataset	MDR	2010	KSP990	South Africa	2.2.2 Beijing Ancestral 1
ERR1873395	4	33	Public dataset	XDR	2008	M 15 A673 9073 F1 Mycobacterium AACCGAG L003	South Africa	2.2.2 Beijing Ancestral 1
ERR1873397	4	33	Public dataset	XDR	2010	M 15 A676 10010 F1 Mycobacterium AAGGTAC L003	South Africa	2.2.2 Beijing Ancestral 1
ERR1873421	4	33	Public dataset	XDR	2009	M 15 A711 F2 R15950 GCTCGGT L004	South Africa	2.2.2 Beijing Ancestral 1
ERR1873422	4	33	Public dataset	XDR	2010	M 15 A712 F2 R14816 GGAGAAC L004	South Africa	2.2.2 Beijing Ancestral 1
ERR1873426	4	33	Public dataset	XDR	2009	M 15 A720 F2 R4775 TCCGTCT L004	South Africa	2.2.2 Beijing Ancestral 1
ERR1873429	4	33	Public dataset	XDR	2010	M 15 A724 F2 R11689 TGGCTTC L004	South Africa	2.2.2 Beijing Ancestral 1
ERR1873430	4	33	Public dataset	XDR	2010	M 15 A725 F2 R15539 TGGTGGT L004	South Africa	2.2.2 Beijing Ancestral 1
ERR1873436	4	33	Public dataset	XDR	2010	M tuberculosis R11121 LFO46Pool106 3312 L6 AACGTGAT L006	South Africa	2.2.2 Beijing Ancestral 1
ERR1873468	4	33	Public dataset	XDR	2011	R13121	South Africa	2.2.2 Beijing Ancestral 1
ERR1873480	4	33	Public dataset	XDR	2011	R15949 LFO46Pool105 3311 L5 GGTGCGAA L005	South Africa	2.2.2 Beijing Ancestral 1
ERR1873486	4	33	Public dataset	MDR	2011	R16787 pool 282 L2 ATCCTGTA L002	South Africa	2.2.2 Beijing Ancestral 1
ERR1873490	4	33	Public dataset	XDR	2012	R17203 pool 282 L2 GACTAGTA L002	South Africa	2.2.2 Beijing Ancestral 1
ERR1873494	4	33	Public dataset	XDR	2012	R17879 pool 282 L2 CAGCGTTA L002	South Africa	2.2.2 Beijing Ancestral 1
ERR1873496	4	33	Public dataset	XDR	2012	R18045 pool 283 L3 AGCACCTC L003	South Africa	2.2.2 Beijing Ancestral 1
ERR1873498	4	33	Public dataset	MDR	2012	R18138 pool 282 L2 CATACCAA L002	South Africa	2.2.2 Beijing Ancestral 1
ERR1873503	4	33	Public dataset	XDR	2012	R18476 pool 282 L2 CCGAAGTA L002	South Africa	2.2.2 Beijing Ancestral 1
ERR1873504	4	33	Public dataset	XDR	2012	R18529	South Africa	2.2.2 Beijing Ancestral 1
ERR1873507	4	33	Public dataset	XDR	2012	R18920 pool 282 L2 CGACTGGA L002	South Africa	2.2.2 Beijing Ancestral 1
ERR1873510	4	33	Public dataset	XDR	2012	R19048 pool 283 L3 CACCTTAC L003	South Africa	2.2.2 Beijing Ancestral 1
ERR1873514	4	33	Public dataset	XDR	2012	R19266	South Africa	2.2.2 Beijing Ancestral 1
ERR1873520	4	33	Public dataset	XDR	2012	R19816	South Africa	2.2.2 Beijing Ancestral 1
ERR1873521	4	33	Public dataset	XDR	2008	R4312	South Africa	2.2.2 Beijing Ancestral 1
ERR1873523	4	33	Public dataset	XDR	2008	R4465	South Africa	2.2.2 Beijing Ancestral 1
ERR1873525	4	33	Public dataset	XDR	2008	R4489 LFO46Pool105 3311 L5 AAACATCG L005	South Africa	2.2.2 Beijing Ancestral 1
ERR1873529	4	33	Public dataset	XDR	2008	R4801 LFO46Pool106 3312 L6 AGGCTAAC L006	South Africa	2.2.2 Beijing Ancestral 1
ERR1873533	4	33	Public dataset	MDR	2009	R4863 LFO46Pool105 3311 L5 CAGATCTG L005	South Africa	2.2.2 Beijing Ancestral 1
ERR1873536	4	33	Public dataset	XDR	2009	R5317 LFO46Pool105 3311 L5 CGCTGATC L005	South Africa	2.2.2 Beijing Ancestral 1
ERR1873538	4	33	Public dataset	XDR	2009	R5354 LFO46Pool105 3311 L5 ACAAGCTA L005	South Africa	2.2.2 Beijing Ancestral 1
ERR1873541	4	33	Public dataset	XDR	2009	R5908	South Africa	2.2.2 Beijing Ancestral 1
ERR1873542	4	33	Public dataset	XDR	2009	R5954	South Africa	2.2.2 Beijing Ancestral 1
ERR1873547	4	33	Public dataset	XDR	2009	R6768 LFO46Pool106 3312 L6 CCATCCTC L006	South Africa	2.2.2 Beijing Ancestral 1

ERR1873553	4	33	Public dataset	XDR	2010	R8247 LFO46Pool105 3311 L5 GAATCTGA L005	South Africa	2.2.2 Beijing Ancestral 1
1296-12	5	30	German dataset	MDR	2012		Germany	2.2.1 Beijing
ERR1633956	5	30	Public dataset	MDR	2010	KSP1150	South Africa	2.2.1 Beijing
ERR1633964	5	30	Public dataset	MDR	2010	KSP1158	South Africa	2.2.1 Beijing
ERR1873390	5	30	Public dataset	XDR	2009	M 15 A665 5917 F1 Mycobacterium ACATTGG L003	South Africa	2.2.1 Beijing
ERR1873399	5	30	Public dataset	XDR	2010	M 15 A678 10282 F1 Mycobacterium ACAGCAG L003	South Africa	2.2.1 Beijing
ERR1873403	5	30	Public dataset	XDR	2011	M 15 A685 13614 F1 Mycobacterium AGTCACT L003	South Africa	2.2.1 Beijing
ERR1873412	5	30	Public dataset	XDR	2010	M 15 A699 F2 R19963 CGACTGG L004	South Africa	2.2.1 Beijing
ERR1873413	5	30	Public dataset	XDR	2010	M 15 A700 F2 R13911 CGCATAC L004	South Africa	2.2.1 Beijing
ERR1873414	5	30	Public dataset	XDR	2010	M 15 A701 F2 R13403 CTCAATG L004	South Africa	2.2.1 Beijing
ERR1873415	5	30	Public dataset	XDR	2011	M 15 A703 F2 R20236 CTGGCAT L004	South Africa	2.2.1 Beijing
ERR1873416	5	30	Public dataset	MDR	2010	M 15 A704 F2 R16888 GAATCTG L004	South Africa	2.2.1 Beijing
ERR1873423	5	30	Public dataset	XDR	2011	M 15 A715 F2 R18095 GTCGTAG L004	South Africa	2.2.1 Beijing
ERR1873431	5	30	Public dataset	XDR	2009	M 15 A726 F2 R14770 TTCACGC L004	South Africa	2.2.1 Beijing
ERR1873450	5	30	Public dataset	XDR	2010	R10819 LFO46Pool105 3311 L5 GCCACATA L005	South Africa	2.2.1 Beijing
ERR1873451	5	30	Public dataset	XDR	2010	R10854	South Africa	2.2.1 Beijing
ERR1873455	5	30	Public dataset	XDR	2010	R10951 pool 283 L3 TAGGATGA L003	South Africa	2.2.1 Beijing
ERR1873457	5	30	Public dataset	XDR	2010	R11138 pool 283 L3 TATCAGCA L003	South Africa	2.2.1 Beijing
ERR1873469	5	30	Public dataset	XDR	2011	R13123	South Africa	2.2.1 Beijing
ERR1873471	5	30	Public dataset	XDR	2011	R13342	South Africa	2.2.1 Beijing
ERR1873478	5	30	Public dataset	XDR	2011	R15692 LFO46Pool106 3312 L6 AACTCACC L006	South Africa	2.2.1 Beijing
ERR1873479	5	30	Public dataset	XDR	2011	R15871	South Africa	2.2.1 Beijing
ERR1873481	5	30	Public dataset	XDR	2011	R16462	South Africa	2.2.1 Beijing
ERR1873484	5	30	Public dataset	XDR	2011	R16642	South Africa	2.2.1 Beijing
ERR1873487	5	30	Public dataset	XDR	2011	R16869 pool 282 L2 ATTGAGGA L002	South Africa	2.2.1 Beijing
ERR1873491	5	30	Public dataset	XDR	2012	R17207 pool 282 L2 CAATGGAA L002	South Africa	2.2.1 Beijing
ERR1873499	5	30	Public dataset	XDR	2012	R18174 pool 283 L3 AGGCTAAC L003	South Africa	2.2.1 Beijing
ERR1873501	5	30	Public dataset	XDR	2012	R18343 pool 282 L2 CCAGTTCA L002	South Africa	2.2.1 Beijing
ERR1873537	5	30	Public dataset	XDR	2009	R5318 LFO46Pool106 3312 L6 ATAGCGAC L006	South Africa	2.2.1 Beijing
ERR1873558	5	30	Public dataset	XDR	2010	R9362	South Africa	2.2.1 Beijing
ERR1873559	5	30	Public dataset	MDR	2010	R9437	South Africa	2.2.1 Beijing
ERR1873402	6	27	Public dataset	XDR	2011	M 15 A684 13608 F1 Mycobacterium AGCAGGA L003	South Africa	2.2.2 Beijing Ancestral 1
ERR1873407	6	27	Public dataset	XDR	2011	M 15 A692 15574 F1 Mycobacterium CAGCGTT L003	South Africa	2.2.2 Beijing Ancestral 1
ERR1873408	6	27	Public dataset	XDR	2011	M 15 A695 F2 R18607 CCGAAGT L004	South Africa	2.2.2 Beijing Ancestral 1
ERR1873434	6	27	Public dataset	XDR	2010	M 15 A729 F2 R9964 AAGGACA L004	South Africa	2.2.2 Beijing Ancestral 1
ERR1873439	6	27	Public dataset	XDR	2011	M tuberculosis R13673 LFO46Pool105 3311 L5 GCTCGGTA L005	South Africa	2.2.2 Beijing Ancestral 1
ERR1873440	6	27	Public dataset	XDR	2011	M tuberculosis R13723 LFO46Pool105 3311 L5 CCAGTTCA L005	South Africa	2.2.2 Beijing Ancestral 1
ERR1873452	6	27	Public dataset	XDR	2010	R10867 pool 283 L3 GTGTTCTA L003	South Africa	2.2.2 Beijing Ancestral 1
ERR1873453	6	27	Public dataset	XDR	2010	R10873 LFO46Pool105 3311 L5 GCGAGTAA L005	South Africa	2.2.2 Beijing Ancestral 1
ERR1873454	6	27	Public dataset	XDR	2010	R10921 LFO46Pool105 3311 L5 GCTAACGA L005	South Africa	2.2.2 Beijing Ancestral 1
ERR1873456	6	27	Public dataset	XDR	2010	R11044	South Africa	2.2.2 Beijing Ancestral 1
ERR1873467	6	27	Public dataset	XDR	2011	R13071	South Africa	2.2.2 Beijing Ancestral 1
ERR1873472	6	27	Public dataset	XDR	2011	R13560	South Africa	2.2.2 Beijing Ancestral 1
ERR1873473	6	27	Public dataset	MDR	2011	R13570 LFO46Pool105 3311 L5 CATACCAA L005	South Africa	2.2.2 Beijing Ancestral 1
ERR1873489	6	27	Public dataset	XDR	2011	R17181 pool 282 L2 CAACCACA L002	South Africa	2.2.2 Beijing Ancestral 1

ERR1873493	6	27	Public dataset	MDR	2012	R17661 pool 283 L3 ACAGATTC L003	South Africa	2.2.2 Beijing Ancestral 1
ERR1873500	6	27	Public dataset	MDR	2012	R18243 pool 283 L3 ATAGCGAC L003	South Africa	2.2.2 Beijing Ancestral 1
ERR1873502	6	27	Public dataset	MDR	2012	R18455 pool 283 L3 ATCATTCC L003	South Africa	2.2.2 Beijing Ancestral 1
ERR1873515	6	27	Public dataset	XDR	2012	R19290 pool 282 L2 CTGGCATA L002	South Africa	2.2.2 Beijing Ancestral 1
ERR1873516	6	27	Public dataset	XDR	2012	R19351	South Africa	2.2.2 Beijing Ancestral 1
ERR1873519	6	27	Public dataset	XDR	2012	R19631 pool 282 L2 GAGCTGAA L002	South Africa	2.2.2 Beijing Ancestral 1
ERR1873524	6	27	Public dataset	XDR	2008	R4488	South Africa	2.2.2 Beijing Ancestral 1
ERR1873526	6	27	Public dataset	XDR	2008	R4577	South Africa	2.2.2 Beijing Ancestral 1
ERR1873528	6	27	Public dataset	XDR	2008	R4731 LFO46Pool105 3311 L5 AGTGGTCA L005	South Africa	2.2.2 Beijing Ancestral 1
ERR1873530	6	27	Public dataset	XDR	2009	R4817	South Africa	2.2.2 Beijing Ancestral 1
ERR1873539	6	27	Public dataset	XDR	2009	R5490	South Africa	2.2.2 Beijing Ancestral 1
ERR1873548	6	27	Public dataset	MDR	2009	R6881 LFO46Pool105 3311 L5 AACGCTTA L005	South Africa	2.2.2 Beijing Ancestral 1
ERR1873551	6	27	Public dataset	XDR	2009	R8081	South Africa	2.2.2 Beijing Ancestral 1
ERR1227527	7	26	Public dataset	MDR	NA	mtb16		4.3.4.2 LAM
ERR1227529	7	26	Public dataset	MDR	NA	mtb18		4.3.4.2 LAM
ERR1227530	7	26	Public dataset	MDR	NA	mtb19		4.3.4.2 LAM
ERR1664619	7	26	Public dataset	XDR	NA	SRMTB1		4.3.4.2 LAM
ERR1664623	7	26	Public dataset	MDR	NA	SRMTB5		4.3.4.2 LAM
ERR1664624	7	26	Public dataset	MDR	NA	SRMTB6		4.3.4.2 LAM
ERR1664625	7	26	Public dataset	MDR	NA	SRMTB7		4.3.4.2 LAM
ERR1664626	7	26	Public dataset	MDR	NA	SRMTB8		4.3.4.2 LAM
ERR1664627	7	26	Public dataset	MDR	NA	SRMTB9		4.3.4.2 LAM
ERR1664628	7	26	Public dataset	MDR	NA	SRMTB10		4.3.4.2 LAM
ERR1664629	7	26	Public dataset	XDR	NA	SRMTB11		4.3.4.2 LAM
ERR1664630	7	26	Public dataset	XDR	NA	SRMTB12		4.3.4.2 LAM
ERR1664631	7	26	Public dataset	XDR	NA	SRMTB13		4.3.4.2 LAM
ERR1664632	7	26	Public dataset	XDR	NA	SRMTB14		4.3.4.2 LAM
ERR1664633	7	26	Public dataset	XDR	NA	SRMTB15		4.3.4.2 LAM
ERR1664634	7	26	Public dataset	XDR	NA	SRMTB16		4.3.4.2 LAM
ERR1664641	7	26	Public dataset	XDR	NA	SRMTB23		4.3.4.2 LAM
ERR1664644	7	26	Public dataset	XDR	NA	SRMTB26		4.3.4.2 LAM
ERR1664648	7	26	Public dataset	MDR	NA	SRMTB30		4.3.4.2 LAM
ERR1664649	7	26	Public dataset	XDR	NA	SRMTB31		4.3.4.2 LAM
ERR1664651	7	26	Public dataset	XDR	NA	SRMTB33		4.3.4.2 LAM
ERR1664654	7	26	Public dataset	XDR	NA	SRMTB36		4.3.4.2 LAM
ERR1664658	7	26	Public dataset	MDR	NA	SRMTB40		4.3.4.2 LAM
ERR1664659	7	26	Public dataset	XDR	NA	SRMTB41		4.3.4.2 LAM
ERR1664661	7	26	Public dataset	XDR	NA	SRMTB43		4.3.4.2 LAM
ERR1815554	7	26	Public dataset	XDR	NA	MTB_PT4		4.3.4.2 LAM
SRR3544718	8	23	Public dataset	XDR	2015	Mycobacterium tuberculosis G-018C_6	Georgia	2.2.1 Beijing Central Asia
SRR3544732	8	23	Public dataset	XDR	2014	Mycobacterium tuberculosis G-014S_1_1	Georgia	2.2.1 Beijing Central Asia
SRR3544739	8	23	Public dataset	XDR	2015	Mycobacterium tuberculosis G-018H_3	Georgia	2.2.1 Beijing Central Asia
SRR3544741	8	23	Public dataset	XDR	2015	Mycobacterium tuberculosis G-018I_2	Georgia	2.2.1 Beijing Central Asia
SRR3743203	8	23	Public dataset	XDR	2015	Mycobacterium tuberculosis G-018N_5	Georgia	2.2.1 Beijing Central Asia
SRR5152897	8	23	Public dataset	MDR	2015	Mycobacterium tuberculosis 15-008957	Georgia	2.2.1 Beijing Central Asia
SRR5152905	8	23	Public dataset	MDR	2015	Mycobacterium tuberculosis 15-003446	Georgia	2.2.1 Beijing Central Asia
SRR5152914	8	23	Public dataset	MDR	2015	Mycobacterium tuberculosis 15-006407	Georgia	2.2.1 Beijing Central Asia
SRR5152956	8	23	Public dataset	XDR	2015	Mycobacterium tuberculosis 15-004340	Georgia	2.2.1 Beijing Central Asia
SRR5152959	8	23	Public dataset	XDR	2015	Mycobacterium tuberculosis 15-005819	Georgia	2.2.1 Beijing Central Asia
SRR5153087	8	23	Public dataset	XDR	2015	Mycobacterium tuberculosis 15-012966	Azerbaijan	2.2.1 Beijing Central Asia
SRR5153092	8	23	Public dataset	XDR	2015	Mycobacterium tuberculosis 15-008981	Georgia	2.2.1 Beijing Central Asia
SRR5153225	8	23	Public dataset	MDR	2014	Mycobacterium tuberculosis G-107	Georgia	2.2.1 Beijing Central Asia
SRR5153227	8	23	Public dataset	MDR	2014	Mycobacterium tuberculosis G-102	Georgia	2.2.1 Beijing Central Asia
SRR5153245	8	23	Public dataset	MDR	2015	Mycobacterium tuberculosis G-053	Georgia	2.2.1 Beijing Central Asia
SRR5153268	8	23	Public dataset	MDR	2014	Mycobacterium tuberculosis G-061	Georgia	2.2.1 Beijing Central Asia
SRR5153275	8	23	Public dataset	XDR	2014	Mycobacterium tuberculosis G-056	Georgia	2.2.1 Beijing Central Asia
SRR5153330	8	23	Public dataset	MDR	2014	Mycobacterium tuberculosis G-127	Georgia	2.2.1 Beijing Central Asia
SRR5153331	8	23	Public dataset	MDR	2014	Mycobacterium tuberculosis G-124	Georgia	2.2.1 Beijing Central Asia
SRR5153509	8	23	Public dataset	MDR	2014	Mycobacterium tuberculosis G-140	Georgia	2.2.1 Beijing Central Asia

SRR5153601	8	23	Public dataset	XDR	2015	Mycobacterium tuberculosis G-039-C	Georgia	2.2.1 Beijing Central Asia
SRR5153614	8	23	Public dataset	XDR	2015	Mycobacterium tuberculosis G-039-I	Georgia	2.2.1 Beijing Central Asia
SRR5153825	8	23	Public dataset	XDR	2015	Mycobacterium tuberculosis complex 15-013728	Azerbaijan	2.2.1 Beijing Central Asia
11883-13	9	18	German dataset	MDR	2013		Germany	2.2.1 Beijing Europe/Russian W148 Outbreak
13344-13	9	18	German dataset	MDR	2013		Germany	2.2.1 Beijing Europe/Russian W148 Outbreak
SRR3544730	9	18	Public dataset	MDR	2014	Mycobacterium tuberculosis G-029_1	Georgia	2.2.1 Beijing Europe/Russian W148 Outbreak
SRR3544737	9	18	Public dataset	MDR	2014	Mycobacterium tuberculosis G-035_1	Georgia	2.2.1 Beijing Europe/Russian W148 Outbreak
SRR5152918	9	18	Public dataset	MDR	2015	Mycobacterium tuberculosis 14-013568	Georgia	2.2.1 Beijing Europe/Russian W148 Outbreak
SRR5152922	9	18	Public dataset	MDR	2015	Mycobacterium tuberculosis 15-008396	Georgia	2.2.1 Beijing Europe/Russian W148 Outbreak
SRR5153079	9	18	Public dataset	MDR	2015	Mycobacterium tuberculosis 15-000459	Georgia	2.2.1 Beijing Europe/Russian W148 Outbreak
SRR5153081	9	18	Public dataset	MDR	2015	Mycobacterium tuberculosis 15-004678	Georgia	2.2.1 Beijing Europe/Russian W148 Outbreak
SRR5153206	9	18	Public dataset	MDR	2015	Mycobacterium tuberculosis 15-008792	Georgia	2.2.1 Beijing Europe/Russian W148 Outbreak
SRR5153221	9	18	Public dataset	MDR	2015	Mycobacterium tuberculosis G-69	Georgia	2.2.1 Beijing Europe/Russian W148 Outbreak
SRR5153224	9	18	Public dataset	MDR	2015	Mycobacterium tuberculosis G-112	Georgia	2.2.1 Beijing Europe/Russian W148 Outbreak
SRR5153232	9	18	Public dataset	MDR	2014	Mycobacterium tuberculosis G-104	Georgia	2.2.1 Beijing Europe/Russian W148 Outbreak
SRR5153237	9	18	Public dataset	MDR	2014	Mycobacterium tuberculosis G-118	Georgia	2.2.1 Beijing Europe/Russian W148 Outbreak
SRR5153256	9	18	Public dataset	MDR	2014	Mycobacterium tuberculosis G-068	Georgia	2.2.1 Beijing Europe/Russian W148 Outbreak
SRR5153259	9	18	Public dataset	MDR	2014	Mycobacterium tuberculosis G-054	Georgia	2.2.1 Beijing Europe/Russian W148 Outbreak
SRR5153263	9	18	Public dataset	MDR	2014	Mycobacterium tuberculosis G-052	Georgia	2.2.1 Beijing Europe/Russian W148 Outbreak
SRR5153265	9	18	Public dataset	MDR	2015	Mycobacterium tuberculosis G-069	Georgia	2.2.1 Beijing Europe/Russian W148 Outbreak
SRR5153424	9	18	Public dataset	MDR	2014	Mycobacterium tuberculosis G-126	Georgia	2.2.1 Beijing Europe/Russian W148 Outbreak
SRR5535892	10	16	Public dataset	MDR	NA	G07483	Tanzania	3.1.1 Delhi-CAS
SRR5535907	10	16	Public dataset	MDR	NA	G08375	Tanzania	3.1.1 Delhi-CAS
SRR5535908	10	16	Public dataset	MDR	NA	G08387	Tanzania	3.1.1 Delhi-CAS
SRR5535909	10	16	Public dataset	MDR	NA	G07484	Tanzania	3.1.1 Delhi-CAS
SRR5535910	10	16	Public dataset	MDR	NA	G08368	Tanzania	3.1.1 Delhi-CAS
SRR5535911	10	16	Public dataset	MDR	NA	G07485	Tanzania	3.1.1 Delhi-CAS
SRR5535912	10	16	Public dataset	MDR	NA	G08384	Tanzania	3.1.1 Delhi-CAS
SRR5535913	10	16	Public dataset	MDR	NA	G08385	Tanzania	3.1.1 Delhi-CAS
SRR5535914	10	16	Public dataset	MDR	NA	G08377	Tanzania	3.1.1 Delhi-CAS
SRR5535915	10	16	Public dataset	MDR	NA	G08381	Tanzania	3.1.1 Delhi-CAS
SRR5535916	10	16	Public dataset	MDR	NA	G08382	Botswana	3.1.1 Delhi-CAS
SRR5535917	10	16	Public dataset	MDR	NA	G08379	Tanzania	3.1.1 Delhi-CAS
SRR5535918	10	16	Public dataset	MDR	NA	G08380	Tanzania	3.1.1 Delhi-CAS
SRR5535920	10	16	Public dataset	MDR	NA	G08369	Tanzania	3.1.1 Delhi-CAS
SRR5535922	10	16	Public dataset	MDR	NA	G08376	Tanzania	3.1.1 Delhi-CAS
SRR5535931	10	16	Public dataset	MDR	NA	G08374	Tanzania	3.1.1 Delhi-CAS
ERR1227528	11	15	Public dataset	XDR	NA	mtb17		4.3.4.2 LAM
ERR1664621	11	15	Public dataset	XDR	NA	SRMTB3		4.3.4.2 LAM
ERR1664635	11	15	Public dataset	XDR	NA	SRMTB17		4.3.4.2 LAM
ERR1664636	11	15	Public dataset	XDR	NA	SRMTB18		4.3.4.2 LAM
ERR1664637	11	15	Public dataset	XDR	NA	SRMTB19		4.3.4.2 LAM
ERR1664638	11	15	Public dataset	XDR	NA	SRMTB20		4.3.4.2 LAM
ERR1664639	11	15	Public dataset	XDR	NA	SRMTB21		4.3.4.2 LAM
ERR1664640	11	15	Public dataset	XDR	NA	SRMTB22		4.3.4.2 LAM
ERR1664642	11	15	Public dataset	XDR	NA	SRMTB24		4.3.4.2 LAM
ERR1664646	11	15	Public dataset	XDR	NA	SRMTB28		4.3.4.2 LAM
ERR1664650	11	15	Public dataset	XDR	NA	SRMTB32		4.3.4.2 LAM
ERR1664652	11	15	Public dataset	XDR	NA	SRMTB34		4.3.4.2 LAM
ERR1664660	11	15	Public dataset	XDR	NA	SRMTB42		4.3.4.2 LAM
ERR1664662	11	15	Public dataset	XDR	NA	SRMTB44		4.3.4.2 LAM
ERR1815555	11	15	Public dataset	XDR	NA	MTB_PT5		4.3.4.2 LAM
1244-13	12	10	German dataset	MDR	2013	NA	Germany	2.2.1 Beijing Europe/Russian W148 Outbreak
12466-13	12	10	German dataset	MDR	2013	NA	Germany	2.2.1 Beijing Europe/Russian W148 Outbreak
12487-13	12	10	German dataset	MDR	2013	NA	Germany	2.2.1 Beijing Europe/Russian W148 Outbreak

3007-13	12	10	German dataset	XDR	2013	NA	Germany	2.2.1 Beijing Europe/Russian W148 Outbreak
4245-13	12	10	German dataset	XDR	2013	NA	Germany	2.2.1 Beijing Europe/Russian W148 Outbreak
5887-13	12	10	German dataset	MDR	2013	NA	Germany	2.2.1 Beijing Europe/Russian W148 Outbreak
6764-13	12	10	German dataset	MDR	2013	NA	Germany	2.2.1 Beijing Europe/Russian W148 Outbreak
ERR1161622	12	10	Public dataset	MDR	2009	EEA200903055		2.2.1 Beijing Europe/Russian W148 Outbreak
ERR1161623	12	10	Public dataset	MDR	2009	EEA200905189		2.2.1 Beijing Europe/Russian W148 Outbreak
ERR1161624	12	10	Public dataset	MDR	2009	EEA200905581		2.2.1 Beijing Europe/Russian W148 Outbreak
7604-12	13	10	German dataset	MDR	2012	NA	Germany	2.2.1 Beijing Central Asia outbreak
ERR1544431	13	10	Public dataset	XDR	2014	Mtb187_2_30_16	Kazakhstan	2.2.1 Beijing Central Asia outbreak
ERR1544432	13	10	Public dataset	XDR	2014	Mtb187_2_45_30	Kazakhstan	2.2.1 Beijing Central Asia outbreak
ERR1544435	13	10	Public dataset	XDR	2014	Mtb187_3_60_51	Kazakhstan	2.2.1 Beijing Central Asia outbreak
ERR1544436	13	10	Public dataset	XDR	2014	Mtb187_4_30_75	Kazakhstan	2.2.1 Beijing Central Asia outbreak
ERR1544437	13	10	Public dataset	XDR	2014	Mtb187_4_60_76	Kazakhstan	2.2.1 Beijing Central Asia outbreak
ERR1544438	13	10	Public dataset	XDR	2014	Mtb187_4_60_77	Kazakhstan	2.2.1 Beijing Central Asia outbreak
ERR1544439	13	10	Public dataset	XDR	2014	Mtb187_5_30_104	Kazakhstan	2.2.1 Beijing Central Asia outbreak
ERR1544440	13	10	Public dataset	XDR	2014	Mtb187_5_60_81.1	Kazakhstan	2.2.1 Beijing Central Asia outbreak
ERR1544441	13	10	Public dataset	XDR	2014	Mtb187_5_60_81.2	Kazakhstan	2.2.1 Beijing Central Asia outbreak
SRR4034747	14	9	Public dataset	MDR	2012	P9-Eta	South Africa	4.4.1.1 S-type
SRR4034749	14	9	Public dataset	MDR	2012	P9-Liver	South Africa	4.4.1.1 S-type
SRR4034750	14	9	Public dataset	MDR	2012	P9-Lung1	South Africa	4.4.1.1 S-type
SRR4034751	14	9	Public dataset	MDR	2012	P9-Lung2	South Africa	4.4.1.1 S-type
SRR4034752	14	9	Public dataset	MDR	2012	P9-Lung3	South Africa	4.4.1.1 S-type
SRR4034753	14	9	Public dataset	MDR	2012	P9-Lung4	South Africa	4.4.1.1 S-type
SRR4034754	14	9	Public dataset	MDR	2012	P9-Lung5	South Africa	4.4.1.1 S-type
SRR4034755	14	9	Public dataset	MDR	2012	P9-Lung6	South Africa	4.4.1.1 S-type
SRR4034756	14	9	Public dataset	MDR	2012	P9-Lymph	South Africa	4.4.1.1 S-type
SRR3205958	15	8	Public dataset	XDR	2007	DS16220 (WBB259)	Thailand	2.1 East-Asian non-Beijing
SRR3205959	15	8	Public dataset	XDR	2008	DS16780 (WBB260)	Thailand	2.1 East-Asian non-Beijing
SRR5114017	15	8	Public dataset	XDR	2012	DS 30971	Thailand	2.1 East-Asian non-Beijing
SRR5114018	15	8	Public dataset	XDR	2011	DS 29366	Thailand	2.1 East-Asian non-Beijing
SRR5114019	15	8	Public dataset	XDR	2007	DS 16220	Thailand	2.1 East-Asian non-Beijing
SRR5114020	15	8	Public dataset	XDR	2008	DS 19109	Thailand	2.1 East-Asian non-Beijing
SRR5114021	15	8	Public dataset	XDR	2012	DS 32449	Thailand	2.1 East-Asian non-Beijing
SRR5114022	15	8	Public dataset	XDR	2008	DS 17841	Thailand	2.1 East-Asian non-Beijing
SRR4033237	16	8	Public dataset	MDR	2012	P12-Eta	South Africa	2.2.1 Beijing
SRR4033248	16	8	Public dataset	MDR	2012	P12-Lung1	South Africa	2.2.1 Beijing
SRR4033260	16	8	Public dataset	MDR	2012	P12-Lung2	South Africa	2.2.1 Beijing
SRR4033271	16	8	Public dataset	MDR	2012	P12-Lung3	South Africa	2.2.1 Beijing
SRR4033283	16	8	Public dataset	MDR	2012	P12-Lung4	South Africa	2.2.1 Beijing
SRR4033294	16	8	Public dataset	MDR	2012	P12-Lung5	South Africa	2.2.1 Beijing
SRR4033305	16	8	Public dataset	MDR	2012	P12-SerousFluid	South Africa	2.2.1 Beijing
SRR4033318	16	8	Public dataset	MDR	2012	P12-Spleen	South Africa	2.2.1 Beijing
SRR3743404	17	7	Public dataset	XDR	2015	Mycobacterium tuberculosis Moldova_9505	Moldova	2.2.1 Beijing Central Asia
SRR3743488	17	7	Public dataset	MDR	2015	Mycobacterium tuberculosis Moldova_21796	Moldova	2.2.1 Beijing Central Asia
SRR5153834	17	7	Public dataset	MDR	2016	Mycobacterium tuberculosis complex 10260	Moldova	2.2.1 Beijing Central Asia
SRR5153841	17	7	Public dataset	MDR	2015	Mycobacterium tuberculosis complex 13165	Moldova	2.2.1 Beijing Central Asia
SRR5153853	17	7	Public dataset	MDR	2015	Mycobacterium tuberculosis complex 21389	Moldova	2.2.1 Beijing Central Asia
SRR5153923	17	7	Public dataset	MDR	2015	Mycobacterium tuberculosis complex 20197	Moldova	2.2.1 Beijing Central Asia
SRR5153926	17	7	Public dataset	MDR	2015	Mycobacterium tuberculosis complex 15312	Moldova	2.2.1 Beijing Central Asia
SRR4037641	18	6	Public dataset	MDR	2013	P28-Eta-A5	South Africa	4.4.1.1 S-type
SRR4037661	18	6	Public dataset	MDR	2013	P28-Liver-B4	South Africa	4.4.1.1 S-type
SRR4037665	18	6	Public dataset	MDR	2013	P28-Liver-C2	South Africa	4.4.1.1 S-type
SRR4037667	18	6	Public dataset	MDR	2013	P28-Liver-C4	South Africa	4.4.1.1 S-type
SRR4037729	18	6	Public dataset	MDR	2013	P28-Lung5-B5	South Africa	4.4.1.1 S-type

SRR4037757	18	6	Public dataset	MDR	2013	P28-Spleen-B3	South Africa	4.4.1.1 S-type
ERR1367667	19	6	Public dataset	MDR	NA	G02421	Switzerland	2.2.1 Beijing
SRR5184975	19	6	Public dataset	MDR	2007	DS 16221	Thailand	2.2.1 Beijing
SRR5184981	19	6	Public dataset	MDR	2005	DS 9862	Thailand	2.2.1 Beijing
SRR5184982	19	6	Public dataset	MDR	2005	DS 9429	Thailand	2.2.1 Beijing
SRR5184983	19	6	Public dataset	MDR	2005	DS 9291	Thailand	2.2.1 Beijing
SRR5184988	19	6	Public dataset	MDR	2003	DS 6156	Thailand	2.2.1 Beijing
SRR2993037	20	6	Public dataset	MDR	NA	4542	Russia	4.3.3 LAM
SRR2993039	20	6	Public dataset	MDR	NA	8279	Russia	4.3.3 LAM
SRR5152943	20	6	Public dataset	MDR	2016	Mycobacterium tuberculosis 15-018511	Azerbaijan	4.3.3 LAM
SRR5153134	20	6	Public dataset	MDR	2015	Mycobacterium tuberculosis 15-013411	Georgia	4.3.3 LAM
SRR5153707	20	6	Public dataset	XDR	2015	Mycobacterium tuberculosis complex 15-013317	Azerbaijan	4.3.3 LAM
SRR5153708	20	6	Public dataset	XDR	2015	Mycobacterium tuberculosis complex 15-016117	Azerbaijan	4.3.3 LAM
12103-13	21	6	German dataset	MDR	2013	NA	Germany	2.2.1 Beijing Central Asia
SRR3544716	21	6	Public dataset	MDR	2014	Mycobacterium tuberculosis G-021S_1_1	Georgia	2.2.1 Beijing Central Asia
SRR3544747	21	6	Public dataset	MDR	2014	Mycobacterium tuberculosis G-021_2	Georgia	2.2.1 Beijing Central Asia
SRR5152941	21	6	Public dataset	MDR	2015	Mycobacterium tuberculosis 15-005770	Georgia	2.2.1 Beijing Central Asia
SRR5153261	21	6	Public dataset	MDR	2014	Mycobacterium tuberculosis G-063	Georgia	2.2.1 Beijing Central Asia
SRR5153308	21	6	Public dataset	MDR	2015	Mycobacterium tuberculosis G-076	Georgia	2.2.1 Beijing Central Asia
SRR3724662	22	6	Public dataset	MDR	NA	G04169	Ivory Cost	4.1 Euro-American
SRR3724803	22	6	Public dataset	MDR	NA	G04037	Ivory Cost	4.1 Euro-American
SRR3724956	22	6	Public dataset	MDR	NA	G04035	Ivory Cost	4.1 Euro-American
SRR3732646	22	6	Public dataset	MDR	NA	G05162	Ivory Cost	4.1 Euro-American
SRR3732647	22	6	Public dataset	MDR	NA	G05166	Ivory Cost	4.1 Euro-American
SRR5535687	22	6	Public dataset	MDR	NA	G04044	Ivory Cost	4.1 Euro-American
SRR5152898	23	6	Public dataset	MDR	2016	Mycobacterium tuberculosis 16-000792	Georgia	2.2.1 Beijing Europe/Russian W148 Outbreak
SRR5152923	23	6	Public dataset	MDR	2015	Mycobacterium tuberculosis 15-005552	Georgia	2.2.1 Beijing Europe/Russian W148 Outbreak
SRR5152937	23	6	Public dataset	MDR	2015	Mycobacterium tuberculosis 15-009019	Georgia	2.2.1 Beijing Europe/Russian W148 Outbreak
SRR5153226	23	6	Public dataset	MDR	2014	Mycobacterium tuberculosis G-117	Georgia	2.2.1 Beijing Europe/Russian W148 Outbreak
SRR5153253	23	6	Public dataset	MDR	2014	Mycobacterium tuberculosis G-055	Georgia	2.2.1 Beijing Europe/Russian W148 Outbreak
SRR5153254	23	6	Public dataset	MDR	2014	Mycobacterium tuberculosis G-058	Georgia	2.2.1 Beijing Europe/Russian W148 Outbreak
4153-13	24	5	German dataset	MDR	2013	NA	Germany	2.2.1 Beijing Central Asia
8017-13	24	5	German dataset	MDR	2013	NA	Germany	2.2.1 Beijing Central Asia
SRR3544728	24	5	Public dataset	MDR	2013	Mycobacterium tuberculosis G-031S_1_1	Georgia	2.2.1 Beijing Central Asia
SRR5152895	24	5	Public dataset	MDR	2015	Mycobacterium tuberculosis 15-009701	Georgia	2.2.1 Beijing Central Asia
SRR5153312	24	5	Public dataset	MDR	2014	Mycobacterium tuberculosis G-038_S-1	Georgia	2.2.1 Beijing Central Asia
SRR3732582	25	5	Public dataset	MDR	NA	G05118	Peru	4.3.3 LAM
SRR3732594	25	5	Public dataset	MDR	NA	G05137	Peru	4.3.3 LAM
SRR3732641	25	5	Public dataset	MDR	NA	G05156	Peru	4.3.3 LAM
SRR3732650	25	5	Public dataset	MDR	NA	G05143	Peru	4.3.3 LAM
SRR3732651	25	5	Public dataset	MDR	NA	G05126	Peru	4.3.3 LAM
SRR3742658	26	5	Public dataset	XDR	2010	D1	China	2.2.1 Beijing
SRR3742659	26	5	Public dataset	XDR	2011	D2	China	2.2.1 Beijing
SRR3742660	26	5	Public dataset	XDR	2011	D3	China	2.2.1 Beijing
SRR3742661	26	5	Public dataset	XDR	2012	D4	China	2.2.1 Beijing
SRR3742662	26	5	Public dataset	XDR	2012	D5	China	2.2.1 Beijing
SRR3743400	27	5	Public dataset	MDR	2015	Mycobacterium tuberculosis Moldova_10933	Moldova	2.2.1 Beijing Central Asia
SRR5153816	27	5	Public dataset	MDR	2015	Mycobacterium tuberculosis complex 19909	Moldova	2.2.1 Beijing Central Asia
SRR5153852	27	5	Public dataset	MDR	2015	Mycobacterium tuberculosis complex 20204	Moldova	2.2.1 Beijing Central Asia
SRR5153919	27	5	Public dataset	MDR	2015	Mycobacterium tuberculosis complex 17184	Moldova	2.2.1 Beijing Central Asia
SRR5153925	27	5	Public dataset	MDR	2009	Mycobacterium tuberculosis complex 5828	Moldova	2.2.1 Beijing Central Asia
SRR4034386	28	5	Public dataset	MDR	2012	P4-Eta	South Africa	4.1.2.1 Haarlem
SRR4034387	28	5	Public dataset	MDR	2012	P4-Lung1	South Africa	4.1.2.1 Haarlem
SRR4034388	28	5	Public dataset	MDR	2012	P4-Lung2	South Africa	4.1.2.1 Haarlem
SRR4034389	28	5	Public dataset	MDR	2012	P4-Lung3	South Africa	4.1.2.1 Haarlem
SRR4034390	28	5	Public dataset	MDR	2012	P4-Lung4	South Africa	4.1.2.1 Haarlem

SRR5152910	29	5	Public dataset	MDR	2015	Mycobacterium tuberculosis 15-009024	Georgia	2.2.1 Beijing Europe/Russian W148 Outbreak
SRR5152924	29	5	Public dataset	MDR	2014	Mycobacterium tuberculosis 14-011961	Georgia	2.2.1 Beijing Europe/Russian W148 Outbreak
SRR5153231	29	5	Public dataset	MDR	2014	Mycobacterium tuberculosis G-115	Georgia	2.2.1 Beijing Europe/Russian W148 Outbreak
SRR5153334	29	5	Public dataset	MDR	2014	Mycobacterium tuberculosis G-132	Georgia	2.2.1 Beijing Europe/Russian W148 Outbreak
SRR5153423	29	5	Public dataset	MDR	2014	Mycobacterium tuberculosis G-128	Georgia	2.2.1 Beijing Europe/Russian W148 Outbreak
ERR1873410	30	4	Public dataset	XDR	2010	M 15 A697 F2 R12967 CCTCCTG L004	South Africa	2.2.1 Beijing Asian/Africa 2
ERR1873470	30	4	Public dataset	XDR	2011	R13319	South Africa	2.2.1 Beijing Asian/Africa 2
ERR1873474	30	4	Public dataset	XDR	2011	R14326 LFO46Pool106 3312 L6 TGGCTTCA L006	South Africa	2.2.1 Beijing Asian/Africa 2
ERR1873556	30	4	Public dataset	XDR	2010	R9261	South Africa	2.2.1 Beijing Asian/Africa 2
ERR1873432	31	4	Public dataset	XDR	2010	M 15 A727 F2 R16838 AACTCAC L004	South Africa	2.2.1 Beijing Asian/Africa 2
ERR1873447	31	4	Public dataset	XDR	2010	R10447 LFO46Pool105 3311 L5 GATAGACA L005	South Africa	2.2.1 Beijing Asian/Africa 2
ERR1873509	31	4	Public dataset	XDR	2012	R19042	South Africa	2.2.1 Beijing Asian/Africa 2
ERR1873562	31	4	Public dataset	XDR	2010	R9771	South Africa	2.2.1 Beijing Asian/Africa 2
SRR3742653	32	4	Public dataset	MDR	2010	A1	China	2.2.1 Beijing
SRR3742654	32	4	Public dataset	MDR	2010	A2	China	2.2.1 Beijing
SRR3742663	32	4	Public dataset	MDR	2011	A3	China	2.2.1 Beijing
SRR3742664	32	4	Public dataset	XDR	2011	A4	China	2.2.1 Beijing
SRR3742655	33	4	Public dataset	XDR	2009	C3	China	2.2.1 Beijing Asian/Africa 2
SRR3742656	33	4	Public dataset	XDR	2011	C4	China	2.2.1 Beijing Asian/Africa 2
SRR3742669	33	4	Public dataset	XDR	2007	C1	China	2.2.1 Beijing Asian/Africa 2
SRR3742670	33	4	Public dataset	XDR	2008	C2	China	2.2.1 Beijing Asian/Africa 2
SRR3743407	34	4	Public dataset	MDR	2015	Mycobacterium tuberculosis Moldova_17996	Moldova	2.2.1 Beijing Central Asia outbreak
SRR3743481	34	4	Public dataset	MDR	2015	Mycobacterium tuberculosis Moldova_21656	Moldova	2.2.1 Beijing Central Asia outbreak
SRR5153911	34	4	Public dataset	MDR	2015	Mycobacterium tuberculosis complex 13800	Moldova	2.2.1 Beijing Central Asia outbreak
SRR5153924	34	4	Public dataset	MDR	2014	Mycobacterium tuberculosis complex 22538	Moldova	2.2.1 Beijing Central Asia outbreak
5190-13	35	4	German dataset	XDR	2013	NA	Germany	2.2.1 Beijing Europe/Russian W148 Outbreak
SRR5152908	35	4	Public dataset	XDR	2015	Mycobacterium tuberculosis 15-010728	Georgia	2.2.1 Beijing Europe/Russian W148 Outbreak
SRR5152921	35	4	Public dataset	XDR	2015	Mycobacterium tuberculosis 15-005821	Georgia	2.2.1 Beijing Europe/Russian W148 Outbreak
SRR5153262	35	4	Public dataset	XDR	2014	Mycobacterium tuberculosis G-065	Georgia	2.2.1 Beijing Europe/Russian W148 Outbreak
SRR5535919	36	4	Public dataset	MDR	NA	G08366	Botswana	4.4.1.1 S-type
SRR5535921	36	4	Public dataset	MDR	NA	G08367	Botswana	4.4.1.1 S-type
SRR5535923	36	4	Public dataset	MDR	NA	G08365	Botswana	4.4.1.1 S-type
SRR5535925	36	4	Public dataset	MDR	NA	G08357	Botswana	4.4.1.1 S-type
SRR5818575	37	4	Public dataset	MDR	NA	120_2015	Djibouti	1.1.2 EAI
SRR5818576	37	4	Public dataset	MDR	NA	119_2015	Djibouti	1.1.2 EAI
SRR5818695	37	4	Public dataset	MDR	NA	96_2015	Djibouti	1.1.2 EAI
SRR5818697	37	4	Public dataset	MDR	NA	94_2015	Djibouti	1.1.2 EAI
SRR3732678	38	4	Public dataset	MDR	NA	G05155	Peru	4.3.3 LAM
SRR3732721	38	4	Public dataset	MDR	NA	G05123	Peru	4.3.3 LAM
SRR5535703	38	4	Public dataset	MDR	NA	G05121	Peru	4.3.3 LAM
SRR5535706	38	4	Public dataset	MDR	NA	G05023	Peru	4.3.3 LAM
SRR5067440	39	4	Public dataset	MDR	2010	HCMC0597	Vietnam	2.2.1 Beijing Asian/Africa 2
SRR5073533	39	4	Public dataset	MDR	2009	HCMC1259	Vietnam	2.2.1 Beijing Asian/Africa 2
SRR5074074	39	4	Public dataset	MDR	2009	HCMC1596	Vietnam	2.2.1 Beijing Asian/Africa 2
SRR5074146	39	4	Public dataset	MDR	2009	HCMC1571	Vietnam	2.2.1 Beijing Asian/Africa 2
ERR1555043	40	3	Public dataset	MDR	NA	37dbcbe0-b5f3-11e5-aed5-3c4a9275d6c6		4.3.4.2 LAM
ERR1555045	40	3	Public dataset	MDR	NA	37f1ebf0-b5f3-11e5-aed5-3c4a9275d6c6		4.3.4.2 LAM
ERR1555054	40	3	Public dataset	MDR	NA	38475ef0-b5f3-11e5-aed5-3c4a9275d6c6		4.3.4.2 LAM
ERR1555058	41	3	Public dataset	MDR	NA	386fce80-b5f3-11e5-aed5-3c4a9275d6c6		4.8 mainly T
ERR1555059	41	3	Public dataset	MDR	NA	3876d360-b5f3-11e5-aed5-3c4a9275d6c6		4.8 mainly T
ERR1555062	41	3	Public dataset	MDR	NA	388db6c0-b5f3-11e5-aed5-3c4a9275d6c6		4.8 mainly T
ERR1664620	42	3	Public dataset	MDR	NA	SRMTB2		4.1.1.1 X-type
ERR1664645	42	3	Public dataset	MDR	NA	SRMTB27		4.1.1.1 X-type
ERR1664655	42	3	Public dataset	MDR	NA	SRMTB37		4.1.1.1 X-type
ERR1664622	43	3	Public dataset	XDR	NA	SRMTB4		4.3.4.2 LAM

ERR1664647	43	3	Public dataset	XDR	NA	SRMTB29		4.3.4.2 LAM
ERR1664657	43	3	Public dataset	XDR	NA	SRMTB39		4.3.4.2 LAM
ERR1679605	44	3	Public dataset	MDR	2012	NG30	Nigeria	4.1 Euro-American
ERR1679609	44	3	Public dataset	MDR	2012	NG34	Nigeria	4.1 Euro-American
ERR1679610	44	3	Public dataset	MDR	2012	NG35	Nigeria	4.1 Euro-American
SRR3544717	45	3	Public dataset	XDR	2014	Mycobacterium tuberculosis G-033_2	Georgia	4.3.3 LAM
SRR3544740	45	3	Public dataset	XDR	2014	Mycobacterium tuberculosis G-033S_1_1	Georgia	4.3.3 LAM
SRR5153332	45	3	Public dataset	XDR	2014	Mycobacterium tuberculosis G-094	Georgia	4.3.3 LAM
SRR3544729	46	3	Public dataset	XDR	2014	Mycobacterium tuberculosis G-034_1	Georgia	2.2.1 Beijing Central Asia
SRR5152909	46	3	Public dataset	XDR	2015	Mycobacterium tuberculosis 15-011925	Georgia	2.2.1 Beijing Central Asia
SRR5761429	46	3	Public dataset	XDR	2013	12-15893	Georgia	2.2.1 Beijing Central Asia
SRR3675262	47	3	Public dataset	MDR	2015	MMMOSAM: 32b66345-5026- 4014-86c6- d18e2f9bdc36	United Kingdom	2.2.1 Beijing Central Asia
SRR3675289	47	3	Public dataset	MDR	2015	MMMOSAM: 4b191af4-5415- 4987-9465- 9a96abaa6e9b	United Kingdom	2.2.1 Beijing Central Asia
SRR3675523	47	3	Public dataset	MDR	2015	MMMOSAM: b3acf5c6-5597- 4e4f-8334- 8cae4be9b61d	United Kingdom	2.2.1 Beijing Central Asia
SRR4423139	48	3	Public dataset	MDR	2008	ITM-083358	Bangladesh	2.2.2 Beijing Ancestral 1
SRR4423146	48	3	Public dataset	MDR	2011	ITM-111346	Bangladesh	2.2.2 Beijing Ancestral 1
SRR4423153	48	3	Public dataset	MDR	2011	ITM-111485	Bangladesh	2.2.2 Beijing Ancestral 1
SRR4423154	49	3	Public dataset	MDR	2007	ITM-072228	Bangladesh	2.2.1 Beijing
SRR4423162	49	3	Public dataset	MDR	2007	ITM-073332	Bangladesh	2.2.1 Beijing
SRR4423171	49	3	Public dataset	MDR	2012	ITM-120718	Bangladesh	2.2.1 Beijing
SRR5065526	50	3	Public dataset	MDR	2011	HCMC0124	Vietnam	2.2.1 Beijing Asian/Africa 1
SRR5065595	50	3	Public dataset	MDR	2011	HCMC0083	Vietnam	2.2.1 Beijing Asian/Africa 1
SRR5067289	50	3	Public dataset	MDR	2010	HCMC0902	Vietnam	2.2.1 Beijing Asian/Africa 1
SRR5486866	51	3	Public dataset	MDR	2015	Romania_24268A	Romania	4.1.2.1 Haarlem
SRR5486884	51	3	Public dataset	MDR	2015	Romania_16280A	Romania	4.1.2.1 Haarlem
SRR5486895	51	3	Public dataset	MDR	2015	Romania_22072A	Romania	4.1.2.1 Haarlem
SRR5486875	52	3	Public dataset	MDR	2015	Romania_18670A	Romania	4.8 mainly T
SRR5486897	52	3	Public dataset	MDR	2016	Romania_1151	Romania	4.8 mainly T
SRR5486906	52	3	Public dataset	MDR	2015	Romania_13787B	Romania	4.8 mainly T
304-13	53	3	German dataset	MDR	2013	NA	Germany	2.2.1 Beijing Central Asia
SRR5486879	53	3	Public dataset	XDR	2015	Romania_23522	Romania	2.2.1 Beijing Central Asia
SRR5486885	53	3	Public dataset	MDR	2016	Romania_2735A	Romania	2.2.1 Beijing Central Asia
10896-12	54	3	German dataset	MDR	2012	NA	Germany	4.8 mainly T
5871-12	54	3	German dataset	MDR	2012	NA	Germany	4.8 mainly T
6364-12	54	3	German dataset	MDR	2012	NA	Germany	4.8 mainly T
ERR1873441	55	3	Public dataset	MDR	2011	M tuberculosis R16758 LFO46Pool105 3311 L5 GTACGCAA L005	South Africa	4.1.1.3 X-type
ERR1873448	55	3	Public dataset	XDR	2010	R10512 LFO46Pool105 3311 L5 AGAGTCAA L005	South Africa	4.1.1.3 X-type
ERR1873550	55	3	Public dataset	XDR	2009	R7895 LFO46Pool106 3312 L6 CGGATTGC L006	South Africa	4.1.1.3 X-type
SRR5065588	56	3	Public dataset	MDR	2010	HCMC0389	Vietnam	2.2.1 Beijing Asian/Africa 1
SRR5065596	56	3	Public dataset	MDR	2011	HCMC0186	Vietnam	2.2.1 Beijing Asian/Africa 1
SRR5073854	56	3	Public dataset	MDR	2009	HCMC1413	Vietnam	2.2.1 Beijing Asian/Africa 1
SRR5153082	57	3	Public dataset	XDR	2015	Mycobacterium tuberculosis 15-001678	Azerbaijan	2.2.1 Beijing Central Asia outbreak
SRR5153716	57	3	Public dataset	MDR	2015	Mycobacterium tuberculosis complex 15-017437	Azerbaijan	2.2.1 Beijing Central Asia outbreak
SRR5153812	57	3	Public dataset	XDR	2016	Mycobacterium tuberculosis complex 16-004615	Azerbaijan	2.2.1 Beijing Central Asia outbreak
833-12	58	3	German dataset	MDR	2012	NA	Germany	2.2.1 Beijing Asian/Africa 2
SRR5341237	58	3	Public dataset	MDR	2014	OU36-FMRHH12	India	2.2.1 Beijing Asian/Africa 2
SRR5341257	58	3	Public dataset	MDR	2005	OU13-FMR310	India	2.2.1 Beijing Asian/Africa 2
10346-12	59	3	German dataset	MDR	2012	NA	Germany	2.2.1 Beijing Ancestral 2
10428-12	59	3	German dataset	MDR	2012	NA	Germany	2.2.1 Beijing Ancestral 2
SRR5153093	59	3	Public dataset	MDR	2015	Mycobacterium tuberculosis 15-008502	Georgia	2.2.1 Beijing Ancestral 2
SRR5153603	60	3	Public dataset	MDR	2015	Mycobacterium tuberculosis complex 15-017447	Azerbaijan	2.2.1 Beijing Central Asia
SRR5153720	60	3	Public dataset	MDR	2015	Mycobacterium tuberculosis complex 15-013696	Azerbaijan	2.2.1 Beijing Central Asia
SRR5153722	60	3	Public dataset	XDR	2015	Mycobacterium tuberculosis complex 15-015398	Azerbaijan	2.2.1 Beijing Central Asia
ERR1633819	61	2	Public dataset	MDR	2010	KSP1013	South Africa	2.2.1.1 Beijing Pacific RD150
ERR1873522	61	2	Public dataset	MDR	2008	R4330 LFO46Pool106 3312 L6 AGCACCTC L006	South Africa	2.2.1.1 Beijing Pacific RD150
ERR1665402	62	2	Public dataset	MDR	NA	43153	Spain	2.2.1 Beijing Europe/Russian W148 Outbreak
ERR1665403	62	2	Public dataset	MDR	NA	43159	Spain	2.2.1 Beijing Europe/Russian W148 Outbreak

ERR1665404	63	2	Public dataset	MDR	NA	43732	Spain	2.2.1 Beijing Central Asia
ERR1665405	63	2	Public dataset	MDR	NA	43736	Spain	2.2.1 Beijing Central Asia
ERR1679614	64	2	Public dataset	MDR	2012	NG39	Nigeria	4.3.4.2 LAM
ERR1679616	64	2	Public dataset	MDR	2012	NG40	Nigeria	4.3.4.2 LAM
ERR1679615	65	2	Public dataset	MDR	2012	NG4	Nigeria	4.1.2.1 Haarlem
ERR1679618	65	2	Public dataset	MDR	2012	NG43	Nigeria	4.1.2.1 Haarlem
ERR1873389	66	2	Public dataset	MDR	2009	M 15 A662 4838 F1 Mycobacterium ATGCCTA L003	South Africa	2.2.2 Beijing Ancestral 1
ERR1873409	66	2	Public dataset	MDR	2009	M 15 A696 F2 R9248 CCGTGAG L004	South Africa	2.2.2 Beijing Ancestral 1
ERR1873392	67	2	Public dataset	MDR	2009	M 15 A669 7202 F1 Mycobacterium ACAAGCT L003	South Africa	4.1.1.3 X-type
ERR1873531	67	2	Public dataset	MDR	2009	R4819	South Africa	4.1.1.3 X-type
ERR1873433	68	2	Public dataset	MDR	2010	M 15 A728 F2 R12931 AAGAGAT L004	South Africa	4.4.1.1 S-type
ERR1873449	68	2	Public dataset	XDR	2010	R10552 pool 283 L3 GTCGTAGA L003	South Africa	4.4.1.1 S-type
ERR1873444	69	2	Public dataset	XDR	2010	R10319	South Africa	4.3.2.1 LAM
ERR1873512	69	2	Public dataset	XDR	2012	R19234 pool 282 L2 CTCAATGA L002	South Africa	4.3.2.1 LAM
ERR1873458	70	2	Public dataset	XDR	2010	R11139	South Africa	2.2.1 Beijing
ERR1873545	70	2	Public dataset	MDR	2009	R6560	South Africa	2.2.1 Beijing
ERR1873464	71	2	Public dataset	XDR	2011	R12966 pool 283 L3 TGAAGAGA L003	South Africa	2.2.1 Beijing
ERR1873535	71	2	Public dataset	XDR	2009	R5166	South Africa	2.2.1 Beijing
ERR1873477	72	2	Public dataset	XDR	2011	R15141 LFO46Pool106 3312 L6 TTCACGCA L006	South Africa	2.2.1 Beijing
ERR1873532	72	2	Public dataset	XDR	2009	R4825 LFO46Pool105 3311 L5 CTCAATGA L005	South Africa	2.2.1 Beijing
SRR3205960	73	2	Public dataset	MDR	2008	DS19048 (WBB270)	Thailand	2.2.1 Beijing
SRR3205961	73	2	Public dataset	MDR	2009	DS21277 (WBB273)	Thailand	2.2.1 Beijing
SRR3205963	74	2	Public dataset	XDR	2011	DS29147 (WBB280)	Thailand	2.2.1 Beijing Asian/Africa 2
SRR3205964	74	2	Public dataset	XDR	2012	DS31231 (WBB284)	Thailand	2.2.1 Beijing Asian/Africa 2
SRR3544724	75	2	Public dataset	XDR	2015	Mycobacterium tuberculosis G-019_2	Georgia	2.2.1 Beijing Central Asia
SRR5153223	75	2	Public dataset	XDR	2014	Mycobacterium tuberculosis G-106	Georgia	2.2.1 Beijing Central Asia
SRR3544731	76	2	Public dataset	MDR	2014	Mycobacterium tuberculosis G-020_1	Georgia	4.1.2.1 Haarlem
SRR5153240	76	2	Public dataset	MDR	2015	Mycobacterium tuberculosis G-120	Georgia	4.1.2.1 Haarlem
SRR3724928	77	2	Public dataset	MDR	NA	G04042	Ivory Cost	4.1 Euro-American
SRR3725718	77	2	Public dataset	MDR	NA	G05036	Ivory Cost	4.1 Euro-American
SRR3732588	78	2	Public dataset	MDR	NA	G05122	Peru	2.2.1 Beijing Asian/Africa 2
SRR3732589	78	2	Public dataset	MDR	NA	G05136	Peru	2.2.1 Beijing Asian/Africa 2
SRR3743391	79	2	Public dataset	MDR	2015	Mycobacterium tuberculosis Moldova_11673	Moldova	4.2.1 Ural
SRR5153835	79	2	Public dataset	MDR	2012	Mycobacterium tuberculosis complex 16647	Moldova	4.2.1 Ural
SRR3743396	80	2	Public dataset	MDR	2014	Mycobacterium tuberculosis Moldova_19095	Moldova	4.2.1 Ural
SRR5153817	80	2	Public dataset	MDR	2015	Mycobacterium tuberculosis complex 22496	Moldova	4.2.1 Ural
SRR3743415	81	2	Public dataset	XDR	2015	Mycobacterium tuberculosis Moldova_3077	Moldova	2.2.1 Beijing Europe/Russian W148 Outbreak
SRR3743438	81	2	Public dataset	XDR	2015	Mycobacterium tuberculosis Moldova_5908	Moldova	2.2.1 Beijing Europe/Russian W148 Outbreak
SRR3743433	82	2	Public dataset	XDR	2015	Mycobacterium tuberculosis Moldova_17123	Moldova	2.2.1 Beijing Central Asia
SRR3743489	82	2	Public dataset	XDR	2015	Mycobacterium tuberculosis Moldova_13381	Moldova	2.2.1 Beijing Central Asia
SRR3743474	83	2	Public dataset	MDR	2014	Mycobacterium tuberculosis Moldova_23493	Moldova	2.2.1 Beijing Central Asia outbreak
SRR5153883	83	2	Public dataset	MDR	2015	Mycobacterium tuberculosis complex 14970	Moldova	2.2.1 Beijing Central Asia outbreak
SRR3743500	84	2	Public dataset	XDR	2015	Mycobacterium tuberculosis Moldova_3343	Moldova	2.2.1 Beijing Central Asia
SRR5153930	84	2	Public dataset	XDR	2012	Mycobacterium tuberculosis complex 6967	Moldova	2.2.1 Beijing Central Asia
SRR4423137	85	2	Public dataset	MDR	2008	ITM-083114	Bangladesh	-
SRR4423142	85	2	Public dataset	MDR	2008	ITM-084090	Bangladesh	-
SRR4423138	86	2	Public dataset	XDR	2008	ITM-083135	Bangladesh	3 Delhi-CAS
SRR4423145	86	2	Public dataset	XDR	2010	ITM-110401	Bangladesh	3 Delhi-CAS
SRR4423152	87	2	Public dataset	MDR	2011	ITM-111354	Bangladesh	4.3.4.2 LAM
SRR4423168	87	2	Public dataset	MDR	2008	ITM-090523	Bangladesh	4.3.4.2 LAM
SRR5007182	88	2	Public dataset	MDR	2015	MMMOSAM: 1c34c976-8a62- 45d3-bb5e- 71499733578e	United Kingdom: England	3 Delhi-CAS
SRR5007187	88	2	Public dataset	MDR	2015	MMMOSAM: 7dd1169f-5b61- 482e-ae55- 690b3927cac4	United Kingdom: England	3 Delhi-CAS
SRR5065488	89	2	Public dataset	MDR	2010	HCMC0379	Vietnam	2.2.1.1 Beijing Pacific RD150

SRR5065608	89	2	Public dataset	MDR	2010	HCMC0386	Vietnam	2.2.1.1 Beijing Pacific RD150
SRR5067292	90	2	Public dataset	MDR	2010	HCMC0756	Vietnam	2.2.1 Beijing Ancestral 3
SRR5067558	90	2	Public dataset	MDR	2010	HCMC0993	Vietnam	2.2.1 Beijing Ancestral 3
SRR5125074	91	2	Public dataset	MDR	2014	PGI_IOB _EPTB_3	India	3 Delhi-CAS
SRR5125078	91	2	Public dataset	MDR	2014	PGI_IOB _EPTB_2	India	3 Delhi-CAS
SRR5152896	92	2	Public dataset	MDR	2015	Mycobacterium tuberculosis 15-004449	Georgia	2.2.1 Beijing Europe/Russian W148 Outbreak
SRR5153234	92	2	Public dataset	MDR	2015	Mycobacterium tuberculosis G-114	Georgia	2.2.1 Beijing Europe/Russian W148 Outbreak
SRR5152906	93	2	Public dataset	MDR	2015	Mycobacterium tuberculosis 15-001598	Georgia	2.2.1 Beijing Central Asia
SRR5152920	93	2	Public dataset	MDR	2015	Mycobacterium tuberculosis 15-012050	Georgia	2.2.1 Beijing Central Asia
SRR5152942	94	2	Public dataset	MDR	2016	Mycobacterium tuberculosis 15-017326	Azerbaijan	2.2.1 Beijing Central Asia
SRR5152944	94	2	Public dataset	MDR	2016	Mycobacterium tuberculosis 15-018949	Azerbaijan	2.2.1 Beijing Central Asia
SRR5152947	95	2	Public dataset	XDR	2016	Mycobacterium tuberculosis 15-017322	Azerbaijan	2.2.1 Beijing Central Asia
SRR5152952	95	2	Public dataset	XDR	2016	Mycobacterium tuberculosis 16-000965	Azerbaijan	2.2.1 Beijing Central Asia
SRR5152950	96	2	Public dataset	XDR	2016	Mycobacterium tuberculosis 15-017787	Azerbaijan	4.3.4.2 LAM
SRR5152973	96	2	Public dataset	XDR	2015	Mycobacterium tuberculosis 15-015957	Azerbaijan	4.3.4.2 LAM
SRR5152954	97	2	Public dataset	MDR	2015	Mycobacterium tuberculosis 15-009715	Georgia	2.2.1 Beijing Central Asia
SRR5153235	97	2	Public dataset	MDR	2015	Mycobacterium tuberculosis G-113	Georgia	2.2.1 Beijing Central Asia
ERR1555049	98	2	Public dataset	MDR	NA	381ec850-b5f3-11e5-aed5-3c4a9275d6c6		4.3.3 LAM
ERR1555056	98	2	Public dataset	MDR	NA	38591230-b5f3-11e5-aed5-3c4a9275d6c6		4.3.3 LAM
SRR5153255	99	2	Public dataset	MDR	2014	Mycobacterium tuberculosis G-122	Georgia	2.2.1 Beijing Central Asia outbreak
SRR5153291	99	2	Public dataset	MDR	2015	Mycobacterium tuberculosis G-077	Georgia	2.2.1 Beijing Central Asia outbreak
SRR5153267	100	2	Public dataset	MDR	2014	Mycobacterium tuberculosis G-066	Georgia	4.3.3 LAM
SRR5153324	100	2	Public dataset	MDR	2014	Mycobacterium tuberculosis G-091	Georgia	4.3.3 LAM
SRR5153271	101	2	Public dataset	MDR	2014	Mycobacterium tuberculosis G-062	Georgia	2.2.1 Beijing Central Asia
SRR5153609	101	2	Public dataset	MDR	2014	Mycobacterium tuberculosis G-131	Georgia	2.2.1 Beijing Central Asia
SRR5153273	102	2	Public dataset	MDR	2014	Mycobacterium tuberculosis G-036_S_2	Georgia	2.2.1 Beijing Europe/Russian W148 Outbreak
SRR5153616	102	2	Public dataset	MDR	2015	Mycobacterium tuberculosis G-125	Georgia	2.2.1 Beijing Europe/Russian W148 Outbreak
12016-13	103	2	German dataset	XDR	2013	NA	Germany	2.2.1 Beijing Central Asia
SRR5153311	103	2	Public dataset	XDR	2014	Mycobacterium tuberculosis G-079	Georgia	2.2.1 Beijing Central Asia
SRR5153839	104	2	Public dataset	MDR	2009	Mycobacterium tuberculosis complex 54	Moldova	4.3.3 LAM
SRR5153844	104	2	Public dataset	MDR	2015	Mycobacterium tuberculosis complex 12590	Moldova	4.3.3 LAM
SRR5341240	105	2	Public dataset	MDR	2006	OU33-FMR473	India	2.2.1 Beijing Ancestral 3
SRR5341255	105	2	Public dataset	MDR	2005	OU15-FMR338	India	2.2.1 Beijing Ancestral 3
SRR5486869	106	2	Public dataset	MDR	2015	Romania_15061A	Romania	4.1.2.1 Haarlem
SRR5486894	106	2	Public dataset	MDR	2015	Romania_11808A	Romania	4.1.2.1 Haarlem
SRR5486883	107	2	Public dataset	MDR	2016	Romania_6010A	Romania	4.8 mainly T
SRR5486900	107	2	Public dataset	MDR	2016	Romania_1138A	Romania	4.8 mainly T
SRR5535857	108	2	Public dataset	MDR	NA	G07455	Congo	4.6.1.2 Uganda
SRR5535858	108	2	Public dataset	MDR	NA	G07412	Congo	4.6.1.2 Uganda
SRR5535861	109	2	Public dataset	MDR	NA	G08378	Tanzania	2.2.1.1 Beijing Pacific RD150
SRR5535924	109	2	Public dataset	MDR	NA	G08386	Tanzania	2.2.1.1 Beijing Pacific RD150
SRR5818581	110	2	Public dataset	MDR	NA	103_2016	Djibouti	4.2.2 Euro-American
SRR5818638	110	2	Public dataset	MDR	NA	124_2015	Djibouti	4.2.2 Euro-American
SRR5818587	111	2	Public dataset	MDR	NA	72_2016	Djibouti	3.1.1 Delhi-CAS
SRR5818637	111	2	Public dataset	MDR	NA	123_2015	Djibouti	3.1.1 Delhi-CAS
SRR5818592	112	2	Public dataset	MDR	NA	73_2016	Djibouti	4.2.2 Euro-American
SRR5818655	112	2	Public dataset	MDR	NA	233_2015	Djibouti	4.2.2 Euro-American
SRR5818617	113	2	Public dataset	MDR	NA	136_2015	Djibouti	3.1.1 Delhi-CAS
SRR5818636	113	2	Public dataset	MDR	NA	126_2015	Djibouti	3.1.1 Delhi-CAS
ERR1952138	114	2	Public dataset	MDR	2004	IEMDR01	Ireland	4.1 Euro-American
SRR5535708	114	2	Public dataset	MDR	NA	G05030	Ivory Cost	4.1 Euro-American
SRR5065537	115	2	Public dataset	MDR	2011	HCMC0218	Vietnam	2.2.1 Beijing
SRR5074155	115	2	Public dataset	MDR	2009	HCMC1564	Vietnam	2.2.1 Beijing
SRR5067504	116	2	Public dataset	MDR	2010	HCMC0545	Vietnam	2.2.1.1 Beijing Pacific RD150
SRR5067543	116	2	Public dataset	MDR	2010	HCMC0614	Vietnam	2.2.1.1 Beijing Pacific RD150
SRR5486893	117	2	Public dataset	MDR	2016	Romania_942	Romania	-
SRR5486896	117	2	Public dataset	MDR	2015	Romania_19609	Romania	4.8 mainly T
ERR1555047	118	2	Public dataset	MDR	NA	38083310-b5f3-11e5-aed5-3c4a9275d6c6		4.8 mainly T

ERR1555048	118	2	Public dataset	MDR	NA	38132f90-b5f3-11e5-aed5-3c4a9275d6c6		4.8 mainly T
8565-12	119	2	German dataset	XDR	2012	NA	Germany	4.3.3 LAM
ERR1193862	119	2	Public dataset	XDR	NA	SAMEA3671272		4.3.3 LAM
ERR1193881	120	2	Public dataset	XDR	NA	SAMEA3671291		4.3.3 LAM
ERR1873435	120	2	Public dataset	XDR	2010	M tuberculosis R10398 L6 TAGCTT L006	South Africa	4.3.3 LAM
ERR1193900	121	2	Public dataset	MDR	NA	SAMEA3671310		4.3.3 LAM
ERR1193903	121	2	Public dataset	MDR	NA	SAMEA3671313		4.3.3 LAM
ERR1465931	122	2	Public dataset	MDR	NA	SAMEA3715562		4.8 mainly T
SRR5153077	122	2	Public dataset	XDR	2016	Mycobacterium tuberculosis 16-005760	Azerbaijan	4.8 mainly T
ERR1555041	123	2	Public dataset	MDR	NA	37cc3b80-b5f3-11e5-aed5-3c4a9275d6c6		4.3.4.2 LAM
ERR1555053	123	2	Public dataset	MDR	NA	38408120-b5f3-11e5-aed5-3c4a9275d6c6		4.3.4.2 LAM
2955-12	124	2	German dataset	XDR	2012	NA	Germany	2.2.1 Beijing Central Asia outbreak
4305-13	124	2	German dataset	MDR	2013	NA	Germany	2.2.1 Beijing Central Asia outbreak
11250-12	125	2	German dataset	MDR	2012	NA	Germany	4.2.2.1 TUR
11686-13	125	2	German dataset	MDR	2013	NA	Germany	4.2.2.1 TUR
10655-12	126	2	German dataset	XDR	2012	NA	Germany	2.2.1 Beijing Central Asia
134-13	126	2	German dataset	XDR	2013	NA	Germany	2.2.1 Beijing Central Asia
10926-12	127	2	German dataset	MDR	2012	NA	Germany	2.2.1 Beijing Central Asia outbreak
11460-12	127	2	German dataset	MDR	2012	NA	Germany	2.2.1 Beijing Central Asia outbreak
1571-12	128	2	German dataset	MDR	2012	NA	Germany	4.8 mainly T
3617-12	128	2	German dataset	MDR	2012	NA	Germany	4.8 mainly T
253-12	129	2	German dataset	MDR	2012	NA	Germany	2.2.1 Beijing Europe/Russian W148 Outbreak
254-12	129	2	German dataset	MDR	2012	NA	Germany	2.2.1 Beijing Europe/Russian W148 Outbreak
SRR5153600	130	2	Public dataset	XDR	2015	Mycobacterium tuberculosis complex 15-013336	Azerbaijan	2.2.1 Beijing Central Asia
SRR5153712	130	2	Public dataset	XDR	2016	Mycobacterium tuberculosis complex 16-000319	Azerbaijan	2.2.1 Beijing Central Asia
SRR5152945	131	2	Public dataset	MDR	2015	Mycobacterium tuberculosis 15-014647	Azerbaijan	2.2.1 Beijing Europe/Russian W148 Outbreak
SRR5152948	131	2	Public dataset	XDR	2016	Mycobacterium tuberculosis 15-018961	Azerbaijan	2.2.1 Beijing Europe/Russian W148 Outbreak
SRR3544733	132	2	Public dataset	XDR	2014	Mycobacterium tuberculosis G-025_1	Georgia	2.2.1 Beijing Central Asia
SRR5153272	132	2	Public dataset	MDR	2014	Mycobacterium tuberculosis G-25_S_1	Georgia	2.2.1 Beijing Central Asia
SRR3544727	133	2	Public dataset	MDR	2014	Mycobacterium tuberculosis G-032_1	Georgia	2.2.1 Beijing Europe/Russian W148 Outbreak
SRR5153276	133	2	Public dataset	MDR	2014	Mycobacterium tuberculosis G-032_S_1	Georgia	2.2.1 Beijing Europe/Russian W148 Outbreak
10162-13		1	German dataset	MDR	2013		Germany	2.2.1 Beijing Central Asia
10284-13		1	German dataset	MDR	2013		Germany	2.2.1 Beijing Central Asia outbreak
10490-13		1	German dataset	MDR	2013		Germany	2.2.1 Beijing Central Asia outbreak
10505-13		1	German dataset	MDR	2013		Germany	4.2.2 Euro-American
10743-13		1	German dataset	MDR	2013		Germany	4.8 mainly T
10759-13		1	German dataset	MDR	2013		Germany	2.2.1 Beijing Central Asia outbreak
10840-13		1	German dataset	MDR	2013		Germany	3.1.2.1 Delhi-CAS
11132-13		1	German dataset	MDR	2013		Germany	4.3.3 LAM
11355-13		1	German dataset	XDR	2013	NA	Germany	4.3.3 LAM
11960-13		1	German dataset	MDR	2013		Germany	2.2.1 Beijing Central Asia outbreak
11987-13		1	German dataset	MDR	2013		Germany	2.2.1 Beijing Central Asia
12009-13		1	German dataset	MDR	2013		Germany	4.6.2.2 Cameroon
12017-13		1	German dataset	MDR	2013		Germany	4.4.1.1 S-type
12018-13		1	German dataset	XDR	2013	NA	Germany	2.2.1 Beijing Central Asia outbreak
12041-13		1	German dataset	MDR	2013		Germany	4.3.3 LAM
12471-13		1	German dataset	MDR	2013		Germany	2.2.1 Beijing Central Asia outbreak
12510-13		1	German dataset	MDR	2013		Germany	2.2.1 Beijing Central Asia
1298-12		1	German dataset	MDR	2012		Germany	4.2.1 Ural
13432-13		1	German dataset	MDR	2013		Germany	2.2.1 Beijing Europe/Russian W148 Outbreak
13739-13		1	German dataset	XDR	2013	NA	Germany	2.2.1 Beijing Central Asia

13898-13		1	German dataset	MDR	2013		Germany	2.2.1 Beijing Europe/Russian W148 Outbreak
14102-13		1	German dataset	not MDR	2013	NA	Germany	2.2.1 Beijing
14217-13		1	German dataset	MDR	2013		Germany	2.2.1 Beijing Europe/Russian W148 Outbreak
14489-13		1	German dataset	MDR	2013		Germany	4.6 Euro-American
1560-12		1	German dataset	MDR	2012		Germany	4.7 mainly T
1635-12		1	German dataset	MDR	2012		Germany	4.4.1.1 S-type
1725-13		1	German dataset	MDR	2013		Germany	2.2.1 Beijing Central Asia
2065-13		1	German dataset	MDR	2013		Germany	2.2.1 Beijing Central Asia
2135-12		1	German dataset	MDR	2012		Germany	2.2.1 Beijing Central Asia outbreak
2303-12		1	German dataset	MDR	2012		Germany	5 West-Africa 1
2378-13		1	German dataset	MDR	2013		Germany	2.2.1 Beijing Europe/Russian W148 Outbreak
2636-13		1	German dataset	MDR	2013		Germany	2.2.1 Beijing Central Asia outbreak
2709-13		1	German dataset	MDR	2013		Germany	4.8 mainly T
2718-13		1	German dataset	MDR	2013		Germany	3 Delhi-CAS
2823-13		1	German dataset	MDR	2013		Germany	4.1.2.1 Haarlem
3106-13		1	German dataset	MDR	2013		Germany	4.2.1 Ural
3116-13		1	German dataset	MDR	2013		Germany	4.3.3 LAM
3125-13		1	German dataset	MDR	2013		Germany	2.2.1 Beijing Central Asia
3201-12		1	German dataset	MDR	2012		Germany	4.3.4.2 LAM
3290-12		1	German dataset	MDR	2012		Germany	4.8 mainly T
3413-12		1	German dataset	MDR	2012		Germany	2.2.1 Beijing Central Asia outbreak
4038-12		1	German dataset	MDR	2012		Germany	4.2.2 Euro-American
4345-12		1	German dataset	MDR	2012		Germany	2.2.1 Beijing Europe/Russian W148 Outbreak
4517-13		1	German dataset	MDR	2013		Germany	2.2.1 Beijing Central Asia outbreak
4556-12		1	German dataset	not MDR	2012	NA	Germany	2.2.1 Beijing Central Asia outbreak
4563-13		1	German dataset	MDR	2013		Germany	-
4751-13		1	German dataset	MDR	2013		Germany	2.2.1 Beijing Europe/Russian W148 Outbreak
479-12		1	German dataset	MDR	2012		Germany	4.8 mainly T
4798-13		1	German dataset	MDR	2013		Germany	2.2.1 Beijing Europe/Russian W148 Outbreak
4839-12		1	German dataset	MDR	2012		Germany	2.2.1 Beijing Central Asia outbreak
4893-12		1	German dataset	MDR	2012		Germany	4.3.2 LAM
4960-13		1	German dataset	XDR	2013	NA	Germany	2.2.1 Beijing Europe/Russian W148 Outbreak
5033-12		1	German dataset	MDR	2012		Germany	2.2.1 Beijing Europe/Russian W148 Outbreak
5096-13		1	German dataset	XDR	2013	NA	Germany	2.2.1 Beijing Central Asia outbreak
5158-12		1	German dataset	MDR	2012		Germany	2.2.1 Beijing Europe/Russian W148 Outbreak
521-14		1	German dataset	MDR	2014		Germany	2.2.1 Beijing Europe/Russian W148 Outbreak
5271-12		1	German dataset	MDR	2012		Germany	4.2.2 Euro-American
5366-12		1	German dataset	MDR	2012		Germany	2.2.1 Beijing Central Asia
5439-12		1	German dataset	MDR	2012		Germany	4.1.2.1 Haarlem
565-12		1	German dataset	MDR	2012		Germany	4.3.3 LAM
5667-13		1	German dataset	MDR	2013		Germany	2.2.1 Beijing Central Asia
5675-12		1	German dataset	MDR	2012		Germany	4.2.1 Ural
6089-13		1	German dataset	MDR	2013		Germany	2.2.1 Beijing Central Asia
6316-13		1	German dataset	XDR	2013	NA	Germany	2.2.1 Beijing Europe/Russian W148 Outbreak
6360-12		1	German dataset	MDR	2012		Germany	4.3.3 LAM
6760-13		1	German dataset	MDR	2013		Germany	3.1.2.1 Delhi-CAS
6934-12		1	German dataset	MDR	2012		Germany	3 Delhi-CAS
72-13		1	German dataset	not MDR	2013	NA	Germany	2.2.1 Beijing Europe/Russian W148 Outbreak
7712-13		1	German dataset	XDR	2013	NA	Germany	2.2.1 Beijing Central Asia outbreak
7854-13		1	German dataset	MDR	2013		Germany	2.2.1 Beijing Europe/Russian W148 Outbreak
7977-12		1	German dataset	MDR	2012		Germany	2.2.1 Beijing Europe/Russian W148 Outbreak
7984-12		1	German dataset	XDR	2012	NA	Germany	4.3.3 LAM
8291-13		1	German dataset	XDR	2013	NA	Germany	2.2.1 Beijing Europe/Russian W148 Outbreak
8300-13		1	German dataset	MDR	2013		Germany	2.2.1 Beijing Central Asia
8305-13		1	German dataset	MDR	2013		Germany	3 Delhi-CAS

8347-13		1	German dataset	MDR	2013		Germany	2.2.1 Beijing Central Asia outbreak
871-13		1	German dataset	MDR	2013		Germany	4.1.2 Euro-American
8847-13		1	German dataset	MDR	2013		Germany	4.1.2.1 Haarlem
886-12		1	German dataset	MDR	2012		Germany	4.4.1.1 S-type
9082-13		1	German dataset	MDR	2013		Germany	4.4.1.1 S-type
9165-12		1	German dataset	not MDR	2012	NA	Germany	2.2.1 Beijing Ancestral 2
9354-12		1	German dataset	MDR	2012		Germany	2.2.1 Beijing Central Asia outbreak
9468-12		1	German dataset	MDR	2012		Germany	2.2.1 Beijing
9498-12		1	German dataset	MDR	2012		Germany	4.1.2.1 Haarlem
9505-13		1	German dataset	MDR	2013		Germany	2.2.1 Beijing Asian/Africa 2
9508-13		1	German dataset	MDR	2013		Germany	2.2.1 Beijing Central Asia
9653-12		1	German dataset	MDR	2012		Germany	2.2.1 Beijing
9771-13		1	German dataset	MDR	2013		Germany	4.8 mainly T
9776-13		1	German dataset	MDR	2013		Germany	4.1.2.1 Haarlem
9777-13		1	German dataset	MDR	2013		Germany	2.2.1 Beijing Europe/Russian W148 Outbreak
9926-13		1	German dataset	MDR	2013		Germany	2.2.2 Beijing Ancestral 1
9927-13		1	German dataset	MDR	2013		Germany	4.8 mainly T
999-13		1	German dataset	MDR	2013		Germany	2.2.1 Beijing Europe/Russian W148 Outbreak
ERR1161619		1	Public dataset	XDR	2008	EEA200801564		2.2.1 Beijing Europe/Russian W148 Outbreak
ERR1161621		1	Public dataset	XDR	2009	EEA200900968		2.2.1 Beijing Europe/Russian W148 Outbreak
ERR1193661		1	Public dataset	MDR	NA	SAMEA3671317		4.3.2 LAM
ERR1193662		1	Public dataset	MDR	NA	SAMEA3671330		4.3.3 LAM
ERR1193674		1	Public dataset	MDR	NA	SAMEA3671341		4.3.4 LAM
ERR1193683		1	Public dataset	MDR	NA	SAMEA3671350		4.3.3 LAM
ERR1193724		1	Public dataset	MDR	NA	SAMEA3671386		4.3.4.2 LAM
ERR1193763		1	Public dataset	MDR	NA	SAMEA3671410		4.3.4 LAM
ERR1193781		1	Public dataset	MDR	NA	SAMEA3671428		4.3.4.2 LAM
ERR1193782		1	Public dataset	XDR	NA	SAMEA3671429		4.3.3 LAM
ERR1193792		1	Public dataset	MDR	NA	SAMEA3671202		4.3.3 LAM
ERR1193842		1	Public dataset	XDR	NA	SAMEA3671252		4.3.1 LAM
ERR1193843		1	Public dataset	MDR	NA	SAMEA3671253		4.3.1 LAM
ERR1193860		1	Public dataset	MDR	NA	SAMEA3671270		4.3.3 LAM
ERR1193861		1	Public dataset	MDR	NA	SAMEA3671271		4.3.3 LAM
ERR1193863		1	Public dataset	MDR	NA	SAMEA3671273		4.3.3 LAM
ERR1193864		1	Public dataset	MDR	NA	SAMEA3671274		4.3.3 LAM
ERR1193865		1	Public dataset	MDR	NA	SAMEA3671275		4.3.3 LAM
ERR1193866		1	Public dataset	MDR	NA	SAMEA3671276		4.3.3 LAM
ERR1193875		1	Public dataset	MDR	NA	SAMEA3671285		4.3.3 LAM
ERR1193877		1	Public dataset	MDR	NA	SAMEA3671287		4.3.4.2 LAM
ERR1193880		1	Public dataset	MDR	NA	SAMEA3671290		4.3.2.1 LAM
ERR1193882		1	Public dataset	MDR	NA	SAMEA3671292		4.3.3 LAM
ERR1193883		1	Public dataset	MDR	NA	SAMEA3671293		4.3.2.1 LAM
ERR1193888		1	Public dataset	MDR	NA	SAMEA3671298		4.3.2.1 LAM
ERR1193892		1	Public dataset	MDR	NA	SAMEA3671302		4.3.4.2.1 LAM
ERR1193897		1	Public dataset	MDR	NA	SAMEA3671307		4.3.2 LAM
ERR1193898		1	Public dataset	MDR	NA	SAMEA3671308		4.3.2.1 LAM
ERR1193901		1	Public dataset	MDR	NA	SAMEA3671311		4.3.3 LAM
ERR1193902		1	Public dataset	MDR	NA	SAMEA3671312		4.3.3 LAM
ERR1193904		1	Public dataset	MDR	NA	SAMEA3671314		4.3.3 LAM
ERR1199093		1	Public dataset	MDR	NA	SAMEA3671461		4.6.1.2 Uganda
ERR1199096		1	Public dataset	MDR	NA	SAMEA3671516		4.6.1.2 Uganda
ERR1199098		1	Public dataset	MDR	NA	SAMEA3671530		4.6.1.2 Uganda
ERR1199099		1	Public dataset	MDR	NA	SAMEA3671531		4.6.1.2 Uganda
ERR1199101		1	Public dataset	MDR	NA	SAMEA3671534		4.6.1.2 Uganda
ERR1199102		1	Public dataset	MDR	NA	SAMEA3671535		4.6.1.2 Uganda
ERR1199103		1	Public dataset	MDR	NA	SAMEA3671536		4.6.1.2 Uganda
ERR1199104		1	Public dataset	MDR	NA	SAMEA3671538		4.6.1.2 Uganda
ERR1199105		1	Public dataset	MDR	NA	SAMEA3671539		4.6.1.2 Uganda
ERR1199106		1	Public dataset	MDR	NA	SAMEA3671545		4.6.1.2 Uganda
ERR1199108		1	Public dataset	MDR	NA	SAMEA3671557		4.6.1.2 Uganda
ERR1199109		1	Public dataset	MDR	NA	SAMEA3671560		4.6.1.2 Uganda
ERR1199110		1	Public dataset	MDR	NA	SAMEA3671563		4.6.1.2 Uganda
ERR1199111		1	Public dataset	MDR	NA	SAMEA3671571		4.6.1.2 Uganda
ERR1199112		1	Public dataset	MDR	NA	SAMEA3671572		4.6.1.2 Uganda
ERR1199113		1	Public dataset	MDR	NA	SAMEA3671573		4.6.1.2 Uganda
ERR1199115		1	Public dataset	MDR	NA	SAMEA3671582		4.6.1.2 Uganda
ERR1199117		1	Public dataset	MDR	NA	SAMEA3671589		4.6.1.2 Uganda
ERR1199120		1	Public dataset	MDR	NA	SAMEA3671594		4.6.1.2 Uganda
ERR1199122		1	Public dataset	MDR	NA	SAMEA3671599		4.6.1.2 Uganda
ERR1199123		1	Public dataset	MDR	NA	SAMEA3671601		4.6.1.2 Uganda
ERR1199126		1	Public dataset	MDR	NA	SAMEA3671614		4.6.1.2 Uganda
ERR1199127		1	Public dataset	MDR	NA	SAMEA3671615		4.6.1.2 Uganda

ERR1199128		1	Public dataset	MDR	NA	SAMEA3671618		4.6.1.2 Uganda
ERR1199130		1	Public dataset	MDR	NA	SAMEA3671465		4.6.1.2 Uganda
ERR1199131		1	Public dataset	MDR	NA	SAMEA3671512		4.6.1.1 Uganda
ERR1199137		1	Public dataset	MDR	NA	SAMEA3671528		4.6.1.2 Uganda
ERR1199145		1	Public dataset	MDR	NA	SAMEA3671581		4.6.1.1 Uganda
ERR1367615		1	Public dataset	MDR	NA	G02429	Switzerland	2.2.1 Beijing Ancestral 3
ERR1367634		1	Public dataset	MDR	NA	G02135	Switzerland	4.1.2.1 Haarlem
ERR1367652		1	Public dataset	MDR	NA	G02131	Switzerland	4.1.2.1 Haarlem
ERR1367666		1	Public dataset	MDR	NA	G02138	Switzerland	4.6.2.2 Cameroon
ERR1413476		1	Public dataset	MDR	2004	4.018		4.6.2.2 Cameroon
ERR1452609		1	Public dataset	MDR	1996	1945		2.2.2 Beijing Ancestral 1
ERR1465864		1	Public dataset	MDR	NA	SAMEA4041351		4.8 mainly T
ERR1465875		1	Public dataset	MDR	NA	SAMEA4041362		4.8 mainly T
ERR1465929		1	Public dataset	MDR	NA	SAMEA3715556		4.8 mainly T
ERR1465930		1	Public dataset	MDR	NA	SAMEA3715559		4.8 mainly T
ERR1465932		1	Public dataset	MDR	NA	SAMEA3715564		4.7 mainly T
ERR1555040		1	Public dataset	MDR	NA	37c33ad0-b5f3-11e5-aed5-3c4a9275d6c6		4.8 mainly T
ERR1555044		1	Public dataset	MDR	NA	37e98780-b5f3-11e5-aed5-3c4a9275d6c6		4.3.4.2 LAM
ERR1555046		1	Public dataset	MDR	NA	37fce870-b5f3-11e5-aed5-3c4a9275d6c6		4.1.2 Euro-American
ERR1555050		1	Public dataset	MDR	NA	382753d0-b5f3-11e5-aed5-3c4a9275d6c6		4.3.4.1 LAM
ERR1555051		1	Public dataset	MDR	NA	38302d70-b5f3-11e5-aed5-3c4a9275d6c6		4.1.1.1 X-type
ERR1555052		1	Public dataset	MDR	NA	383891e0-b5f3-11e5-aed5-3c4a9275d6c6		4.3.4.1 LAM
ERR1555055		1	Public dataset	MDR	NA	384f4e30-b5f3-11e5-aed5-3c4a9275d6c6		4.3.4.1 LAM
ERR1555057		1	Public dataset	MDR	NA	38654730-b5f3-11e5-aed5-3c4a9275d6c6		4.3.3 LAM
ERR1555060		1	Public dataset	MDR	NA	387e4d70-b5f3-11e5-aed5-3c4a9275d6c6		4.3.4.2 LAM
ERR1555061		1	Public dataset	MDR	NA	388615a0-b5f3-11e5-aed5-3c4a9275d6c6		4.3.4.2 LAM
ERR1633780		1	Public dataset	MDR	2010	KSP974	South Africa	4.3.3 LAM
ERR1633839		1	Public dataset	MDR	2010	KSP1033	South Africa	4.3.2.1 LAM
ERR1633881		1	Public dataset	MDR	2010	KSP1075	South Africa	2.2.1 Beijing
ERR1633939		1	Public dataset	MDR	2010	KSP1133	South Africa	2.2.1.1 Beijing Pacific RD150
ERR1679585		1	Public dataset	MDR	2012	NG1	Nigeria	4.6.2.2 Cameroon
ERR1679586		1	Public dataset	MDR	2012	NG10	Nigeria	4.1 Euro-American
ERR1679587		1	Public dataset	MDR	2012	NG12	Nigeria	4.6.2.2 Cameroon
ERR1679600		1	Public dataset	MDR	2012	NG25	Nigeria	4.6.2.2 Cameroon
ERR1679607		1	Public dataset	MDR	2012	NG32	Nigeria	4.6.2.2 Cameroon
ERR1679611		1	Public dataset	MDR	2012	NG36	Nigeria	4.6.2.2 Cameroon
ERR1679617		1	Public dataset	MDR	2012	NG42	Nigeria	4.1.2.1 Haarlem
ERR1679619		1	Public dataset	MDR	2012	NG44	Nigeria	4.3.3 LAM
ERR1679621		1	Public dataset	MDR	2012	NG46	Nigeria	4.6.2.2 Cameroon
ERR1679628		1	Public dataset	MDR	2012	NG54	Nigeria	4.6.2.2 Cameroon
ERR1679629		1	Public dataset	MDR	2012	NG55	Nigeria	4.2.2.1 TUR
ERR1679633		1	Public dataset	MDR	2012	NG64	Nigeria	4.6.2.2 Cameroon
ERR1679636		1	Public dataset	MDR	2012	NG68	Nigeria	4.6.2.2 Cameroon
ERR1679656		1	Public dataset	MDR	2012	NG9	Nigeria	4.6.2.2 Cameroon
ERR1750880		1	Public dataset	MDR	NA	004-2		4.4.1.1 S-type
ERR1815553		1	Public dataset	MDR	NA	MTB_PT3		4.3.4.2 LAM
ERR1815556		1	Public dataset	MDR	NA	MTB_PT6		4.3.4.2 LAM
ERR1873405		1	Public dataset	XDR	2011	M 15 A690 15441 F1 Mycobacterium CAATGGA L003	South Africa	4.1.1.3 X-type
ERR1873475		1	Public dataset	XDR	2011	R14852	South Africa	4.1.2.1 Haarlem
ERR1873483		1	Public dataset	MDR	2011	R16583	South Africa	4.3.2.1 LAM
ERR1873488		1	Public dataset	XDR	2012	R17085 pool 283 L3 ACACGACC L003	South Africa	2.2.1 Beijing
ERR1873505		1	Public dataset	MDR	2012	R18832 pool 282 L2 CCTCCTGA L002	South Africa	2.2.2 Beijing Ancestral 1
ERR1873506		1	Public dataset	MDR	2012	R18913 pool 282 L2 CGAACTTA L002	South Africa	4.1.1.3 X-type
ERR1873518		1	Public dataset	XDR	2012	R19366	South Africa	4.1.1.3 X-type
ERR1873527		1	Public dataset	XDR	2008	R4674 LFO46Pool106 3312 L6 AGCCATGC L006	South Africa	2.2.1 Beijing
ERR1873534		1	Public dataset	XDR	2009	R5065	South Africa	4.1.2.1 Haarlem
ERR1873540		1	Public dataset	XDR	2009	R5847 LFO46Pool105 3311 L5 CTGTAGCC L005	South Africa	4.1.1.3 X-type
ERR1873544		1	Public dataset	XDR	2009	R6297 LFO46Pool105 3311 L5 AACAACCA L005	South Africa	2.2.2 Beijing Ancestral 1
ERR1873552		1	Public dataset	XDR	2009	R8194 LFO46Pool106 3312 L6 CTAAGGTC L006	South Africa	2.2.2 Beijing Ancestral 1
ERR1952140		1	Public dataset	MDR	2004	IEMDR03	Ireland	1.2.1 EAI Manila

ERR1952141		1	Public dataset	XDR	2004	IEMDR04	Ireland	2.2.1 Beijing Europe/Russian W148 Outbreak
ERR1952142		1	Public dataset	MDR	2004	IEMDR05	Ireland	1.1.2 EAI
ERR1988847		1	Public dataset	XDR	2010	2	South Africa	2.2.2 Beijing Ancestral 1
ERR2027265		1	Public dataset	MDR	2016	Mtb__2274		2.2.1 Beijing Europe/Russian W148 Outbreak
ERR2027285		1	Public dataset	MDR	2016	Mtb__2298		4.6.2.2 Cameroon
ERR2145493		1	Public dataset	MDR	NA	MTB Saudi 1459 MTB-Pool59 3499 L3 CGCTGATC L003		3 Delhi-CAS
ERR2145508		1	Public dataset	MDR	NA	MTB Saudi 1865 MTB-Pool59 3499 L3 AGATCGCA L003		3 Delhi-CAS
ERR2145512		1	Public dataset	MDR	NA	MTB Saudi 1910 MTB-Pool59 3499 L3 ATTGAGGA L003		4.1.2.1 Haarlem
ERR2145520		1	Public dataset	MDR	NA	MTB Saudi 1991 MTB-Pool59 3499 L3 CCGAAGTA L003		4.2.2 Euro-American
ERR2145524		1	Public dataset	MDR	NA	MTB Saudi 2057 MTB-Pool59 3499 L3 CGCATACA L003		4.2.2 Euro-American
SRR3205962		1	Public dataset	MDR	2009	DS21644 (WBB274)	Thailand	2.2.1 Beijing Asian/Africa 2
SRR3544725		1	Public dataset	MDR	2014	Mycobacterium tuberculosis G-022__1	Georgia	2.2.1 Beijing Central Asia outbreak
SRR3544734		1	Public dataset	MDR	2013	Mycobacterium tuberculosis G-001S__1__1	Georgia	2.2.1 Beijing Central Asia outbreak
SRR3544735		1	Public dataset	MDR	2014	Mycobacterium tuberculosis G-019S__1__1	Georgia	2.2.1 Beijing Central Asia outbreak
SRR3544748		1	Public dataset	MDR	2014	Mycobacterium tuberculosis G-016S__1__1	Georgia	2.2.1 Beijing Central Asia
SRR3618864		1	Public dataset	XDR	2013	M67	Myanmar: Yangon	2.2.1 Beijing Asian/Africa 2
SRR3647353		1	Public dataset	MDR	2010	22103	China	4.2.2 Euro-American
SRR3647359		1	Public dataset	MDR	2006	2242	China	2.2.1 Beijing Asian/Africa 2
SRR3647360		1	Public dataset	MDR	2006	2279	China	2.2.1 Beijing Ancestral 3
SRR3675217		1	Public dataset	MDR	2015	MMMOSAM: 0975de12- 90b8-4cc5-800a -1c3dba1920b8	United Kingdom	3 Delhi-CAS
SRR3675285		1	Public dataset	MDR	2015	MMMOSAM: 49693fa3- 1486-4e97-9229 -ba7691f7a3d3	United Kingdom	2.2.1 Beijing Central Asia
SRR3724660		1	Public dataset	MDR	NA	G04030	Congo	4.3.2 LAM
SRR3724791		1	Public dataset	MDR	NA	G04170	Ivory Cost	4.1 Euro-American
SRR3724794		1	Public dataset	MDR	NA	G04319	Thailand	1.1.1.1 EAI
SRR3724799		1	Public dataset	MDR	NA	G04032	Congo	4.3.4.2.1 LAM
SRR3724802		1	Public dataset	MDR	NA	G04043	Ivory Cost	4.1 Euro-American
SRR3724807		1	Public dataset	MDR	NA	G04033	Ivory Cost	4.1 Euro-American
SRR3724819		1	Public dataset	MDR	NA	G04040	Ivory Cost	4.1 Euro-American
SRR3724950		1	Public dataset	MDR	NA	G04031	Congo	4.6.1.2 Uganda
SRR3724951		1	Public dataset	MDR	NA	G04057	Congo	4.3.2 LAM
SRR3724965		1	Public dataset	MDR	NA	G04039	Ivory Cost	4.1 Euro-American
SRR3724971		1	Public dataset	MDR	NA	G04041	Ivory Cost	4.1 Euro-American
SRR3724998		1	Public dataset	MDR	NA	G04171	Ivory Cost	4.1 Euro-American
SRR3725008		1	Public dataset	MDR	NA	G04108	Thailand	2.2.1 Beijing Asian/Africa 2
SRR3725012		1	Public dataset	MDR	NA	G04111	Thailand	2.2.1 Beijing
SRR3725013		1	Public dataset	MDR	NA	G04114	Thailand	2.2.1 Beijing
SRR3725693		1	Public dataset	MDR	NA	G05033	Ivory Cost	4.1 Euro-American
SRR3725703		1	Public dataset	MDR	NA	G05025	Ivory Cost	4.1 Euro-American
SRR3725708		1	Public dataset	MDR	NA	G05032	Ivory Cost	4.1 Euro-American
SRR3725712		1	Public dataset	MDR	NA	G05028	Ivory Cost	4.1 Euro-American
SRR3725714		1	Public dataset	MDR	NA	G05026	Ivory Cost	4.1 Euro-American
SRR3725717		1	Public dataset	MDR	NA	G05035	Ivory Cost	4.6 Euro-American
SRR3725719		1	Public dataset	MDR	NA	G05034	Ivory Cost	4.1 Euro-American
SRR3725722		1	Public dataset	MDR	NA	G05037	Ivory Cost	4.1 Euro-American
SRR3732570		1	Public dataset	MDR	NA	G05107	Peru	4.3.3 LAM
SRR3732576		1	Public dataset	MDR	NA	G05145	Peru	4.3.3 LAM
SRR3732578		1	Public dataset	MDR	NA	G05133	Peru	4.3.3 LAM
SRR3732579		1	Public dataset	MDR	NA	G05138	Peru	4.3.4.1 LAM
SRR3732580		1	Public dataset	MDR	NA	G05147	Peru	4.3.3 LAM

Appendix Table 3.2: **Characteristics of the 131 multi- and extensively drug resistant *M. tuberculosis* isolates from Germany analyzed in Chapter 4.** We could find demographic and epidemiological information in the national TB notification system for 129 isolates. Four isolates were classified as not multi-resistent according to the molecular drug resistance prediction. MDR: multidrug resistant; XDR: extensively drug-resistant; NA: not available; S: Streptomycin; E: Ethambutol; Z: Pyrazinamide; R: Rifampicin; H: isoniazid; Mfx: Moxifloxacin, Lfx: Levofloxacin, Ofx: Ofloxacin, Amk: Amikacin; Cm: Capreomycin, Km: Kanamycin; Res: resistant; Sus: susceptible

Isolate ID	Molecular drug resistance prediction	Patient country of birth	Patient nationality	Federal state of isolation	Antibiotic susceptibility test results											Phenotypic drug resistance prediction
					S	E	Z	R	H	Mfx	Lfx	Ofx	Amk	Cm	Km	
10162-13	MDR	Kyrgyzstan	Kyrgyzstan	Baden-Württemberg	Res	Sus	Res	Res	Res	Sus	NA	Sus	Sus	Sus	Sus	MDR
10284-13	MDR	Russia	Russia	Saxony	Res	Res	Res	Res	Res	NA	NA	Sus	Sus	Sus	NA	MDR
10346-12	MDR	Georgia	Georgia	Saxony	Res	Res	Sus	Res	Res	NA	NA	NA	NA	NA	NA	MDR
10428-12	MDR	Ukraine	Ukraine	Hesse	Res	Sus	Sus	Res	Res	NA	NA	NA	NA	NA	NA	MDR
10490-13	MDR	Kazakhstan	Kazakhstan	Lower Saxony	Res	Res	Res	Res	Res	NA	NA	NA	NA	NA	NA	MDR
10505-13	MDR	Somalia	Somalia	Hesse	Res	Sus	Res	Res	Res	NA	NA	NA	NA	NA	NA	MDR
10655-12	XDR	Lebanon	Lebanon	North Rhine- Westphalia	Res	Res	Res	Res	Res	Res	NA	Res	Res	Res	NA	XDR
10743-13	MDR	Pakistan	Pakistan	North Rhine- Westphalia	Res	Res	Res	Res	Res	Res	NA	Res	Sus	Sus	NA	MDR
10759-13	MDR	Vietnam	Vietnam	Hamburg	Res	Res	Res	Res	Res	NA	NA	Sus	Res	Res	NA	MDR
10840-13	MDR	India	India	Saxony-Anhalt	Sus	Sus	Sus	Res	Res	Sus	Res	Res	Sus	Sus	NA	MDR
10896-12	MDR	Germany	Germany	North Rhine- Westphalia	Res	Res	Res	Res	Res	NA	NA	Sus	Sus	Sus	NA	MDR
10926-12	MDR	Ukraine	Ukraine	North Rhine- Westphalia	Res	Sus	Res	Res	Res	NA	NA	Sus	Sus	Sus	NA	MDR
10962-13	MDR	Romania	Romania	Lower Saxony	Res	Sus	Sus	Res	Res	NA	NA	Sus	Sus	Sus	NA	MDR
11132-13	MDR	Ukraine	Ukraine	Saxony-Anhalt	Res	Sus	Sus	Res	Res	NA	NA	NA	NA	NA	NA	MDR
11250-12	MDR	Bulgaria	Bulgaria	Rhineland-Palatinate	Sus	Sus	Sus	Res	Res	NA	NA	Sus	Sus	Sus	Sus	MDR
11355-13	XDR	Poland	Poland	North Rhine- Westphalia	Res	Sus	Res	Res	Res	Sus	NA	NA	Sus	Sus	NA	MDR
11460-12	MDR	Nepal	Nepal	North Rhine- Westphalia	Res	Res	Res	Res	Res	NA	NA	Sus	Sus	Sus	NA	MDR
11686-13	MDR	Kosovo	Kosovo	Hesse	Res	Res	Res	Res	Res	NA	NA	Sus	Sus	Sus	NA	MDR
11883-13	MDR	Georgia	Georgia	Bavaria	Res	Res	Res	Res	Res	NA	NA	NA	NA	NA	NA	MDR
11960-13	MDR	Russia	Russia	Brandenburg	Res	Res	Res	Res	Res	NA	NA	Sus	Sus	Sus	NA	MDR
11987-13	MDR	Russia	Russia	NA	Res	Res	Res	Res	Res	NA	NA	NA	NA	NA	NA	MDR
12009-13	MDR	Nigeria	Nigeria	North Rhine- Westphalia	Res	Sus	Sus	Res	Res	NA	NA	NA	NA	NA	NA	MDR
12016-13	XDR	Georgia	Georgia	Bavaria	Res	Res	Res	Res	Res	NA	NA	NA	NA	NA	NA	MDR
12017-13	MDR	Romania	Romania	Bavaria	Res	Sus	Sus	Res	Res	NA	NA	Sus	Sus	Sus	NA	MDR
12018-13	XDR	Azerbaijan	Azerbaijan	Bavaria	Res	Sus	Res	Res	Res	Sus	NA	Res	Res	Sus	NA	XDR
12041-13	MDR	Germany	Germany	Rhineland-Palatinate	Res	Res	Res	Res	Res	NA	NA	NA	NA	NA	NA	MDR
12103-13	MDR	Syria	Syria	Schleswig-Holstein	Res	Res	Res	Res	Res	Res	Res	Res	Sus	Sus	NA	MDR
1244-13	MDR	Germany	Germany	North Rhine- Westphalia	Res	Res	Res	Res	Res	NA	NA	Sus	Sus	Sus	NA	MDR
12466-13	MDR	Lithuania	Lithuania	North Rhine- Westphalia	Res	Res	Res	Res	Res	NA	NA	Sus	Sus	Sus	NA	MDR
12471-13	MDR	Germany	Kasachstan	Schleswig-Holstein	Res	Res	Res	Res	Res	Sus	Res	Res	Sus	Sus	NA	MDR
12487-13	MDR	Poland	Germany	Berlin	Res	Sus	Res	Res	Res	NA	NA	Sus	Sus	Sus	NA	MDR
12510-13	MDR	Abroad	Abroad	Hamburg	Res	Res	Res	Res	Res	Res	Res	Res	Sus	Sus	NA	MDR
1296-12	MDR	Abroad	Abroad	North Rhine- Westphalia	Res	Res	Res	Res	Res	NA	NA	NA	NA	NA	NA	MDR
1298-12	MDR	Kazakhstan	Kazakhstan	North Rhine- Westphalia	Res	Sus	Res	Res	Res	NA	NA	NA	NA	NA	NA	MDR
13344-13	MDR	Romania	Germay	Bavaria	Res	Res	Res	Res	Res	NA	NA	Sus	Sus	Sus	NA	MDR
134-13	XDR	Lebanon	Lebanon	North Rhine- Westphalia	Res	Res	Res	Res	Res	Res	NA	Res	Res	Res	NA	XDR
13432-13	MDR	Kazakhstan	Kazakhstan	Schleswig-Holstein	Res	Res	Res	Res	Res	Sus	NA	Sus	Res	Res	NA	MDR
13739-13	XDR	Armenia	Armenia	Schleswig-Holstein	Res	Res	Res	Res	Res	Res	Res	NA	Res	Res	NA	XDR
13898-13	MDR	Azerbaijan	Azerbaijan	North Rhine- Westphalia	Res	Sus	Res	Res	Res	Res	NA	Res	Sus	Sus	NA	MDR
14102-13	not MDR	Thailand	Thailand	NA	Res	Res	Res	Res	Res	Sus	NA	Res	Sus	Sus	NA	MDR
14217-13	MDR	Sri Lanka	Sri Lanka	North Rhine- Westphalia	Res	Sus	Sus	Res	Res	NA	NA	Sus	Sus	Sus	NA	MDR
14489-13	MDR	Camerun	Camerun	Bremen	Res	Res	Res	Res	Res	NA	NA	Sus	Sus	Sus	NA	MDR
1560-12	MDR	Ukraine	Deutschland	Lower Saxony	Res	Res	Res	Res	Res	NA	NA	Sus	Sus	Sus	NA	MDR
1571-12	MDR	Germany	Germany	Lower Saxony	Res	Sus	Sus	Res	Res	NA	NA	NA	NA	NA	NA	MDR
1635-12	MDR	Germany	Germany	Lower Saxony	Res	Sus	Sus	Res	Res	NA	NA	NA	NA	NA	NA	MDR
1725-13	MDR	Russia	Russia	Baden-Württemberg	Res	Res	Res	Res	Res	NA	NA	NA	NA	NA	NA	MDR
2065-13	MDR	Russia	Germany	Hamburg	Res	Res	Sus	Res	Res	NA	NA	NA	NA	NA	NA	MDR
2135-12	MDR	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
2303-12	MDR	Camerun	Camerun	Lower Saxony	Res	Res	Res	Res	Res	NA	NA	NA	NA	NA	NA	MDR
2378-13	MDR	Russia	Russia	North Rhine- Westphalia	Res	Sus	Sus	Res	Res	NA	NA	Sus	Sus	Sus	NA	MDR
253-12	MDR	Germany	Germany	Baden-Württemberg	Res	Res	Res	Res	Res	NA	NA	Sus	Sus	Sus	NA	MDR
254-12	MDR	Germany	Germany	Baden-Württemberg	Res	Res	Res	Res	Res	NA	NA	Sus	Sus	Sus	NA	MDR
2636-13	MDR	India	India	Saxony-Anhalt	Res	Res	Res	Res	Res	NA	NA	NA	NA	NA	NA	MDR
2709-13	MDR	Romania	Romania	Baden-Württemberg	Res	Res	Res	Res	Res	NA	NA	NA	NA	NA	NA	MDR
2718-13	MDR	Afghanistan	Deutschland	North Rhine- Westphalia	Sus	Sus	Sus	Res	Res	NA	NA	NA	NA	NA	NA	MDR
2823-13	MDR	Georgia	Georgia	Berlin	Res	Sus	Sus	Res	Res	NA	NA	NA	NA	NA	NA	MDR
2955-12	XDR	Russia	Russia	Hesse	Res	Res	Res	Res	Res	NA	NA	NA	NA	NA	NA	MDR
3007-13	XDR	Algeria	Algeria	North Rhine- Westphalia	Res	Res	Res	Res	Res	Res	NA	Res	Sus	Sus	NA	MDR
304-13	MDR	Romania	Romania	Bavaria	Res	Res	Sus	Res	Res	NA	NA	Sus	Res	Sus	NA	MDR
3106-13	MDR	Russia	Russia	Hesse	Res	Sus	Sus	Res	Res	NA	NA	Sus	Sus	Sus	NA	MDR
3116-13	MDR	India	India	North Rhine- Westphalia	Res	Sus	Res	Res	Res	NA	NA	Sus	Sus	Sus	NA	MDR

3125-13	MDR	Abroad	Abroad	Hamburg	Res	Res	Res	Res	Res	Res	Res	Res	Sus	Sus	NA	MDR
3201-12	MDR	China	China	Saxony	Res	Sus	Sus	Res	Res	NA	NA	NA	NA	NA	NA	MDR
3290-12	MDR	Kenia	Kenia	Saarland	Sus	Sus	Sus	Res	Res	NA	NA	NA	NA	NA	NA	MDR
3413-12	MDR	Kazakhstan	Deutschland	Brandenburg	Res	Sus	Sus	Res	Res	NA	NA	NA	NA	NA	NA	MDR
3593-12	MDR	Germany	Germany	Schleswig-Holstein	Res	Sus	Sus	Res	Res	NA	NA	NA	NA	NA	NA	MDR
3617-12	MDR	Germany	Germany	Lower Saxony	Res	Sus	Sus	Res	Res	NA	NA	NA	NA	NA	NA	MDR
4038-12	MDR	Germany	Germany	North Rhine- Westphalia	Res	Sus	Sus	Res	Res	NA	NA	NA	NA	NA	NA	MDR
4153-13	MDR	Georgia	Georgia	Berlin	Res	Res	NA	Res	Res	NA	NA	NA	NA	NA	NA	MDR
4245-13	XDR	Lithuania	Lithuania	Berlin	Res	Res	Res	Res	Res	Res	NA	Res	Sus	Sus	NA	MDR
4305-13	MDR	Russia	Russia	Berlin	Res	Res	Res	Res	Res	NA	NA	Sus	Sus	Sus	NA	MDR
4345-12	MDR	Russia	Russia	Brandenburg	Res	Res	Res	Res	Res	NA	NA	Sus	Res	Res	NA	MDR
4517-13	MDR	Russia	Russia	Saxony	Res	Res	Res	Res	Res	NA	NA	Sus	Sus	Sus	NA	MDR
4556-12	not MDR	NA	NA	Saxony	Res	Sus	Sus	Res	Res	NA	NA	Sus	Sus	Sus	NA	MDR
4563-13	MDR	Bulgaria	Bulgaria	Hesse	Sus	Sus	Sus	Res	Res	NA	NA	NA	NA	NA	NA	MDR
4751-13	MDR	Russia	Russia	Bavaria	Res	Sus	Res	Res	Res	NA	NA	Sus	Sus	Sus	NA	MDR
479-12	MDR	Romania	Germay	Bavaria	Res	Res	Sus	Res	Res	NA	NA	Sus	Sus	Sus	NA	MDR
4798-13	MDR	Russia	Russia	Saxony	Res	Res	Sus	Res	Res	NA	NA	Sus	Sus	Sus	NA	MDR
4839-12	MDR	Germany	Germany	Saxony-Anhalt	Res	Sus	Sus	Res	Res	NA	NA	Sus	Sus	Sus	NA	MDR
4893-12	MDR	Peru	Peru	North Rhine- Westphalia	Res	Sus	Sus	Res	Res	NA	NA	Sus	Sus	Sus	NA	MDR
4960-13	XDR	Russia	Russia	Berlin	Res	Res	Res	Res	Res	Res	NA	Res	Res	Res	NA	XDR
5033-12	MDR	Russia	NA	Berlin	Res	Sus	Sus	Res	Res	NA	NA	NA	NA	NA	NA	MDR
5096-13	XDR	Russia	Russia	Saxony	Res	Sus	Sus	Res	Res	Res	NA	Res	Sus	Sus	NA	MDR
5158-12	MDR	Russia	Russia	North Rhine- Westphalia	Res	Sus	Res	Res	Res	NA	NA	Sus	Sus	Sus	NA	MDR
5190-13	XDR	Georgia	Georgia	Baden-Württemberg	Res	Sus	Sus	Res	Res	Res	Res	Res	Sus	Sus	NA	MDR
521-14	MDR	Russia	Russia	Mecklenburg-Vorpommern	NA	Res	NA	Res	Res	NA	NA	NA	NA	NA	NA	MDR
5271-12	MDR	Ethiopia	Ethiopia	North Rhine- Westphalia	Sus	Res	Res	Res	Res	Sus	NA	Sus	Sus	Sus	Sus	MDR
5366-12	MDR	Ukraine	Ukraine	Bavaria	Res	Res	Sus	Res	Res	NA	NA	Sus	Sus	Sus	NA	MDR
5439-12	MDR	Romania	Germay	Bavaria	Sus	Sus	Sus	Res	Res	NA	NA	NA	NA	NA	NA	MDR
565-12	MDR	Kazakhstan	Kazakhstan	North Rhine- Westphalia	Res	Sus	Sus	Res	Res	NA	NA	Sus	Res	Sus	NA	MDR
5667-13	MDR	Russia	Russia	Bavaria	Res	Sus	Res	Res	Res	NA	NA	Sus	Sus	Sus	NA	MDR
5675-12	MDR	Germany	Germany	Rhineland-Palatinate	Res	Sus	Res	Res	Res	NA	NA	Sus	Sus	Sus	NA	MDR
5871-12	MDR	Romania	Romania	North Rhine- Westphalia	Res	Res	Res	Res	Res	NA	NA	Sus	Sus	Sus	NA	MDR
5887-13	MDR	Germany	Germany	North Rhine- Westphalia	Res	Res	Res	Res	Res	NA	NA	Sus	Sus	Sus	NA	MDR
6089-13	MDR	Kazakhstan	Deutschland	Hamburg	Res	Sus	Sus	Res	Res	NA	NA	Sus	Sus	Sus	NA	MDR
6316-13	XDR	Russia	Russia	Saxony	Res	Res	Res	Res	Res	Res	Res	Res	Res	Res	NA	XDR
6360-12	MDR	India	India	Bremen	Sus	Sus	Sus	Res	Res	NA	NA	NA	NA	NA	NA	MDR
6364-12	MDR	Romania	Romania	North Rhine- Westphalia	Res	Res	Res	Res	Res	NA	NA	Sus	Sus	Sus	NA	MDR
6760-13	MDR	India	India	Hesse	Sus	Sus	Sus	Res	Res	NA	NA	NA	NA	NA	NA	MDR
6764-13	MDR	Lithuania	Lithuania	North Rhine- Westphalia	Res	Res	Res	Res	Res	NA	NA	Sus	Sus	Sus	NA	MDR
6934-12	MDR	Eritrea	Eritrea	Baden-Württemberg	Sus	Sus	Sus	Res	Res	NA	NA	NA	NA	NA	NA	MDR
72-13	not MDR	Russia	Russia	Saxony	Res	Sus	Res	Res	Res	NA	NA	Sus	Sus	Sus	NA	MDR
7604-12	MDR	Kazakhstan	Deutschland	Bavaria	Res	Sus	Sus	Res	Res	NA	NA	NA	NA	NA	NA	MDR
7712-13	XDR	Russia	Russia	Brandenburg	Res	Res	Res	Res	Res	NA	NA	NA	NA	NA	NA	MDR
7854-13	MDR	Ukraine	Deutschland	Lower Saxony	Res	Res	Sus	Res	Res	Sus	NA	Sus	Sus	Sus	NA	MDR
7977-12	MDR	Russia	Russia	Bavaria	Res	Res	Res	Res	Res	NA	NA	Sus	Res	Sus	NA	MDR
7984-12	XDR	Russia	Russia	North Rhine- Westphalia	Res	Res	Res	Res	Res	Res	NA	Res	Res	Res	NA	XDR
8017-13	MDR	Georgia	Georgia	Saxony	Res	Res	Res	Res	Res	NA	NA	Res	Sus	Sus	NA	MDR
8291-13	XDR	Azerbaijan	Azerbaijan	Berlin	Res	Res	Res	Res	Res	NA	NA	NA	NA	NA	NA	MDR
8300-13	MDR	Lithuania	Lithuania	North Rhine- Westphalia	NA	NA	NA	Res	Res	NA	NA	Sus	Sus	Sus	NA	MDR
8305-13	MDR	India	India	Saxony-Anhalt	Sus	Sus	Sus	Res	Res	Sus	NA	Res	Sus	Sus	NA	MDR
833-12	MDR	India	India	Baden-Württemberg	Res	Res	Res	Res	Res	NA	NA	NA	NA	NA	NA	MDR
8347-13	MDR	Kazakhstan	Kazakhstan	Berlin	Res	Sus	Res	Res	Res	Sus	NA	Sus	Sus	Sus	NA	MDR
8565-12	XDR	Germany	Germany	Lower Saxony	Res	Res	Res	Res	Res	NA	NA	NA	NA	NA	NA	MDR
871-13	MDR	Germany	Germany	Baden-Württemberg	Res	Res	Res	Res	Res	NA	NA	Sus	Sus	Sus	NA	MDR
8847-13	MDR	Russia	Russia	Berlin	Res	Res	Res	Res	Res	NA	NA	NA	NA	NA	NA	MDR
886-12	MDR	Germany	Germany	Hamburg	Res	Sus	Sus	Res	Res	NA	NA	NA	NA	NA	NA	MDR
9082-13	MDR	Somalia	Somalia	Baden-Württemberg	Res	Res	Res	Res	Res	NA	NA	Sus	Sus	Sus	NA	MDR
9165-12	not MDR	Kyrgyzstan	Kyrgyzstan	Hamburg	Res	Sus	Sus	Res	Res	NA	NA	NA	NA	NA	NA	MDR
9354-12	MDR	Ukraine	Ukraine	North Rhine- Westphalia	Res	Res	Sus	Res	Res	Res	Res	Res	Sus	Sus	NA	MDR
9468-12	MDR	Germany	Germany	North Rhine- Westphalia	NA	Sus	NA	Res	Res	NA	NA	Sus	Sus	Sus	NA	MDR
9498-12	MDR	Turkey	Turkey	North Rhine- Westphalia	Res	Sus	Sus	Res	Res	NA	NA	Sus	Sus	Sus	NA	MDR
9505-13	MDR	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
9508-13	MDR	Russia	Russia	Baden-Württemberg	Res	Sus	Sus	Res	Res	NA	NA	Sus	Res	Sus	NA	MDR
9653-12	MDR	India	India	Baden-Württemberg	Res	Res	Res	Res	Res	NA	NA	NA	NA	NA	NA	MDR
9771-13	MDR	Germany	Germany	Bavaria	Res	Res	Res	Res	Res	NA	NA	Sus	Res	Res	NA	MDR
9776-13	MDR	Romania	Romania	Bavaria	Res	Sus	Sus	Res	Res	NA	NA	NA	NA	NA	NA	MDR
9777-13	MDR	Ukraine	Deutschland	Lower Saxony	Sus	Sus	Sus	Res	Res	NA	NA	NA	NA	NA	NA	MDR
9926-13	MDR	North Korea	North Korea	Hesse	Sus	Res	Res	Res	Res	NA	NA	Sus	Sus	Sus	NA	MDR
9927-13	MDR	Kosovo	Kosovo	Baden-Württemberg	Sus	Sus	Sus	Res	Res	NA	NA	NA	NA	NA	NA	MDR
999-13	MDR	Russia	Russia	Hamburg	Res	Res	Res	Res	Res	NA	NA	Sus	Res	Res	NA	MDR

Appendix Table 3.3: **Characteristics of the molecular clusters identified among the 1339 multi- and extensively drug resistant Mycobacterium tuberculosis isolates analyzed in Chapter 4.** The classification of the isolates in MDR or XDR was based on the molecular drug resistance prediction. MDR: multidrug resistant; XDR: extensively drug-resistant; NA: not available. * four isolates were classified as not multi-resistent according to the molecular drug resistance prediction

Cluster name	Cluster size	No. of MDR	Country of isolation of MDR (n)	No. of XDR	Country of isolation of XDR (n)	Year of isolation (n)
1	79	79	South Africa (79)	0	0	2013 (79)
2	56	54	Moldova (48) Germany (2) Georgia (1) NA (3)	2	Moldova (2)	2016 (6) 2015 (43) 2014 (1) 2013 (1) 2012 (1) 2009 (1) NA (3)
3	54	54	South Africa (54)	0	0	2013 (54)
4	33	4	South Africa (4)	29	South Africa (29)	2012 (10) 2011 (3) 2010 (7) 2009 (8) 2008 (5)
5	30	5	South Africa (4) Germany (1)	25	South Africa (25)	2012 (4) 2011 (10) 2010 (13) 2009 (3)
6	27	5	South Africa (5)	22	South Africa (22)	2012 (6) 2011 (9) 2010 (5) 2009 (4) 2008 (3)
7	26	11	NA (11)	15	NA (15)	NA (26)
8	23	10	Georgia (10)	13	Georgia (11) Azerbaijan (2)	2014 (8) 2015 (15)
9	18	18	Georgia (16) Germany (2)	0	0	2015 (8) 2014 (8) 2013 (2)
10	16	16	Tanzania (15) Botswana (1)	0	0	NA (16)
11	15	0	0	15	NA (15)	NA (15)
12	10	8	Germany (5) NA (3)	2	Germany (2)	2013 (7) 2009 (3)
13	10	1	Germany (1)	9	Kazakhstan (9)	2014 (9)2012 (1)
14	9	9	South Africa (9)	0	0	2012 (9)
15	8	0	0	8	Thailand (8)	2012 (2) 2011 (1) 2008 (3) 2007 (2)
16	8	8	South Africa (8)	0	0	2012 (8)
17	7	6	Moldova (6)	1	Moldova (1)	2016 (1) 2015 (6)
18	6	6	South Africa (6)	0	0	2013 (6)
19	6	6	Thailand (5) Switzerland (1)	0	0	2007 (1) 2005 (3) 2003 (1) NA (1)
20	6	4	Russia (2) Azerbaijan (1) Georgia (1)	2	Azerbaijan (2)	2016 (1) 2015 (3) NA (2)
21	6	6	Georgia (5) Germany (1)	0	0	2015 (2) 2014 (3) 2013 (1)
22	6	6	Ivory Coast (6)	0	0	NA (6)
23	6	6	Georgia (6)	0	0	2014 (3) 2015 (2) 2016 (1)
24	5	5	Georgia (3) Germany (2)	0	0	2015 (1) 2014 (1) 2013 (3)
25	5	5	Peru (5)	0	0	NA (5)
26	5	0	0	5	China (5)	2012 (2) 2011 (2) 2010 (1)
27	5	5	Moldova (5)	0	0	2015 (4)

						2009 (1)
28	5	5	South Africa (5)	0	0	2012 (5)
29	5	5	Georgia (5)	0	0	2015 (1) 2014 (4)
30	4	0	0	4	South Africa (4)	2011 (2) 2010 (2)
31	4	0	0	4	South Africa (4)	2012 (1) 2010 (3)
32	4	3	China (3)	1	China (1)	2011 (2) 2010 (2)
33	4	0	0	4	China (4)	2011 (1) 2009 (1) 2008 (1) 2007 (1)
34	4	4	Moldova (4)	0	0	2015 (3) 2014 (1)
35	4	0	0	4	Georgia (3) Germany (1)	2015 (2) 2014 (1) 2013 (1)
36	4	4	Botswana (4)	0	0	NA (4)
37	4	4	Djibouti (4)	0	0	NA (4)
38	4	4	Peru (4)	0	0	NA (4)
39	4	4	Vietnam (4)	0	0	2010 (1) 2009 (3)
40	3	3	NA (3)	0	0	NA (3)
41	3	3	NA (3)	0	0	NA (3)
42	3	3	NA (3)	0	0	NA (3)
43	3	0	0	3	NA (3)	NA (3)
44	3	3	Nigeria (3)	0	0	2012 (3)
45	3	0	0	3	Georgia (3)	2014
46	3	0	0	3	Georgia (3)	2015 (1) 2014 (1) 2013 (1)
47	3	3	UK Oxford (3)	0	0	2015 (3)
48	3	3	Bangladesh (3)	0	0	2011 (2) 2008 (1)
49	3	3	Bangladesh (3)	0	0	2012 (1) 2007 (2)
50	3	3	Vietnam (3)	0	0	2011 (2) 2010 (1)
51	3	3	Romania (3)	0	0	2015 (3)
52	3	3	Romania (3)	0	0	2016 (1) 2015 (2)
53	3	2	Romania (1) Germany (1)	1	Romania (1)	2016 (1) 2015 (1) 2013 (1)
54	3	3	Germany (3)	0	0	2012 (3)
55	3	1	South Africa (1)	2	South Africa (2)	2011 (1) 2010 (1) 2009 (1)
56	3	3	Viet Nam (3)	0	0	2011 (1) 2010 (1) 2009 (1)
57	3	1	Azerbaijan (1)	2	Azerbaijan (2)	2016 (1) 2015 (2)
58	3	3	India (2) Germany (1)	0	0	2014 (1) 2012 (1) 2005 (1)
59	3	3	Georgia (1) Germany (2)	0	0	2015 (1) 2012 (2)
60	3	2	Azerbaijan (2)	1	Azerbaijan (1)	2015 (3)
61	2	2	South Africa (2)	0	0	2010 (1) 2008 (1)
62	2	2	Spain (2)	0	0	NA (2)
63	2	2	Spain (2)	0	0	NA (2)
64	2	2	Nigeria (2)	0	0	2012 (2)
65	2	2	Nigeria (2)	0	0	2012 (2)
66	2	2	South Africa (2)	0	0	2009 (2)
67	2	2	South Africa (2)	0	0	2009 (2)
68	2	1	South Africa (1)	1	South Africa (1)	2010 (2)
69	2	0	0	2	South Africa (2)	2012 (1) 2010 (1)
70	2	1	South Africa (1)	1	South Africa (1)	2010 (1) 2009 (1)
71	2	0	0	2	South Africa (2)	2011 (1)

						2009 (1)
72	2	0	0	2	South Africa (2)	2011 (1) 2009 (1)
73	2	2	Thailand (2)	0	0	2009 (1) 2008 (1)
74	2	0	0	2	Thailand (2)	2012 (1) 2011 (1)
75	2	0	0	2	Georgia (2)	2015 (1) 2014 (1)
76	2	2	Georgia (2)	0	0	2015 (1) 2014 (1)
77	2	2	Ivory Coast (2)	0	0	NA (2)
78	2	2	Peru (2)	0	0	NA (2)
79	2	2	Moldova (2)	0	0	2015 (1) 2012 (1)
80	2	2	Moldova (2)	0	0	2015 (1) 2014 (1)
81	2	0	0	2	Moldova (2)	2015 (2)
82	2	0	0	2	Moldova (2)	2015 (2)
83	2	2	Moldova (2)	0	0	2015 (1) 2014 (1)
84	2	0	0	2	Moldova (2)	2015 (1) 2012 (1)
85	2	2	Bangladesh (2)	0	0	2008 (2)
86	2	0	0	2	Bangladesh (2)	2010 (1) 2008 (1)
87	2	2	Bangladesh (2)	0	0	2011 (1) 2008 (1)
88	2	2	England (2)	0	0	2015 (2)
89	2	2	Vietnam (2)	0	0	2010 (2)
90	2	2	Vietnam (2)	0	0	2010 (2)
91	2	2	India (2)	0	0	2014 (2)
92	2	2	Georgia (2)	0	0	2015 (2)
93	2	2	Georgia (2)	0	0	2015 (2)
94	2	2	Azerbaijan (2)	0	0	2016 (2)
95	2	0	0	2	Azerbaijan (2)	2016 (2)
96	2	0	0	2	Azerbaijan (2)	2016 (1) 2015 (1)
97	2	2	Georgia (2)	0	0	2015 (2)
98	2	2	NA (2)	0	0	NA (2)
99	2	2	Georgia (2)	0	0	2015 (1) 2014 (1)
100	2	2	Georgia (2)	0	0	2014 (2)
101	2	2	Georgia (2)	0	0	2014 (2)
102	2	2	Georgia (2)	0	0	2015 (1) 2014 (1)
103	2	0	0	2	Georgia (1) Germany (1)	2014 (1) 2013 (1)
104	2	2	Moldova (2)	0	0	2015 (1) 2009 (1)
105	2	2	India (2)	0	0	2005 (2)
106	2	2	Romania (2)	0	0	2015 (2)
107	2	2	Romania (2)	0	0	2016 (2)
108	2	2	Congo (2)	0	0	NA (2)
109	2	2	Tanzania (2)	0	0	NA (2)
110	2	2	Djibouti (2)	0	0	NA (2)
111	2	2	Djibouti (2)	0	0	NA (2)
112	2	2	Djibouti (2)	0	0	NA (2)
113	2	2	Djibouti (2)	0	0	NA (2)
114	2	2	Ireland (1) Ivory Coast (1)	0	0	2004 (1) NA (1)
115	2	2	Vietnam (2)	0	0	2011 (1) 2009 (1)
116	2	2	Vietnam (2)	0	0	2010 (2)
117	2	2	Romania (2)	0	0	2016 (1) 2015 (1)
118	2	2	NA (2)	0	0	NA (2)
119	2	0	0	2	Germany (1) NA (1)	2012 (1) NA (1)
120	2	0	0	2	South Africa (1) NA (1)	2010 (1) NA (1)
121	2	2	NA (2)	0	0	NA (2)
122	2	1	NA (1)	1	Azerbaijan (1)	2016 (1) NA (1)
123	2	2	NA (2)	0	0	NA (2)

124	2	1	Germany (1)	1	Germany (1)	2013 (1) 2012 (1)
125	2	2	Germany (2)	0	0	2013 (1) 2012 (1)
126	2	0	0	2	Germany (2)	2013 (1) 2012 (1)
127	2	2	Germany (2)	0	0	2012 (2)
128	2	2	Germany (2)	0	0	2012 (2)
129	2	2	Germany (2)	0	0	2012 (2)
130	2	0	0	2	Azerbaijan (2)	2015 (1) 2016 (1)
131	2	1	Azerbaijan (1)	1	Azerbaijan (1)	2015 (1) 2016 (1)
132	1	1	Georgia (1)	1	Gergia (1)	2014 (2)
133	2	2	Georgia (2)	0	0	2014 (2)
All other	1 (n=595)*	507	Azerbaijan (17) Bangladesh (34) Botswana (1) Canada (2) China (7) Congo (5) Djibouti (19) Georgia (56) Germany (84) India (22) Ireland (2) Ivory Coast (20) Moldova (38) Nigeria (21) Peru (12) Romania (18) South Africa (7) Switzerland (4) Thailand (7) UK (3) Vietnam (50) NA (82)	84	Azerbaijan (19) China (3) Georgia (13) Germany (9) Ireland (1) Moldova (17) Myanmar (2) Romania (6) South Africa (10) NA (4)	1996 (1) 2004 (5) 2005 (9) 2006 (5) 2007 (5) 2008 (8) 2009 (29) 2010 (37) 2011 (19) 2012 (58) 2013 (68) 2014 (47) 2015 (120) 2016 (30) NA (154)

Appendix Table 3.4: Statistics for distances of the molecular clusters identified in the 1339 isolates analyzed in Chapter 4.

Cluster name	No. of isolates in the cluster	Min. distance	1st Qu. distance	Median distance	Mean distance	3rd Qu. distance	Max. distance	Max. cluster distance
1	79	0	0	0	0.0384	0	1.013	3.057156273
2	56	0	2.022	5.042	5.263	8.318	12.1	28.2808796
3	54	0	0	0	0.09361	0	2.02	6.150290122
4	33	0	1.016	3.036	3.223	5.042	8.084	26.38442151
5	30	0	1.276	3.032	3.96	6.093	12.12	24.57137224
6	27	0	1.009	4.038	3.772	6.067	10.09	21.20895554
7	26	0	0	0	0.7377	0	8.07	18.28616526
8	23	0	0	2.019	2.593	3.034	8.077	22.32021608
9	18	0	2.023	3.028	3.201	5.052	6.08	14.2067409
10	16	0	0	0	0	0	0	2.023471048
11	15	0	0	0	0.7423	0	9.109	15.19133342
12	10	1.008	6.112	7.584	6.889	8.897	12.09	21.35428151
13	10	0	0	0	1.516	1.006	12.15	16.22793222
14	9	0	0	0	0.2243	0	1.01	2.025796376
15	8	0	0	0	0	0	0	13.18884877
16	8	0	0	0	0.3786	0	3.029	4.052700175
17	7	9.085	9.588	10.09	10.38	11.11	12.12	21.1984447
18	6	0	0	0	0.1688	0	1.013	2.02753764
19	6	1.012	1.262	2.013	1.847	2.014	3.019	12.14704801
20	6	0	1.772	7.089	5.235	7.858	9.116	14.17014257
21	6	2.021	2.277	3.562	3.394	4.081	5.117	8.087446989
22	6	4.064	5.079	8.124	7.613	8.887	12.16	16.26158603
23	6	0	0	0.5057	2.866	4.054	11.12	14.14966393
24	5	2.044	2.044	9.166	6.508	9.166	10.12	15.2871022
25	5	0	0	0	1.014	2.027	3.044	5.074593272
26	5	0	0	1.008	1.615	3.04	4.026	8.097812921
27	5	1.011	1.011	6.056	5.654	9.083	11.11	18.15923512
28	5	0	0	0	0	0	0	0
29	5	0	0	2.017	1.211	2.018	2.021	5.046757154
30	4	0	0	1.015	3.03	4.045	10.09	12.19771673
31	4	6.125	6.125	8.125	8.369	10.37	11.1	21.22629168

32	4	1.006	1.006	1.512	1.763	2.269	3.021	5.047357973
33	4	0	0	0	2.273	2.273	9.091	12.13501466
34	4	0	0	4.539	5.042	9.581	11.09	20.17565658
35	4	6.082	6.082	6.087	6.088	6.093	6.096	8.086253523
36	4	2.028	2.028	3.034	3.287	4.293	5.051	7.09895851
37	4	6.091	6.091	6.602	6.602	7.112	7.112	15.15739149
38	4	1.012	1.012	2.024	2.024	3.035	3.035	13.12900207
39	4	8.08	8.08	9.603	9.851	11.37	12.12	19.19595401
40	3	0	0	0	0	0	0	5.03089688
41	3	0	0	0	3.026	4.539	9.079	10.09109431
42	3	0	0	0	0.3372	0.5058	1.012	1.011640313
43	3	0	0	0	0	0	0	0
44	3	0	0	0	0	0	0	0
45	3	0	0	0	0.3364	0.5045	1.009	1.016143906
46	3	1.009	1.009	1.009	2.692	3.533	6.057	6.069907387
47	3	0	0	0	0.6755	1.013	2.026	2.026567623
48	3	3.033	3.033	3.033	4.045	4.551	6.069	7.077201974
49	3	4.04	4.04	4.04	4.383	4.555	5.07	9.125076402
50	3	7.079	7.079	7.079	8.755	9.593	12.11	15.14912629
51	3	3.021	3.021	3.021	4.027	4.53	6.04	7.04796942
52	3	2.017	2.017	2.017	2.017	2.018	2.019	2.018999692
53	3	6.078	6.078	6.078	7.765	8.608	11.14	14.14915994
54	3	0	0	0	0	0	0	0
55	3	10.08	10.08	10.08	10.42	10.59	11.1	13.11956476
56	3	8.076	8.076	8.076	8.747	9.083	10.09	12.10643743
57	3	11.13	11.13	11.13	11.46	11.63	12.13	17.1762409
58	3	4.056	4.056	4.056	6.748	8.095	12.13	12.17467231
59	3	10.16	10.16	10.16	10.82	11.15	12.13	14.22472827
60	3	6.058	6.058	6.058	7.07	7.576	9.094	12.11396659
61	2	2.027	2.027	2.027	2.027	2.027	2.027	2.027426003
62	2	5.054	5.054	5.054	5.054	5.054	5.054	5.05361068
63	2	0	0	0	0	0	0	0
64	2	0	0	0	0	0	0	0
65	2	0	0	0	0	0	0	0
66	2	8.066	8.066	8.066	8.066	8.066	8.066	8.06562748
67	2	0	0	0	0	0	0	0
68	2	0	0	0	0	0	0	0
69	2	3.027	3.027	3.027	3.027	3.027	3.027	3.026657042
70	2	11.09	11.09	11.09	11.09	11.09	11.09	11.08763435
71	2	1.008	1.008	1.008	1.008	1.008	1.008	1.008321702
72	2	5.044	5.044	5.044	5.044	5.044	5.044	5.044447093
73	2	2.027	2.027	2.027	2.027	2.027	2.027	2.026908846
74	2	1.011	1.011	1.011	1.011	1.011	1.011	1.011461827
75	2	0	0	0	0	0	0	0
76	2	7.07	7.07	7.07	7.07	7.07	7.07	7.069841622
77	2	4.037	4.037	4.037	4.037	4.037	4.037	4.037379951
78	2	11.13	11.13	11.13	11.13	11.13	11.13	11.12975949
79	2	0	0	0	0	0	0	0
80	2	6.101	6.101	6.101	6.101	6.101	6.101	6.10061254
81	2	3.029	3.029	3.029	3.029	3.029	3.029	3.028773342
82	2	12.11	12.11	12.11	12.11	12.11	12.11	12.11005164
83	2	1.01	1.01	1.01	1.01	1.01	1.01	1.010298885
84	2	9.089	9.089	9.089	9.089	9.089	9.089	9.088819984
85	2	11.11	11.11	11.11	11.11	11.11	11.11	11.11348411
86	2	0	0	0	0	0	0	0
87	2	6.054	6.054	6.054	6.054	6.054	6.054	6.054024077
88	2	0	0	0	0	0	0	0
89	2	6.048	6.048	6.048	6.048	6.048	6.048	6.048411812
90	2	9.085	9.085	9.085	9.085	9.085	9.085	9.085285817
91	2	0	0	0	0	0	0	0
92	2	6.105	6.105	6.105	6.105	6.105	6.105	6.104561037
93	2	5.047	5.047	5.047	5.047	5.047	5.047	5.047092604
94	2	0	0	0	0	0	0	0
95	2	1.013	1.013	1.013	1.013	1.013	1.013	1.012699552
96	2	0	0	0	0	0	0	0
97	2	8.071	8.071	8.071	8.071	8.071	8.071	8.070668529
98	2	0	0	0	0	0	0	0
99	2	3.034	3.034	3.034	3.034	3.034	3.034	3.033742147
100	2	0	0	0	0	0	0	0
101	2	0	0	0	0	0	0	0
102	2	3.024	3.024	3.024	3.024	3.024	3.024	3.024079401
103	2	4.047	4.047	4.047	4.047	4.047	4.047	4.04661729
104	2	1.012	1.012	1.012	1.012	1.012	1.012	1.012043976
105	2	0	0	0	0	0	0	0
106	2	9.065	9.065	9.065	9.065	9.065	9.065	9.064913792
107	2	1.01	1.01	1.01	1.01	1.01	1.01	1.009584824
108	2	9.078	9.078	9.078	9.078	9.078	9.078	9.078266857
109	2	1.012	1.012	1.012	1.012	1.012	1.012	1.012189473

110	2		3.055	3.055	3.055	3.055	3.055	3.055	3.055240456
111	2		4.054	4.054	4.054	4.054	4.054	4.054	4.053501895
112	2		7.08	7.08	7.08	7.08	7.08	7.08	7.079531158
113	2		0	0	0	0	0	0	0
114	2		12.16	12.16	12.16	12.16	12.16	12.16	12.16095963
115	2		11.1	11.1	11.1	11.1	11.1	11.1	11.10496898
116	2		12.11	12.11	12.11	12.11	12.11	12.11	12.10506517
117	2		10.11	10.11	10.11	10.11	10.11	10.11	10.10987531
118	2		0	0	0	0	0	0	0
119	2		0	0	0	0	0	0	0
120	2		8.125	8.125	8.125	8.125	8.125	8.125	8.124976419
121	2		11.1	11.1	11.1	11.1	11.1	11.1	11.09912356
122	2		11.16	11.16	11.16	11.16	11.16	11.16	11.16087906
123	2		0	0	0	0	0	0	0
124	2		10.1	10.1	10.1	10.1	10.1	10.1	10.10190631
125	2		11.21	11.21	11.21	11.21	11.21	11.21	11.21107428
126	2		0	0	0	0	0	0	0
127	2		0	0	0	0	0	0	0
128	2		5.051	5.051	5.051	5.051	5.051	5.051	5.05145359
129	2		3.033	3.033	3.033	3.033	3.033	3.033	3.032689063
130	2		1.01	1.01	1.01	1.01	1.01	1.01	1.010227003
131	2		0	0	0	0	0	0	0
132	2		0	0	0	0	0	0	0
133	2		0	0	0	0	0	0	0

Appendix Table 3.5: Comparison of MTBseq Pipeline (Kohl et al., 2018b) 2018 and PANPASCO Pipeline (Jandrasits et al.) for the German dataset in Chapter 4. The MTBseq pipeline detected 13 clusters while PANPASCO detected 11 clusters in the dataset of 131 samples. 32/131 samples were clustered by one of the methods. All clusters detected by PANPASCO were identical to clusters from MTBseq.

SampleID	MTBseq cutoff <13	PANPASCO cutoff <13
10926-12	group_1	cluster_5
11460-12	group_1	cluster_5
253-12	group_10	cluster_9
254-12	group_10	cluster_9
1244-13	group_11	cluster_1
12466-13	group_11	cluster_1
12487-13	group_11	cluster_1
3007-13	group_11	cluster_1
4245-13	group_11	cluster_1
5887-13	group_11	cluster_1
6764-13	group_11	cluster_1
11250-12	group_12	cluster_6
11686-13	group_12	cluster_6
4153-13	group_13	cluster_11
8017-13	group_13	cluster_11
10346-12	group_2	cluster_3
10428-12	group_2	cluster_3
2955-12	group_3	cluster_10
4305-13	group_3	cluster_10
4839-12	group_4	ungrouped
5096-13	group_4	ungrouped
10962-13	group_5	ungrouped
3593-12	group_5	ungrouped
10655-12	group_6	cluster_4
134-13	group_6	cluster_4
10896-12	group_7	cluster_2
5871-12	group_7	cluster_2
6364-12	group_7	cluster_2
11883-13	group_8	cluster_7
13344-13	group_8	cluster_7
1571-12	group_9	cluster_8
3617-12	group_9	cluster_8

Bibliography

Ahmad S. New approaches in the diagnosis and treatment of latent tuberculosis infection. *Respiratory Research*, 11(1):169, 2010.

Al-Humadi H. W., Al-Saigh R. J., and Al-Humadi A. W. Addressing the challenges of tuberculosis: A brief historical account. *Frontiers in Pharmacology*, 8:689, 2017.

Andrés M., Göhring-Zwacka E., Fiebig L., Priwitzer M., Richter E., Rüsch-Gerdes S., Haas W., Niemann S., and Brodhun B. Integration of molecular typing results into tuberculosis surveillance in Germany—A pilot study. *PLOS ONE*, 12(11), 2017.

Andrés M., Werf M. J. v. d., Ködmön C., Albrecht S., Haas W., Fiebig L., and Group S. s. Molecular and genomic typing for tuberculosis surveillance: A survey study in 26 European countries. *PLOS ONE*, 14(3):e0210080, 2019.

Angiuoli S. V. and Salzberg S. L. Mugsy: fast multiple alignment of closely related whole genomes. *Bioinformatics*, 27(3):334–342, 2011.

Ayabina D., Ronning J. O., Alfsnes K., Debech N., Brynildsrud O. B., Arnesen T., Norheim G., Mengshoel A.-T., Rykkvin R., Dahle U. R., Colijn C., and Eldholm V. Genome-based transmission modelling separates imported tuberculosis from recent transmission within an immigrant population. *Microbial Genomics*, 4(10), 2018.

Baier U., Beller T., and Ohlebusch E. Graphical pan-genome analysis with compressed suffix trees and the Burrows-Wheeler transform. *Bioinformatics*, 32(4):497–504, 2016.

Begun M., Newall A. T., Marks G. B., and Wood J. G. Contact Tracing of Tuberculosis: A Systematic Review of Transmission Modelling Studies. *PLOS ONE*, 8(9), 2013.

Beller T. and Ohlebusch E. A representation of a compressed de Bruijn graph for pan-genome analysis that enables search. *Algorithms for Molecular Biology*, 11(1):20, 2016.

Bjorn-Mortensen K., Lillebaek T., Koch A., Soborg B., Ladefoged K., Sørensen H. C. F., Kohl T. A., Niemann S., and Andersen A. B. Extent of transmission captured by contact tracing in a tuberculosis high endemic setting. *European Respiratory Journal*, 49(3): 1601851, 2017.

Blanchette M., Kent W. J., Riemer C., Elnitski L., Smit A. F., Roskin K. M., Baertsch R., Rosenbloom K., Clawson H., Green E. D., et al. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Research*, 14(4):708–715, 2004.

Bloom B. R., Atun R., Cohen T., Dye C., Fraser H., Gomez G. B., Knight G., Murray M., Nardell E., Rubin E., Salomon J., Vassall A., Volchenkov G., White R., Wilson D., and Yadav P. Tuberculosis. In Holmes K. K., Bertozzi S., Bloom B. R., and Jha P., editors, *Major Infectious Diseases*, chapter 11. The International Bank for Reconstruction and Development / The World Bank, Washington (DC), 3rd edition, 2017.

Bolger A. M., Lohse M., and Usadel B. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, 30(15):2114–2120, 2014.

Bouckaert R., Heled J., Kühnert D., Vaughan T., Wu C.-H., Xie D., Suchard M. A., Rambaut A., and Drummond A. J. BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLOS Computational Biology*, 10(4), 2014.

Bradley P., Gordon N. C., Walker T. M., Dunn L., Heys S., Huang B., Earle S., Pankhurst L. J., Anson L., de Cesare M., Piazza P., Votintseva A. A., Golubchik T., Wilson D. J., Wyllie D. H., Diel R., Niemann S., Feuerriegel S., Kohl T. A., Ismail N., Omar S. V., Smith E. G., Buck D., McVean G., Walker A. S., Peto T. E. A., Crook D. W., and Iqbal Z. Rapid antibiotic-resistance predictions from genome sequence data for Staphylococcus aureus and Mycobacterium tuberculosis. *Nature Communications*, 6: 10063, 2015.

Brent M. R. Genome annotation past, present, and future: How to define an ORF at each locus. *Genome Research*, 15(12):1777–1786, 2005.

Broad Institute. Picard tools. `http://broadinstitute.github.io/picard/`, Accessed: 2018-02-21.

Brudey K., Driscoll J. R., Rigouts L., Prodinger W. M., Gori A., Al-Hajoj S. A., Allix C., Aristimuño L., Arora J., Baumanis V., et al. Mycobacterium tuberculosis complex genetic diversity: mining the fourth international spoligotyping database (spoldb4) for classification, population genetics and epidemiology. *BMC Microbiology*, 6(1):23, 2006.

Brudno M., Do C. B., Cooper G. M., Kim M. F., Davydov E., Green E. D., Sidow A., Batzoglou S., Program N. C. S., et al. Lagan and multi-lagan: efficient tools for large-scale multiple alignment of genomic dna. *Genome Research*, 13(4):721–731, 2003.

Bryant J. M., Schürch A. C., van Deutekom H., Harris S. R., de Beer J. L., de Jager V., Kremer K., van Hijum S. A., Siezen R. J., Borgdorff M., et al. Inferring patient to patient transmission of mycobacterium tuberculosis from whole genome sequencing data. *BMC Infectious Diseases*, 13(1):110, 2013.

Campbell F., Didelot X., Fitzjohn R., Ferguson N., Cori A., and Jombart T. outbreaker2: a modular platform for outbreak reconstruction. *BMC Bioinformatics*, 19(11):363, 2018.

Campbell F., Cori A., Ferguson N., and Jombart T. Bayesian inference of transmission chains using timing of symptoms, pathogen genomes and contact data. *PLOS Computational Biology*, 15(3), 2019.

Carriço J. A., Rossi M., Moran-Gilad J., Domselaar G. V., and Ramirez M. A primer on microbial bioinformatics for nonbioinformaticians. *Clinical Microbiology and Infection*, 24 (4):342–349, 2018.

Casali N., Nikolayevskyy V., Balabanova Y., Harris S. R., Ignatyeva O., Kontsevaya I., Corander J., Bryant J., Parkhill J., Nejentsev S., et al. Evolution and transmission of drug-resistant tuberculosis in a russian population. *Nature Genetics*, 46(3):279, 2014.

Caws M., Thwaites G., Dunstan S., Hawn T. R., Lan N. T. N., Thuong N. T. T., Stepniewska K., Huyen M. N. T., Bang N. D., Loc T. H., et al. The influence of host and bacterial genotype on the development of disseminated disease with mycobacterium tuberculosis. *PLOS Pathogens*, 4(3):e1000034, 2008.

Chee M., Yang R., Hubbell E., Berno A., Huang X. C., Stern D., Winkler J., Lockhart D. J., Morris M. S., and Fodor S. P. A. Accessing Genetic Information with High-Density DNA Arrays. *Science*, 274(5287):610–614, 1996.

Cock P. J., Antao T., Chang J. T., Chapman B. A., Cox C. J., Dalke A., Friedberg I., Hamelryck T., Kauff F., Wilczynski B., and De Hoon M. J. L. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25 (11):1422–1423, 2009.

Cohen K. A., Manson A. L., Desjardins C. A., Abeel T., and Earl A. M. Deciphering drug resistance in Mycobacterium tuberculosis using whole-genome sequencing: progress, promise, and challenges. *Genome Medicine*, 11(1):45, 2019.

Coll F., McNerney R., Guerra-Assunção J. A., Glynn J. R., Perdigão J., Viveiros M., Portugal I., Pain A., Martin N., and Clark T. G. A robust SNP barcode for typing *Mycobacterium tuberculosis* complex strains. *Nature Communications*, 5:4812, 2014.

Coll F., McNerney R., Preston M. D., Guerra-Assunção J. A., Warry A., Hill-Cawthorne G., Mallard K., Nair M., Miranda A., Alves A., Perdigão J., Viveiros M., Portugal I., Hasan Z., Hasan R., Glynn J. R., Martin N., Pain A., and Clark T. G. Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. *Genome Medicine*, 7(1):51, 2015.

Comas I., Homolka S., Niemann S., and Gagneux S. Genotyping of Genetically Monomorphic Bacteria: DNA Sequencing in Mycobacterium tuberculosis Highlights the Limitations of Current Methodologies. *PLOS ONE*, 4(11):e7815, 2009.

Comas I., Chakravartti J., Small P. M., Galagan J., Niemann S., Kremer K., Ernst J. D., and Gagneux S. Human T cell epitopes of Mycobacterium tuberculosis are evolutionarily hyperconserved. *Nature Genetics*, 42(6):498–503, 2010.

Computational Pan-Genomics Consortium. Computational pan-genomics: status, promises and challenges. *Briefings in Bioinformatics*, 19:118—-135, 2018.

Couvin D., David A., Zozio T., and Rastogi N. Macro-geographical specificities of the prevailing tuberculosis epidemic as seen through SITVIT2, an updated version of the Mycobacterium tuberculosis genotyping database. *Infection, Genetics and Evolution*, 72: 31–43, 2019.

Critical Path Institute. Global Health Partners Accelerate Uptake of Genetic Sequencing for Surveillance And Diagnosis Of Drug-Resistant Tuberculosis. `https://c-path.org/global-health-partners-accelerate-uptake-of-genetic-sequencing-for-surveillance-and-diagnosis-of-drug-resistant-tuberculosis/`, 2018. Accessed: 2019-09-04.

CRyPTIC Consortium and the 100,000 Genomes Project. Prediction of susceptibility to first-line tuberculosis drugs by dna sequencing. *New England Journal of Medicine*, 379 (15):1403–1415, 2018.

Darling A. E. the darling lab | computational (meta)genomics. `http://darlinglab.org/mauve/user-guide/files.html#the-alignment-file-and-the-xmfa-file-format`, 2015. Accessed: 2017-07-20.

Darling A. E., Mau B., and Perna N. T. progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement. *PLOS ONE*, 5(6):e11147, 2010.

Dawson E. T. svaha - generate variation graphs for structural variants. `https://github.com/edawson/svaha`, 2016. Accessed: 2017-23-01.

DePristo M. A., Banks E., Poplin R., Garimella K. V., Maguire J. R., Hartl C., Philippakis A. A., Del Angel G., Rivas M. A., Hanna M., et al. A framework for variation discovery and genotyping using next-generation dna sequencing data. *Nature Genetics*, 43(5):491, 2011.

Dheda K., Gumbo T., Gandhi N. R., Murray M., Theron G., Udwadia Z., Migliori G. B., and Warren R. Global control of tuberculosis: from extensively drug-resistant to untreatable tuberculosis. *The Lancet Respiratory medicine*, 2(4):321–338, 2014.

Di Tommaso P., Moretti S., Xenarios I., Orobitg M., Montanyola A., Chang J.-M., Taly J.-F., and Notredame C. T-coffee: a web server for the multiple sequence alignment of protein and rna sequences using structural information and homology extension. *Nucleic Acids Research*, 39(suppl_2):W13–W17, 2011.

Didelot X., Fraser C., Gardy J., and Colijn C. Genomic Infectious Disease Epidemiology in Partially Sampled and Ongoing Outbreaks. *Molecular Biology and Evolution*, 34(4): 997–1007, 2017.

Dilthey A., Cox C., Iqbal Z., Nelson M. R., and McVean G. Improved genome inference in the MHC using a population reference graph. *Nature Genetics*, 47(6):682–688, 2015.

Doyle R. M., Burgess C., Williams R., Gorton R., Booth H., Brown J., Bryant J. M., Chan J., Creer D., Holdstock J., Kunst H., Lozewicz S., Platt G., Romero E. Y., Speight G., Tiberi S., Abubakar I., Lipman M., McHugh T. D., and Breuer J. Direct Whole-Genome Sequencing of Sputum Accurately Identifies Drug-Resistant Mycobacterium tuberculosis Faster than MGIT Culture Sequencing. *Journal of Clinical Microbiology*, 56(8), 2018.

Dutheil J. Y., Gaillard S., and Stukenbrock E. H. Maffilter: a highly flexible and extensible multiple genome alignment files processor. *BMC genomics*, 15(1):53, 2014.

Earl D., Paten B., and Diekhans M. evolverSimControl. `https://github.com/dentearl/evolverSimControl`, 2012. Accessed: 2017-24-04.

Earl D., Nguyen N., Hickey G., Harris R. S., Fitzgerald S., Beal K., Seledtsov I., Molodtsov V., Raney B. J., Clawson H., Jaebum K., Kemena C., Chang J.-M., Erb I., Alexander P., Hou M., Herrero J., Kent W. J., Solovyev V., E. D. A., Ma J., Notredame C., Brudno M., Dubchak I., Haussler D., and Paten B. Alignathon: a competitive assessment of whole-genome alignment methods. *Genome Research*, 24(12):2077–2089, 2014.

ECDC. Molecular typing for surveillance of multidrug-resistant tuberculosis in the EU/EEA. `http://ecdc.europa.eu/en/publications/Publications/MDR-TB-molecular-typing-surveillance-mar-2017.pdf`, 2017. Accessed: 2019-09-10.

Edgar R. C., Asimenos G., Batzoglou S., and Sidow A. EVOLVER. `http://www.drive5.com/evolver`, 2006. Accessed: 2017-24-04.

Ei P. W., Aung W. W., Lee J. S., Choi G.-E., and Chang C. L. Molecular Strain Typing of Mycobacterium tuberculosis: a Review of Frequently Used Methods. *Journal of Korean Medical Science*, 31(11):1673–1683, 2016.

Ernst C. and Rahmann S. PanCake: A Data Structure for Pangenomes. In Beißbarth T., Kollmar M., Leha A., Morgenstern B., Schultz A.-K., Waack S., and Wingender E., editors, *German Conference on Bioinformatics 2013*, volume 34 of *OASICS*, pages 35–45, Dagstuhl, Germany, 2013.

European Centre for Disease Prevention and Control/WHO Regional Office for Europe. Tuberculosis surveillance and monitoring in europe 2018. `https://ecdc.europa.eu/en/publications-data/tuberculosis-surveillance-and-monitoring-europe-2018`, 2018. Accessed: 2019-08-09.

Feuerriegel S., Köser C. U., and Niemann S. Phylogenetic polymorphisms in antibiotic resistance genes of the Mycobacterium tuberculosis complex. *The Journal of Antimicrobial Chemotherapy*, 69(5):1205–1210, 2014.

Feuerriegel S., Schleusener V., Beckert P., Kohl T. A., Miotto P., Cirillo D. M., Cabibbe A. M., Niemann S., and Fellenberg K. PhyResSE: a Web Tool Delineating Mycobacterium tuberculosis Antibiotic Resistance and Lineage from Whole-Genome Sequencing Data. *Journal of Clinical Microbiology*, 53(6):1908–1914, 2015.

Fiebig L., Kohl T. A., Popovici O., Mühlenfeld M., Indra A., Homorodean D., Chiotan D., Richter E., Rüsch-Gerdes S., Schmidgruber B., et al. A joint cross-border investigation of a cluster of multidrug-resistant tuberculosis in austria, romania and germany in 2014 using classic, genotyping and whole genome sequencing methods: lessons learnt. *Eurosurveillance*, 22(2), 2017.

Firestone S. M., Hayama Y., Bradhurst R., Yamamoto T., Tsutsui T., and Stevenson M. A. Reconstructing foot-and-mouth disease outbreaks: a methods comparison of transmission network models. *Scientific Reports*, 9(1), 2019.

Ford C., Yusim K., Ioerger T., Feng S., Chase M., Greene M., Korber B., and Fortune S. Mycobacterium tuberculosis–heterogeneity revealed through whole genome sequencing. *Tuberculosis*, 92(3):194–201, 2012.

Ford C. B., Lin P. L., Chase M. R., Shah R. R., Iartchouk O., Galagan J., Mohaideen N., Ioerger T. R., Sacchettini J. C., Lipsitch M., et al. Use of whole genome sequencing to estimate the mutation rate of mycobacterium tuberculosis during latent infection. *Nature Genetics*, 43(5):482, 2011.

Ford C. B., Shah R. R., Maeda M. K., Gagneux S., Murray M. B., Cohen T., Johnston J. C., Gardy J., Lipsitch M., and Fortune S. M. Mycobacterium tuberculosis mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis. *Nature Genetics*, 45(7):784, 2013.

Frith J. et al. History of tuberculosis. part 1-phthisis, consumption and the white plague. *Journal of Military and Veterans Health*, 22(2):29, 2014.

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

Gardy J. L., Johnston J. C., Sui S. J. H., Cook V. J., Shah L., Brodkin E., Rempel S., Moore R., Zhao Y., Holt R., et al. Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. *New England Journal of Medicine*, 364(8):730–739, 2011.

Garrison E., Sirén J., Novak A. M., Hickey G., Eizenga J. M., Dawson E. T., Jones W., Garg S., Markello C., Lin M. F., Paten B., and Durbin R. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature Biotechnology*, 36:875, 2018.

Ghodbane R., Raoult D., and Drancourt M. Dramatic reduction of culture time of Mycobacterium tuberculosis. *Scientific Reports*, 4:4236, 2014.

Gilbert D. G. Phylodendron. `http://iubio.bio.indiana.edu/treeapp/treeprint-form.html`, 1999. Accessed: 2017-24-04.

Guerra-Assunção J., Crampin A., Houben R., Mzembe T., Mallard K., Coll F., Khan P., Banda L., Chiwaya A., Pereira R., et al. Large-scale whole genome sequencing of m. tuberculosis provides insights into transmission in a high prevalence area. *eLife*, 4, 2015.

Gurjav U., Outhred A. C., Jelfs P., McCallum N., Wang Q., Hill-Cawthorne G. A., Marais B. J., and Sintchenko V. Whole genome sequencing demonstrates limited transmission within identified mycobacterium tuberculosis clusters in new south wales, australia. *PLOS ONE*, 11(10):e0163612, 2016.

Hatherell H., Colijn C., Stagg H. R., Jackson C., Winter J. R., and Abubakar I. Interpreting whole genome sequencing for investigating tuberculosis transmission: a systematic review. *BMC Medicine*, 14(1), 2016.

Herbig A., Jäger G., Battke F., and Nieselt K. GenomeRing: alignment visualization based on SuperGenome coordinates. *Bioinformatics*, 28(12):i7–i15, 2012.

Homolka S., Projahn M., Feuerriegel S., Ubben T., Diel R., Nübel U., and Niemann S. High Resolution Discrimination of Clinical Mycobacterium tuberculosis Complex Strains Based on Single Nucleotide Polymorphisms. *PLOS ONE*, 7(7):e39855, 2012.

Huang L., Popic V., and Batzoglou S. Short read alignment with populations of genomes. *Bioinformatics*, 29(13):i361–i370, 2013.

Hubisz M. J., Pollard K. S., and Siepel A. Phast and rphast: phylogenetic analysis with space/time models. *Briefings in Bioinformatics*, 12(1):41–51, 2010.

Iqbal Z., Caccamo M., Turner I., Flicek P., and McVean G. De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nature Genetics*, 44(2):226–232, 2012.

Iqbal Z., Turner I., and McVean G. High-throughput microbial population genomics using the Cortex variation assembler. *Bioinformatics*, 29(2):275–276, 2013.

Jagielski T., van Ingen J., Rastogi N., Dziadek J., Mazur P. K., and Bielecki J. Current Methods in the Molecular Typing of Mycobacterium tuberculosis and Other Mycobacteria. *BioMed Research International*, 2014.

Jandrasits C. and Renard B. Y. Inferring transmission chains of tuberculosis from genetic and epidemiological data. manuscript in preparation.

Jandrasits C., Kröger S., Haas W., and Renard B. Y. Computational Pan-genome Mapping and pairwise SNP-distance improve Detection of Mycobacterium tuberculosis Transmission Clusters. *PLOS Computational Biology*. (in revision).

Jandrasits C., Dabrowski P. W., Fuchs S., and Renard B. Y. seq-seq-pan: Building a computational pan-genome data structure on whole genome alignment. *BMC Genomics*, 19(1):47, 2018.

Jobin M., Schurz H., and Henn B. M. IMPUTOR: Phylogenetically Aware Software for Imputation of Errors in Next-Generation Sequencing. *Genome Biology and Evolution*, 10 (5):1248–1254, 2018.

Jombart T., Eggo R. M., Dodd P. J., and Balloux F. Reconstructing disease outbreaks from genetic data: a graph approach. *Heredity*, 106(2):383–390, 2011.

Jombart T., Cori A., Didelot X., Cauchemez S., Fraser C., and Ferguson N. Bayesian Reconstruction of Disease Outbreaks by Combining Epidemiologic and Genomic Data. *PLOS Computational Biology*, 10(1):e1003457, 2014.

Kantorovitz M. R., Robinson G. E., and Sinha S. A statistical method for alignment-free comparison of regulatory sequences. *Bioinformatics*, 23(13):i249–i255, 2007.

Kato-Maeda M., Ho C., Passarelli B., Banaei N., Grinsdale J., Flores L., Anderson J., Murray M., Rose G., Kawamura L. M., et al. Use of whole genome sequencing to determine the microevolution of mycobacterium tuberculosis during an outbreak. *PLOS ONE*, 8(3): e58235, 2013.

Kearse M., Moir R., Wilson A., Stones-Havas S., Cheung M., Sturrock S., Buxton S., Cooper A., Markowitz S., Duran C., Thierer T., Ashton B., Mentjies P., and Drummond A. Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, 28(12):1647–1649, 2012.

Kent W. J. Blat—the blast-like alignment tool. *Genome Research*, 12(4):656–664, 2002.

Kim J. and Ma J. Psar-align: improving multiple sequence alignment using probabilistic sampling. *Bioinformatics*, 30(7):1010–1012, 2013.

Klinkenberg D., Backer J. A., Didelot X., Colijn C., and Wallinga J. Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. *PLOS Computational Biology*, 13(5):1–32, 2017.

Koch A., Cox H., and Mizrahi V. Drug-resistant tuberculosis: challenges and opportunities for diagnosis and treatment. *Current Opinion in Pharmacology*, 42:7–15, 2018.

Kohl T. A., Diel R., Harmsen D., Rothgänger J., Walter K. M., Merker M., Weniger T., and Niemann S. Whole-genome-based mycobacterium tuberculosis surveillance: a standardized, portable, and expandable approach. *Journal of Clinical Microbiology*, 52(7): 2479–2486, 2014.

Kohl T. A., Harmsen D., Rothgänger J., Walker T., Diel R., and Niemann S. Harmonized genome wide typing of tubercle bacilli using a web-based gene-by-gene nomenclature system. *EBioMedicine*, 34:131 – 138, 2018a.

Kohl T. A., Utpatel C., Schleusener V., Filippo M. R. D., Beckert P., Cirillo D. M., and Niemann S. MTBseq: a comprehensive pipeline for whole genome sequence analysis of Mycobacterium tuberculosis complex isolates. *PeerJ*, 6:e5895, 2018b.

Köster J. and Rahmann S. Snakemake - a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522, 2012.

Kurtz S., Phillippy A., Delcher A. L., Smoot M., Shumway M., Antonescu C., and Salzberg S. L. Versatile and open software for comparing large genomes. *Genome Biology*, 5(2):R12, 2004.

Land M., Hauser L., Jun S.-R., Nookaew I., Leuze M. R., Ahn T.-H., Karpinets T., Lund O., Kora G., Wassenaar T., et al. Insights from 20 years of bacterial genome sequencing. *Functional and Integrative Genomics*, 15(2):141–161, 2015.

Lee R. S. and Behr M. A. Does choice matter? reference-based alignment for molecular epidemiology of tuberculosis. *Journal of Clinical Microbiology*, pages JCM–00364, 2016.

Leggett R. M. and MacLean D. Reference-free SNP detection: dealing with the data deluge. *BMC Genomics*, 15(4):S10, 2014.

Leinonen R., Sugawara H., Shumway M., and on behalf of the International Nucleotide Sequence Database Collaboration. The Sequence Read Archive. *Nucleic Acids Research*, 39:D19–D21, 2010.

Lew J. M., Kapopoulou A., Jones L. M., and Cole S. T. Tuberculist – 10 years after. *Tuberculosis*, 91(1):1–7, 2011.

Li H. Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. *arXiv:1303.3997v1 [q-bio.GN]*, 2013.

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

Lieberman T. D., Wilson D., Misra R., Xiong L. L., Moodley P., Cohen T., and Kishony R. Genomic diversity in autopsy samples reveals within-host dissemination of HIV-associated Mycobacterium tuberculosis. *Nature Medicine*, 22:1470, 2016.

Lipworth S., Jajou R., Neeling A. d., Bradley P., Hoek W. v. d., Maphalala G., Bonnet M., Sanchez-Padilla E., Diel R., Niemann S., Iqbal Z., Smith G., Peto T., Crook D., Walker T., and Soolingen D. v. SNP-IT Tool for Identifying Subspecies and Associated Lineages of Mycobacterium tuberculosis Complex. *Emerging Infectious Diseases*, 25(3), 2019.

Liu L., Li Y., Li S., Hu N., He Y., Pong R., Lin D., Lu L., and Law M. Comparison of Next-Generation Sequencing Systems. *Journal of Biomedicine and Biotechnology*, 2012, 2012.

Lopez B., Aguilar D., Orozco H., Burger M., Espitia C., Ritacco V., Barrera L., Kremer K., HERNANDEZ-PANDO R., Huygen K., et al. A marked difference in pathogenesis and immune response induced by different mycobacterium tuberculosis genotypes. *Clinical & Experimental Immunology*, 133(1):30–37, 2003.

Magoč T. and Salzberg S. L. Flash: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, 27(21):2957–2963, 2011.

Maiden M. C., Van Rensburg M. J. J., Bray J. E., Earle S. G., Ford S. A., Jolley K. A., and McCarthy N. D. Mlst revisited: the gene-by-gene approach to bacterial genomics. *Nature Reviews Microbiology*, 11(10):728, 2013.

Maio N. D., Wu C.-H., and Wilson D. J. SCOTTI: Efficient Reconstruction of Transmission within Outbreaks with the Structured Coalescent. *PLOS Computational Biology*, 12(9): e1005130, 2016.

Maio N. D., Worby C. J., Wilson D. J., and Stoesser N. Bayesian reconstruction of transmission within outbreaks using genomic variants. *PLOS Computational Biology*, 14(4): e1006117, 2018.

Manson A. L., Cohen K. A., Abeel T., Desjardins C. A., Armstrong D. T., Barry III C. E., Brand J., TBResist Global Genome Consortium, Brand J., Jureen P., Malinga L., Nordenberg D., Velayati A. A., Cassell G. H., Farnia P., Homorodean D., Van der Walt M., Hoffner S., Chapman S. B., Cho S.-N., Gabrielian A., Gomez J., Jodals A. M., Joloba M., Jureen P., Lee J. S., Malinga L., Maiga M., Nordenberg D., Noroc E., Romancenco E., Salazar A., Ssengooba W., Velayati A. A., Winglee K., Zalutskaya A., Via L. E., Cassell G. H., Dorman S. E., Ellner J., Farnia P., Galagan J. E., Rosenthal A., Crudu V., Homorodean D., Hsueh P.-R., Narayanan S., Pym A. S., Skrahina A., Swaminathan S., Van der Walt M., Alland D., Bishai W. R., Cohen T., Hoffner S., Birren B. W., and Earl A. M. Genomic analysis of globally diverse Mycobacterium tuberculosis strains provides insights into the emergence and spread of multidrug resistance. *Nature Genetics*, 49:395, 2017.

Marcus S., Lee H., and Schatz M. C. SplitMEM: a graphical algorithm for pan-genome analysis with suffix skips. *Bioinformatics*, 30(24):3476–3483, 2014.

Mardis E. R. The impact of next-generation sequencing technology on genetics. *Trends in Genetics*, 24(3):133–141, 2008.

Martin M. A., Lee R. S., Cowley L. A., Gardy J. L., and Hanage W. P. Within-host Mycobacterium tuberculosis diversity and its utility for inferences of transmission. *Microbial Genomics*, 4(10):e000217, 2018.

Matteelli A., Rendon A., Tiberi S., Al-Abri S., Voniatis C., Carvalho A. C. C., Centis R., D'Ambrosio L., Visca D., Spanevello A., and Battista Migliori G. Tuberculosis elimination: where are we now? *European Respiratory Review*, 27(148), 2018.

Mazariegos-Canellas O., Do T., Peto T., Eyre D. W., Underwood A., Crook D., and Wyllie D. H. Bugmat and findneighbour: command line and server applications for investigating bacterial relatedness. *BMC Bioinformatics*, 18(1):477, 2017.

Meehan C. J., Moris P., Kohl T. A., Pečerska J., Akter S., Merker M., Utpatel C., Beckert P., Gehre F., Lempens P., Stadler T., Kaswa M. K., Kühnert D., Niemann S., and de Jong B. C. The relationship between transmission time and clustering methods in Mycobacterium tuberculosis epidemiology. *EBioMedicine*, 37:410–416, 2018.

Meehan C. J., Goig G. A., Kohl T. A., Verboven L., Dippenaar A., Ezewudo M., Farhat M. R., Guthrie J. L., Laukens K., Miotto P., Ofori-Anyinam B., Dreyer V., Supply P., Suresh A., Utpatel C., van Soolingen D., Zhou Y., Ashton P. M., Brites D., Cabibbe A. M., de Jong B. C., de Vos M., Menardo F., Gagneux S., Gao Q., Heupink T. H., Liu Q., Loiseau C., Rigouts L., Rodwell T. C., Tagliani E., Walker T. M., Warren R. M., Zhao Y., Zignol M., Schito M., Gardy J., Cirillo D. M., Niemann S., Comas I., and Van Rie A. Whole genome sequencing of Mycobacterium tuberculosis: current standards and open issues. *Nature Reviews Microbiology*, 17:533–545, 2019.

Mehaffy C., Guthrie J. L., Alexander D. C., Stuart R., Rea E., and Jamieson F. B. Marked Microevolution of a Unique Mycobacterium tuberculosis Strain in 17 Years of Ongoing Transmission in a High Risk Population. *PLOS ONE*, 9(11), 2014.

Merker M., Blin C., Mona S., Duforet-Frebourg N., Lecher S., Willery E., Blum M. G. B., Rüsch-Gerdes S., Mokrousov I., Aleksic E., Allix-Béguec C., Antierens A., Augustynowicz-Kopeć E., Ballif M., Barletta F., Beck H. P., Barry III C. E., Bonnet M., Borroni E., Campos-Herrero I., Cirillo D., Cox H., Crowe S., Crudu V., Diel R., Drobniewski F., Fauville-Dufaux M., Gagneux S., Ghebremichael S., Hanekom M., Hoffner S., Jiao W.-w., Kalon S., Kohl T. A., Kontsevaya I., Lillebæk T., Maeda S., Nikolayevskyy V., Rasmussen M., Rastogi N., Samper S., Sanchez-Padilla E., Savic B., Shamputa I. C., Shen A., Sng L.-H., Stakenas P., Toit K., Varaine F., Vukovic D., Wahl C., Warren R., Supply P., Niemann S., and Wirth T. Evolutionary history and global spread of the Mycobacterium tuberculosis Beijing lineage. *Nature Genetics*, 47:242, 2015.

Metzker M. L. Sequencing technologies — the next generation. *Nature Reviews Genetics*, 11(1):31–46, 2010.

Minkin I., Pham S., and Medvedev P. TwoPaCo: An efficient algorithm to build the compacted de Bruijn graph from many complete genomes. *Bioinformatics*, 33:4024–4032, 2016.

Miotto P., Tessema B., Tagliani E., Chindelevitch L., Starks A. M., Emerson C., Hanna D., Kim P. S., Liwski R., Zignol M., Gilpin C., Niemann S., Denkinger C. M., Fleming J., Warren R. M., Crook D., Posey J., Gagneux S., Hoffner S., Rodrigues C., Comas I., Engelthaler D. M., Murray M., Alland D., Rigouts L., Lange C., Dheda K., Hasan R., Ranganathan U. D. K., McNerney R., Ezewudo M., Cirillo D. M., Schito M., Koser C. U., and Rodwell T. C. A standardised method for interpreting the association between mutations and phenotypic drug resistance in Mycobacterium tuberculosis. *European Respiratory Journal*, 50(6), 2017.

Muir P., Li S., Lou S., Wang D., Spakowicz D. J., Salichos L., Zhang J., Weinstock G. M., Isaacs F., Rozowsky J., and Gerstein M. The real cost of sequencing: scaling computation to keep pace with data generation. *Genome Biology*, 17:53, 2016.

Murray M. and Alland D. Methodological problems in the molecular epidemiology of tuberculosis. *American Journal of Epidemiology*, 155(6):565–71, 2002.

Nakato R. and Gotoh O. Cgaln: fast and space-efficient whole-genome alignment. *BMC Bioinformatics*, 11(1):224, 2010.

Ngo T.-M. and Teo Y.-Y. Genomic prediction of tuberculosis drug-resistance: benchmarking existing databases and prediction algorithms. *BMC Bioinformatics*, 20(1):68, 2019.

Nielsen R., Paul J. S., Albrechtsen A., and Song Y. S. Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, 12(6):443–451, 2011.

Niemann S., Köser C. U., Gagneux S., Plinke C., Homolka S., Bignell H., Carter R. J., Cheetham R. K., Cox A., Gormley N. A., et al. Genomic diversity among drug sensitive and multidrug resistant isolates of mycobacterium tuberculosis with identical dna fingerprints. *PLOS ONE*, 4(10):e7407, 2009.

Nikolayevskyy V., Kranzer K., Niemann S., and Drobniewski F. Whole genome sequencing of mycobacterium tuberculosis for detection of recent transmission and tracing outbreaks: a systematic review. *Tuberculosis*, 98:77–85, 2016.

Nimmo C., Shaw L. P., Doyle R., Williams R., Brien K., Burgess C., Breuer J., Balloux F., and Pym A. S. Whole genome sequencing Mycobacterium tuberculosis directly from sputum identifies more genetic diversity than sequencing from culture. *BMC Genomics*, 20(1):389, 2019.

Odone A., Tillmann T., Sandgren A., Williams G., Rechel B., Ingleby D., Noori T., Mladovsky P., and McKee M. Tuberculosis among migrant populations in the European Union and the European Economic Area. *European Journal of Public Health*, 25(3): 506–12, 2015.

Ohta T., Nakazato T., and Bono H. Calculating the quality of public high-throughput sequencing data to obtain a suitable subset for reanalysis from the Sequence Read Archive. *Gigascience*, 6(6):1–8, 2017.

Pankhurst L. J., del Ojo Elias C., Votintseva A. A., Walker T. M., Cole K., Davies J., Fermont J. M., Gascoyne-Binzi D. M., Kohl T. A., Kong C., et al. Rapid, comprehensive, and affordable mycobacterial diagnosis with whole-genome sequencing: a prospective study. *The Lancet Respiratory Medicine*, 4(1):49–58, 2016.

Paten B., Earl D., Nguyen N., Diekhans M., Zerbino D., and Haussler D. Cactus: Algorithms for genome multiple sequence alignment. *Genome Research*, 21(9):1512, 2011.

Pérez-Lago L., Comas I., Navarro Y., González-Candelas F., Herranz M., Bouza E., and García-de Viedma D. Whole genome sequencing analysis of intrapatient microevolution in mycobacterium tuberculosis: potential impact on the inference of tuberculosis transmission. *The Journal of Infectious Diseases*, 209(1):98–108, 2013.

Periwal V., Patowary A., Vellarikkal S. K., Gupta A., Singh M., Mittal A., Jeyapaul S., Chauhan R. K., Singh A. V., Singh P. K., et al. Comparative whole-genome analysis of clinical isolates reveals characteristic architecture of mycobacterium tuberculosis pangenome. *PLOS ONE*, 10(4):e0122979, 2015.

Peterlongo P., Riou C., Drezen E., and Lemaitre C. DiscoSnp++: de novo detection of small variants from raw unassembled read set(s). *bioRxiv*, page 209965, 2017.

Poliakov A., Foong J., Brudno M., and Dubchak I. Genomevista—an integrated software package for whole-genome alignment and visualization. *Bioinformatics*, 30(18):2654–2655, 2014.

Quainoo S., Coolen J. P. M., Hijum S. A. F. T. v., Huynen M. A., Melchers W. J. G., Schaik W. v., and Wertheim H. F. L. Whole-Genome Sequencing of Bacterial Pathogens: the Future of Nosocomial Outbreak Analysis. *Clinical Microbiology Reviews*, 30(4):1015–1063, 2017.

Quinlan A. R. and Hall I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010.

R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2014. URL `http://www.R-project.org/`.

Rahn R., Weese D., and Reinert K. Journaled string tree - a scalable data structure for analyzing thousands of similar genomes on your laptop. *Bioinformatics*, 30(24):3499–3505, 2014.

Rand K. D., Grytten I., Nederbragt A. J., Storvik G. O., Glad I. K., and Sandve G. K. Coordinates and intervals in graph-based reference genomes. *BMC Bioinformatics*, 18(1): 263, 2017.

Ratan A., Zhang Y., Hayes V. M., Schuster S. C., and Miller W. Calling SNPs without a reference sequence. *BMC Bioinformatics*, 11(1):130, 2010.

Roetzer A., Schuback S., Diel R., Gasau F., Ubben T., di Nauta A., Richter E., Rusch-Gerdes S., and Niemann S. Evaluation of Mycobacterium tuberculosis typing methods in a 4-year study in Schleswig-Holstein, Northern Germany. *Journal of Clinical Microbiology*, 49(12):4173–8, 2011.

Roetzer A., Diel R., Kohl T. A., Rückert C., Nübel U., Blom J., Wirth T., Jaenicke S., Schuback S., Rüsch-Gerdes S., et al. Whole genome sequencing versus traditional genotyping for investigation of a mycobacterium tuberculosis outbreak: a longitudinal molecular epidemiological study. *PLOS Medicine*, 10(2):e1001387, 2013.

Rosenthal A., Gabrielian A., Engle E., Hurt D. E., Alexandru S., Crudu V., Sergueev E., Kirichenko V., Lapitskii V., Snezhko E., Kovalev V., Astrovko A., Skrahina A., Taaffe J., Harris M., Long A., Wollenberg K., Akhundova I., Ismayilova S., Skrahin A., Mammadbayov E., Gadirova H., Abuzarov R., Seyfaddinova M., Avaliani Z., Strambu I., Zaharia D., Muntean A., Ghita E., Bogdan M., Mindru R., Spinu V., Sora A., Ene C., Vashakidze S., Shubladze N., Nanava U., Tuzikov A., and Tartakovsky M. The TB Portals: an Open-Access, Web-Based Platform for Global Drug-Resistant-Tuberculosis Data Sharing and Analysis. *Journal of Clinical Microbiology*, 55(11):3267–3282, 2017.

Saelens J. W., Viswanathan G., and Tobin D. M. Mycobacterial Evolution Intersects With Host Tolerance. *Frontiers in Immunology*, 10, 2019.

Salmonière Y.-O. L. G. d. l., Kim C. C., Tsolaki A. G., Pym A. S., Siegrist M. S., and Small P. M. High-Throughput Method for Detecting Genomic-Deletion Polymorphisms. *Journal of Clinical Microbiology*, 42(7):2913–2918, 2004.

Sanchini A., Andrés M., Fiebig L., Albrecht S., Hauer B., and Haas W. Assessment of the use and need for an integrated molecular surveillance of tuberculosis: an online survey in Germany. *BMC Public Health*, 19(1):321, 2019.

Sanchini* A., Jandrasits* C., Tembrockhaus J., Kohl T. A., Utpatel C., Maurer F., Niemann S., Haas W., Renard B. Y., and Kröger S. Improving tuberculosis surveillance by detecting international transmission using publicly available whole genome sequencing data. submission in preparation.

Sandgren A., Strong M., Muthukrishnan P., Weiner B. K., Church G. M., and Murray M. B. Tuberculosis Drug Resistance Mutation Database. *PLOS Medicine*, 6(2):e1000002, 2009.

Sanger F. and Coulson A. R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology*, 94(3):441–448, 1975.

Sanger F., Nicklen S., and Coulson A. R. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12):5463–5467, 1977.

Schleusener V., Köser C. U., Beckert P., Niemann S., and Feuerriegel S. *Mycobacterium tuberculosis* resistance prediction and lineage classification from genome sequencing: comparison of automated analysis tools. *Scientific Reports*, 7:46327, 2017.

Schneeberger K., Hagmann J., Ossowski S., Warthmann N., Gesing S., Kohlbacher O., and Weigel D. Simultaneous alignment of short reads against multiple genomes. *Genome Biology*, 10(9):R98, 2009.

Schwartzman K. and Menzies D. How long are TB patients infectious? *CMAJ: Canadian Medical Association Journal*, 163(2):157–158, 2000.

Schön T., Miotto P., Köser C. U., Viveiros M., Böttger E., and Cambau E. Mycobacterium tuberculosis drug-resistance testing: challenges, recent developments and perspectives. *Clinical Microbiology and Infection*, 23(3):154–160, 2017.

Schürch A. C., Kremer K., Daviena O., Kiers A., Boeree M. J., Siezen R. J., and Soolingen D. v. High-Resolution Typing by Integration of Genome Sequencing Data in a Large Tuberculosis Cluster. *Journal of Clinical Microbiology*, 48(9):3403–3406, 2010.

Shaik F., Bezawada S., and Goveas N. Cyspanningtree: Minimal spanning tree computation in cytoscape. *F1000Research*, 4, 2015.

Shannon P., Markiel A., Ozier O., Baliga N. S., Wang J. T., Ramage D., Amin N., Schwikowski B., and Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11):2498–2504, 2003.

Shendure J. and Ji H. Next-generation DNA sequencing. *Nature Biotechnology*, 26(10): 1135–1145, 2008.

Shih A. C.-C., Lee D., Lin L., Peng C.-L., Chen S.-H., Wu Y.-W., Wong C.-Y., Chou M.-Y., Shiao T.-C., and Hsieh M.-F. Sinicview: a visualization environment for comparisons of multiple nucleotide sequence alignment tools. *BMC Bioinformatics*, 7(1):103, 2006.

Shiloh M. U. Mechanisms of mycobacterial transmission: how does Mycobacterium tuberculosis enter and escape from the human host. *Future Microbiology*, 11(12):1503–1506, 2016.

Shrivastava S. R., Shrivastava P. S., and Ramasamy J. Assessing the utility of contact tracing in reducing the magnitude of tuberculosis. *Infection Ecology & Epidemiology*, 4, 2014.

Sievers F. and Higgins D. G. Clustal omega, accurate alignment of very large numbers of sequences. In Russell D., editor, *Multiple Sequence Alignment Methods*, volume 1079 of *Methods in Molecular Biology (Methods and Protocols)*, pages 105–116, Totowa, NJ, 2014. Humana Press.

Sims D., Sudbery I., Ilott N. E., Heger A., and Ponting C. P. Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics*, 15:121, 2014.

Sirén J. Indexing variation graphs. In Fekete S. and Ramachandran V., editors, *2017 Proceedings of the Ninteenth Workshop on Algorithm Engineering and Experiments (ALENEX)*, pages 13–27, Philadelphia, USA, 2017. SIAM.

Sirén J., Välimäki N., and Mäkinen V. Indexing Graphs for Path Queries with Applications in Genome Research. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 11(2):375–388, 2014.

Starks A. M., Avilés E., Cirillo D. M., Denkinger C. M., Dolinger D. L., Emerson C., Gallarda J., Hanna D., Kim P. S., Liwski R., Miotto P., Schito M., and Zignol M. Collaborative Effort for a Centralized Worldwide Tuberculosis Relational Sequencing Data Platform. *Clinical Infectious Diseases*, 61(suppl_3):S141–S146, 2015.

Steiner A., Stucki D., Coscolla M., Borrell S., and Gagneux S. KvarQ: targeted and direct variant calling from fastq reads of bacterial genomes. *BMC Genomics*, 15(1):881, 2014.

Stephens Z. D., Hudson M. E., Mainzer L. S., Taschuk M., Weber M. R., and Iyer R. K. Simulating next-generation sequencing datasets from empirical mutation and sequencing models. *PLOS ONE*, 11(11):e0167047, 2016.

Stimson J., Gardy J., Mathema B., Crudu V., Cohen T., and Colijn C. Beyond the snp threshold: identifying outbreak clusters using inferred transmissions. *Molecular Biology and Evolution*, 36(3):587–603, 2019.

Stop TB Partnership. Open Letter to the WHO to put TB on the List. `http://www.stoptb.org/news/stories/2017/ns17_014.asp`, 2017. Accessed: 2018-07-19.

Struelens M. and Brisse S. From molecular to genomic epidemiology: transforming surveillance and control of infectious diseases. *Eurosurveillance*, 18(4):20386, 2013.

Stucki D., Malla B., Hostettler S., Huna T., Feldmann J., Yeboah-Manu D., Borrell S., Fenner L., Comas I., Coscollà M., and Gagneux S. Two New Rapid SNP-Typing Methods for Classifying Mycobacterium tuberculosis Complex into the Main Phylogenetic Lineages. *PLOS ONE*, 7(7), 2012.

Stucki D., Ballif M., Egger M., Furrer H., Altpeter E., Battegay M., Droz S., Bruderer T., Coscolla M., Borrell S., et al. Standard genotyping overestimates transmission of mycobacterium tuberculosis among immigrants in a low incidence country. *Journal of Clinical Microbiology*, pages JCM–00126, 2016.

Tagini F. and Greub G. Bacterial genome sequencing in clinical microbiology: a pathogen-oriented review. *European Journal of Clinical Microbiology & Infectious Diseases*, 36(11): 2007–2020, 2017.

Tettelin H., Masignani V., Cieslewicz M. J., Donati C., Medini D., Ward N. L., Angiuoli S. V., Crabtree J., Jones A. L., Durkin A. S., DeBoy R. T., Davidsen T. M., Mora M., Scarselli M., Ros I. M. y., Peterson J. D., Hauser C. R., Sundaram J. P., Nelson W. C., Madupu R., Brinkac L. M., Dodson R. J., Rosovitz M. J., Sullivan S. A., Daugherty S. C., Haft D. H., Selengut J., Gwinn M. L., Zhou L., Zafar N., Khouri H., Radune D., Dimitrov G., Watkins K., O'Connor K. J. B., Smith S., Utterback T. R., White O., Rubens C. E., Grandi G., Madoff L. C., Kasper D. L., Telford J. L., Wessels M. R., Rappuoli R., and Fraser C. M. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences of the United States of America*, 102(39):13950, 2005.

Thierry D., Brisson-Noël A., Vincent-Lévy-Frébault V., Nguyen S., Guesdon J. L., and Gicquel B. Characterization of a Mycobacterium tuberculosis insertion sequence, IS6110, and its application in diagnosis. *Journal of Clinical Microbiology*, 28(12):2668–2673, 1990.

Thwaites G., Caws M., Chau T. T. H., D'Sa A., Lan N. T. N., Huyen M. N. T., Gagneux S., Anh P. T. H., Tho D. Q., Torok E., et al. Relationship between mycobacterium tuberculosis genotype and the clinical phenotype of pulmonary and meningeal tuberculosis. *Journal of Clinical Microbiology*, 46(4):1363–1368, 2008.

UCSC Genome Bioinformatics Group. Frequently asked questions: Data file formats. `https://genome.ucsc.edu/FAQ/FAQformat.html#format5`, 2017. Accessed: 2017-12-29.

Uplekar M., Weil D., Lonnroth K., Jaramillo E., Lienhardt C., Dias H. M., Falzon D., Floyd K., Gargioni G., Getahun H., et al. WHO's new end TB strategy. *The Lancet*, 385 (9979):1799–1801, 2015.

Valenzuela D., Välimäki N., Pitkänen E., and Mäkinen V. On enhancing variation detection through pan-genome indexing. *bioRxiv*, 2015. doi: https://doi.org/10.1101/021444.

van der Werf M. J. and Ködmön C. Whole-Genome Sequencing as Tool for Investigating International Tuberculosis Outbreaks: A Systematic Review. *Frontiers in Public Health*, 7(87), 2019.

Walker T. M., Ip C. L., Harrell R. H., Evans J. T., Kapatai G., Dedicoat M. J., Eyre D. W., Wilson D. J., Hawkey P. M., Crook D. W., et al. Whole-genome sequencing to delineate mycobacterium tuberculosis outbreaks: a retrospective observational study. *The Lancet Infectious Diseases*, 13(2):137–146, 2013.

Walker T. M., Kohl T. A., Omar S. V., Hedge J., Elias C. D. O., Bradley P., Iqbal Z., Feuerriegel S., Niehaus K. E., Wilson D. J., et al. Whole-genome sequencing for prediction of mycobacterium tuberculosis drug susceptibility and resistance: a retrospective cohort study. *The Lancet Infectious Diseases*, 15(10):1193–1202, 2015.

Walker T. M., Merker M., Knoblauch A. M., Helbling P., Schoch O. D., van der Werf M. J., Kranzer K., Fiebig L., Kröger S., Haas W., et al. A cluster of multidrug-resistant mycobacterium tuberculosis among patients arriving in europe from the horn of africa: a molecular epidemiological study. *The Lancet Infectious Diseases*, 18(4):431–440, 2018.

Wetterstrand K. A. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). `www.genome.gov/sequencingcostsdata`, 2019. Accessed: 2019-09-03.

Wheeler N. We are falling behind on TB elimination targets: can whole-genome sequencing guide our efforts? *Thorax*, 74(9):833–834, 2019.

WHO. *Global tuberculosis report 2017*. World Health Organization, 2017.

WHO. The top 10 causes of death. `https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death`, 2018. Accessed: 2019-08-25.

WHO. *Global tuberculosis report 2018*. World Health Organization, 2018a.

WHO. BCG vaccines: WHO position, February 2018. *Weekly Epidemiological Record*, 93 (08), 2018b.

WHO. Ten threats to global health in 2019. `https://www.who.int/emergencies/ten-threats-to-global-health-in-2019`, 2019. Accessed: 2019-09-01.

Wiens K. E., Woyczynski L. P., Ledesma J. R., Ross J. M., Zenteno-Cuevas R., Goodridge A., Ullah I., Mathema B., Siawaya J. F. D., Biehl M. H., et al. Global variation in bacterial strains that cause tuberculosis disease: a systematic review and meta-analysis. *BMC Medicine*, 16(1):196, 2018.

Wirth T., Hildebrand F., Allix-Béguec C., Wölbeling F., Kubica T., Kremer K., van Soolingen D., Rüsch-Gerdes S., Locht C., Brisse S., Meyer A., Supply P., and Niemann S. Origin, spread and demography of the mycobacterium tuberculosis complex. *PLOS Pathogens*, 4 (9):1–10, 2008.

Witney A. A., Gould K. A., Arnold A., Coleman D., Delgado R., Dhillon J., Pond M., Pope C. F., Planche T. D., Stoker N. G., et al. Clinical application of whole genome sequencing to inform treatment for multi-drug resistant tuberculosis cases. *Journal of Clinical Microbiology*, pages JCM–02993, 2015.

Witney A. A., Bateson A. L. E., Jindani A., Phillips P. P. J., Coleman D., Stoker N. G., Butcher P. D., McHugh T. D., and RIFAQUIN Study Team. Use of whole-genome sequencing to distinguish relapse from reinfection in a completed tuberculosis clinical trial. *BMC Medicine*, 15(1):71, 2017.

Wyllie D. H., Davidson J. A., Smith E. G., Rathod P., Crook D. W., Peto T. E. A., Robinson E., Walker T., and Campbell C. A Quantitative Evaluation of MIRU-VNTR Typing Against Whole-Genome Sequencing for Identifying Mycobacterium tuberculosis Transmission: A Prospective Observational Cohort Study. *EBioMedicine*, 34:122–130, 2018.

Xu Y., Liu F., Chen S., Wu J., Hu Y., Zhu B., and Sun Z. In vivo evolution of drug-resistant Mycobacterium tuberculosis in patients during long-term treatment. *BMC Genomics*, 19 (1):640, 2018.

Yang C., Luo T., Shen X., Wu J., Gan M., Xu P., Wu Z., Lin S., Tian J., Liu Q., Yuan Z., Mei J., DeRiemer K., and Gao Q. Transmission of multidrug-resistant mycobacterium tuberculosis in shanghai, china: a retrospective observational study using whole-genome sequencing and epidemiological investigation. *The Lancet Infectious Diseases*, 17(3):275–284, 2017.

Zusammenfassung

Tuberkulose ist eine große Bedrohung für die globale Gesundheit, die jedes Jahr weltweit für über eine Million Todesfälle verantwortlich ist. Es ist wichtig, Übertragungen zu erkennen und zu unterbrechen, um die Ausbreitung dieser Infektionskrankheit zu stoppen. Mit dem zunehmenden Einsatz von NGS hat ihre Anwendung in der Überwachung von *M. tuberculosis* in den letzten Jahren an Bedeutung gewonnen. Das Hauptziel der molekularen Überwachung ist die Identifizierung von Patienten-Patienten-Übertragungen. Distanzberechnung basierend auf Vollgenomsequenzierung sind zu einer integralen Ergänzung von epidemiologischen Untersuchungen von Ausbrüchen von Infektionskrankheiten geworden. Aktuelle Ansätze basieren auf einzelnen Referenzsequenzen und verursachen daher eine Verzerrung in Richtung der gewählten Referenz. Außerdem liefern sie unzureichende Ergebnisse für den Vergleich von Isolaten, da ihre Auflösung zu begrenzt ist.

In dieser Arbeit stelle ich bioinformatische Methoden zur Verbesserung der molekularen Überwachung von *M. tuberculosis* vor. Ich stelle Seq-Seq-Pan vor, ein Framework für das Hinzufügen oder Entfernen neuer Genome aus einem Set alignierter Genome und deren Verwendung zur Konstruktion eines Pan-Genoms. Diese Methode basiert auf der sequentiellen Alignierung der gesamten Genome und ist optimiert für die Erstellung einer linearen Darstellung des Sets von alignierten Genomen, die dessen Verwendung für die Annotation in nachfolgenden Analysen ermöglicht. Ich stelle PANPASCO vor, eine Methode zur Distanzberechnung basierend auf einem Pan-genom, die qualitativ hochwertige Varianten für jedes einzelne Probenpaar vergleicht. Die Methode ist sehr empfindlich gegenüber Unterschieden zwischen Fällen, einschließlich Varianten, die sich in Regionen von linienspezifischen Referenzgenomen befinden. Dieser Ansatz ermöglicht den Vergleich einer großen Anzahl verschiedener Proben. Ich wende diese Methoden auf einen großen internationalen Datensatz von medikamentenresistenten Proben zur Detektion von Übertragungsclustern an. Ich zeige die Verbesserung der Erkennung von internationalen Übertragungen und die Vorteile der Einbeziehung von öffentlich zugänglichen whole genome sequencing von *M. tuberculosis* zur Interpretation der nationalen Überwachungsergebnisse. Darüber hinaus vergleiche ich Übertragungsinferenzmethoden, um eine wichtige Frage bei *M. tuberculosis*-Ausbrüchen zu beantworten: "Wer hat wen angesteckt?"

Eigenständigkeitserklärung

Ich versichere, dass ich die hier vorgelegte Dissertation selbstständig angefertigt habe und die benutzten Quellen und Hilfsmittel vollständig angegeben sind. Ein Promotionsverfahren wurde zu keinem früheren Zeitpunkt an einer anderen in- oder ausländischen Hochschule oder bei einem anderen Fachbereich beantragt. Die Bestimmungen der Promotionsordnung sind mir bekannt.

Christine Jandrasits, Berlin, 18. September 2019