Transmission and evolution of *Mycobacterium tuberculosis* studied by whole genome sequencing and single nucleotide polymorphism-typing

INAUGURALDISSERTATION
zur
Erlangung der Würde eines Doktors der Philosophie

vorgelegt der
Philosophisch-Naturwissenschaftlichen Fakultät
der Universität Basel

von
David Paul Stucki
aus Blumenstein BE

Basel, 2015

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät auf Antrag von Herrn Prof. Dr. Sebastien Gagneux und Herrn Prof. Dr. Stefan Niemann.

Basel, den 24. Juni 2014

Prof. Dr. Jörg Schibler
Dekan

Summary

Each year close to 9 million cases of tuberculosis (TB) are caused by bacteria of the *Mycobacterium tuberculosis* complex (MTBC). The genetic and phenotypic diversity of the pathogen has been underestimated for a long time. Recently, large-scale whole genome sequencing (WGS) became available and has revealed thousands of single nucleotide polymorphisms (SNPs). In contrast to other molecular markers, these SNPs can be used to construct robust phylogenies and to characterize the population structure of MTBC. WGS also shows great promise as the new method of choice for molecular epidemiology of TB. WGS has a higher discriminatory power than classical genotyping methods, and additionally allows for the studying of MTBC micro-evolution during chains of TB transmission.

However, for both research and clinical applications, the costs of WGS are still high and the analytical challenges numerous. Routine application of WGS to large collections of MTBC isolates is not yet feasible. Particularly in settings where the burden of TB is highest, the capacities to generate and analyse WGS data are limited. Hence, innovative approaches are needed to identify the subset of MTBC isolates for which WGS brings the highest level of added value. Due to the strictly clonal nature of the MTBC, we can often use single mutations (i.e. SNPs) to identify isolates of a specific genotype.

In this thesis, we first aimed at developing new, cost-effective strategies for MTBC strain classification based on SNP-typing. We then aimed at applying a combination of SNP-screening and targeted WGS to study the transmission and micro-evolution of MTBC in a local TB outbreak during 20 years. Third, we aimed at using a similar combination of SNP-typing and WGS to infer the global evolutionary scenario of one particular lineage of MTBC, Lineage 4, during historical times.

The first three chapters of this thesis introduce the current knowledge in MTBC research and the objectives of this thesis, and the following four chapters represent primary research work.

In *Chapter 1*, the global burden of TB, the species concept, the genetic diversity and the genotyping methods of MTBC are introduced. *Chapter 2* lists the objectives of this thesis. In *Chapter 3*, we reviewed the nature and consequences of SNPs in MTBC and the

SNP-discovery by WGS. Also, we discuss the potential of SNPs to serve as phylogenetic markers and the use of SNPs for the detection of antibiotic resistance.

In *Chapter 4*, we developed two new SNP-typing assays for cost-effective classification of clinical MTBC isolates into the main phylogenetic lineages. These assays provide a solid basis to study phenotypical and clinically relevant differences between MTBC lineages.

In *Chapter 5*, a new, user-friendly tool for *in silico* SNP-genotyping in MTBC is described. We developed the open-source software `KvarQ` to scan raw WGS reads in `fastq`-format. Drug resistance, phylogenetic, or any other allelic information can be obtained in a matter of minutes. Using a large set of 880 bacterial genome sequences, we showed the high accuracy of the software.

In *Chapter 6*, we aimed at resolving the dynamics of a TB outbreak in the Canton of Bern, Switzerland. This outbreak was caused by an MTBC strain initially described in 1993, and that is still circulating today. Using WGS data generated from three historical "Bernese outbreak" isolates, we developed a *strain-specific* SNP-genotyping assay to screen 1642 isolates from between 1991 and 2011. We identified 68 patients with the same MTBC strain. The majority of patients were from a social "milieu" of homeless and substance abusers. We then applied WGS to all isolates of the "Bernese outbreak" and resolved transmission events to the single patient level. Simultaneously, we also revealed the limits of WGS.

In *Chapter 7* we zoomed out of the *micro*-evolutionary level and looked at the *macro*-evolutionary level. We used a combination of SNP-typing and WGS to track the global dispersal of the "Euro-American" lineage of MTBC (Lineage 4). Using WGS data of 72 MTBC isolates, we defined 10 Lineage 4 sublineages. We screened more than 3,000 clinical isolates with *sublineage-specific* markers and mapped sublineage proportions to countries. Five sublineages were restricted to specific geographical areas, indicating a clonal expansion outside of Europe. Three sublineages were observed frequently among patient isolates born in Europe and were also found globally distributed. Focusing on one of the "global" sublineages, the "Latin-American Mediterranean" strain family, we found evidence that the hypothetical origin of the "LAM" sublineage was in Europe.

Contents

Acknowledgements **x**

1. Introduction **1**

1.1. The global burden of tuberculosis 1
1.2. TB is an ancient disease 2
1.3. Current genetic diversity in the MTBC 5
1.4. Genotyping and molecular markers of MTBC 6
1.5. Phenotypic consequences of genetic diversity 12
1.6. Treatment of TB and drug resistance 13

2. Objectives and outline **15**

2.1. Aims of this thesis 15
2.2. Specific objectives 15
2.3. Outline 16

3. Single nucleotide polymorphisms in *Mycobacterium tuberculosis* **17**

3.1. Summary 19
3.2. Why are SNPs important for our understanding of TB? 20
3.3. What are SNPs and how many do we observe? 21
3.4. SNPs are phylogenetically informative in MTBC 23
3.5. The functional consequences of SNPs 26
3.6. How do we discover new SNPs in MTBC? 30
3.7. The need for a new SNP database for MTBC 32
3.8. Features of a new MTBC SNP database 35
3.9. Conclusions 38
3.10. Acknowledgements 39

4. Two new rapid SNP-typing methods for classifying *Mycobacterium tuberculosis* complex into the main phylogenetic lineages **41**

4.1. Abstract 43

4.2. Introduction . . . 44
4.3. Methods . . . 46
4.4. Results . . . 57
4.5. Discussion . . . 62
4.6. Acknowledgement . . . 65
4.7. Supplementary information . . . 65

5. KvarQ: Targeted and direct variant calling from FastQ reads of bacterial genomes 67
5.1. Abstract . . . 69
5.2. Background . . . 70
5.3. Implementation . . . 71
5.4. Results . . . 74
5.5. Discussion . . . 83
5.6. Conclusion . . . 85
5.7. Materials and methods . . . 85
5.8. Availability and requirements . . . 87

6. Tracking a tuberculosis outbreak over 21 years: strain-specific single nucleotide polymorphism-typing combined with targeted whole genome sequencing 89
6.1. Abstract . . . 91
6.2. Introduction . . . 92
6.3. Methods . . . 93
6.4. Results . . . 98
6.5. Discussion . . . 107
6.6. Acknowledgements . . . 109
6.7. Financial support . . . 110
6.8. Potential conflicts of interest . . . 110

7. The global spread of the Euro-American lineage of *M. tuberculosis* 111
7.1. Abstract . . . 113
7.2. Introduction . . . 115
7.3. Methods . . . 117
7.4. Results . . . 124
7.5. Discussion . . . 139

8. General discussion **147**
8.1. The role of SNP-typing and whole genome sequencing 147
8.2. Technical and analytical limitations of WGS 148
8.3. Future improvements of WGS 150
8.4. Micro-evolutionary aspects 151
8.5. Macro-evolutionary aspects 152
8.6. Evolutionary dating 153
8.7. Public health relevance 154
8.8. Conclusions 155

9. Bibliography **157**

List of Figures **191**

List of Tables **193**

A. Appendix **195**
A.1. Appendix to Chapter 5 (KvarQ) 195
A.2. Appendix to Chapter 6 (Transmission of MTBC in an outbreak in Bern) . 198

B. List of publications **209**

Acknowledgements

This PhD thesis was made possible only with the help of several people. I had the great pleasure to be involved in projects combining microbiology, molecular and evolutionary biology, bioinformatics, epidemiology, public health and medicine. I highly appreciated the commitment and the efforts of all the persons involved.

First and foremost, I am indebted to my supervisor and mentor during this PhD, Prof. Sebastien Gagneux. Sebastien, your guidance through these four years were of greatest value. I learned how to accomplish a project successfully, how to identify which aspects are the important ones, how to write and present a story, and also how good leadership can be. The time and effort you dedicated to each project and to each of us was just priceless. And I also highly appreciated the efforts for social interactions, which kept the group spirit so high. This included the weekly meetings, the yearly retreats and the fantastic dinners. Thank you for all that.

I would also like to specially thank Dr. Lukas Fenner for getting me involved in several projects and for making me part of the TB research community in Switzerland. Your input and the skills I obtained during these collaborations are invaluable for my future work.

I thank Prof. Stefan Niemann for being part of my PhD thesis committee, and Prof. Gerd Pluschke for joining as external expert. I'd like to give special thanks to Prof. Hans-Peter Beck, who made my work at the Swiss TPH possible in the first place, and who kindly agreed to chair my defense at the end.

I would also like to thank the director of the Swiss TPH, Prof. Marcel Tanner. On the one hand for the financial support to attend a workshop and a conference, but even more for his enthusiasm that translates into each project, and that makes the institute the most inspiring place.

I also thank the numerous collaborators within the research projects, especially for sharing data, ideas and isolates. In particular, I very much appreciated the possibility to work with Prof. Dorothy Yeboah-Manu and her team at the Noguchi Memorial Institute in Accra, Ghana.

I was most lucky to be part of the TB research group in Basel. The enthusiasm, the contribution of everybody, the support and the friendly interactions just make a difference. I would especially like to mention Mireia and Daniela for great support and patience in the genomics work. Sonia, for your advice and support in many areas, especially the training in the BSL3-laboratory, but also the many constructive debates over the years, and not less a lot of fun. Julia, for the great daily support in the lab, for everything we set up together successfully, and for the good times we shared. Marie, for always reminding me what is relevant, and for the fun moments that sometimes make a day. Liliana, for all that we achieved during the year of your MSc thesis. Andreas, for an incredibly productive collaboration on KvarQ and for the patience with technical explanations. The later members of the group; Andrej, Sebastian, Miriam and Rhastin, who supported me whenever necessary. Most important, my PhD- and third-floor office mates whom I can now call friends: Adwoa, Bijaya, Damien, Mohamad and Serej—for a most enjoyable time and for keeping the feet on the ground.

My deepest gratitude goes to my family, my friends and specially Carolin for all the support during these years of study. Without you, I would not have made it. Thank you all!

Funding for the work in this thesis was provided by the Swiss National Science Foundation and the Swiss Lung Association (Lungenliga). The participation at several conferences was supported by the “Reisefonds für den akademischen Nachwuchs der Universität Basel”.

1. Introduction

Tuberculosis is driven by interacting environmental, host and pathogen factors (Comas *et al.*, 2009a). This thesis is focused on understanding the pathogen, members of the *Mycobacterium tuberculosis* complex (MTBC). The origin, the species concept and the diversity of the pathogen are presented in this introductory chapter. First, however, the importance of the disease and the global burden of TB are introduced.

1.1. The global burden of tuberculosis

The World Health Organization (WHO) estimated that there were 8.6 million cases and 1.3 million deaths caused by tuberculosis (TB) in 2012, making TB (together with HIV and malaria) one of the top three global killers among infectious diseases (WHO, 2013). For a long time dangerously neglected, TB came into the focus of the WHO in the beginning of the 1990s, when it was recognized as a re-emerging disease (Dye *et al.*, 2010). Since 1990, TB mortality has dropped by 45% thanks to the renewed efforts in TB control and research. After peaking between 2000 and 2005, the global TB incidence rate has now also started to slowly decrease (WHO, 2013). However, in at least five of the 22 high-burden countries that carry 80% of the TB burden (WHO, 2013), the incidence is not decreasing. In South Africa, as a particular case, the incidence rate has more than doubled in the last 15 years, and South Africa has now, together with Swaziland, the highest yearly incidence in the world with over 1,000 new cases per 100,000 population (WHO, 2013). Indeed, a large number of the countries with the highest incidence rates are in sub-Saharan Africa (Figure 1.1).

The largest absolute numbers of TB cases occur in Asian countries, with India and China together carrying more than a third of the global burden (2.2 million and 1 million cases, respectively, in 2012) (WHO, 2013). The Central Asian region also has the highest number of drug-resistant TB cases, with more than 18% of new TB cases and 50% of previously treated TB cases having multi-drug resistant TB (MDR-TB) in some countries. The Central Asian region is followed by Eastern Europe with between 10% and 20% MDR-TB cases among new cases.

TB is a poverty disease, reflected in the large majority of cases in developing countries and in the poor population of newly industrialized countries (Figure 1.1). Although in high-income countries, particularly in Northern America, Europe and Australia / New Zealand, the incidence and mortality of TB could be massively reduced in the last decades, elimination of the disease is in no reach, and likewise, no country has ever eliminated TB (WHO, 2013). In Europe, the remaining cases are mainly found in immigrants (in Switzerland as an example, approximately 75% of cases) and in populations of homeless, drug addicts and alcoholics, who share some of the strongest risk factors for TB. These include bad living and housing conditions, malnutrition, smoking and often HIV co-infection (Palomino *et al.*, 2007).

In 1999, WHO has estimated that nearly one third of the world population is latently infected with bacteria of the MTBC (Figure 1.2) (Dye C *et al.*, 1999). This population represents a huge reservoir for the disease, and around 5-10% of these persons will progress to active disease at any point in their life (Barry *et al.*, 2009). However, the risk of progressing to active TB, the infectious form of the disease, depends on host factors, which have been well established and include immunological factors (e.g. immunosuppression), co-infections (e.g. HIV, diabetes), physical condition and other factors (Barry *et al.*, 2009).

The role of the causative agent of the disease, MTBC, is less well understood. In particular, the diversity of the pathogen has been underestimated for a long time (Hershberg *et al.*, 2008). In the last two decades, molecular methods have revealed genetic diversity within the MTBC, phenotypic differences between MTBC strains, and the role of co-evolution with humans (Gagneux, 2012). In the next sections, the origin of the pathogen, the genetic and phenotypic diversity and the genotyping methods of MTBC are introduced.

1.2. TB is an ancient disease

Tuberculosis has co-evolved with humans for thousands of years (Donoghue, 2011). Evidence for TB being an ancient disease includes morphological changes associated with TB in skeletal remains, molecular data from ancient DNA, and phylogenetic data from contemporary isolates.

In 2005, Gutierrez *et al.* (2005) proposed that a progenitor species, ancestral to today's MTBC, could have caused a form of TB as early as 3 million years ago. The oldest morphological indications of TB were found in a 500,000 year old remains of *Homo erectus* in Turkey (Kappelman *et al.*, 2008), but are controversially discussed. Later indications date back to Neolithic times (Donoghue, 2011). Molecular studies using ancient DNA

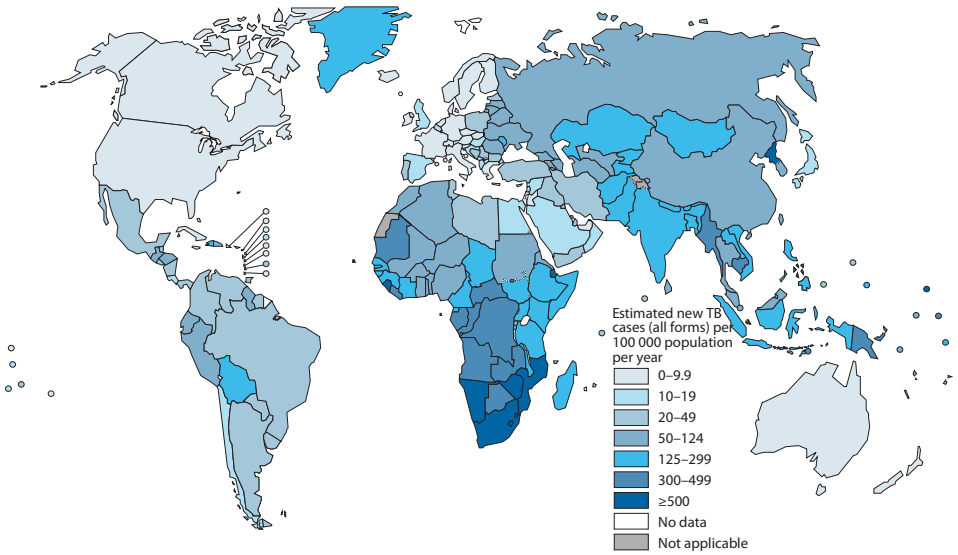


Figure 1.1.: **TB incidence rates per country in 2012 as estimated by WHO.** Figure from WHO (2013).

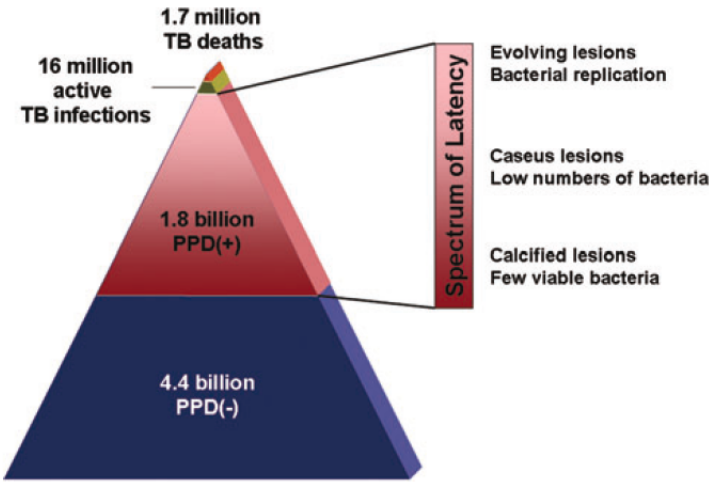


Figure 1.2.: **The global burden of TB as estimated by WHO (Dye C *et al.*, 1999).** Each section of the triangle is roughly in scale. Figure from Rustad *et al.* (2009).

(aDNA) were then able to demonstrate the presence of MTBC DNA in bones of mummified humans, and even to uncover the MTBC genotypes of the historic MTBC strains (Donoghue, 2009). Hershkovitz *et al.* (2008) found DNA in bodies of a woman and an infant in the region of today's Haifa (Israel), dating to 9250 to 8150 BP, for which spoligotyping revealed a pattern consistent with contemporary Lineage 4, the "Euro-American" lineage (see below) (Hershkovitz *et al.*, 2008). Zink *et al.* (2003) investigated Egyptian mummies from 2500 to 500 BC, and found spoligotypes also consistent with contemporary Lineage 4. Ancient DNA was also isolated from several locations in Europe, reflecting the large burden of TB during the last centuries. This is discussed later in this thesis (Chapter 7). Recently, the first whole genome sequences of historic isolates were reported (Bouwman *et al.*, 2012; Chan *et al.*, 2013).

The question as to which genotypes of Mycobacteria caused TB in the pre-Columbian times in the Americas is debated. The fact that "European" genotypes, i.e. MTBC Lineage 4 strains, are pre-dominant from North to South America indicates a (recent) introduction from Europe. Molecular evidence, however, has indicated the presence of MTBC DNA in much older mummified remains. The first molecular evidence of MTBC in the Americas was found in a mummified Bison from 17,870 BP in Wyoming. The spoligotyping pattern however, was inconclusive and only allowed the exclusion of *M. bovis* (Rothschild *et al.*, 2001). Salo *et al.* (1994) and Arriaza *et al.* (1995) had identified MTBC DNA in mummies before, from 1000-1300 AD in Peru and from 1000 AD in Chile, but did not report genotypic data.

The evolutionary age calculated with molecular data is the subject of ongoing discussions. For a long time, it was thought that during the Neolithic Revolution (10,000 years BP), human TB emerged as a zoonosis with the advent of domestication of animals (Daniel, 2006). But using population genetic methods, Wirth *et al.* (2008) and Comas *et al.* (2013) suggested an origin of the MTBC as early as 40,000 years and 70,000 years ago, respectively. Pepperell *et al.* (2013) recently estimated a much higher evolutionary mutation rate, and less than 5,000 years for the origin of the MTBC. The divergence might result from different calibration points to estimate mutation rates.

Despite the increasing knowledge about the origin of the MTBC and the proposed "out-of-and-back-to-Africa" scenario (see also below) (Hershberg *et al.*, 2008), no studies have used robust genomic data to address the question as to how the "modern" MTBC Lineage 4 evolved to be globally dispersed. This will be discussed in Chapter 7.

1.3. Current genetic diversity in the MTBC

Tuberculosis is today caused by closely related members of the MTBC. The MTBC represents a group of species or *ecotypes* (Smith *et al.*, 2006) within the genus of *Mycobacterium* (Figure 1.3) (Brosch *et al.*, 2002; Achtman, 2008). *M. tuberculosis sensu stricto* (s.s.) and *M. africanum* are responsible for the large majority of human TB cases. The MTBC further includes the animal-associated pathogens *M. bovis* (infecting and causing disease mainly in cows), *M. caprae* (goats), *M. microti* (voles) and *M. pinnipedii* (seals) (Palomino *et al.*, 2007) (Figure 1.5). On a side note, *M. bovis* was an important cause of TB in humans before the introduction of milk pasteurization, but the cases have since decreased to low numbers, in particular where bovine TB incidence is low or pasteurization is standard (Müller *et al.*, 2013). Additional members of the MTBC include the "dassie bacillus" (Mostowy *et al.*, 2004), *M. mungi* (infecting mongoose) (Alexander *et al.*, 2010), *M. orygis* (antelopes) (Gey van Pittius *et al.*, 2012) and the chimp bacillus (Coscolla *et al.*, 2013), all of which have only been described anecdotally. The "smooth tubercle bacteria", including *M. canettii*, are traditionally also considered part of the MTBC and cause sporadic cases of TB, but are only isolated from patients from the Horn of Africa (or with a connection to the Horn of Africa). These microbes are much more diverse than the other members of the MTBC, have a larger genome size, and show a clear evidence of ongoing horizontal gene exchange (Supply *et al.*, 2013). The categories and species names within the MTBC were originally assigned based on biochemical characteristics (Collins *et al.*, 1982). The differentiation between *M. bovis* and *M. tuberculosis* was mainly based on the presence of nitrate reductase and the production of niacin. The identification of *M. africanum* based on the chemical tests is ambiguous, but colony morphology resembles *M. bovis* (Jong *et al.*, 2010). Brosch *et al.* (2002) then provided a comprehensive classification scheme based on large sequence polymorphism.

Until the 1990s, MTBC was considered as genetically uniform. First genotyping methods, including *IS*6110-RFLP (see below) and spoligotyping (Kamerbeek *et al.*, 1997) revealed major global families (Embden *et al.*, 1993; Brudey *et al.*, 2006; Demay *et al.*, 2012) (methods see below). However, these genotyping methods based on repetitive or mobile elements (rather than DNA sequencing data) both had limitations for evolutionary inference (Comas *et al.*, 2009b). Sequencing of the genes *katG* and *gyrA* identified "principal genetic groups" with two polymorphisms (Sreevatsan *et al.*, 1997). Only multi-locus sequence typing (MLST) (Maiden *et al.*, 1998), i.e. the sequencing of a small number of housekeeping genes (traditionally by Sanger sequencing) revealed robust lineages (115 polymorphic positions in seven genes) (Baker *et al.*, 2004), but could not resolve all

phylogenetic lineages known today. The first complete genome sequence (H37Rv) (Cole *et al.*, 1998) and three additional genome sequences enabled various SNP-typing studies (Gutacker *et al.*, 2002; Filliol *et al.*, 2006; Gutacker *et al.*, 2006) and allowed robust groupings (further discussed in Chapter 3). Gagneux *et al.* (2006b) used microarrays to identify main phylogenetic lineages based on large deletions, and showed an association of these lineages with human populations. Hershberg *et al.* (2008) found 488 SNP by sequencing 1.5% of the 4.4 Mb genome in 108 MTBC strains, and not only confirmed the phylogenetic lineages, but in particular discovered a strikingly reduced purifying selection and likely a large part of polymorphisms having functional consequences (Hershberg *et al.*, 2008). The data was also consistent with an "out-of-and-back-to-Africa" evolutionary scenario of the MTBC. Wirth *et al.* (2008) proposed a similar grouping of strains and a congruent evolutionary scenario using MIRU-VNTR, and dated the age of the MTBC to around 40,000 years. Population genomic analyses using whole genome sequences (see below) proposed 70,000 years for the origin of the MTBC, and found a striking congruence of co-evolution between humans and the MTBC (Comas *et al.*, 2013). The seven main phylogenetic lineages (Figure 1.4) are being confirmed repeatedly (Casali *et al.*, 2014; Farhat *et al.*, 2013). However, a confusing of genetic groupings *within* these lineages and a number of not necessarily overlapping definitions is complicating the situation. A comprehensive nomenclature, including main phylogenetic lineages and (what we refer to as) "sublineages" is now needed, which ideally should be based on whole genome data.

"Zooming" to the tips of the phylogeny, i.e. MTBC genotypes isolated from individual patients, WGS has also reaveled new diversity. Recent studies have applied WGS to MTBC isolates in transmission chains that were indistinguishable by classical genotyping methods (*IS*6110-RFLP, MIRU-VNTR, spoligotyping) (Schürch *et al.*, 2010; Gardy *et al.*, 2011; Walker *et al.*, 2013b; Roetzer *et al.*, 2013; Bryant *et al.*, 2013b; Walker *et al.*, 2014). SNP-analyses showed that a considerable number of SNPs are found between isolates from different patients, and even within patients. These aspects are further discussed below and in Chapter 6. First, however, the different genotyping methods for MTBC are introduced in chronological order.

1.4. Genotyping and molecular markers of MTBC

Strain classification and nomenclature within the MTBC has a long history (Schürch *et al.*, 2012). After the discovery of the aetiological agent in 1882 (Koch, 1932), the bacterium

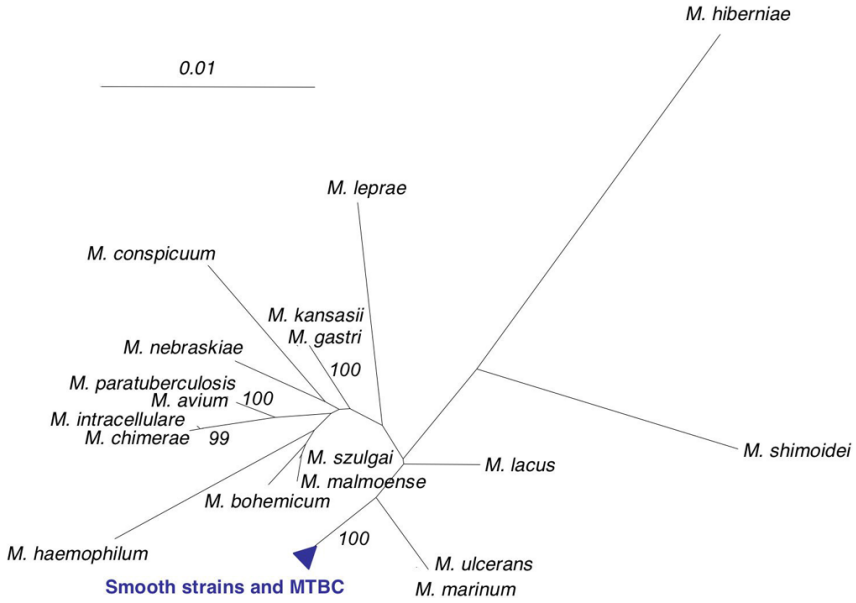


Figure 1.3.: **Phylogeny of the genus *Mycobacterium* using 16S rRNA sequences.** Unrooted neighbor-joining tree using 1,325 nucleotides. The blue triangle corresponds to *Mycobacterium tuberculosis* and the smooth tubercle bacilli, which are identical or differing by only one single nucleotide. Figure from Gutierrez *et al.* (2005).

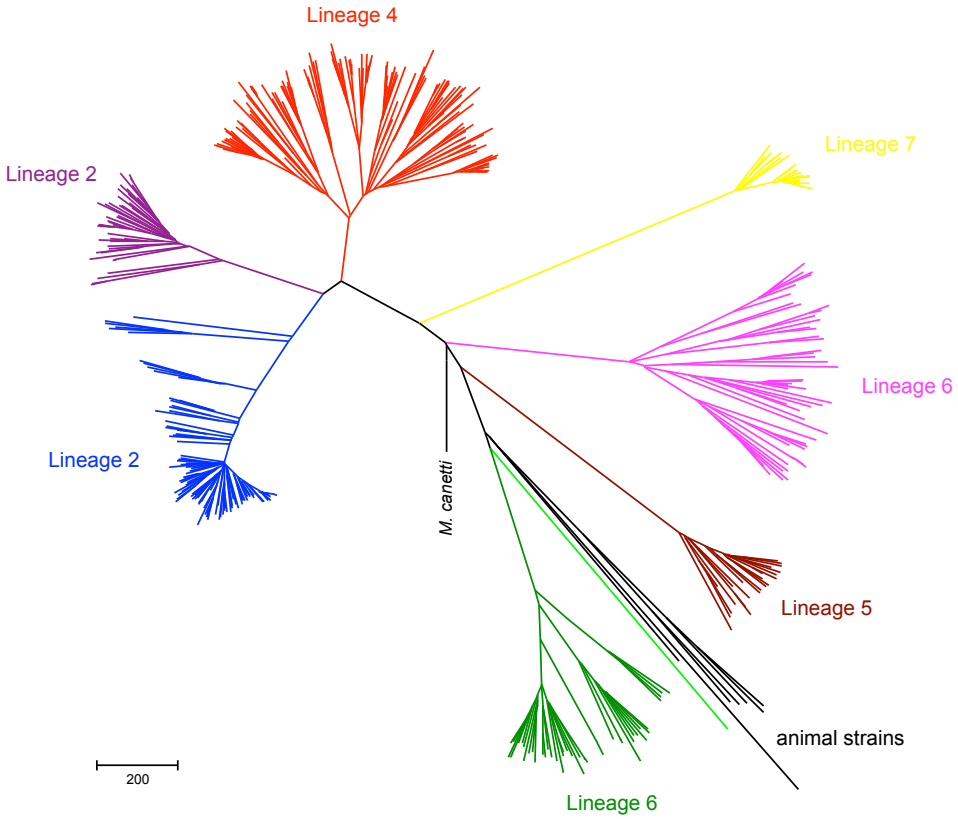


Figure 1.4.: **The current phylogeny of 420 whole genome sequences of the MTBC reveals a pronounced substructure within phylogenetic lineages.** Figure modified from Mireia Coscollà (manuscript in preparation). The scale bar indicates number of SNPs.

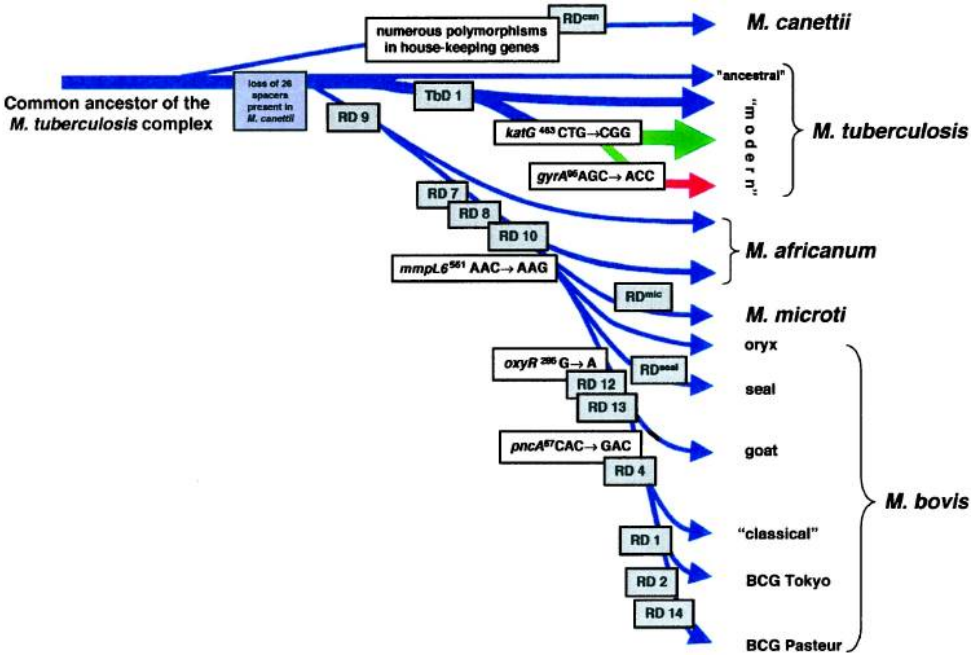


Figure 1.5.: **Proposed evolutionary pathway of the tubercle bacilli using Regions of Difference (RD).** Figure from Brosch *et al.* (2002).

was named *M. tuberculosis* (Lehmann *et al.*, 1896), but it was not until 1976 that the "TB complex" was defined (Tsukamura, 1976). At this time, *M. bovis* and *M. africanum* were recognized as closely related species (Castets *et al.*, 1969; Collins *et al.*, 1982). The three species could be differentiated by biochemical tests (Collins *et al.*, 1982). Some of these tests are still performed today, but have limitations (Jong *et al.*, 2010). With the availability of DNA- and especially PCR-based methods, genotyping became the new gold standard for species and strain differentiation.

1.4.1. *IS*6110 RFLP

The first genotyping method to be used was a restriction fragment length polymorphism (RFLP) method based on insertion sequence *IS*6110 (Embden *et al.*, 1993). RFLP has routinely been used in molecular epidemiology of TB to identify outbreaks and chains of transmission (Small *et al.*, 1994). *IS*6110-RFLP has a high discriminatory power, but is labour-intensive, time consuming and needs a lot of high-quality DNA. The results are difficult to compare between laboratories (Comas *et al.*, 2009b). Nevertheless, *IS*6110-RFLP paved the way for large epidemiologic and phylogenetic studies and revolutionized the understanding of the transmission of the disease. Chains of transmission were identified (same *IS*6110-RFLP pattern), re-infection could be discriminated from relapse, and contaminations could be detected (Barnes *et al.*, 2003; Genewein *et al.*, 1993). On the

other hand, *IS*6110-RFLP was not ideally suited for strain *classification*, and also lacked a format to be easily exchanged between laboratories.

1.4.2. Spoligotyping

Kamerbeek *et al.* then developed a CRISPR-based (Clustered Regularly Interspaced Short Palindromic Repeats) method and named it "spoligotyping" (Kamerbeek *et al.*, 1997). Spoligotyping takes advantage of the Direct Repeat (DR) locus of the MTBC genome, a region consisting of identical 36 bp direct repeats, interspersed with unique 35-41 bp spacer sequences. The latter are amplified and visualized on a membrane, which allows testing for the presence or absence of each of the spacers (usually, a set of 43 spacers is used). Strains have different patterns of absence and presence of these spacers, and common patterns among clades can be found. However, this method has limited value for constructing phylogenies, as convergent evolution among strains from different lineages can occur. Low statistical support for phylogenetic groupings is usually obtained, and not all lineages can be detected that are found with other methods (Comas *et al.*, 2009b). Nevertheless, spoligotyping has been an invaluable tool for the TB research community to identify genotyping families and phylogeographical distributions, and also paved the way for the next generation of molecular markers. Thanks to its ease of use and high portability, spoligotyping is still used extensively in many laboratories. The largest database of MTBC genotyping data is based on spoligotyping (SITVITWEB, formerly SpolDB4) (Demay *et al.*, 2012), and tools to extract spoligotyping patterns from WGS data have been developed (Coll *et al.*, 2012) (and Chapter 5).

1.4.3. MIRU-VNTR

The first complete whole genome sequence, generated in 1998 with Sanger-sequencing (Cole *et al.*, 1998), allowed the development of whole genome-based methods. Mycobacterial Interspersed Repetitive Units (MIRU) - Multiple Loci Variable Number of Tandem Repeats (VNTR) Analysis (MLVA) is today's accepted gold standard for molecular epidemiology of TB (Supply *et al.*, 2006). MIRUs are tandemly repeated DNA elements which are dispersed in inter- and intragenic regions of the genome. The number of tandem repeats per locus varies from strain to strain, and can be presented in a numeric way (as number of copies per locus). MIRU-VNTR can be automatized with capillary electrophoresis, but is still labour-intensive due to a high number of individual PCRs required. A database was introduced for MIRU-VNTR, known as MIRU-VNTRplus, to globally compare among samples (Allix-Béguec *et al.*, 2008; Weniger *et al.*, 2010; Weniger

et al., 2012). In addition to epidemiological applications, MIRU-VNTR data have been used for the definition of MTBC clades and for the dating of the MTBC phylogeny (Wirth *et al.*, 2008).

1.4.4. Large Sequence Polymorphisms (LSP)

Due to the limitations of spoligotyping and MIRU-VNTR, robust alternatives were needed as markers for phylogenetic applications. Large sequence polymorphisms (LSP) and single nucleotide polymorphisms (SNPs) were proposed as robust phylogenetic markers. LSPs in form of deletions in mycobacteria represent unidirectional events, i.e. they are irreversible, as ongoing horizontal gene transfer does not occur in MTBC (Gagneux *et al.*, 2006b). One of the first deletion analyses was performed on different strains of the vaccine strain BCG (Behr, 2001). Hirsh *et al.* (2004) identified 68 genomic regions that were present in H37Rv but absent in other strains. Gagneux *et al.* (2006b) later identified six main lineages of MTBC based on LSPs, also known as regions of difference (RD). Tsolaki *et al.* (2004) and Gagneux *et al.* (2006a) have used these robust markers to assess phenotypic differences between genotypes. In the study by Gagneux *et al.* (2006b), each MTBC lineage was associated with specific, sympatric human populations, i.e. humans from a certain geographic region were more likely to be infected with a strain phylogeographically associated with this region. RD deletions have also unravelled the evolutionary scenario of the animal-associated species or ecotypes of MTBC (Figure 1.5, Brosch *et al.* (2002)).

1.4.5. Single Nucleotide Polymorphisms (SNP)

The large deletions discussed above reflect unidirectional events and are therefore not prone to homoplasy. However, they do not allow the calculation of genetic distances and also cannot completely resolve all deep-rooting branches of the MTBC phylogeny, and are therefore of limited phylogenetic value (Comas *et al.*, 2009a). With the availability of sequencing technologies, large numbers of SNPs have been discovered in MTBC and have been questioning the notion of MTBC as a pathogen with very restricted genetic diversity (Achtman, 2008). Because of the absence of recombination and lateral gene transfer, SNPs are perfect phylogenetic and epidemiological markers. They are unique events and show almost no homoplasy. The discovery of SNPs in MTBC, as well as the SNP-genotyping past and present, are discussed in more detail in Chapter 3.

1.4.6. Whole genome sequencing as a typing method

SNPs are nowadays discovered by sequencing the complete genome of MTBC isolates. WGS has become increasingly available, and prices have exponentially decreased in the last years. Barcoded DNA sequencing libraries for pooled sequencing of up to 96 isolates, as well as new benchtop DNA sequencing devices have lead to a free fall of sequencing costs in the last decade (although slowing down recently, `http://www.genome.gov/sequencingcosts/`). It can be foreseen that WGS will be implemented in daily research and routine clinical work and replace previous genotyping methods as well as genotypic drug resistance testing. The first researchers to use WGS to investigate a chain of transmission in the Netherlands (the infamous Haarlem cluster) were Schürch *et al.* (2010) that resolved three isolates indistinguishable by classical DNA fingerprinting. They found 8 polymorphic positions. The first comprehensive study using WGS for molecular epidemiology was published in 2011 (Gardy *et al.*, 2011). The authors sequenced 32 recent and 4 historical isolates of a MIRU-VNTR cluster and combined the SNP-data with social network analysis. Two distinct lineages, a superspreader behaviour, and a correlation with crack cocaine use were identified. In a landmark study, Walker *et al.* (2013b) sequenced 390 MTBC isolates, estimated a mutation rate of 0.5 SNPs/genome/year in longitudinal isolates, and found usually five or fewer SNPs between epidemiologically linked cases. A study by Bryant *et al.* (2013b) found a comparable mutation rate (0.3 SNPs/genome/year), but a large variability and no clear SNP-threshold for linked cases. Roetzer *et al.* (2013) used WGS to decipher an outbreak of 86 patients in Germany, and found again a similar mutation of 0.4 SNPs/genome/year and several subclusters, i.e. distinct clades in the genomic network. Ford *et al.* (2011) used WGS of MTBC from infected macaques and inferred MTBC mutation rates comparable between latent infection and active disease. Several studies have applied large-scale WGS to identify drug-resistance associated mutations (drug resistance in TB is discussed below). Comas *et al.* (2011a), Casali *et al.* (2012) and Casali *et al.* (2014) have found mutations compensating for the fitness defect associated with rifampicin resistance. Köser *et al.* (2013) report the use of WGS to rapidly identify drug resistance mutations of an XDR-TB patient.

These studies demonstrate the potential for future routine applications of WGS in research and molecular epidemiology. However, large-scale applications are not yet possible as cost and the analyses remain important hurdles (also discussed in Chapter 3). *Targeted* sequencing of selected isolates, on the other hand, is feasible in many settings. Hence, innovative solutions are needed to identify subsets of MTBC isolates for which targeted WGS can be applied. The subsets would consist of the isolates where the benefit of WGS is highest compared to other molecular markers. For molecular epidemiology, these

subsets would be molecular clusters, which classical markers can not resolve due to the lack of discriminatory power (Niemann *et al.*, 2009) (see Chapter 6). For evolutionary applications, genotypes of MTBC can be identified with clade-specific SNPs, and only the uncategorised samples subjected to WGS. This approach is further elaborated in Chapter 7.

1.5. Phenotypic consequences of genetic diversity

The genetic diversity of MTBC identified in the last 20 years has also lead to increased efforts to study phenotypic consequences of this genetic diversity. However, already in the 1960s, studies in guinea pigs found differences in virulence between MTBC isolates from India and from the UK (Mitchison *et al.*, 1960). Since then, many studies have sought associations between genotype and phenotype (Homolka *et al.*, 2010). They were recently reviewed by Coscolla *et al.* (2010). MTBC Lineage 2 strains (in particular the "Beijing" sublineage) have repeatly been associated with increased virulence and drug resistance (Hanekom *et al.*, 2011). For example, Lan *et al.* (2003) found an association of Beijing strains with treatment failure and disease relapse, Caws *et al.* (2006) found Beijing strains associated with HIV co-infection, resistance to any drug and multi drug-resistance in TB meningitis cases, and Hanekom *et al.* (2007) found one specific Beijing sublineage more abundant than other MTBC strains in Cape Town, South Africa. Strikingly, a higher mutation rate towards antibiotic resistance of MTBC Lineage 2 strains compared to MTBC Lineage 4 was recently found by Ford *et al.* (2013). However, results are not always congruent (Yuan *et al.*, 2014), and might also depend on the geographical region (i.e. the human population). Furthermore, not only lineage-specific, but also strain-specific characteristics influence the phenotypic differences and need to be considered (Reiling *et al.*, 2013).

Many of these phenotypic differences have been observed between main phylogenetic lineages (as defined by LSPs or SNP data). However, it is also increasingly appreciated that there is diversity *within* these main lineages. We refer to sub-clades of the 7 main human-associated MTBC lineages as "sublineages". In the last years, a few studies have looked at phenotypic differences between isolates of sublineages. In fact, the "Beijing" clade is a sublineage of Lineage 2 (defined by RD207), and has recently been associated with a higher pathogenicity than other Lineage 2 strains in guinea-pigs (Kato-Maeda *et al.*, 2012). Nahid *et al.* (2010) have found Lineage 4 sublineage RD724 (see also Chapter 7) associated with more severe disease of TB at baseline compared to non-RD724 strains. Anderson *et al.* (2013) have compared isolates from the RD-defined sublineages of Gag-

neux *et al.* (2006a) in an epidemiological study, and found one sublineage, defined by RD183-deleted, associated with clustering and homelessness. However, the comparison of these sublineage-definitions between studies is often difficult. Various, often incompatible markers and typing schemes are used, and groupings are not always based on robust markers. For the future analysis of genotype-phenotype associations between sublineages (and main lineages), a classification based on WGS data is needed (see also Chapter 7). The population structure based on robust groupings can then to be considered for the development of new diagnostic methods and new treatment options. Specific members of the MTBC could be naturally resistant to new antituberculosis drugs (such as *M. bovis* to pyrazinamide), or genetic background mutations could lead to false-positive results in genotypic drug resistance diagnostic assays (Köser *et al.*, 2012a; Feuerriegel *et al.*, 2014).

1.6. Treatment of TB and drug resistance

TB can be treated with antituberculosis drugs since the 1940s, when Streptomycin and para-aminosalicylic acid were discovered (Schatz *et al.*, 1944; Lehmann, 1946). Several other drugs have become available since then (Zhang *et al.*, 2009). Today, the standard TB treatment regimen is defined by the WHO (WHO, 2010). It is part of the DOTS strategy (Direct Observed Treatment, Short-course) that was launched in 1994. The standard regimen for new cases of TB consists of 2 months of isoniazid, rifampicin, pyrazinamid and ethambutol, followed by 4 months of isoniazid and rifampicin (WHO, 2010). In the case of drug resistance, the WHO defined multidrug-resistant strains (MDR) as strains resistant to at least the two major first line drugs, isoniazid and rifampicin. Further drugs are available for MDR-TB, categorized in five groups; 1) pyrazinamide, ethambutol and rifabutin (as first-line oral agents); 2) kanamycin, amikacin, capreomycin and streptomycin (injectables/aminoglycosides); 3) levofloxacin, moxifloxacin and ofloxacin (fluoroquinolones); 4) para-aminosalicylic acid, cycloserine, terizidone, ethionamide and prototionamide (oral bacteriostatic second-line agents) 5) clofazimine, linezolid, amoxicillin/clavulanate, thiacetazone, imipenem/cilastatin, high-dose isoniazid and carithromycin (with unclear role). Drug resistance to these second-line drugs can also occur. The definition of extensively drug resistant strains (XDR) is defined as MDR plus the resistance to any fluoroquinolone and to at least one of the injectable second-line drugs. XDR strains represent a major concern for global TB control, but other drug resistances and combinations can occur as well, and reports of totally drug resistant strains (TDR) have been published (Migliori *et al.*, 2007).

Resistance to all the available antituberculosis drugs can occur. To a large part, resistance is conferred by single nucleotide mutations in the chromosome of MTBC. No resistance plasmids exist in MTBC and, as mentioned above, horizontal gene transfer does not occur. Due to the clonal nature of MTBC, single SNPs can therefore be used as resistance markers. This is discussed in the following chapter (Chapter 3). In Chapter 5, we make use of the most important drug resistance markers in the software `KvarQ` to rapidly identify drug resistance patterns from WGS data. A list of the currently known most important drug resistance mutations in MTBC is found in Table A.2 in Appendix A.1. A public database, TBDReaMDB serves as an (unofficial) reference for known drug resistance associated mutations (Sandgren *et al.*, 2009).

2. Objectives and outline

2.1. Aims of this thesis

The overarching aims of this thesis were i) to develop a framework for WGS- and SNP-based classification of MTBC isolates into main phylogenetic lineages and sublineages, ii) to study the transmission and micro-evolution of MTBC in a local outbreak, and iii) to infer the global evolutionary scenario of MTBC Lineage 4.

2.2. Specific objectives

The corresponding six specific objectives were:

- **Objective 1**. To review the knowledge about SNPs in MTBC and to identify the needs of the community (Chapter 3).
- **Objective 2**. To extract MTBC lineage-specific SNPs from WGS data and to develop two laboratory SNP-typing assays (Chapter 4).
- **Objective 3**. To develop a software tool for rapid *in silico* SNP-typing of MTBC raw genome sequencing data (Chapter 5).
- **Objective 4**. To study the transmission of a specific MTBC strain in an outbreak in Switzerland (Chapter 6).
- **Objective 5**. To define sublineages of MTBC Lineage 4 using WGS data and to develop a SNP-based classification scheme for Lineage 4 sublineages (Chapter 7).
- **Objective 6**. To study the evolutionary trajectory of MTBC Lineage 4 on a global scale (Chapter 7).

2.3. Outline

In the following chapter *(Chapter 3)*, we reviewed the current knowledge about SNPs in MTBC. We summarized how SNPs in MTBC are identified, how they are used as molecular markers for genotyping, and what we would need to sustainably store SNP information in the public domain.

In *Chapter 4*, we developed two new laboratory SNP-typing assays to unambiguously classify clinical isolates into the main phylogenetic lineages of MTBC.

To facilitate the analysis of WGS data, we also developed a software for *in silico* SNP-typing. The new software `KvarQ` allows the targeted extraction of phylogenetically informative SNPs and drug resistance mutations from raw genome sequencing data. `KvarQ` is described in *Chapter 5*.

In *Chapter 6*, we describe the application of combined SNP-typing and WGS to identify a large TB outbreak in Switzerland during 21 years. We tracked the transmission with WGS and social network analysis.

In *Chapter 7*, we expanded our focus to the macro-evolutionary scale. We aimed at inferring the historical dispersal of MTBC Lineage 4 (previously called the "Euro-American" lineage). To that end, we again applied a combination of SNP-typing and targeted WGS to a collection of more than 3,000 clinical isolates to generate a phylogeographical distribution. Focusing on one of the globally most frequent genotypes, the "Latin American Mediterranean" (LAM) family, we studied the global spread with phylogenetic and population genomic methods.

In *Chapter 8* the key findings are summarized, and general points addressed that not were discussed in the individual chapters.

3. Single nucleotide polymorphisms in *Mycobacterium tuberculosis* and the need for a curated database

David Stucki [1,2] and Sebastien Gagneux [1,2,*]

[1] Swiss Tropical and Public Health Institute, Basel, Switzerland

[2] University of Basel, Switzerland

[*] Corresponding author

This paper has been published in *Tuberculosis (Edinburgh, Scotland)* 2013, 93(1):30–39.

3.1. Summary

Recent advances in DNA sequencing have lead to the discovery of thousands of single nucleotide polymorphisms (SNPs) in clinical isolates of *Mycobacterium tuberculosis* complex (MTBC). This genetic variation has changed our understanding of the differences and phylogenetic relationships between strains. Many of these mutations can serve as phylogenetic markers for strain classification, while others cause drug resistance. Moreover, SNPs can affect the bacterial phenotype in various ways, which may have an impact on the outcome of tuberculosis (TB) infection and disease. Despite the importance of SNPs for our understanding of the diversity of MTBC populations, the research community is currently lacking a comprehensive, well-curated and user-friendly database dedicated to SNP data. First attempts to catalogue and annotate SNPs in MTBC have been made, but more work is needed. In this review, we discuss the biological and epidemiological relevance of SNPs in MTBC. We then review some of the analytical challenges involved in processing SNP data, and end with a list of features, which should be included in a new SNP database for MTBC.

3.2. Why are SNPs important for our understanding of TB?

The declaration of tuberculosis (TB) as a global public health emergency in 1993 (WHO, 2011) lead to renewed efforts to study the biology of the *Mycobacterium tuberculosis* complex (MTBC). For many years, the main research focus was on individual genes and proteins, but the generation of the first *M. tuberculosis* genome sequence in 1998 (Cole *et al.*, 1998) opened the door for more comprehensive approaches. In particular, comparative genomics studies have helped us gain a better insight into the genetic diversity and phylogenetic relationships in MTBC (Mostowy *et al.*, 2002; Comas *et al.*, 2010; Brosch *et al.*, 2002). These studies showed that the different members of MTBC primarily associated with human disease (i.e. *M. tuberculosis* sensu stricto and *M. africanum*) are more genetically diverse than previously appreciated (Hershberg *et al.*, 2008; Bentley *et al.*, 2012). Increasingly, various "omics" approaches in TB research are being combined into what is generally known as Systems Biology (Comas *et al.*, 2011b). Systems Biology tries to understand complex biological systems by integrating data from various disciplines; in TB for example the comprehensive data from human, animal, and computational model systems (Breitling, 2010; Kirschner *et al.*, 2010). There is increasing evidence that, in addition to environmental factors and human genetics, strain variation in MTBC plays a role in the outcome of TB infection and disease (Coscolla *et al.*, 2010). Hence, there is a need to better understand the global diversity of MTBC, and determine if and how this diversity has relevance for global TB control (Gagneux *et al.*, 2007; Comas *et al.*, 2009a). The advent of next-generation DNA sequencing (NGS) methods is likely to facilitate this task, and indeed, many genome-sequencing projects of MTBC clinical isolates are currently underway (Sanger Institute, 2012a). More than 3,800 raw genome sequences of MTBC strains have already been deposited on public sequence read archives (Figure 3.1), and it is safe to assume that this number will continue to grow rapidly as sequencing costs keep decreasing (Stein, 2010; Wetterstrand, 2012).

In contrast to the relative ease with which DNA sequencing data can be generated today, extracting useful information and compiling these in a user-friendly manner is less straightforward. In particular, thousands of genetic polymorphisms have been extracted from whole genome sequences, but the TB research community currently lacks a centralized database, which would allow accessing and handling these data more efficiently. Several TB-specific databases have been created over the last years, including genome browsers, genotyping- and drug resistance databases (Sharma *et al.*, 2011), but despite these existing platforms, we lack a centralized and comprehensive repository for data on

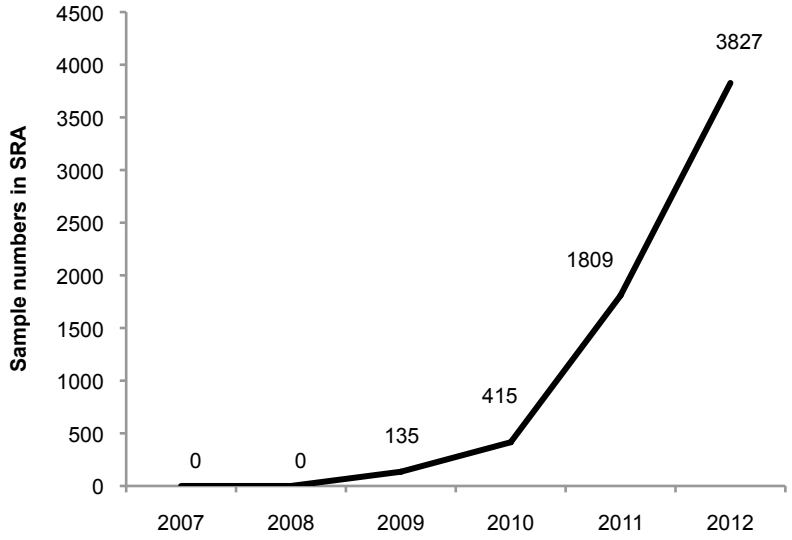


Figure 3.1.: **Number of MTBC samples with raw genome sequences available in the Sequence Read Archive of the National Center for Biotechnology Information (NCBI SRA).** Search query was "*Mycobacterium tuberculosis* complex" in the NCBI Biosample database, and results were extracted from filter "Used in SRA". Y-axis represents cumulative numbers of entries on October 15 of each year.

strain-specific genetic variation in MTBC. Particularly, the field would benefit greatly from a new database compiling all known single nucleotide polymorphisms (SNPs) in MTBC. Ideally, such a database should include proper annotation of these SNPs as well as all relevant metadata. Considering the increasing number of MTBC genomes becoming available, the number of MTBC SNPs identified in the coming years will surely increase by one or more orders of magnitude.

In this review, we start by summarizing the nature of SNPs in MTBC, how SNPs can be used to define phylogenetic relationships between strains, and how they might impact on the phenotype of particular MTBC variants. We then elaborate on how new SNP data can be obtained with NGS technologies, and discuss some of the analytical challenges involved. We end by advocating for a new, user-friendly SNP database for MTBC.

3.3. What are SNPs and how many do we observe?

SNPs are the most common form of genetic variation in MTBC, after insertions and deletions (InDels). A total of 9,037 SNPs were discovered by sequencing 21 clinical strains of MTBC (Comas *et al.*, 2010). Generally, SNPs represent single nucleotide differences between at least two DNA sequences. The term "SNP" is often used interchangeably with "mutation", "polymorphism" or "substitution". Strictly speaking, a change in a single base

pair is generally referred to as a (point) mutation, and happens through errors during DNA replication or as a consequence of DNA damage. Such a mutation represents a relatively rare change from the "normal" base to a mutant form, and is most likely to be neutral or (slightly) deleterious; beneficial mutations can of course occur, but they are generally much less likely. Mutations that are highly deleterious will be rapidly removed by purifying selection, whereas beneficial mutations will increase because of positive selection. In addition, any mutation can increase in frequency as a consequence of random genetic drift. When an allele reaches a certain frequency in the population, we refer to it as a polymorphism (i.e. the co-existence at a specific locus of two or more alleles in a given population). The threshold for defining a new variant as a "polymorphism" as opposed to a "mutation" is arbitrary, and e.g. is set at 1% for human variants (i.e. the minority allele has to be present at a frequency of at least 1% in a given human population for the corresponding nucleotide position to be referred to as "polymorphic"). Below this threshold frequency, this new variant would be referred to as a single nucleotide "mutation". If a new variant becomes fixed in a given population (i.e. 100% of the members of a given population have the new variant), this variant will be referred to as a "substitution" rather than a "polymorphism" (Hartl *et al.*, 2007).

In addition to the difference between single nucleotide mutation, polymorphism, and substitution, biologist often differentiate between "natural polymorphisms" and "adaptive mutations" such as those conferring drug resistance. Natural polymorphisms can be thought of pre-existing variation which defines the overall genomic diversity of a population or a particular strain background, while adaptive mutations represent de novo acquired changes driven by a particular selective force (e.g. exposure to antibiotics, further discussed below). However, it is not always straightforward to discriminate between these different types of mutations. For example, the intrinsic resistance of *Mycobacterium bovis* to pyrazinamide is due to an amino acid change in pncA (Rv2043c) (Huard *et al.*, 2006). This drug resistance-conferring mutation clearly predates the introduction of pyrazinamide and can therefore be considered a natural polymorphism. Moreover, as all strains of the classical *M. bovis* clade harbour this mutation, one could also refer to this mutation as a natural substitution when comparing all classical *M. bovis* to the rest of MTBC. For the sake of simplicity, we will use the term "SNP" for the remainder of this article when referring to any form of single nucleotide variant.

Depending on their position in the genome, SNPs can be either coding or non-coding. With a coding density of 90-96% (Namouchi *et al.*, 2012), most of the SNPs in MTBC are in coding regions of the genome. Coding SNPs can be further divided into synonymous (sSNP) and non-synonymous (nSNP) depending on whether they lead to changes in the

corresponding amino acid sequence. While in average nSNPs are likely to have a stronger effect on the organism's fitness (either beneficial or deleterious), and will therefore be under a stronger selective pressure than sSNPs, the latter are not necessarily selectively neutral. Phenomena such as the codon bias in MTBC and the general mutational bias in bacteria (Hershberg *et al.*, 2010; Namouchi *et al.*, 2012) suggest that sSNPs, too, can be affected by natural selection. Similarly, non-coding SNPs are often considered selectively "neutral", but increasingly the importance of non-coding (i.e. un-translated) regions of the bacterial genome for gene regulation is becoming evident (Arnvig *et al.*, 2011).

SNPs are relatively rare events in MTBC. They occur approximately every 3 kb of DNA sequence (Comas *et al.*, 2010). Hence, there is about three times less genetic variation in MTBC than in humans (Wheeler *et al.*, 2008). Together with other bacterial pathogens such as *Yersinia pestis*, *Bacillus anthracis*, *Mycobacterium leprae* or *Salmonella enterica* serovar Typhi, MTBC has been referred to as "genetically monomorphic", even though MTBC harbours significantly more variation than other monomorphic bacteria (Achtman, 2008). One interesting observation is that the large majority of SNPs in MTBC occur as singletons (variants that only occur in a single strain). No clear explanation currently exists for this phenomenon, but a possible effect of background selection has been proposed (Hershberg *et al.*, 2008; Pepperell *et al.*, 2010).

3.4. SNPs are phylogenetically informative in MTBC

The comparably low frequency of SNPs and limited ongoing horizontal gene transfer in MTBC result in low levels of homoplasy (i.e. the independent occurrence of the same SNP in phylogenetically unrelated strains) (Hershberg *et al.*, 2008; Comas *et al.*, 2010). Hence, SNPs represent robust markers for inferring phylogenies and for strain classification (Gagneux *et al.*, 2007). SNPs can also be used to measure evolutionary distances between strains, i.e. to estimate the time of divergence of strains from their genetic distance, if a mutation rate is known (Ford *et al.*, 2011).

The first step in generating SNP data is referred to as SNP discovery and usually involves comparative sequencing of multiple genes or whole genomes in two or more strains of interest. Once a set of SNPs has been identified, these can then be used to screen additional strains using one of many available SNP-typing platforms (Kim *et al.*, 2007; Wang *et al.*, 2007). Importantly, the usefulness of a given SNP-set is dependent on the amount of effort put into the initial identification of these SNPs, in particular on the number of strains included at the discovery stage. Poor representation of the relevant strain diversity during the discovery process will result in a biased set of SNPs which

can lead to erroneous phylogenetic inferences; this phenomenon is known as "phylogenetic discovery bias" and has been discussed in detail elsewhere (Pearson *et al.*, 2004; Alland *et al.*, 2003; Achtman, 2008; Smith *et al.*, 2009).

In 1997, Sreevatsan and colleagues sequenced 26 drug resistance-associated genes in 842 clinical isolates and identified two nSNPs which were unrelated to drug resistance (Sreevatsan *et al.*, 1997). Based on these two SNPs, a classification scheme into three Principle Genetic Groups was developed, which has been widely used in the past. The whole genome sequence of H37Rv published in 1998 (Cole *et al.*, 1998) established a first reference against which other genomes could be compared. In 2002, CDC1551 was sequenced (Fleischmann *et al.*, 2002) allowing for a first insight into the genome-wide SNP-diversity in M. tuberculosis; 1,075 SNP differences were found between the two strains. The whole genome sequence of *Mycobacterium bovis* AF2122/97 (Garnier *et al.*, 2003) and the partial genome of the "Beijing" strain 210 became available shortly thereafter, generating an increased collection of SNPs for genotyping purposes. Two studies took advantage of the availability of these four genome sequences and indentified phylogenetically informative SNPs to genotype large strain collections and identify phylogenetic groups within MTBC. However, as outlined above, both of these studies suffered from a phylogenetic discovery bias due to the low number of genomes used for SNP discovery. Hence, the resulting phylogenies presented by these groups were similarly affected by this problem (Achtman, 2008; Smith *et al.*, 2009). By contrast, three other studies used de novo sequencing of six genes (Baker *et al.*, 2004), 56 genes (Dos Vultos *et al.*, 2008) and 89 genes (Hershberg *et al.*, 2008), respectively, in large strain collections and produced unbiased phylogenies which were more congruent with those inferred using other methods, i.e. genomic deletions (Gagneux *et al.*, 2006b). However, even the phylogeny by Hershberg *et al.* based on 89 whole gene sequences was unable to completely resolve all the branches within the tree. In 2010, Comas *et al.* published the first whole-genome based global phylogeny of human-associated MTBC (Comas *et al.*, 2010). As expected given MTBC's largely clonal population structure, this genome-based phylogeny was highly congruent with those published previously, but all main lineages were now clearly resolved (Figure 3.2). This phylogeny has since then been used as a reference for phylogenetic studies including an increasing number of MTBC strains (Bentley *et al.*, 2012; Firdessa *et al.*, 2013).

The growing number of individual gene sequences and whole genomes in MTBC has already resulted in the identification of thousands of SNPs, which in recent years have been incorporated into various SNP-typing schemes. Because MTBC is largely clonal, for each of the main lineages "diagnostic" or "canonical" SNPs can be extracted and used as markers to assign unknown isolates to a particular phylogenetic group or lineage. Various

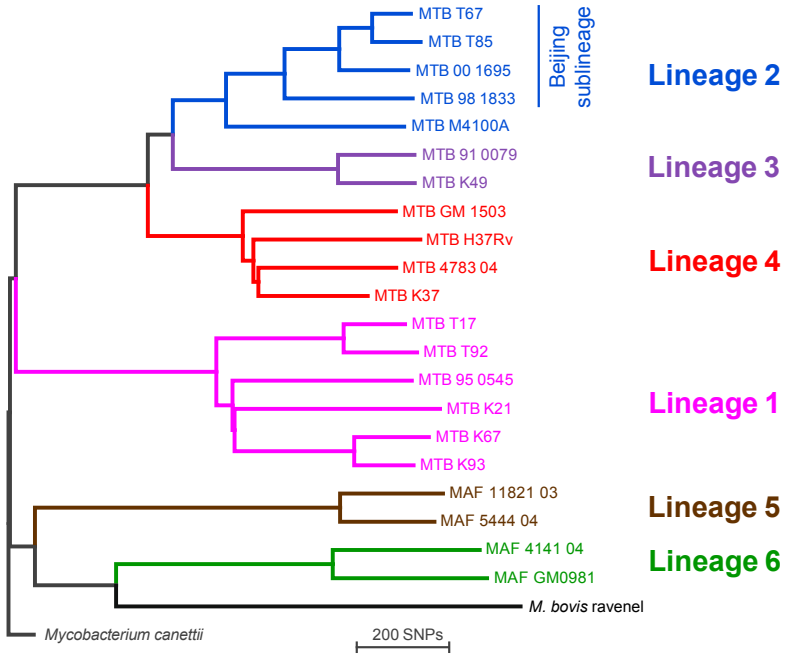


Figure 3.2.: **Phylogenetic tree of 22 whole genome sequences of MTBC and *M. canettii* used as the outgroup.** Modified from (Stucki *et al.*, 2012; Bentley *et al.*, 2012).

SNP-typing assays have been developed on different platforms (Sreevatsan *et al.*, 1997; Baker *et al.*, 2004; Filliol *et al.*, 2006; Gutacker *et al.*, 2006; Bergval *et al.*, 2008; Abadia *et al.*, 2010; Bouakaze *et al.*, 2010; Bouakaze *et al.*, 2011; Bergval *et al.*, 2012; Stucki *et al.*, 2012), and the latest assays can interrogate multiple SNPs simultaneously in one reaction. Examples include assays designed for the MassArray and the Luminex platform (Bouakaze *et al.*, 2011; Stucki *et al.*, 2012; Bergval *et al.*, 2012). These methods provide the robust phylogenetic framework necessary for genotype-phenotype- and other association studies (Comas *et al.*, 2009a). For example, there is increasing evidence for phenotypic differences between strains, and studies need to be conducted to assess if these differences are due to genetic differences between MTBC clades (Coscolla *et al.*, 2010).

Even though a lot of progress has been made in our understanding of the global phylogenetic diversity of MTBC, much remains unknown with respect to both human- and animal-associated MTBC diversity (Gagneux, 2012). For example, in addition to the six main human-associated MTBC lineages, a seventh lineage was recently discovered in TB patients from Ethiopia (Firdessa *et al.*, 2013). Similarly, new animal-associated lineages have been identified in several African mammal species, indicating that more MTBC diversity exists (Mostowy *et al.*, 2004; Alexander *et al.*, 2010). Moreover, in addition to focusing on differences between the main lineages of MTBC, we also need to look deeper into the diversity within individual lineages. For example, the Beijing family of strains is

a sublineage of Lineage 2 (Figure 3.2) and is currently a strong focus of research because of its association with drug resistance (Borrell *et al.*, 2009), hypervirulence in animal models (Caws *et al.*, 2006; Parwati *et al.*, 2010), and recent expansion in human populations (Cowley *et al.*, 2008; Spuy *et al.*, 2009). Moreover, even though phenotypic diversity has been associated with the different MTBC lineages (e.g. in the elicited innate immune responses), individual strains within these lineages exhibit a wide range of phenotypes (Portevin *et al.*, 2011), suggesting that in addition to strain-specific effects, sub-lineage structure should also be considered (Schürch *et al.*, 2011; Wada *et al.*, 2012). Only with a full understanding of the nature and phenotypic consequence of MTBC diversity will we be able to properly evaluate new diagnostics, vaccines and treatment options (Gagneux *et al.*, 2007).

To achieve a more comprehensive view of the global diversity of MTBC, we propose an iterative process, in which, first, genome sequencing of the most diverse strains is performed to identify new phylogenetically informative SNPs. These SNPs are then used to genotype large collections of strains, whereby some strains will be classified into known lineages and others identified as novel. Genome sequencing of these unclassified strains will then define their phylogenetic positions, identify new lineages and corresponding signature SNPs, which can be used in a following round of genotyping.

In the future, genome sequencing is likely to replace any sort of genotyping in MTBC, including SNP-typing. While SNP-typing is an ideal tool for phylogenetic strain classification in MTBC, it does not have the necessary resolution for standard molecular epidemiological investigation such as defining chains of transmission or differentiating cases of relapse from re-infection; MIRU-VNTR in combination with spoligotyping is still the gold standard for these applications (Supply *et al.*, 2006). Genome sequencing, on the other hand, generates a complete "barcode" of a strain, including the evolutionary background, drug resistance mutations or virulence-associated polymorphisms, and at the same time provides high resolution for transmission studies (Schürch *et al.*, 2010; Gardy *et al.*, 2011). Yet, until large-scale genome-sequencing becomes more readily available in standard laboratories, SNP-typing in MTBC will continue to play an important role in TB research and control.

3.5. The functional consequences of SNPs

In addition to being useful phylogenetic markers, SNPs carry functional information. The best-characterized "SNPs"‘ in MTBC are drug resistance-conferring mutations. Drug resistance in MTBC is largely caused by single nucleotide mutations (Musser, 1995; Telenti,

1997; Ramaswamy *et al.*, 1998; Riska *et al.*, 2000). Many drug resistance-conferring mutations have been identified, and are publicly available in the TBDReamDB database (Sandgren *et al.*, 2009) (currently containing information on 1447 mutations relevant for most anti-TB drugs (Table 3.1). This kind of molecular information is crucial for the development of new and faster diagnostic methods to detect drug resistance. While for the first-line anti-TB drugs, the most important drug resistance-conferring mutations have been identified and incorporated into promising new diagnostic tools (Hillemann *et al.*, 2005; Boehme *et al.*, 2010; *Hain Lifescience - Mycobacteria* 2012), many mutations remain unknown, including many of those causing resistance to the 2nd-line drugs. Besides the mutations causing drug resistance, other associated mutations could also be targeted in the future. For example, two recent studies have shown that compensatory mutations in the RNA polymerase of MTBC can contribute to the fitness of rifampicin-resistant strains (Comas *et al.*, 2011a; Casali *et al.*, 2012). While the molecular mechanisms involved remain to be determined, other mutations associated with drug resistance (e.g. compensatory mutations) could be used for molecular diagnostics even if they are not directly responsible for the drug resistance phenotype.

Table 3.1.: **Relevant SNP databases for MTBC.** Databases already containing SNP data of MTBC, and examples of relevant SNP-databases of human variation, which could serve as examples for a future MTBC SNP-database.

MTBC SNP-databases

Name	Species	# of SNPs (# of genomes[1])	Features	URL (Reference)
dbSNP	*M. tuberculosis*	40,303 MTBC[2] (3827 MTBC samples in SRA)	NCBI curated relational SNP-database for all organisms, MTBC SNPs are not annotated	`http://www.ncbi.nlm.nih.gov/projects/SNP/` (Sherry *et al.*, 2001)
TBDB	*M. tuberculosis*	23,795 (25 MTBC[3])	Relational database with various MTBC data sets such as expression, diversity, proteins, ChiPSeq, publications etc. SNPs are well annotated and interlinked with other tables, but not updated.	`http://www.tbdb.org/` (Reddy *et al.*, 2009)
PATRIC	*M. tuberculosis*	0 (75 MTBC)	Extensive relational database for various bacterial pathogens linking genomic data with NIH disease, epitopes etc. SNP database in preparation.	`http://www.patricbrc.org/portal/portal/patric/Home` (Gillespie *et al.*, 2011)
MGDD	*M. tuberculosis*	n.a. (6 MTBC)	One-by-one comparison of 6 MTBC strains; not updated since 2008	`http://mirna.jnu.ac.in/mgdd/` (Vishnoi *et al.*, 2008)
MTCID	*M. tuberculosis*	n.a.	List of mainly drug resistance conferring mutations	`http://ccbb.jnu.ac.in/Tb/` (Bharti *et al.*, 2012)
TBDReaMDB	*M. tuberculosis*	1447 (0)	Drug resistance conferring mutations	`http://www.tbdreamdb.com/` (Sandgren *et al.*, 2009)

Relevant SNP-databases from other organisms that could serve as example databases

Name	Species	# of SNPs (# of genomes[1])	Features	URL (Reference)
dbSNP	Various	53,558,214 human SNPs[4] (34'970 human samples in SRA)	NCBI curated SNP database, interlinked with many other databases; can also contain indels, IS sequences, microsatellites	`http://www.ncbi.nlm.nih.gov/projects/SNP/` (Sherry *et al.*, 2001)
ENSEMBL	Various	synchronized with dbSNP	SNP database with extensive (graphical) links to other features (genomic context, genes, population genetics, phylogenetic context, flanking sequence, etc.)	`http://www.ensembl.org/index.html` (Flicek *et al.*, 2011)
PASNP	*H. sapiens*	55,998 SNPs (from 1719 individuals)	Pan-Asian SNP database with extensive graphical browsing	`http://www4a.biotec.or.th/PASNP` (Ngamphiw *et al.*, 2011)
JSNP	*H. sapiens*	197,195 human SNPs (n.a.)	Japan SNP database with SNP data from genotyping	urlhttp://snp.ims.u-tokyo.ac.jp/ (Hirakawa *et al.*, 2002)
HapMap genome browser	*H. sapiens*	1,440,616 genotyped SNPs in 1184 individuals[5]	Haplotype database	`http://hapmap.ncbi.nlm.nih.gov/`("Integrating common and rare genetic variation in diverse human populations" 2010)
GWAS central	*H. sapiens*	62,322,744 entries (n.a.) from dbSNP build 135	Former HGVbase, links human genetic association studies with dbSNP rs#	`https://www.gwascentral.org/` (Thorisson *et al.*, 2009)
topoSNP	*H. sapiens*	27,417 SNPs (publication 2004)	Mapping of human non-synonymous SNPs from OMIM and dbSNP to protein structures	`http://gila.bioengr.uic.edu/snp/toposnp/` (Stitziel *et al.*, 2004)

[1] Complete genomes or resequenced

[2] Number found in dbSNP for keyword "*Mycobacterium tuberculosis*", but rs#s are invalid links

[3] MTBC genomes under "Diversity sequencing" on `tbdb.org`

[4] Number of RefSNP Clusters (rs#s) in build 137 as found in `http://www.ncbi.nlm.nih.gov/SNP/snp_summary.cgi`

[5] As in HapMap3 ("Integrating common and rare genetic variation in diverse human populations" 2010)

Although drug resistance-conferring mutations are most important as far as TB control is concerned, most SNPs in MTBC are unrelated to drug resistance. Indeed, thousands of SNPs have already been identified across MTBC strains and lineages. Many of these might translate into cellular and/or clinically relevant phenotypic changes. This is particularly true given the fact that up to two thirds of SNPs in MTBC are non-synonymous, which is unlike most other organisms in which sSNPs predominate (Hershberg *et al.*, 2008). The reason for the high proportion of nSNPs in MTBC is unclear but has been proposed to be the consequence of the relatively short evolutionary age of MTBC (i.e. purifying selection has not had time to remove nSNPs which on average are slightly deleterious) (Rocha *et al.*, 2006). Alternatively, it might reflect a reduction in the efficacy of purifying selection as a consequence of increased genetic drift in MTBC (Hershberg *et al.*, 2008). While intragenic and sSNPs are not necessarily neutral (e.g. they can affect regulatory regions), the effect of nSNPs is likely to be more pronounced. Yet, in contrast to drug resistance-conferring mutations, the effects of these nSNPs are less evident and likely to be more subtle, rendering the study of the biological and epidemiological consequences of these SNPs difficult.

One possible way forward is to use *in silico* predictions of the effects of SNPs. There are a number of freely available tools that can be used for this purpose. A prominent example is Sorting Intolerant From Tolerant (SIFT), a tool that, based on sequence homology and amino acid properties, predicts how much a given polymorphism might affect the function of the corresponding protein (Kumar *et al.*, 2009). Other tools include ANNOVAR (Wang *et al.*, 2010) and SVA (`http://www.svaproject.org`). In MTBC, such an approach has recently been used to look at variation in mammalian cell entry (mce) operons (Pasricha *et al.*, 2011). Ultimately however, polymorphisms predicted to effect protein function will have to be experimentally confirmed using other tools (Lamrabet *et al.*, 2012).

Adding another level of complexity, which so far has been largely ignored, are the possible epistatic interactions between SNPs in the same genome. A good example of this phenomenon are compensatory mutations that interact with corresponding drug resistance-conferring mutations (Comas *et al.*, 2011a; Casali *et al.*, 2012). Similar effects might be occurring among other mutations (Borrell *et al.*, 2011). Hence, the phenotypic characteristics of a given strain genetic background will depend on both the individual mutations as well as on the interactions between these mutations. Another important characteristic of MTBC is the fact that horizontal gene exchange is comparably rare and as a consequence, MTBC exhibits a largely clonal population structure (Supply *et al.*, 2003; Liu *et al.*, 2006; Hirsh *et al.*, 2004). Hence, all SNPs in a particular strain's genome are linked (i.e. they are in linkage disequilibrium), which makes attributing the phenotypic behaviour of a

given strain to a particular mutation non-trivial.

In summary, thousands of SNPs are being identified in MTBC thanks to the new DNA sequencing technologies. We can use these SNPs for phylogenetic and population genetic analyses and to study drug resistance. As for the large majority of SNPs, we do not know what their functional impact might be (partially also because we do not know the functions of the relevant genes). The functional consequences of these SNPs, and their potential for driving phenotypic differences between strains should be studied in the future. In the meantime, most ongoing SNP work in MTBC largely consists of discovering and cataloguing new SNPs. Let us now discuss some of the technicalities involved in these processes.

3.6. How do we discover new SNPs in MTBC?

In the upcoming years, we expect whole genome sequencing to at least partially replace all previous genotyping methods for MTBC. So far, large-scale DNA sequencing projects have usually been performed by specialized Sequencing Centres, but new benchtop sequencing devices increasingly allow for "do-it-yourself" approaches in the standard laboratory (Köser *et al.*, 2012b). In this section, we elaborate on some of the technical aspects of NGS genome analysis, with a particular focus on the workflow during the bioinformatics analysis. The DNA sequencing technologies themselves have been covered by several recent reviews (Ansorge, 2009; Metzker, 2010; Lee *et al.*, 2012).

Panel A of Figure 3.3 shows one possible data analysis pipeline for Illumina short read data, as currently used in our laboratory (Comas *et al.*, 2010; Comas *et al.*, 2011a). Starting with a `fastq` file with millions of short sequence reads (50-200 bp in length), different software tools are used to align each read to a reference genome and to call variant positions (*Maq* 2012; Li *et al.*, 2009; Sanger Institute, 2012b; *Samtools* 2012). As a reference genome we generally use the H37Rv genome with all known variant positions replaced by the hypothetical ancestral allelic states, representing a putative reconstructed ancestor of all MTBC (Comas *et al.*, 2010). Each nucleotide position differing from the reference, i.e. each SNP, can be annotated with gene, amino acid change and several other features (Wang *et al.*, 2010). Throughout the workflow, a number of filtering steps are used to remove SNPs showing low confidence. These include SNPs in repetitive regions of the genome, including *PE/PPE* genes and insertion sequences (panel B in Figure 3.3). A list of SNPs per strain is given as an output. Similar to single nucleotide variants, other polymorphisms such as small insertions and deletions can be analyzed, but will not be further discussed here.

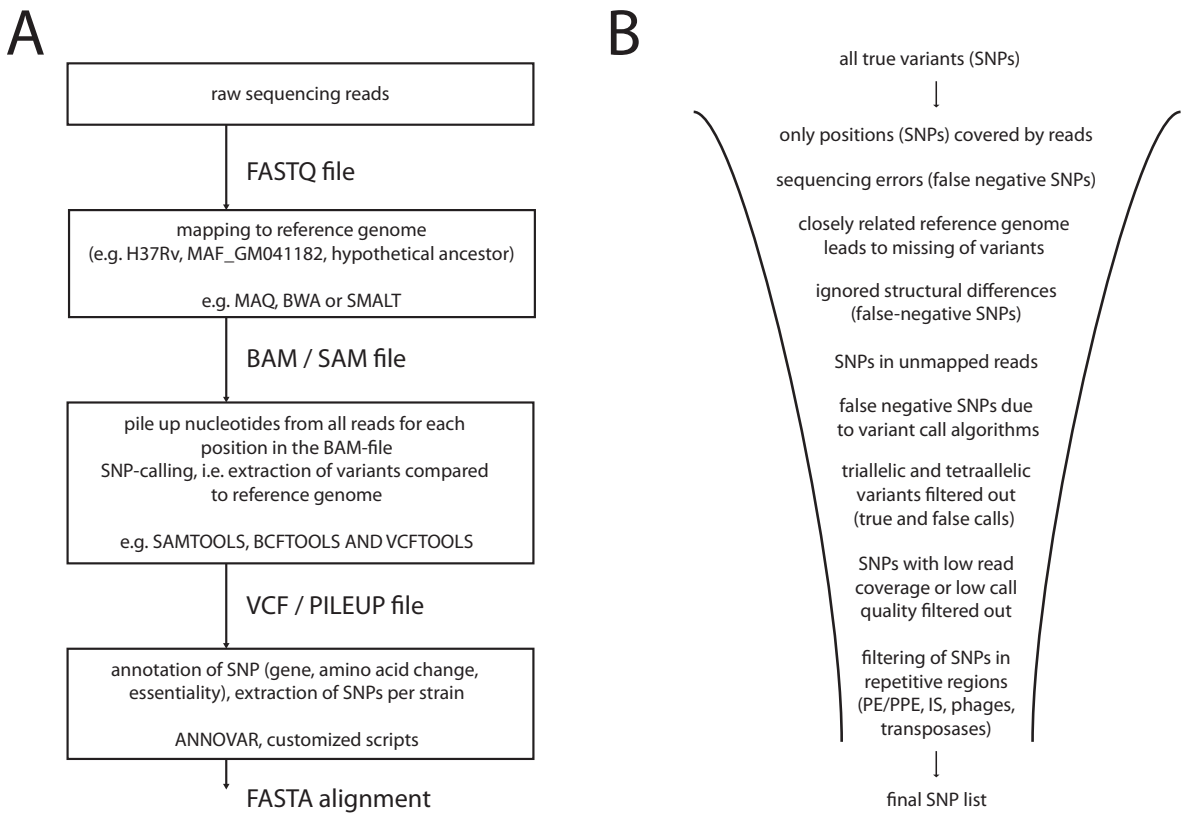


Figure 3.3.: **Example of a genome (re-) sequencing analysis pipeline for MTBC and reduction of the number of SNPs due to filtering. A**. The whole-genome sequencing data analysis pipeline for SNP calling as currently used in our laboratory. Input data are Illumina short reads in `fastq`-format. Outputs are single nucleotide variants per strain compared to the hypothetical ancestor of MTBC. **B**. Schematic illustration of steps where (true or false-positive) SNPs remain undiscovered or are lost due to filtering.

The number of called SNPs is therefore a result of combining different algorithms. Panel B of Figure 3.3 shows schematically where true- or false-positive SNPs are lost in the filtering process (i.e. they are filtered out by the software according to the particular parameter settings). This particular approach aims at reducing the number of false-positive calls. Often, the question as to what SNPs are true (as opposed to false-positives) remains, and ideally, SNPs should be confirmed by other methods such as Sanger sequencing. The issue of data quality is key in all sequencing projects, and there is a general trade-off between data quality (i.e. the stringency of filter settings) and the number of true SNPs identified. Moving forward, there is a need for minimum requirements for SNP data generated by NGS. One possible approach could be that a particular SNP has to be confirmed by at least one independent method. However, such an approach might become increasingly difficult as the number of SNPs increases beyond a few dozen. Also, filtering thresholds will have to be defined, and difficulties resulting from ambiguous base calls or multiallelic variants discussed.

The NGS data pipeline shown in Figure 3.3 consists of a combination of several command-line tools and customized scripts. Even though scripting can automatize the processing of one or many genomes, running these tools requires a certain level of bioinformatic expertise. But if whole genome sequencing is to be used more broadly in clinical settings, we need good software packages with automatic SNP-calling and SNP-annotation (Köser *et al.*, 2012b). There are several (not MTBC specific) public platforms for customized and semi-automated NGS data analysis, with the most prominent and feature-rich among them known as "Galaxy" (Goecks *et al.*, 2010). Galaxy allows graphically guided analysis of NGS data. The DNA data bank of Japan (DDJB) also features a NGS analysis platform, which is publicly available and makes automatic integration into the DDBJ-archive possible for publication (`http://p.ddbj.nig.ac.jp/pipeline/Login.do`). These platforms are designed as customized combinations of tools that can be controlled with a graphical interface. However, what is actually needed for the TB-community is a "one-click" tool to upload NGS data in `fastq` format, and to get a list of polymorphisms as an output (see Box 1). Moreover, the rapidly increasing amount of SNP data generated through all the ongoing and future NGS projects should ideally be centralized in a user-friendly and well-curated database.

3.7. The need for a new SNP database for MTBC

So far, most SNP data in MTBC have been computed and stored on local workstations, and only made available upon publication. For the raw NGS reads, NCBI, EBI and DDJB

have created Sequence Read Archives (SRA) where these data can be deposited (`http://www.ncbi.nlm.nih.gov/sra`, `http://www.ebi.ac.uk/ena/home`, `http://trace.ddbj.nig.ac.jp/dra/index_e.shtml`). These archives contain the raw sequencing reads in SRA format, which can be downloaded and converted to `fastq` files (`http://www.ncbi.nlm.nih.gov/books/NBK47537/`). Raw sequence reads are required whenever a re-analyses of variants (e.g. with new software algorithms) are performed.

Once the polymorphisms have been extracted from the raw NGS data, they have to be made available upon publication. With previous projects based on classical Sanger sequencing, it was often possible to present SNP data as tables or spreadsheets, but the need for an online centralized database becomes obvious when considering the large amount of SNP- and associated metadata coming out of current and future MTBC NGS projects. The number of SNPs identified in a particular NGS project will likely be between several hundreds and several tens of thousands, depending on the number of MTBC genomes analyzed. This makes the use of lists or spreadsheets problematic and error-prone. Furthermore, to make use of this SNP information, researchers need to have access to the biological context, and to be able to interlink SNP data with other metadata as well as with other databases containing e.g. information on protein structure.

At least five existing databases harbour information on polymorphisms in MTBC (Table 3.1). Four of them were designed to contain data on MTBC SNPs only. The most important and multifunctional among these is probably the Tuberculosis Database (TBDB) (Reddy *et al.*, 2009). It currently contains 23,795 SNPs extracted from 25 MTBC genomes (Table 3.1). These SNPs can be viewed as pair-wise or multiple comparisons of genomes with a variety of display options. SNPs can also be viewed in form of sequence alignments and downloaded as text files. SNPs can be separated by coding- and non-coding regions, or clustered by functional enrichment (e.g. all polymorphisms per COG-category gene). Drug resistance mutations can be specifically sought for. Each SNP is linked to the annotated gene it falls in, which is then linked to other databases. Phylogenetically relevant SNPs, i.e. potential markers, can be extracted by generating lists of SNPs shared by members of one lineage, and excluding SNPs shared with members of other lineages.

Another database is TBDReamDB which focuses on drug resistance-conferring mutations (Sandgren *et al.*, 2009). It is probably the most complete repository for drug resistance mutations in MTBC. It currently features 1447 mutations and is regularly updated with information from new publications. MGDD (Vishnoi *et al.*, 2008) is a database to compare SNPs across 6 MTBC genomes by gene name, nucleotide position or base change. MTCID (Bharti *et al.*, 2012) is another database comparable in function to MGDD, and has in addition geographical mapping of SNPs implemented. Both MGDD and MTCID

focus on drug resistance-associated mutation, but not exclusively. However, they have limited search functions and it is unclear if these databases are still being updated. The SNP database with the largest number of MTBC SNPs is dbSNP of NCBI (Sherry *et al.*, 2001). Currently about 40,000 SNPs are obtained using the search query "Mycobacterium tuberculosis". Unfortunately, these SNPs are not annotated, and the origin of the data is unclear. But dbSNP is likely to be the most powerful database for SNP data in other organisms, and serves as reference database for the known genetic variants in *Homo sapiens*. Several features of dbSNP established for human variation could inform the establishment of a new SNP database for MTBC.

Several large databases of human variants are available (Table 3.1). dbSNP comprises currently 53,558,214 human variants. Most other databases reference their SNPs to dbSNP by the respective rs# number. This is an established annotation system that functions as follows: whenever a SNP of any species is uploaded to dbSNP (from publications or directly from the user) it is assigned a unique and position-independent submitted SNP ID number (ss#) and mapped to the reference genome (position on the contig). In a next step, the ss# is linked to a unique new RefSNP ID number (rs#), or is assigned to an existing rs# if a SNP at this genomic position was found and uploaded before. An rs# can therefore contain multiple ss# numbers, which are found in a table shown in each rs# record. The rs# is also position-independent, but each record contains the genomic position of the SNP. The SNP, i.e. the rs#, is then fully annotated, and linked to other NCBI resources such as gene context or publication source. Any attribute field or database can be linked to a specific rs#. In the next "build" of the database, the SNP is then incorporated (Kitts *et al.*, 2011). Other databases containing human data can use the dbSNP data to build, link and annotate their own SNP information. A selection of other large SNP databases potentially relevant for MTBC is listed in Table 3.1. These databases have different contents (i.e. only a subset of the dbSNP entries or refer to restricted populations as in PASNP or JSNP), different structures and are tailored to match different requirements, such as the HapMap genome browser (`http://hapmap.ncbi.nlm.nih.gov/whatishapmap.html`), or the topoSNP database (Stitziel *et al.*, 2004), which maps SNPs to protein structures. All of these databases have specific fields, which could be included in a future MTBC SNP database. In addition to all the descriptive information of a given SNP, human dbSNP entries contain data about the frequency of the SNP in different human populations. Drawing a line to a future MTBC SNP database, frequency data could be important for drug resistance-conferring mutations, and could be calculated e.g. by lineage (e.g. the frequency of *rpoB* mutations in each lineage, calculated and updated automatically). To

allow calculating frequencies of mutations, a database needs to allow for multiple uploads of the same SNP. In the final chapter of this review we discuss some other features that a new MTBC SNP database should include (Box 1).

3.8. Features of a new MTBC SNP database

Given the existing features of TBDB (discussed above), this database represents an ideal starting point for an extended SNP database for MTBC (Box 1). TBDB already includes important aspects such as the relational tables and annotations. Unique and highly valuable modules such as the phylogenetic context could be extended to deal with larger numbers of taxa (strains) and characters (SNPs). So far, SNPs in TBDB are identified based on their position in the reference genome, but with larger numbers of genome sequences and characters, as well as alternative reference genomes used by different researchers, a unique SNP ID will become necessary, similar to what has been implemented in dbSNP. This would allow for unambiguous communication across the research community, solve the problem of SNPs whose positions are not found in a particular reference genome (e.g. in regions where large deletions in H37Rv occur), and allow for specifying a position of interest on more than one genome (e.g. when referring to the recently finished genome of *Mycobacterium africanum* GM041182). Moreover, a new MTBC SNP database should be frequently updated. Regular builds are also important in case of new annotations in the reference genome(s). The database should include as many informative fields as possible, but should not store any redundant information available in other databases. Many important fields are already present in TBDB and could be adopted into an extended version (Figure 3.4). These fields include the "standard" data on position, nucleotide change, gene annotation, and amino acid change (synonymous/non-synonymous). Additional fields could include "essentiality" of the corresponding gene (based on experimental evidence (Sassetti *et al.*, 2003a)), validation (methods used to confirm polymorphism and to exclude sequencing errors), source of data (publication, upload, diversity sequencing project, etc.), associations with drug resistance, phylogenetic context (e.g. "Lineage 4 marker"), and data quality scores (e.g. phred score or coverage depth from the corresponding sequencing project). Other metadata such as clinical associations (e.g. virulence, transmission, vaccine efficacy, drug treatment) could be annotated in attribute fields as they become available. Other fields could be developed based on functional predictions of SNPs, and experimental confirmation of any phenotypic effects. Here, the SIFT algorithm (see above) and other related approaches could be implemented more widely. The way by which a particular SNP was discovered has to be defined (sequencing method,

confirmation by other methods), and the adjacent sequence up- and down-stream of the SNP position should be shown for unambiguous identification.

Regarding the original source of the SNP data, the ideal database should harbour both the raw sequencing data and the corresponding polymorphism data. This would allow a direct link to the source data, and – if needed – the possibility to retrieve the original sequencing reads for quality assessment or application of alternative mapping algorithms or other analytical approaches. For example, `fastq` files could be uploaded by users, with SNPs called automatically and put into the context of the existing SNP data (see also Box 1). Following a set of quality criteria, the new SNPs would then be automatically merged into the main SNP database. SNPs should also be linked to relevant entries in other databases. There are a variety of databases that store MTBC-specific information (Sharma *et al.*, 2011), and potential links are shown in Figure 3.4. Some of these databases include existing and highly accessed databases such as gene annotations on TubercuList or expression data on TBDB. Others could include new functions such as visualizing the spatial location of a SNP on the corresponding protein structure. Moreover, two large Systems Biology consortia are currently working on TB (`http://www.systemtb.org/` and `http://www.broadinstitute.org/annotation/tbsysbio/home.html` (Aderem *et al.*, 2011)). Linking MTBC SNP information with transcriptional, proteomic, and metabolomic data generated through these consortia should allow for a more comprehensive understanding of TB biology.

As already mentioned, frequency data on SNP distribution should be included. As the number of available whole-genome sequences increase, we will need sophisticated tools to visualize allele frequencies in a geography-dependent manner. The high clonality of MTBC and the strong genome-wide linkage disequilibrium between individual SNPs in a given genome will have to be considered in this respect. Finally, the SNP database should be easily searchable by position, reference number, keywords, or free text. All data contained in the database should also be downloadable as a bulk. In dbSNP, this is achieved by building a local copy of the database. For MBTC, as the number of SNP will be considerably smaller, this could be achieved by a bulk download in text format.

The requirements and opportunities for a new SNP database for MTBC are manifold but will not be easy to implement (Box 1). Ideally, joint efforts between NCBI, TBDB and PATRIC are needed to achieve this goal. PATRIC (Gillespie *et al.*, 2011) is a platform for storage and exchange of bacterial data, featuring a variety of tools including genome browsing, BLAST, comparative pathways, and protein annotations. PATRIC has already compiled a lot of mycobacterial data, and there are plans to host bacterial SNP data as well (B. Sobral, personal communication). Thanks to the large amount of information

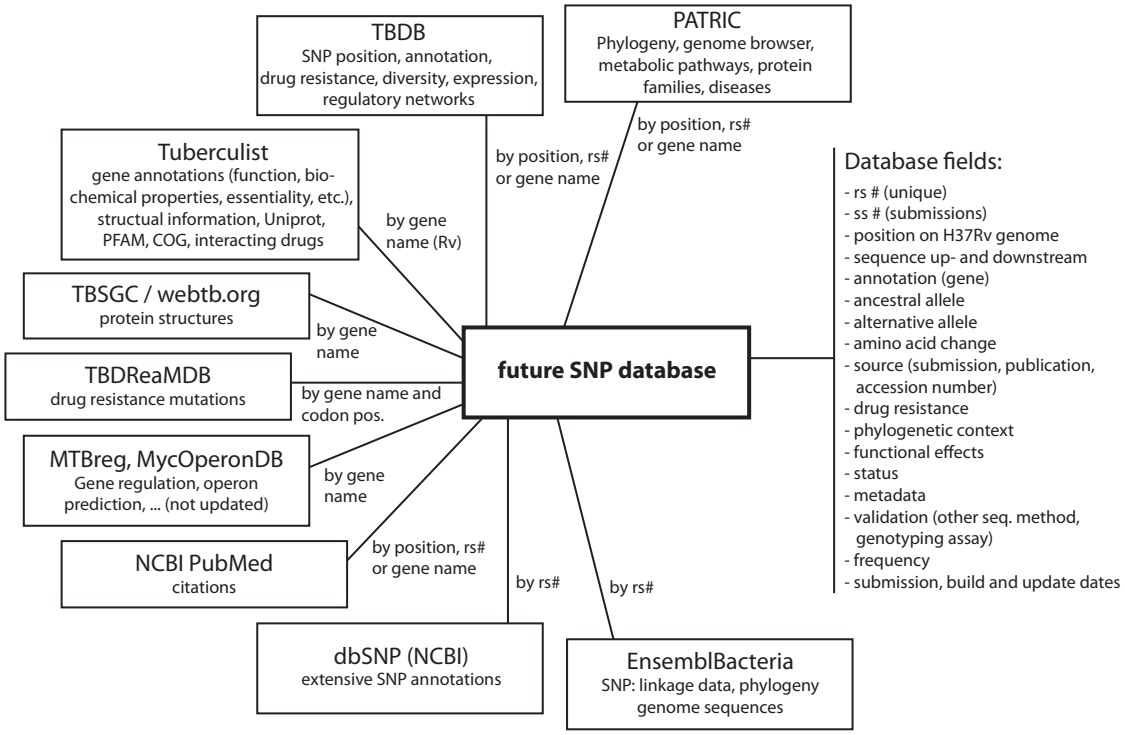


Figure 3.4.: **A simplified scheme of a future relational SNP-database for MTBC SNPs, including links to other databases.** "rs" numbers are unique IDs from dbSNP (reference cluster). This proposal for a new MTBC SNP database is independent of the platform hosting the central database.

already included, a joint venture between TBDB and PATRIC seems like an ideal way forward to establish a new SNP database for MTBC.

Box 1. "Wish list" for a new SNP database for MTBC. A future database for SNP annotation and genome analysis to serve the needs of the TB research community should include (amongst other):

- Position-independent reference numbers for SNPs for unambiguous communication between laboratories and data management. The rs# numbers of NCBI provide a suitable starting point.
- Implementation of a "one-click" genome analysis pipeline to extract SNPs and indels from uploaded `fastq`-files. Fastq file would be automatically processed, all SNPs given as output, compared with the existing ones in the SNP-database, and annotated. The phylogenetic position of the user-genome would be given (see below). The SNP data from the user genome could then be shared with the database to be included in a next build. Multiple genomes sequences (i.e. multiple `fastq`-files) can be uploaded.
- Phylogenetic context: MTBC strains harbouring a particular SNP can be shown as a list. Strains of interest (e.g. all Lineage 4 strains) can be selected and shared SNPs shown ("SNPs per selected clade").
- Generation of a phylogenetic tree with all available genomes or SNPs, and possibility of mapping a genome uploaded by the user.
- Visualization of SNPs as an alignment of all strains piled up (as in TBDB).
- A genome map highlighting SNP positions.
- Visualization of sequence upstream- and downstream of the SNP.
- Genotyping assays for selected SNPs: either published, or automatically calculated (e.g. primer pairs for PCR amplification of the SNP-relevant genomic region).

3.9. Conclusions

NGS studies of MTBC clinical isolates are discovering thousands of SNPs. Studying the functional effects of these SNPs and their association with phylogenetic clades should become an increasing part of the research portfolio. MTBC consist of a diverse population of strains, and this diversity should be considered when developing new tools and strategies to combat TB. A new, extended, and well-curated database is necessary to accommodate these rapidly accumulating SNP data in a user-friendly and integrated format. Ideally, dbSNP/NCBI, TBDB and PATRIC should join forces and play major roles in the development of such a database. After establishing the basic framework of such a database,

specific needs and wishes of the community can be discussed and incorporated (Box 1). The aim of this review is to contribute to the discussion.

3.10. Acknowledgements

We thank Mireia Coscollà and Iñaki Comas as well as the other members of our group for the inspiring discussions and comments on the manuscript. The work in our laboratory is supported by the Swiss National Science Foundation (grant number PP0033-119205) and the National Institutes of Health (AI090928 and HSN266200700022C).

4. Two new rapid SNP-typing methods for classifying *Mycobacterium tuberculosis* complex into the main phylogenetic lineages

David Stucki [1,2], Bijaya Malla [1,2], Simon Hostettler [1,2], Thembela Huna [3], Julia Feldmann [1,2], Dorothy Yeboah-Manu [4], Sonia Borrell [1,2], Lukas Fenner [5], Iñaki Comas [6,7], Mireia Coscollà [1,2], Sebastien Gagneux [1,2,*]

[1] Swiss Tropical and Public Health Institute, Basel, Switzerland

[2] University of Basel, Switzerland

[3] Division of Mycobacterial Research, National Institute for Medical Research, London, United Kingdom

[4] Noguchi Memorial Institute for Medical Research, University of Ghana, Legon, Ghana

[5] Institute of Social and Preventive Medicine, University of Bern, Bern, Switzerland

[6] Genomics and Health Unit, Centre for Public Health Research, Valencia, Spain

[7] CIBER Epidemiologìa y Salud Publica, Madrid, Spain

[*] Corresponding author

This paper has been published in *PLoS ONE* 2012, 7(7):p.e41253.

4.1. Abstract

There is increasing evidence that strain variation in *Mycobacterium tuberculosis* complex (MTBC) might influence the outcome of tuberculosis infection and disease. To assess genotype-phenotype associations, phylogenetically robust molecular markers and appropriate genotyping tools are required. Most current genotyping methods for MTBC are based on mobile or repetitive DNA elements. Because these elements are prone to convergent evolution, the corresponding genotyping techniques are suboptimal for phylogenetic studies and strain classification. By contrast, single nucleotide polymorphisms (SNP) are ideal markers for classifying MTBC into phylogenetic lineages, as they exhibit very low degrees of homoplasy. In this study, we developed two complementary SNP-based genotyping methods to classify strains into the six main human-associated lineages of MTBC, the "Beijing" sublineage, and the clade comprising *Mycobacterium bovis* and *Mycobacterium caprae*. Phylogenetically informative SNPs were obtained from 22 MTBC whole-genome sequences. The first assay, referred to as MOL-PCR, is a ligation-dependent PCR with signal detection by fluorescent microspheres and a Luminex flow cytometer, which simultaneously interrogates eight SNPs. The second assay is based on six individual TaqMan real-time PCR assays for singleplex SNP-typing. We compared MOL-PCR and TaqMan results in two panels of clinical MTBC isolates. Both methods agreed fully when assigning 36 well-characterized strains into the main phylogenetic lineages. The sensitivity in allele-calling was 98.67% and 98.8% for MOL-PCR and TaqMan, respectively. Typing of an additional panel of 78 unknown clinical isolates revealed 99.2% and 100% sensitivity in allele-calling, respectively, and 100% agreement in lineage assignment between both methods. While MOL-PCR and TaqMan are both highly sensitive and specific, MOL-PCR is ideal for classification of isolates with no previous information, whereas TaqMan is faster for confirmation. Furthermore, both methods are rapid, flexible and comparably inexpensive.

4.2. Introduction

Genotyping of human-associated *Mycobacterium tuberculosis* complex (MTBC) plays an increasing role for understanding the epidemiology and biology of tuberculosis (TB) (Kato-Maeda *et al.*, 2011b; Coscolla *et al.*, 2010). On the one hand, genotyping techniques are key for standard molecular epidemiological investigations of TB such as defining chains of ongoing transmission and differentiating patient relapse from exogenous re-infection (Kato-Maeda *et al.*, 2011b). On the other hand, strain genotyping in MTBC is important for defining the evolutionary background of clinical isolates. There is mounting evidence that the strain genetic background might influence the outcome of TB infection and disease. Experimental studies have shown that clinical strains differ in immunogenicity and virulence, but whether this variation translates into comparable clinical phenotypes is unclear (Caws *et al.*, 2008; Nicol *et al.*, 2008; Thwaites *et al.*, 2008; Jong *et al.*, 2008; Coscolla *et al.*, 2010; Homolka *et al.*, 2010; Portevin *et al.*, 2011; Click *et al.*, 2012). To study the effect of strain variation and detect relevant genotype-phenotype associations, suitable phylogenetic markers and appropriate genotyping methods are required (Comas *et al.*, 2009b).

Several methods for genotyping of MTBC have been developed over the past years (Schürch *et al.*, 2012). Two of the most popular methods, spoligotyping (Kamerbeek *et al.*, 1997) and MIRU-VNTR (Supply *et al.*, 2006), rely on repetitive DNA elements, and are currently considered the gold standard for TB transmission studies (Kato-Maeda *et al.*, 2011b; Supply *et al.*, 2006). However, due to their propensity for convergent evolution and resulting homoplasies (Comas *et al.*, 2009b), inferring robust phylogenies using these methods can be problematic. Moreover, because of these homoplasies, classification of MTBC strains into robust phylogenetic lineages is not always possible and misclassification can occur (Fenner *et al.*, 2011; Flores *et al.*, 2007). In addition to mobile genetic elements (e.g. IS6110 (McEvoy *et al.*, 2007)), repetitive DNA (DR region (Groenen *et al.*, 1993), MIRU-VNTR (Supply *et al.*, 2000)), and large sequence polymorphisms (Tsolaki *et al.*, 2004; Gagneux *et al.*, 2006b), single nucleotide polymorphisms (SNPs) have become available as phylogenetic markers for MTBC. SNPs are ideal markers for genotyping of MTBC, as they represent unique events and show virtually no homoplasy (Comas *et al.*, 2009b). Several studies have reported the use of SNPs for phylogenetic classification of MTBC (Sreevatsan *et al.*, 1997; Baker *et al.*, 2004; Filliol *et al.*, 2006; Gutacker *et al.*, 2006; Bergval *et al.*, 2008; Hershberg *et al.*, 2008; Abadia *et al.*, 2010; Bouakaze *et al.*, 2010; Bouakaze *et al.*, 2011). However, no consensus has yet been reached as to what set of SNPs should be used for standard phylotyping of MTBC. Importantly, not all phylogenetically informative SNPs will be equally amendable to all possible SNP-typing platforms

(Wang *et al.*, 2007; Kim *et al.*, 2007). In MTBC, some of the methods that have been proposed are allele-specific PCR (Espinosa de los Monteros *et al.*, 1998), real-time PCR (Halse *et al.*, 2011), Sanger sequencing (Mestre *et al.*, 2011), SNaPshot minisequencing (Bouakaze *et al.*, 2010), iPlex Gold (SEQUENOM Inc.) (Bouakaze *et al.*, 2011) and MLPA (Bergval *et al.*, 2008). All of these methods differ in their technical requirements. Hence, depending on the particular technique used, different, but phylogenetically equivalent SNP sets will be required. In addition, these typing methods vary largely in their throughput, cost, and flexibility. With respect to the latter, we identified a lack of methods, which are at the same time flexible and rapid ($<$ 1 day / 96 samples), and amendable to a variable number of strains and a limited number of SNPs (5-50)

Recently, whole genome sequences of a global collection of 21 MTBC strains representing the six main lineages of human-associated MTBC became available (Comas *et al.*, 2010). The 9,037 SNPs identified in these genomes represent an ideal starting point for extracting diagnostic or "canonical" SNPs, which are lineage-specific, to design phylogenetically robust genotyping assays (Keim *et al.*, 2004). In this study, we present two new and complementary SNP-typing methods for MTBC. These methods are based on two phylogenetically equivalent sets of SNP markers that are specific for the 6 main human-associated lineages of MTBC. Additionally, we present SNPs specific for the clade comprising *Mycobacterium bovis* and *Mycobacterium caprae* (Hershberg *et al.*, 2008), and for the "Beijing"-sublineage of Lineage 2 (East-Asian) (Figure 1). "Beijing" strains are of special epidemiological interest because of their repeated association with drug resistance and hypervirulence in infection models (Caws *et al.*, 2006; Parwati *et al.*, 2010). The *M. bovis* / *M. caprae* specific SNP does not capture the other animal-associated strains, i.e. *M. pinnipedii*, *M. microti*, *M. orygis*, *M. mungi* and the Dassie bacillus.

The first SNP-typing method is MOL-PCR (multiplexed oligonucleotide ligation PCR) (Deshpande *et al.*, 2010), which uses allele-specific ligation for allele discrimination, singleplex PCR for signal amplification and fluorescent microspheres (beads) for the signal detection on a flow cytometer (Figure 2, see Methods). A Luminex device is needed; i.e. a flow cytometric platform for various nucleic acid and immunological assays (Dunbar, 2006; Lee *et al.*, 2004; Clotilde *et al.*, 2011), which allows for simultaneous interrogation of 50 biallelic SNPs. MOL-PCR is flexible for both the number of SNPs as well as for the number of strains tested (individual tubes to 96-well plates). We developed an 8-plex SNP-typing assay for the identification of the main phylogenetic lineages of MTBC. As a second method, we present SNP-typing with TaqMan real-time PCR. As a commercially available system, TaqMan has proved to be sensitive and specific in numerous studies for various applications and species (Yesilkaya *et al.*, 2006; Ben Shabat *et al.*, 2010; Milner

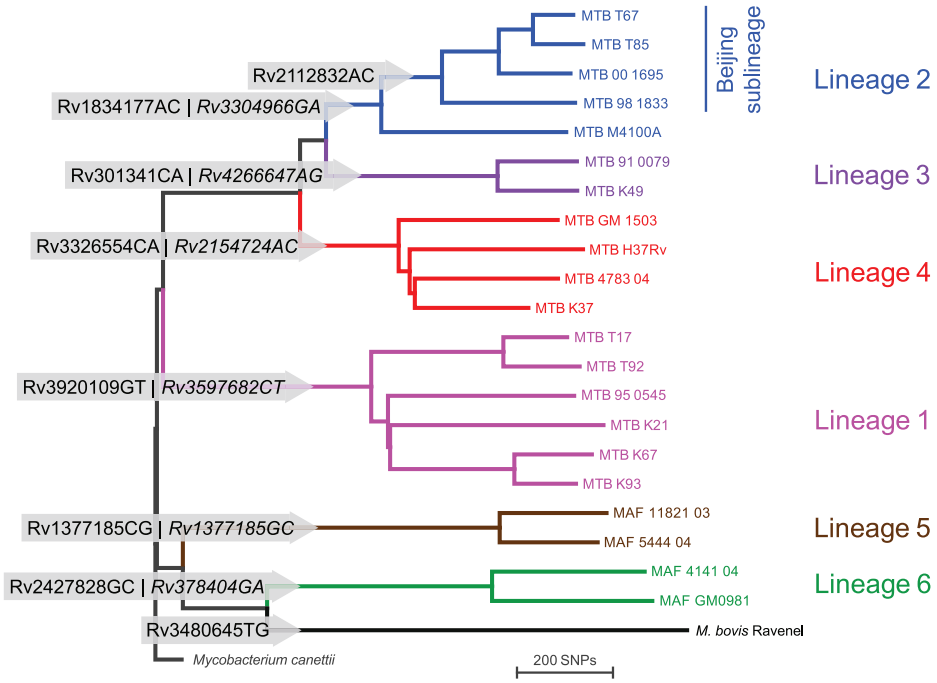


Figure 4.1.: **Phylogenetic tree of 22 whole genome sequences of MTBC plus *Mycobacterium canettii* as outgroup, and canonical SNPs used for MOL-PCR and TaqMan assays.** Rv-numbers in grey arrows represent SNPs used for MOL-PCR and TaqMan in this study, respectively, and nucleotide change at the position in the annotated reference sequence of H37Rv (Cole *et al.*, 1998). The first number indicates the SNP used for MOL-PCR, and the italic number the SNP used for TaqMan. Assays for the "Beijing" sublineage of Lineage 2 and for *M. bovis* / *M. caprae* were developed only for MOL-PCR (modified from (Bentley *et al.*, 2012)).

et al., 2012; Nübel *et al.*, 2012), and is therefore considered the standard for SNP-typing in our laboratory (Gagneux *et al.*, 2006b; Fenner *et al.*, 2011). Reactions are performed in a 96-well format in less than 2 hours. In contrast to MOL-PCR, only one SNP is assessed per reaction.

4.3. Methods

4.3.1. Ethics statement

All MTBC strains used in this study were from published reference collections (Gagneux *et al.*, 2006b; Allix-Béguec *et al.*, 2008; Hershberg *et al.*, 2008; Yeboah-Manu *et al.*, 2011; Fenner *et al.*, 2011; Fenner *et al.*, 2012b) or from an ongoing molecular epidemiology study on tuberculosis in Nepal. This study was ethically approved by the Nepal Health Research Council (NHRC), Kathmandu, Nepal, and the Ethics Committee of Basel, Switzerland (EKBB). Written informed consent was obtained from all Nepalese patients.

4.3.2. Strain panels

Two different strain panels were used in this study. A "training panel" of 46 MTBC strains was first chosen for establishing the MOL-PCR and TaqMan assays. These strains had been characterized previously by spoligotyping and were available as crude lysates and purified DNA. This training panel contained clinical strains from published reference collections (Gagneux *et al.*, 2006b; Allix-Béguec *et al.*, 2008; Hershberg *et al.*, 2008; Yeboah-Manu *et al.*, 2011; Fenner *et al.*, 2011; Fenner *et al.*, 2012b) and some clinical isolates from Nepal (Malla *et al.*, 2012). The training panel was composed of five to seven isolates of the 6 main phylogenetic lineage of MTBC, and additionally, *Mycobacterium bovis*, *Mycobacterium caprae*, *Mycobacterium pinnipedii* and *Mycobacterium microti* as controls for the *M. bovis / M. caprae*-specific SNP in the MOL-PCR assay. For each of these strains, SNP-typing was done by MOL-PCR and TaqMan for both crude lysates and purified DNA. Allele calls, i.e. lineage assignments, were obtained for all data points, and results between the two methods compared. As it was not possible to use the same SNPs for both MOL-PCR and TaqMan (due to probe design restrictions), we compared the lineage assignment rather than the locus-specific alleles.

A "test panel" was then chosen for the validation of MOL-PCR and TaqMan under "blind" conditions. This test panel consisted of 78 clinical MTBC isolates recently obtained from a molecular epidemiology study in Nepal. No previous genotyping had been done for these samples. The samples were cultured on Löwenstein-Jenssen (LJ) medium, frozen in tryptic soy agar and glycerol, and heat-inactivated for 1 h at 90°C. Purified DNA was available for fifteen of these strains.

4.3.3. Culturing and DNA extraction

MTBC strains were grown in standard Middlebrook 7H9 medium supplemented with ADC (Becton Dickinson, Allschwil, Switzerland), Tween80 (Sigma-Aldrich) and glycerol (Sigma-Aldrich). Liquid medium and agar plates for culturing Mycobacterium africanum (i.e. MTBC Lineage 5 and 6) were supplemented with 40 mM sodium pyruvate (Sigma-Aldrich). All bacterial cultures were incubated at 37°C. After two weeks, 500 μL culture were heat-inactivated (1 hour at 90°C) and centrifuged. The supernatant was used as input for TaqMan and MOL-PCR. Alternatively, 100 μL of a frozen stock culture was thawed and heat-inactivated (100 μL culture added to 100 μL TE buffer, 1 hour at 90°C). From all cultures, purified genomic DNA was extracted with the CTAB method (Embden *et al.*, 1993). DNA concentration was determined with a Nanodrop ND-1000 device (Thermo Scientific NanoDrop Products, Wilmington, USA).

4.3.4. Whole genome sequencing and SNP identification

Informative SNPs were obtained from whole genome sequencing as described before (Comas *et al.*, 2010; Bentley *et al.*, 2012). In brief, short reads from Illumina-sequencing of 22 MTBC strains were mapped to the reference genome of H37Rv and strain-specific nucleotide differences were extracted as SNPs. A neighbor-joining phylogenetic tree was constructed with MEGA (Tamura *et al.*, 2011), and SNPs mapped to the tree using Mesquite (Maddison *et al.*, 2011). Clade-specific SNPs were compiled in OpenOffice spreadsheets (`openoffice.org`). Position in reference to H37Rv, codon change, essentiality (Sassetti *et al.*, 2003a; Sassetti *et al.*, 2003b), and annotated function of the gene (Cole *et al.*, 1998) were collected for each SNP. This list was used as a starting point for oligonucleotide design. In the course of the study, 151 additional genome sequences became available (Comas *et al.*, 2013). With the total of 172 genome sequences representing the global MTBC diversity, we confirmed that the SNPs chosen were specific for all members of a lineage. In reference to the first MTBC genome sequenced and annotated (Cole *et al.*, 1998) (Genbank AL123456.2), every SNP was named with the prefix "Rv", followed by 7 digits that represent the position of this SNP in the genome of H37Rv, and the nucleotide change (i.e. "Rv2154724AC" for the Lineage 4-specific SNP in katG, codon 463).

4.3.5. Assay description Luminex MOL-PCR

The assay described here is conceptually related to Multiplex Ligation-dependent Probe Amplification (MLPA) that was also described for SNP-typing of *M. tuberculosis* (Schouten *et al.*, 2002; Bergval *et al.*, 2008). MLPA uses allele-specific ligation and stuffer sequences of variable lengths, which allow multiplexed typing with a capillary electrophoresis device. The high level of multiplexing is made possible by the PCR-amplification of the ligated oligonucleotides rather than the amplification of template DNA. This allows the use of a universal primer-pair for the singleplex-amplification of all ligated oligonucleotides and avoids multiplex-PCR. Allele-specific ligation was recently combined with the Luminex platform and named MOL-PCR, and used for pathogen detection and SNP-typing of *Bacillus anthracis*, *Yersinia pestis*, and *Francisella tularensis* (Deshpande *et al.*, 2010; Song *et al.*, 2010). MOL-PCR uses fluorescent beads and coupled tag-sequences instead of stuffer sequences for the multiplexed analysis. This allows the signal detection on a flow cytometer, i.e. a Luminex device. We adapted and modified MOL-PCR for typing of an 8-plex set of MTBC lineage-specific SNPs. In our assay, the tag-sequences complementary to antiTag-sequences on microspheres were included on the left-probe-oligonucleotide

(LPO), i.e. preceding the 5'-end of the allele-discriminating sequence (left-hybridizing sequence, LHS), contrary to the original publication of MOL-PCR (Deshpande *et al.*, 2010), where tag-sequences were included on the right-probe-oligonucleotide (RPO). With this modification in oligonucleotide design, two competing LPO could be used, differing only by the 3'-end (the SNP of interest) and the tag-sequence. This change allowed the calculation of an allelic ratio rather than a signal-to-noise ratio, and improved the sensitivity by discriminating background signal from uncalled allele signal. An additional change was made in separating the hybridization/ligation step from the PCR step. We obtained high background signal levels when combining these two steps. The assay therefore consisted of three steps (Figure 2): 1. Competitive hybridization of allele-specific probes and ligation to universal, 3'-adjacent probes. This step provides the allele-specificity of the assay. 2. Signal amplification by PCR and a fluorescently labeled primer, which guarantees the sensitivity. 3. Detection of allele-specific signals by hybridizing amplicons to allele-specific beads and read-out by flow cytometry.

4.3.6. SNP selection and MOL-PCR oligonucleotide design

The selection of suitable SNPs is crucial for the success of a typing assay. In MOL-PCR, probes of 20-45 bp are designed to anneal directly adjacent to the SNP of interest. The need for a suitable sequence up- and downstream of the SNP requires a number of SNPs to choose from as a starting point. Eligible SNPs for MOL-PCR were chosen using following criteria:

1. SNP is specific for the lineage of interest
2. SNP is not in a region where a genomic deletion has been described (to ensure binding of oligonucleotides and subsequent allele-calls for all strains)
3. No other SNP less than 50 bp up- or downstream (to avoid mismatches in oligonucleotide annealing)
4. SNP is not a transition (i.e. no nucleotide change A>G, G>A, C>T or T>C; lower unspecific signal because no alternate base pairing possible)
5. SNP is in a coding sequence (deletions or insertions less likely)
6. SNP is in an essential gene (deletions less likely)
7. SNP change is synonymous (no amino acid change; therefore lower selective pressure acting on SNP compared to nonsynonymous SNP)

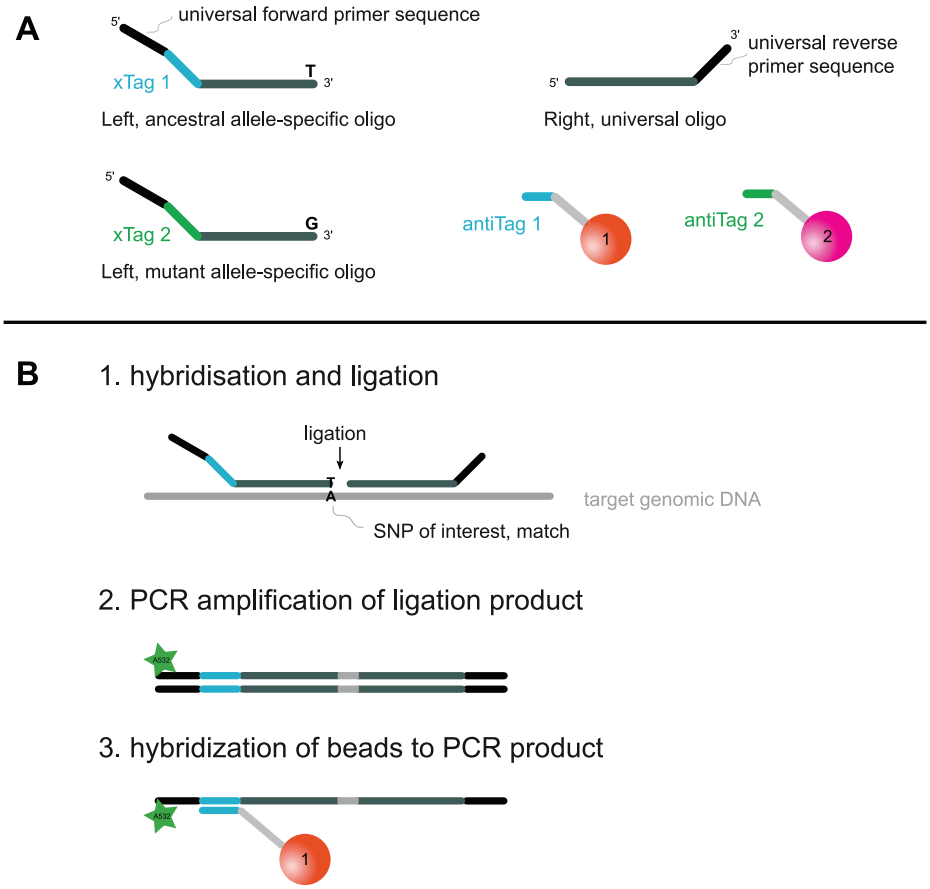


Figure 4.2.: **Schematic illustration of MOL-PCR. A.** Three oligonucleotides and two fluorescently labelled beads are used for interrogation of one SNP. **B.** MOL-PCR consists of three steps. 1. One of two competitive allele-specific left-hand probe oligonucleotides (LPO) and one universal right-hand probe oligonucleotide (RPO) are hybridized to the template DNA and ligated. 2. With PCR and a reporter-labelled forward primer, the ligated oligonucleotides (LPO + RPO) are amplified. 3. After denaturation of the PCR product, allele-specific fluorescent beads carrying an allele-specific antiTag sequence are hybridized to the amplicons. This will result in beads carrying reporter fluorescence (bead 1 in example) and beads not carrying fluorescence (bead 2). Reporter fluorescence (Alexa532) per bead is measured with a flow cytometric device (Luminex). For the 8-plex assay, a total of 24 oligonucleotides and 16 beads are used in the same reaction.

For each eligible SNP, 100 bp of enclosing genomic sequence of H37Rv up- and downstream of the SNP were obtained using Ugene (Okonechnikov *et al.*, 2012) and saved as multifasta file. Sequences and SNPs were introduced in AlleleID (Premier Biosoft, Palo Alto, USA) and suitable oligonucleotides searched using the MLPA module. As MLPA uses stuffer sequences rather than tag-sequences (for hybridizing to microspheres with coupled antitag-oligonucleotides), the latter were introduced manually and according to sequences provided by Luminex Corporation (Austin, USA). We used the following criteria for oligonucleotide search in AlleleID: 1. GC-content between 35% and 65%, 2. Tm between 72°C and 90°C, 3. Minimum left and right hybridization sequence (LHS and RHS) length 21 bp. Probes with high scores in AlleleID were exported from AlleleID to OpenOffice spreadsheets. BLAST search against *M. tuberculosis* H37Rv was run to exclude cross-homology. The 24 oligonucleotides for each lineage-specific SNP were pooled (3 oligonucleotides each) and assessed for heterodimer formation with SBEprimer (Kaderali *et al.*, 2003), AutoDimer (Vallone *et al.*, 2004) and Oligoanalizer (`www.idtdna.com`). Probes with potential dimer formation (dG $<$ -11 kcal/mol) were excluded, a new SNP chosen and new probes designed, until a compatible set was found. The final set of 24 oligonucleotides was synthesized by Sigma-Aldrich (Buchs, Switzerland). The 5'-end of each RPO was phosphorylated to enable ligation.

4.3.7. Luminex MOL-PCR procedure

All SNPs were interrogated in one tube (multiplexed). Our modified MOL-PCR assay consisted of three laboratory steps. 1. Oligonucleotides were hybridized to target DNA and ligated. 2. The ligation product was amplified by singleplex PCR (one primer pair for all SNPs). 3. Beads were hybridized to the amplification products and run on a flow cytometric device.

Hybridization and ligation: In a 10 μL volume with H_2O, 0.5 μL of oligonucleotide-mix (40 nM each oligonucleotide; see Table 4.1) were mixed with 1 μL of ligation buffer (NEB; New England Biolabs, Ipswich, USA), 0.1 μL of thermostable ligase (NEB) and 3 μL of pure DNA (10 ng / μL) or 4 μL heat-inactivated, crude extract. In a thermocycler, the following protocol was run: 4 min 94°C, 30 cycles 25 sec 94°C and 30 sec 50°C.

PCR: In a 10 μL volume, 3 μL of the ligation product was added to the PCR mix of 5.095 μL of H_2O, 1 μL PCR buffer without MgCl2 (Roche, Basel, Switzerland), 0.2 μL dNTPs (10 mM, Sigma-Aldrich), 0.5 μL forward primer (10 μM; Sigma-Aldrich, 5'-Alexa532-GGGTTCCCTAAGGGTTGGA), 0.125 μL reverse primer (10 μM; Sigma-Aldrich, GT-GCCAGCAAGATCCAATCTAGA) and 0.08 μL FastStart Taq polymerase (Roche). The following protocol was run in a thermocycler: 2 min 96°C and 43 cycles of 20 sec 94°C,

20 sec 58°C, 20 sec 72°C.

Hybridization of beads and signal detection: A bead mix containing 2 μL of beads (16 polystyrene bead regions of MagPlex-Tag beads, Luminex Corp., Austin, USA; Table 4.1) and 3 μL of buffer (1.33 M NaCl (Merck Chemicals, Darmstadt, Germany), 83.33 mM MES (Sigma-Aldrich)) was added to the 10 μL PCR product. In a thermocycler, the following protocol was run to hybridize beads (antiTags) to PCR products (tags): 1 min 94°C, ramp-down 0.1°C / sec decrement to 25°C, 5 min 25°C. 80 μL of running buffer (10 mM Tris-HCl, 0.1 mM EDTA (Sigma-Aldrich), 90 mM NaCl, 0.04% (v/v) Triton X-100 (Sigma-Aldrich), pH 8) were added and the reaction transferred to flat-bottom plates (BioRad, Hercules, USA). A BioPlex 200 device (BioRad) was used to discriminate beads and measure reporter fluorescence (Alexa532) for each bead, i.e. each allele. Results were obtained as reporter median fluorescence intensity (MFI) per bead region and sample, and exported from BioPlex Manager software into spreadsheets.

Table 4.1.: **Oligonucleotides used for Luminex MOL-PCR.**

MTBC Lineage[1]	SNP[2]	Gene	Ess / Syn[3]	Strand[4]	Allele[5]	Bead[6]	Sequence[7]
1	Rv3920109GT	yrbE4A	ess / syn	antisense	LPO_ancestral	30	GGGTTCCCTAAGGGTTGGActtaacatttaacttctataacacGTAGTCATGACCGCTATAGGTGTGCGAAT**C**
					LPO_mutant	35	GGGTTCCCTAAGGGTTGGAcatcttcatatcaattctcttattGTAGTCATGACCGCTATAGGTGTGCGAAT**A**
					RPO		P-TCGGCGCAGTAACAAACGACGGTGATCTAGATTGGATCTTGCTGGCAC
2	Rv1834177AC	rpsA	ess / syn	sense	LPO_ancestral	12	GGGTTCCCTAAGGGTTGGAcataatcaatttcaactttctactGCGAGTTCCTGAATAACTTGCAAAAAGGCACCATCCG**A**
					LPO_mutant	13	GGGTTCCCTAAGGGTTGGAcaaatacataatcttacattcactGCGAGTTCCTGAATAACTTGCAAAAAGGCACCATCCG**C**
					RPO		P-AAGGGTGTCGTGTCCTCGATCGTCAACTTCGGTCTAGATTGGATCTTGCTGGCAC
3	Rv301341CA	Rv0249c	ess / syn	sense	LPO_ancestral	18	GGGTTCCCTAAGGGTTGGAacacttatctttcaattcaattacTGAGCCGGAATCCAAGCAGGAA**C**
					LPO_mutant	19	GGGTTCCCTAAGGGTTGGAatactttacaaacaaataacacacTGAGCCGGAATCCAAGCAGGAA**A**
					RPO		P-GGTAATACCATCGCTCCCAACGGAATCTCTAGATTGGATCTTGCTGGCAC
4	Rv3326554CA	Rv2971	ess / nonsyn.	antisense	LPO_ancestral	20	GGGTTCCCTAAGGGTTGGActttctcatactttcaactaatttGACGAATGTGAGGTCGATAAGGTTTTCGATGT**G**
					LPO_mutant	21	GGGTTCCCTAAGGGTTGGAtcaaactctcaattcttacttaatGACGAATGTGAGGTCGATAAGGTTTTCGATGT**T**
					RPO		P-CTCCGCGGTGAAGTTGGACACGCCGATTCTAGATTGGATCTTGCTGGCAC
5	Rv1377185CG	Rv1234	noness / syn	sense	LPO_ancestral	36	GGGTTCCCTAAGGGTTGGAattaaacaactcttaactacacaaGCAGGTGACCATCGTTGGCGTGGACCT**C**
					LPO_mutant	37	GGGTTCCCTAAGGGTTGGAtacaacatctcattaacatatacaGCAGGTGACCATCGTTGGCGTGGACCT**G**
					RPO		P-ATGCAGGTTGAACGGGTCACAGGCCTCTAGATTGGATCTTGCTGGCAC
6	Rv2427828GC	Rv2164c	ess / syn	sense	LPO_ancestral	33	GGGTTCCCTAAGGGTTGGAactacttattctcaaactctaataCATCGGTGGACAACCACAGTGTGAG**G**
					LPO_mutant	34	GGGTTCCCTAAGGGTTGGAacttatttcttcactactatatcaCATCGGTGGACAACCACAGTGTGAG**C**
					RPO		P-CCTAGTCCGACGCCGAGCGAACCGATAATCTAGATTGGATCTTGCTGGCAC
M. bovis / *M. caprae*	Rv3480645TG	Rv3113	ess / nonsyn	sense	LPO_ancestral	28	GGGTTCCCTAAGGGTTGGAcacttaattcattctaaatctatcTTTCCGTCTTGGTCACCGGAGCTTCAC**T**
					LPO_mutant	29	GGGTTCCCTAAGGGTTGGAtactacttctataactcacttaaaTTTCCGTCTTGGTCACCGGAGCTTCAC**G**
					RPO		P-GCAGCCACCCGACCGAGTCATGCTCAAATCTAGATTGGATCTTGCTGGCAC
Beijing sublineage	Rv2112832AC	Rv1865c	noness / syn	sense	LPO_ancestral	14	GGGTTCCCTAAGGGTTGGAaatttcttctctttctttcacaatTTTCGCCATCGCCTCATCGATGTCGCCGAT**A**
					LPO_mutant	25	GGGTTCCCTAAGGGTTGGActttcttaatacattacaacatacTTTCGCCATCGCCTCATCGATGTCGCCGAT**C**
					RPO		P-GCGACCTTGGCTCCCAAGTTGTGCATCTAGATTGGATCTTGCTGGCAC

[1] Nomenclature according to Coscolla *et al.* (2010)
[2] Position of SNP in reference to H37Rv
[3] SNP in an essential gene ("ess"); synonymous SNP ("syn")
[4] Strand orientation of oligonucleotides
[5] LPO, left probe oligonucleotide, annealing upstream of SNP of interest and including it; RPO, right probe oligonucleotide, annealing downstream of SNP of interest
[6] xTag bead with coupled antiTag-sequence
[7] Universal primer sequences (upper case, not underlined), allele-specific tag-sequences (lower case), and sequences hybridizing to the genomic sequence (upper case right of tag-sequences). RPO are 5'-phosphorylated.

4.3.8. Allele calling MOL-PCR

For each sample, the allelic state of each SNP was assessed with the following algorithm:

1. Bead counts > 40 beads per bead region
2. Allelic ratio (AR) < 0.4; ancestral allele called
3. Allelic ratio (AR) > 0.6; mutant allele called
4. MFI > threshold MFI (see below)

$$AR = \frac{\frac{MFI_{mutant}}{MFI_{H_2Omutant}}}{\frac{MFI_{mutant}}{MFI_{H_2Omutant}} + \frac{MFI_{ancestral}}{MFI_{H_2Oancestral}}}$$

MFI = median fluorescence intensity of reporter-dye (Alexa532) of the corresponding allele / bead region

A threshold MFI for each allele/bead was used to exclude low signal calls and to avoid SNP-calls for H_2O signals (false-positives). This value was different for each SNP and each allele (i.e. each bead region) and had to be calculated for each run, as we observed significant differences of H_2O signal levels between different runs, and occasionally H_2O signal levels higher than uncalled allele signals. The pragmatic approach that we used to calculate the signal threshold for allele was the following. For each allele, the threshold was set higher than each of the corresponding H_2O values (a minimum of three H_2O samples per 96-well-plate were included). The threshold was then lowered so far that no allele was called for any of the H_2O-sample, which is made possible by the range of the allelic ratio. Furthermore, an absolute minimal threshold value was used.

4.3.9. Assay description TaqMan PCR

TaqMan real-time PCR was originally described in 1991 (Holland *et al.*, 1991) and was since then optimized for commercial use. The PCR based method makes use of fluorescently labeled allele-specific probes, the 5'-3'-exonuclease activity of Taq polymerase and a non-fluorescent quencher. During PCR amplification of the region of interest, fluorescence is detected in a real-time PCR thermocycler.

4.3.10. SNP selection and probe design for TaqMan

Primers and probes for Lineage 2, Lineage 3 and Lineage 4 were described before (Table 4.2). SNP for Lineage 1 was chosen from the list of lineage-specific SNPs recently published (Comas *et al.*, 2009b), and SNPs for Lineage 5 and 6 from the list of genome sequences as described above. For the design of probes and primers, the software Primer Express (Applied Biosystems, Carlsbad, USA) was used. As TaqMan assays were singleplex, oligonucleotide design was less complex, as heterodimer formation is limited to the two primers and two probes of one assay. Furthermore, the design of the primer pair can be varied over a large sequence range.

Table 4.2.: **Primer and probe sequences for TaqMan assays.**

MTBC Lineage[1]	LSP name[2]	Spoligotype name[3]	SNP[4]	Forward primer	Reverse primer	Ancestral allele probe[5]	Mutant allele probe[5]	Reference
1	Indo-oceanic	EAI, MANU1	Rv3597682CT	TGTCAACGAAGGCGATCAGA	GACCGTTCCGGCAGCTT	6FAM-ACAAGGGCGACGTC-MGBNFQ	VIC-ACAAGGGCGACATC-MGBNFQ	This study
2	East Asian	Beijing	Rv3304966GA	CCTTCGATGTTGTGCTCAATGT	CATGCGGCGATCTCATTGT	6FAM-CCCAGGAGGGTAC-MGBNFQ	VIC-CCCAGGAAGGTACT-MGBNFQ	(Fenner *et al.*, 2011)
3	East-African-Indian	CAS	Rv4266647AG	GCATGGATGCGTTGAGATGA	CGAGTCGACGCGACATACC	VIC-AAGAATGCAGCTTGTTGA-MGBNFQ	6FAM-AAGAATGCAGCTTGTCGA-MGBNFQ	(Fenner *et al.*, 2011)
4	Euro-American	X, Haarlem, LAM, Uganda	Rv2154724AC	CCGAGATTGCCAGCCTTAAG	GAAACTAGCTGTGAGACAGTCAATCC	VIC-CCAGATCCTGGCATC-MGBNFQ	6FAM-CAGATCCGGGCATC-MGBNFQ	(Gagneux *et al.*, 2006b)
5	*M. africanum* West African I	AFRI 2	Rv1377185GC	TCCAGCAGGTGACCATCGT	GGCCTGTGACCCGTTCAAC	VIC-CGTGGACCTCATG-MGBNFQ	6FAM-CGTGGACCTGATGC-MGBNFQ	This study
6	*M. africanum* West African II	AFRI 1	Rv378404GA	CGGCCGACAGCGAGAA	CCATCACGACCGAATGCTT	6FAM-CTGCAAATCCCGCAGTA-MGBNFQ	VIC-CTGCAAATCCCACAGT-MGBNFQ	This study

[1] Nomenclature according to (Coscolla *et al.*, 2010)
[2] Nomenclature according to (Gagneux *et al.*, 2006b)
[3] Nomenclature according to (Filliol *et al.*, 2006)
[4] Position of SNP in reference to the H37Rv genome
[5] 6FAM and VIC, fluorescent dyes at the 5'-end of probes; MGB-NFQ, MinorGrooveBinder-NonFluorescentQuencher at the 3'-end

4.3.11. TaqMan real-time PCR

TaqMan real-time PCR was performed according to standard protocols (Applied Biosystems, Carlsbad, USA). Briefly, in a 12 μL volume, 2 μL DNA sample were added to a mix of 5 μL TaqMan Universal MasterMix II (Applied Biosystems), 5 μL H_2O containing forward and reverse primer (0.21 μM each; Sigma-Aldrich), probe A for ancestral allele and probe B for mutant allele (0.83 μM each; Applied Biosystems). Primer and probe sequences are listed in Table 4.2. Reactions were run in a StepOnePlus thermocycler (Applied Biosystems; 60°C 30 sec; 95°C 10 min; 95°C 15 sec and 60°C 1 min for 40 cycles; 60°C 30 sec) and fluorescence intensity in the VIC and FAM channels measured at the end of every cycle. Results were analyzed with StepOne software (Applied Biosystems) and alleles called with the default algorithm.

4.3.12. Spoligotyping

43-spacer spoligotyping was performed following standard protocols on a membrane (Kamerbeek *et al.*, 1997).

4.3.13. PCR and sequencing of putative deletion in Rv3113

As we did not obtain MOL-PCR signal for the *M. bovis* / *M. caprae*-specific SNP for two strains (N1007 and N1032, see Table S1), the locus was assessed for a potential deletion in Rv3113. PCR was run in a 25 μL volume with 12.3 μL H_2O, 2.5 μL PCR buffer with $MgCl_2$ (Roche), 5 μL GC buffer, 0.75 μL forward primer K-297 (10 μM, Sigma-Aldrich, CCATGATGCTGGCAGAACTGA), 0.75 μL reverse primer K-298 (10 μM, Sigma-Aldrich, CCTGCGTACCTTCGTCGTCA), 0.5 μL dNTPs (10 mM, Qiagen, Hombrechtikon, Switzerland) and 0.2 μL Fast Start Taq Polymerase (Roche). Reaction was run in a standard thermocycler (96°C, 7 min; 35 cycles of 96°C 30 sec, 62°C 30 sec, 72°C 30 sec; 72°C 4 min). PCR product was analyzed using standard gel electrophoresis and stained with ethidium bromide.

4.4. Results

Eight SNPs were used in the multiplexed MOL-PCR assay (Table 1). Six SNPs were interrogated serially with TaqMan assays (Table 4.2). Forty-six MTBC strains were run as a "training panel" (Table S1). Most samples were available as 10 ng / μL purified genomic DNA (CTAB-method) and as crude lysates from 7H9 culture. Both methods,

MOL-PCR and TaqMan, were run in parallel for all samples and all SNPs. Table S1 shows allelic state for each SNP for both purified DNAs and crude extracts, called by MOL-PCR and TaqMan, respectively.

For MOL-PCR, 662 of 672 (98.6%) total data points (alleles) resulted in a successful call of either the ancestral or the mutant allele (Table S1 and Table 4.3). Eight data points, for which no allele call was obtained, involved crude lysates, and two were purified DNA samples. For two strains, N1007 and N1032, no allele was called for the *M. bovis / M. caprae*-specific SNP, Rv3480645TG, for both crude and purified DNA. To test for a potential deletion in the corresponding genomic region, we designed PCR primers and amplified the region up- and downstream of this SNP. Gel electrophoresis showed an expected amplicon length of approximately 250 bp for H37Rv, but no band for the two strains in question, suggesting that this region was deleted in strains N1007 and N1032 (Figure S1).

For TaqMan PCR, 498 alleles of 504 (98.8%) total data points were successfully called (Table S1 and Table 4.3). The remaining six data points without allele calls were two crude lysates and four purified DNA samples. There was no case where TaqMan PCR failed to call an allele for both purified and heat-inactivated sample. Both methods agreed 100% in lineage assignment of the 36 strains comprising the training panel (excluding *M. bovis*, *M. caprae*, *M. microti* and *M. pinnipedi*, as no SNP for *M. bovis / M. caprae* was used in the TaqMan assay) (Table 4.4). Using the prototype spacer definitions (Kato-Maeda *et al.*, 2011a; Demay *et al.*, 2012), spoligotyping data confirmed the lineage-assignments of TaqMan and MOL-PCR for 39 of 42 strains (excluding *M. microti / M. pinnipedii*). Of the three remaining strains, N0153 and N0051 could not be classified into any of the known spoligotype families. Strain N1024 showed a "Beijing" spoligopattern and was confirmed by TaqMan and MOL-PCR as a Lineage 3 (also known as CAS/Dehli) strain (Table 4.3). The spoligotype of this strain was described as "Pseudo-Beijing" previously (Fenner *et al.*, 2011).

Table 4.3.: **Comparison of MOL-PCR and TaqMan for allele calling and lineage assignment.**

		MOL-PCR	TaqMan
Training panel	Total data points	672 (100%)[1]	504 (100%)[1]
	Data points with allele call	662 (98.6%)	498 (98.8%)
	Total number of strains	42 (100%)[2]	36 (100%)[3]
	Strains assigned to a lineage	42 (100%)	36 (100%)
	Congruence in lineage assignment	100%	
Test panel	Total data points	624 (100%)	78 (100%)
	Data points with allele call	619 (99.2%)	78 (100%)
	Total number of strains	78	78
	Strains assigned to a lineage	76 (97.4%)	78 (100%)
	Congruence in lineage assignment	100%	

[1] excluding N/A as indicated in Table S1

[2] excluding *M. microti*, *M. pinnipedi*

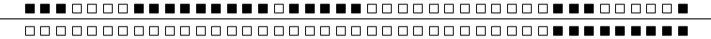
[3] excluding *M. bovis*, *M. caprae*, *M. microti*, *M. pinnipedii*

Next, we used a test panel of 78 clinical isolates from Nepal to validate our methods (Table S2). These samples were mainly crude, heat-inactivated extracts, and no genotyping had been done before. To follow the procedure that would be used as routine SNP-typing, we first performed MOL-PCR, as information for all SNPs is obtained in parallel. Of a total of 624 data points (i.e. 78 strains and eight SNPs), 619 (99.2%) allele calls were obtained (Table S2 and Table 4.3). The five data points with no calls were from three strains (1671A, 3052B and 3074B). Thereof, two strains (1671A, 3074B) failed to be assigned to a lineage by MOL-PCR (97.4% sensitivity in lineage assignment). Nevertheless, we obtained ancestral allele calls for seven (1671B) and six (3074B) other SNPs for these strains. We then performed the corresponding TaqMan assay for each sample. TaqMan called alleles for all 78 samples. Again, lineage-assignment was 100% congruent between MOL-PCR and TaqMan.

To evaluate the signal detection limit of MOL-PCR, we used different DNA concentrations as input material. Samples with known DNA concentration were diluted appropriately for different amounts of input DNA. One strain of each lineage was used and MOL-PCR performed with 60 ng, 10 ng, 1 ng, 0.1 ng purified DNA, and 4 μL of crude, heat-inactivated extract. Figures S2-S8 show a decrease of signal intensities with de-

creasing amount of DNA. Nevertheless, the use of 0.1 ng purified DNA (CTAB) or 4 μL crude extract still resulted in signals high enough for allelic ratio determination. Signal intensities of the complementary allele were constantly low.

Table 4.4.: **Lineage assignments of MOL-PCR, TaqMan and spoligotyping for the 46 MTBC strains as in Table S1.**

MTBC strain	MOL-PCR Lineage	TaqMan Lineage	Spoligotyping Lineage[1]	Spoligotype	SITVIT lineage[2]	SIT[2]
N1004	1	1	1		EAI6-BGD1	292
N1006	1	1	1		EAI5	ORPHAN
N0153	1	1	- [3]		ZERO	405
N1011	1	1	1		EAI5	ORPHAN
N1030	1	1	1		EAI5	138
N1068	1	1	1		EAI3-IND	ORPHAN
N1053	2 / Beijing	2	2		BEIJING	1
N1002	2 / Beijing	2	2		BEIJING	1
N1010	2 / Beijing	2	2		BEIJING	1
N1012	2 / Beijing	2	2		BEIJING	1
N0051	2 / non-Beijing	2	- [x]		-	955
N1069	2 / Beijing	2	2		BEIJING	1
N1013	2 / Beijing	2	2		BEIJING	1
N1144	3	3	3		CAS1-DELHI	26
N1007	3	3	3		CAS	142
N1082	3	3	3		-	-
N1024	3	3	- [4]		BEIJING	1
N1032	3	3	3		CAS1-DELHI	26
N1021	3	3	3		CAS	357
N1022	3	3	3		-	2485
H37Rv	4	4	4		H37RV	451
N1008	4	4	4		H3	655
N1015	4	4	4		T1	102
N1019	4	4	4		-	-
N1052	4	4	4		T1	53
N1057	4	4	4		T1	244
N1034	5	5	5		-	-
N1035	5	5	5		-	ORPHAN
N1063	5	5	5		-	-
N1064	5	5	5		AFRI_2	101
N1286	5	5	5		-	-
N1201	6	6	6		-	3476
N1292	6	6	6		-	-
N0091	6	6	6		AFRI_1	181
N0092	6	6	6		AFRI_1	181
N0115	6	6	6		AFRI_1	181
M. bovis BCG Pasteur	*M. bovis* / *M. caprae*	(N/A)	*M. bovis*		BOV_1	482
M. bovis BCG Montreal	*M. bovis* / *M. caprae*	(N/A)	*M. bovis*		BOV_1	482
M. bovis 1290-03	*M. bovis* / *M. caprae*	(N/A)	*M. bovis*		BOV_1	665
M. bovis 8217-02	*M. bovis* / *M. caprae*	(N/A)	*M. bovis*		BOV_1	1037
M. caprae 7140-99	*M. bovis* / *M. caprae*	(N/A)	(N/A)		BOV4_CAPRAE	647
M. caprae 8986-99	*M. bovis* / *M. caprae*	(N/A)	(N/A)		BOV4_CAPRAE	818
M. microti 1479-00	-	(N/A)	(N/A)		MICROTI	539
M. microti 8753-00	-	(N/A)	(N/A)		PINI1	642
M. pinnipedii 7011-02	-	(N/A)	(N/A)		-	-
M. pinnipedii 7739-01	-	(N/A)	(N/A)		-	-

[1] Lineage-classification using spoligotype prototypes as described in (Comas *et al.*, 2009b; Kato-Maeda *et al.*, 2011a; Demay *et al.*, 2012)

[2] As described in (Demay *et al.*, 2012)

[3] No lineage-classification possible based on spoligotyping

[4] "Pseudo-Beijing" as described in (Fenner *et al.*, 2011)

4.5. Discussion

Informative SNPs are valuable markers for deep phylogenetic typing of clinical MTBC isolates (Schürch *et al.*, 2012). Therefore, SNP-typing will likely play an important role for genotype-phenotype association studies in the coming years. Even though routine whole-genome-sequencing, i.e. "genometyping", might replace most other typing methods eventually, current limitations such as costs, bioinformatics capacity, the lack of user-friendly software packages, and data storage has been hindering the broad implementation of "genometyping" (Kato-Maeda *et al.*, 2011b; Schürch *et al.*, 2012). Nevertheless, thousands of informative SNPs have become available from whole genome sequences, which can be used for SNP-typing.

In this study, we presented 14 canonical SNPs, comprising two sets of markers for the six main phylogenetic lineages of MTBC, the *M. bovis / M. caprae* clade and the "Beijing" sublineage. In contrast to other published SNP sets (Gutacker *et al.*, 2006; Filliol *et al.*, 2006), our SNPs were obtained from many whole genome sequences representing the global diversity of MTBC, and thus do not suffer from phylogenetic discovery bias (Pearson *et al.*, 2004). Phylogenetic discovery bias occurs when SNPs used for subsequent genotyping are initially identified by comparing only a few whole genome sequences (for example when these genomes represent only a subset of all MTBC lineages). As a result, strains belonging to lineages not represented in the initial genome set will falsely appear as intermediates in the corresponding phylogenies (Smith *et al.*, 2009). In additional advantage of our SNP sets is that they also identify *Mycobacterium africanum* (Jong *et al.*, 2010) and discriminate between the West-African I and the West-African II subtypes (i.e. MTBC Lineages 5 and 6). Moreover, the addition of a "Beijing"-specific SNP makes the differentiation between this sublineage and other strains of Lineage 2 (East-Asian) possible.

For the rapid typing of the SNPs described, we have successfully developed a multiplexed assay adapted from MOL-PCR (Deshpande *et al.*, 2010). Our assay uses three oligonucleotides per SNP instead of two, with the LPOs competing for annealing to the template DNA. This allows an allele-call based on allelic ratio rather than signal-to-noise ratio, which increases sensitivity. We further separated the hybridization/ligation from the PCR step, which reduced unspecific signal. We compared the new MOL-PCR assay with TaqMan real-time PCR which was used as standard SNP-typing method in our laboratory. We found an agreement of 100% in classifying strains into the main phylogenetic lineages and an allele-calling sensitivity of 98.6% and 98.8%, respectively, when typing two panels of clinical MTBC strains.

MOL-PCR was sensitive enough to be used with low concentrations of DNA as well as

heat-inactivated samples (crude extracts) (Figures S2-S8). We found that the detection limit of MOL-PCR was as low as 0.1 ng of input DNA. Heat-inactivated samples resulted in similarly high signal levels. This is especially valuable as performing MTBC cultures and DNA extraction is laborious and time-consuming. Even though the signal levels of H_2O control reactions were at time relatively high (Figures S2-S8), the MFI threshold determination we applied allowed us to control for this, and water samples could not be called false-positive alleles. We recommend to include at least three water samples per 96-well plate, and to use a mean value for the calculation of allelic ratios. Additionally, we always included H_2O control samples for the PCR step to detect potential contamination.

We observed a putative deletion in the region of SNP Rv3480645TG, as we did not obtain signal for neither allele for two samples (N1007 and N1032). To our knowledge, no genomic deletion has been described to span this SNP (Tsolaki *et al.*, 2004). PCR amplification of this region resulted in no visible band for these two samples (Figure S1), which suggested that there might be a deletion present in some Lineage 3 (CAS/Delhi) strains. This highlights the fact that genomic deletions can affect a SNP call, even though the corresponding gene was originally described as essential (Sassetti *et al.*, 2003a; Sassetti *et al.*, 2003b). However, with MOL-PCR we always generated complete allelic information for all other SNPs of the assay. As the canonical SNPs defining different MTBC lineages are mutually exclusive (except Lineage 2 and Beijing), the lack of an allelic call for one SNP can be neglected if the alleles at other loci are called successfully.

A limitation in the development of the MOL-PCR assay was the low efficiency of design and validation of oligonucleotides. For each lineage, we had to design up to 5 sets of probes targeting different SNPs, because many had to be rejected during validation, most often because of unfavorable allelic or signal-to-noise ratios or lack of signal. The use of long (i.e. 40-140 bp) oligonucleotides makes hairpin and homodimer formations likely, and multiplexing probes adds a level of complexity due to a high combinatorial number of possible heterodimer formations. We used different software for homo- and heterodimer analysis, but still were not able to fully predict the success of probes. Recently, an online tool for MOL-PCR specific oligonucleotide design and evaluation became available (moligodesigner.lanl.gov) (Song *et al.*, 2010), which could strongly facilitate the design of such oligonucleotides. Furthermore, an oligonucleotide design tool for copy-number-variation analysis of human, mouse and rat genomes with MLPA and MOL-PCR has been described (Zhi *et al.*, 2008; Zhi, 2010), and might be extended to SNP analysis and bacterial genomes.

Four other methods have been described for multiplexed SNP-typing on the Luminex platform, namely Direct Hybridization, Allele-Specific Primer Extension, Single-Base-

Extension and Oligonucleotide Ligation (Dunbar, 2006; Lee *et al.*, 2004). In contrast to MOL-PCR, these methods use multiplex-PCR for amplification of template DNA. This makes high multiplex-levels difficult, as a multiplex PCR is usually the limiting factor in multiplexed assays. In our assay, 24 oligonucleotides of 48 to 80 bp lengths were included for eight SNPs. A number of additional SNPs could be interrogated in the same reaction, i.e. for sublineages, drug resistance mutations (Bergval *et al.*, 2008) or outbreak-specific SNPs (Schürch *et al.*, 2010).

In addition to MOL-PCR, we also developed TaqMan SNP-assays for robust and rapid singleplex-typing of all main MTBC lineages. TaqMan SNP-typing is a well-established method for genotyping and available commercially. This facilitates the design and validation of probes, and provides standardized reaction conditions. Furthermore, criteria of TaqMan primer and probe design are less strict than in MOL-PCR. TaqMan makes use of quantitative PCR, which renders the method particularly sensitive. On the other hand, multiplexing of TaqMan assays is demanding and the degree of multiplexing is limited. Hence, we recommend using MOL-PCR for strain collections with no a priori information. TaqMan SNP-typing, on the other hand, is most suitable for confirmation of MOL-PCR results, or for strain collections for which previous information about genotypes is available.

Costs for reagents and consumables are another important consideration when implementing a new technique. Similar to other methods, running costs of MOL-PCR per SNP decrease with increasing number of multiplexed SNPs. We have estimated the minimal reagent costs per SNP to be 0.8 Euro when run in a singleplex reaction. Reagent costs drop to 0.15 Euro per SNP when running an 8-plex reaction. The cost of one TaqMan reaction was estimated at 0.9 Euro per SNP. For both assays, the initial setup costs are relatively high due to purchase of customized and labeled oligonucleotides. Furthermore, both methods require an expensive piece of equipment in form of a Luminex or a real-time-PCR machine. However, both of these devices can be used for a wide variety of applications in addition to SNP-typing (e.g. spoligotyping on the Luminex platform (Cowan *et al.*, 2004)), and many laboratories already have access to such equipment.

In conclusion, we propose a new set of canonical SNPs specific for the main phylogenetic lineages of MTBC. When combined with multiplexed MOL-PCR or singleplex TaqMan SNP-typing, these SNPs are ideal for phylotyping of strain collections from a small to a large scale. These assays provide a new basis for phylogenetically robust classification of clinical isolates.

4.6. Acknowledgement

We would like to thank Bhawana Shrestha, Director of the German Nepal Tuberculosis Project, Kathmandu, Nepal, for help with sample collection, Alina Deshpande for helpful technical advice, and Jean Pieters and Stefan Niemann for providing MTBC strains.

4.7. Supplementary information

Supplementary Figures and Tables are available online under `http://dx.doi.org/10.1371/journal.pone.0041253`.

Figure S1. Agarose gel of PCR product of Rv3114 covering the SNP specific for *M. bovis* / *M. caprae* (Rv3480645TG). H37Rv was used as positive control, whereas N1007 and N1032 were the samples that did not result in any signal in MOL-PCR. Ladder is Hyperladder II (Bioline).

Figure S2 to S8. Dilution series of DNA from previously characterized clinical isolates of Lineages 1 to 6, and *M. bovis*.

Table S1. MOL-PCR and TaqMan allele calls for the training panel of 46 well-characterized MTBC strains.

Table S2. MOL-PCR and TaqMan allele calls for the test panel of 78 MTBC strains with no previous genotyping information.

5. KvarQ: Targeted and direct variant calling from FastQ reads of bacterial genomes

Andreas Steiner [1,2,§], David Stucki [1,2,§], Mireia Coscollà [1,2], Sonia Borrell [1,2] and Sebastien Gagneux [1,2,*]

[1] Swiss Tropical and Public Health Institute, Basel, Switzerland

[2] University of Basel, Switzerland

[§] Contributed equally

[*] Corresponding author

This paper has been published in *BMC Genomics* 2014, 15:881.

5.1. Abstract

Background: High-throughput DNA sequencing produces vast amounts of data, with millions of short reads that usually have to be mapped to a reference genome or newly assembled. Both reference-based mapping and de novo assembly are computationally intensive, often generating large intermediary data files. These types of analyses also require bioinformatics skills that are often lacking in the laboratories producing the data. Moreover, many research questions and practical applications require only a small fraction of the whole genome data of interest.

Results: We developed `KvarQ`, a new tool that directly scans `fastq` files for known variants, such as single nucleotide polymorphisms (SNP), bypassing the need of mapping all sequencing reads to a reference genome and *de novo* assembly. Instead, `KvarQ` loads "testsuites" that define specific SNPs or short regions of interest in a reference genome, and directly synthesizes the relevant results based on the occurrence of these markers in the `fastq` files. `KvarQ` has a versatile command line interface and a graphical user interface. `KvarQ` currently ships with two "testsuites" for *Mycobacterium tuberculosis*, but new "testsuites" for other organisms can easily be created and distributed. In this article, we demonstrate how `KvarQ` can be used to successfully detect all main drug resistance mutations and phylogenetic markers in 880 bacterial genome sequences. The average scanning time per genome sequence was two minutes. The variant calls of a subset of these genomes were validated with a standard bioinformatics pipeline and revealed >99% congruency.

Conclusions: `KvarQ` is a user-friendly tool that directly extracts relevant information from `fastq` files. This enables researchers and laboratory technicians with limited bioinformatics expertise to scan and analyze raw sequencing data in a matter of minutes. `KvarQ` is open-source, and pre-compiled packages with a graphical user interface are available at `http://www.swisstph.ch/kvarq`.

5.2. Background

Large scale whole genome sequencing (WGS) is revolutionizing microbiology and public health (Olsen *et al.*, 2012; Bertelli *et al.*, 2013; Branco, 2013). Thanks to the latest technological advances, bacterial genomes can now be sequenced in less than 48 hours for less than €100 per genome. New benchtop devices are providing WGS capability to routine microbiological laboratories, and WGS is now becoming an important part of clinical microbiology and molecular epidemiology. This emerging field of "genomic epidemiology" (Robinson *et al.*, 2013; Parkhill *et al.*, 2011; Loman *et al.*, 2012; Le *et al.*, 2013; Croucher *et al.*, 2013; Gilmour *et al.*, 2013; Gardy, 2013; Walker *et al.*, 2013a) already includes numerous studies that have evaluated the potential of WGS to trace the transmission of pathogens such as *Clostridium difficile*, *Pseudomonas aeruginosa*, *Klebsiella pneumonia*, *Neisseria meningitidis*, *Staphylococcus aureus* and *Mycobacterium tuberculosis* (Didelot *et al.*, 2012; Snyder *et al.*, 2013; Snitkin *et al.*, 2012; Lavezzo *et al.*, 2013; Köser *et al.*, 2013; Castillo-Ramirez *et al.*, 2012; Gardy *et al.*, 2011; Bryant *et al.*, 2013b; Roetzer *et al.*, 2013; Walker *et al.*, 2013b). Moreover, WGS is playing an increasing role in molecular drug susceptibility testing (DST) and the identification of drug resistance mutations (Mather *et al.*, 2013; Drobniewski *et al.*, 2013; Chen *et al.*, 2013; Köser *et al.*, 2013; Paterson *et al.*, 2013; Zankari *et al.*, 2013; Török *et al.*, 2012). It is expected that WGS will eventually replace all other genotyping methods (Pallen *et al.*, 2010).

In contrast to the increasing availability of WGS platforms, the capability to analyse WGS data is lagging behind, partially because of the lack of rapid and user-friendly analysis tools. Extracting single nucleotide polymorphisms (SNPs), the most widely studied form of genetic variation, from WGS data requires substantial bioinformatic expertise, as well as dedicated bioinformatic analysis pipelines often based on customized scripts. Moreover, this standard approach of analysing WGS data is time-consuming and computationally expensive (Stucki *et al.*, 2013), and as a consequence, many biological questions can often not be addressed by the people generating the data. Furthermore, large amounts of data are produced when analyzing complete genomes, although often only a small proportion of these data are actually relevant for the study question; this is particularly true when screening for drug resistance determinants. Identifying individual SNPs of interest usually requires mapping all sequencing reads in a `fastq` file to a given reference genome, despite the fact that only a few nucleotide positions might be relevant. Tools have been developed to speed up the extraction of relevant information. These include *in silico* multi-locus sequence typing (MLST) from draft genomes or contigs, but all these methods rely on previous assembly (Kruczkiewicz *et al.*, 2012; Jolley *et al.*, 2010; Inouye

et al., 2012). One of these tools, SRST, facilitates the process by accepting `fastq` files for a given typing scheme. SRST uses BWA and SAMtools and is highly configurable, but is aimed at the bioinformatically skilled users and lacks a graphical interface. However, a graphical interface is crucial for making these kind of genome analyses accessible to a broader group of microbiologists, including the ones primarily interested in routine clinical applications. Consequently, there is a need for software able to extract allelic information directly from the raw sequencing reads without the need of previous mapping or *de novo* assembly.

Here we present `KvarQ`, a software that enables rapid screening of short sequence reads for mutations at multiple nucleotide positions of interest. `KvarQ` uses as input a `fastq` file, and generates the output in form of a textfile in JavaScript Object Notation (`json`) format. Using a completed bacterial genome sequence as a reference, `KvarQ` can interrogate any single nucleotide position or short DNA sequence of interest for known or new mutations, respectively.

`KvarQ` is universally applicable to any short read dataset and corresponding reference sequence, and is highly adaptable with respect to target polymorphisms. Hence, `KvarQ` may benefit many users, including clinical microbiological laboratories, where time from generating data to results is crucial, as well as laboratories in resource-limited settings, where access to both computational power and skilled staff remains limiting. Even advanced users will save time by avoiding the need for formal mapping or assembly to answer specific questions, particularly when studying large numbers of genomes.

`KvarQ` is available for Linux, OS X and Windows, and runs on any portable or desktop system. A command line interface and a simple graphical user interface are available. We aimed at a self-explaining application, short run times (<20 minutes on a standard workstation for a `fastq` file of 1000 MB) and a high flexibility for application to a variety of organisms. `KvarQ` can be downloaded (source code as well as precompiled packages) from `www.swisstph.ch/kvarq`.

5.3. Implementation

5.3.1. Overview

`KvarQ` analyzes a `fastq` file and detects the allelic state of known polymorphisms, and compiles all the relevant information about the genome sequence. In contrast to current genetic analysis software (Kruczkiewicz *et al.*, 2012; Inouye *et al.*, 2012), `KvarQ` extracts the relevant information directly from the sequence reads, without the need to map every

read to a reference genome (hence `KvarQ` is a "mapping free" tool). The software can be used via the command line or the graphical user interface, and is split into one part that scans the `fastq` files and saves the results to a `json` file, and a second part that extracts and illustrates the information contained in the `json` file. Although results reported in this article mainly use sequencing data from the *Mycobacterium tuberculosis* complex (MTBC), the software can be used with any short read sequencing data from any organism. Currently, `KvarQ` is optimized for haploid organisms, but could be extended to analyze diploid data sets.

5.3.2. Algorithm and parameters

The following paragraphs describe the flow of information inside `KvarQ` from the `fastq` files to the final results that are displayed to the user (Figure 5.1).

Before the scanning process, target sequences are generated from known mutations or regions of interest. `KvarQ` uses a modular approach, where different "testsuites" (realized as python modules) define known single nucleotide polymorphisms (SNPs) or small regions of interest which harbour potential mutations (such as the Rifampicin Resistance-Determining Region, RRDR, in MTBC (Ramaswamy *et al.*, 1998)). The target genomic sequence is extracted from a reference genome and flanked on both sides with additional bases. This flanking increases the length of the sequence and thereby increases the coverage over the region of interest (see Figure 5.2). For each of these "target sequences", the corresponding base-sequence on the complementary strand is also generated.

The actual scanning algorithm is implemented in C and splits the `fastq` file into small parts that are distributed to multiple threads and scanned simultaneously. First, every read is trimmed by a user-specified minimum quality score. The trimmed part is then shifted along each of the target sequences and every match (not exceeding a specified number of single nucleotide differences) is recorded in the "hit-list" if a minimum overlap is warranted. Currently, no indexing/caching is used for comparison of the reads with the target sequences. This works fast enough for the short target sequences used in our analysis. The scanning process continues until the whole file is scanned or a specified minimum coverage is reached.

Next, the data gathered from the `fastq` files is translated into intermediary data structures ("coverages"). These data structures combine the reads from the original as well as the complementary strand, position them within the genome, and summarize non-matching bases in reads (these can be true mutations in the genome or sequencing errors).

Finally, the testsuites determine the results based on the information conveyed by the intermediary data structures created in the last step. Currently, mutations are extracted

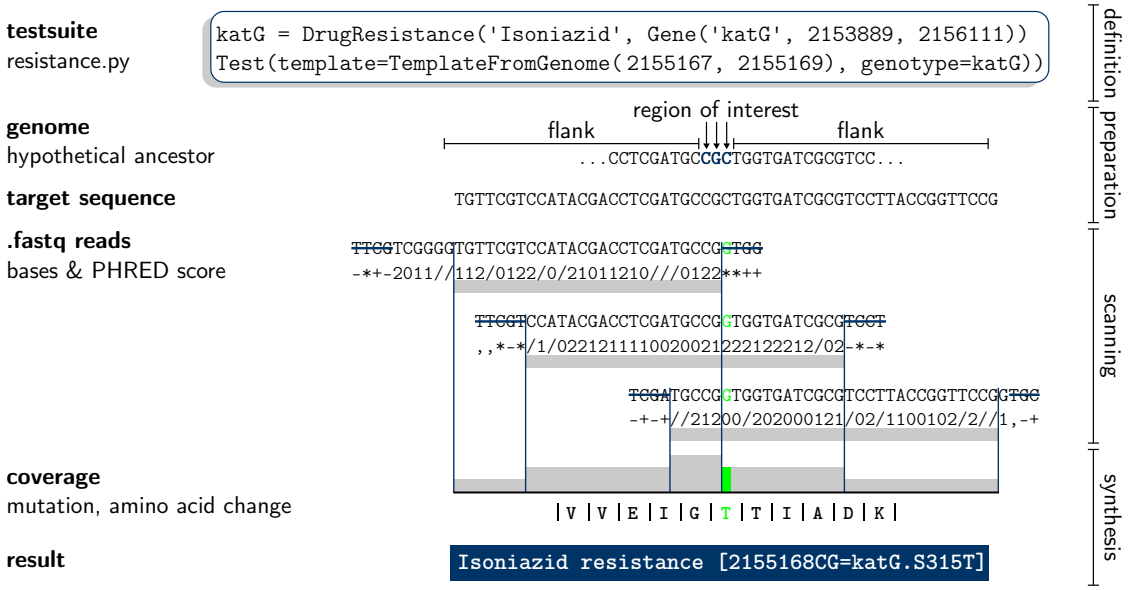


Figure 5.1.: **Simplified overview of scanning process and preparation.** "Testsuites" are python source files that define the SNPs of interest (or in this case a 3 base pair long region) as well as other relevant genetic information (in this case the *katG* gene in which mutations can confer isoniazid resistance). This information is used to extract a "target sequence" from a reference MTBC genome: on both sides, additional bases ("flanks") are concatenated to avoid border effects within the sequence of interest. During the scanning process, every read is trimmed depending on its PHRED score (in this case, a quality cutoff of `Q=13` was defined which corresponds to the ASCII character "/"). After the scanning, the part of the reads that matched the target sequence and exceeded the minimum quality score (represented with gray bars) are assembled to a "coverage" that indicates the overall coverage depth as well as all detected mutations (green). In a further step, additional information is generated from this coverage (such as the resulting amino acid sequence) and finally a short "result" string is generated that summarizes the result of the scanning process.

by using a simple threshold that separates them from sequencing errors – this is possible because `KvarQ` neglects bases that do not satisfy the minimum read quality requirements set by the user. How this mutation information is processed further varies depending on the various testsuites. For example, the phylogenetic testsuite we applied for MTBC uses three phylogenetically informative SNPs for every lineage and imposes logical constraints on *sub*lineages. For example, any genome belonging to the MTBC "Beijing" sublineage must also have the SNPs characteristic of "Lineage 2". The resistance testsuite checks for every mutation whether a change in amino acid sequence follows the nucleotide change (non-synonymous mutations) and only reports those to the user. If the results have a low confidence due to low coverage of the sequences, this is also added as a remark to the result output.

5.3.3. Data format and analysis

`KvarQ` uses `fastq` files with Solexa, Sanger and different Illumina quality scores as input files. The quality format is determined by a heuristic search. Only one input file per genome sequence is accepted, but paired-end data files can be merged into a single file. The output is structured data in `json` format, a human readable file that contains information about the scanning process (parameters used, `fastq` file size, statistics about read number, length and quality, and scan time), the final results, and the intermediary data structures that were used to calculate these results. The information contained in the `json` file can be used to re-calculate the results without the need to re-scan the `fastq` file, because most of the relevant information is contained in the intermediary data structures that are saved along with the results in the `json` file.

To facilitate the extraction of relevant results, `KvarQ` provides data analysis tools to inspect the data interactively from the command line or with a hierarchical menu-driven graphical user interface. The "`json` explorer" shows the summarized test results for every testsuite (`KvarQ`'s main goal is to be user-friendly) as well as detailed information about the coverage of every target sequence that was used by the different testsuites (Figures 5.2, 5.3). The interactive exploration of this wealth of information is intended for the advanced operator to get, for example, a better impression of the usefulness of newly designed target sequences or the `fastq` quality. Additionally, the "`json` explorer" displays overall number of reads in a `fastq` file and the length of the reads.

5.4. Results

We validated `KvarQ` with 880 `fastq` files of predominantly *Mycobacterium tuberculosis* complex (MTBC) isolates, for which we interrogated 206 genomic positions where polymorphisms have been previously described (Additional Information 1 and 2 in Appendix A.1).

In MTBC, single nucleotide mutations are reliable drug resistance markers, as most phenotypic drug resistance is conferred by single amino acid changes (Ramaswamy *et al.*, 1998; Riska *et al.*, 2000; Rodwell *et al.*, 2013). Hence, molecular DST is becoming increasingly widespread in the diagnosis of TB, and various genotyping platforms make use of these markers. Moreover, SNPs are powerful phylogenetic markers in MTBC, because similar to other genetically monomorphic bacteria, MTBC exhibits limited horizontal gene exchange and as a consequence, SNP homoplasies are rare (Hershberg *et al.*, 2008; Comas *et al.*, 2010). Hence, SNPs can be used to study the evolutionary history of MTBC, as well as classify clinical strains associated with variable degrees of virulence or

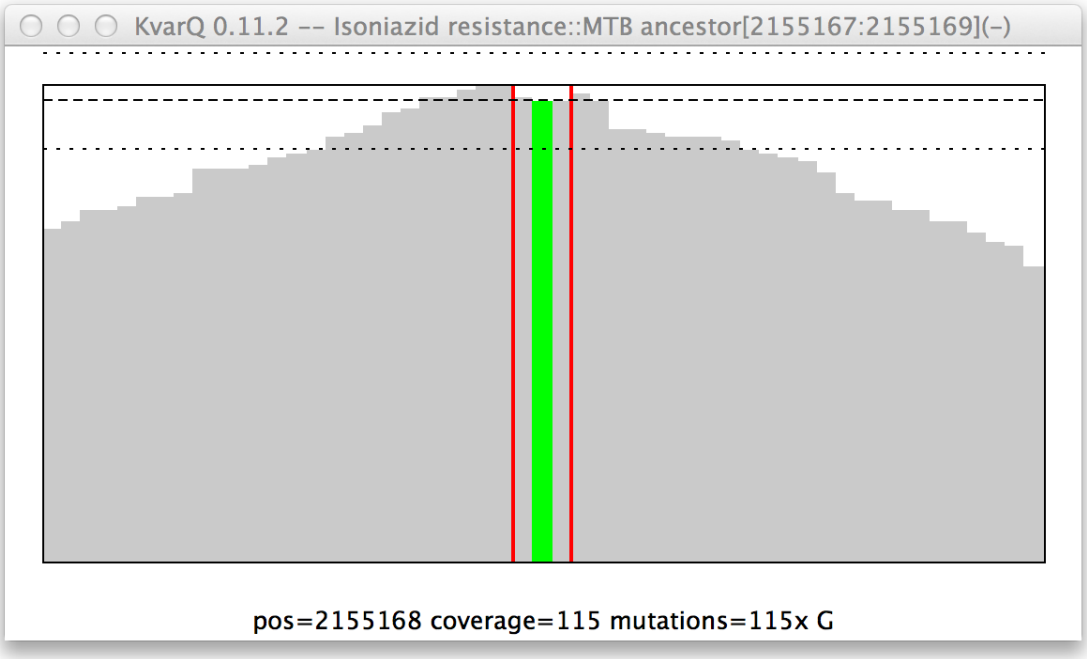


Figure 5.2.: **Interactive inspection of "coverage".** This screenshot shows the coverage over a short region of interest (3 basepairs within vertical red lines, corresponding to codon 315 in *katG*) and the surrounding flanking sequence. Note the mutation in the middle of the three base pair long sequence (in green, text description at bottom of window). The decrease of coverage on both sides is due to the minimum overlap that is set to a value equal to the length of the two flanking regions.

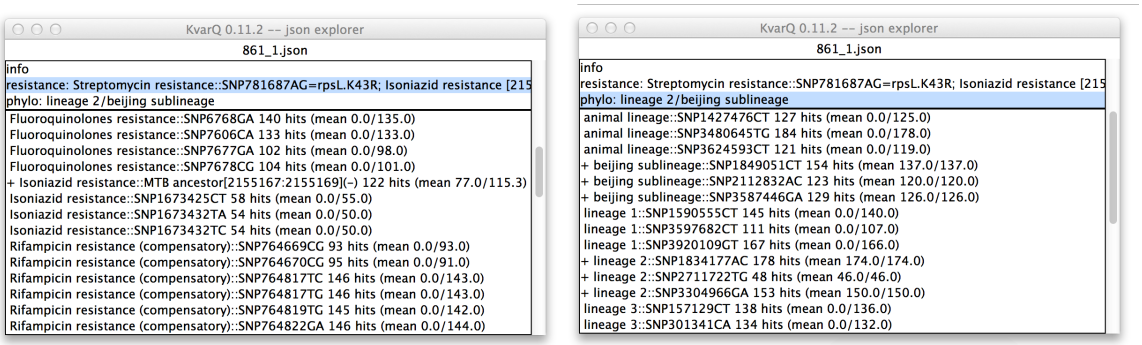


Figure 5.3.: **Interactive inspection of `json` file.** The upper pane of each window in these screenshots shows the main categories of data contained in the `json` file. In the left window, the drug resistance section is selected and the lower pane shows details about all target sequences in this testsuite (the "+" in front of the "Isoniazide resistance" indicates a non-synonymous mutation). In the right window, the phylogenetic section is selected, showing that all SNPs for "lineage 2" and "beijing sublineage" were found.

different clinical outcomes (Coscolla *et al.*, 2010; Stucki *et al.*, 2012). MTBC comprises several phylogenetic lineages, and SNPs represent ideal markers to classify clinical MTBC isolates into these lineages (Comas *et al.*, 2013; Comas *et al.*, 2009b).

In this study, we included 27 canonical SNPs as phylogenetic markers defining MTBC lineages and sublineages, as well as 32 single nucleotide mutations and three genomic regions (63 base pairs (bp) in *gyrA*, 81 bp in *rpoB* and 3 bp in *katG*) associated with drug resistance for the detection of variants with `KvarQ` (Tables A.1 and A.2 in Appendix A.1).

We tested `KvarQ` with a set of "in-house" generated whole genome sequences of clinical MTBC isolates and additionally downloaded sequences from public sequence read archives (Additional Information 2 in Appendix A.1). Different, overlapping subsets of `fastq` files were used for *i)* the comparison of SNP-calls with a standard BWA-based mapping pipeline (N=206), *ii)* detecting drug resistance mutations (N=19) and iii) comparison of `KvarQ` phylogenetic classification with laboratory based classification (N=321)—see Figure 5.4. In addition to these well-characterized genome subsets, we used a previously published "blind" subset of 388 `fastq` files downloaded from a public source to extract phylogenetic information and drug resistance mutations, which were not reported in the original publication (Walker *et al.*, 2013b).

5.4.1. `KvarQ` scanning times and overall performance

First, we applied `KvarQ` to all 880 `fastq` files and calculated scanning times for the set of SNPs described. The scanning times were found to be correlated with the average genome coverage (or the total base output) in the `fastq` file. A genome sequence file with 100-fold coverage of the MTBC genome required approximately 2 minutes of `KvarQ` scanning time, and increased linearly with sequence coverage (Figure 5.5).

To assess overall performance, we looked at phylogenetic classifications in all 880 `fastq` files. These files included whole genome sequences generated on different sequencing devices using different laboratory procedures (i.e. library preparations), and of unknown quality (Additional Information 2 in Appendix A.1). Successful phylogenetic classification was obtained for 865/880 `fastq` files (98.2%) (Figure 5.6). The remaining 15 included the chimpanzee bacillus (Coscolla *et al.*, 2013), *Mycobacterium canettii* and eight confirmed non-MTBC isolates, for all of which no MTBC lineage classification would be expected. For five other strains, no MTBC lineage-specific SNP was detected, which can be due to low coverage or because isolates were not MTBC.

Furthermore, we compared the number of `KvarQ` hits per position with overall base output in the `fastq` file (data not shown). The correlation varied strongly between positions

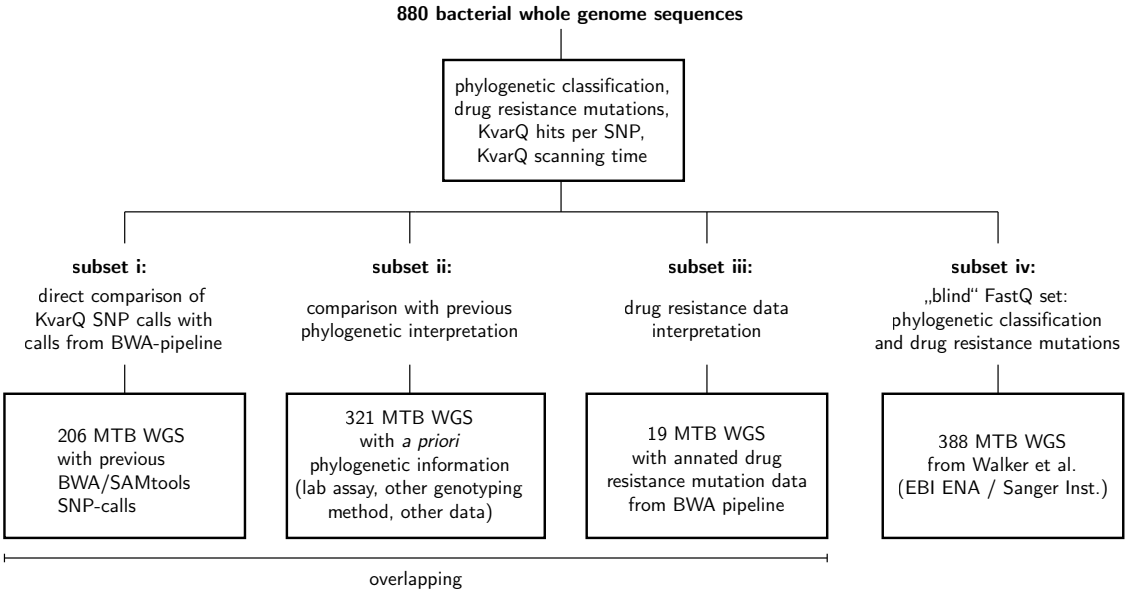


Figure 5.4.: **FastQ dataset used to validate KvarQ.** A total of 880 whole genome sequences in `fastq` format from various sources were used in this study. All 880 genome sequences were scanned for phylogenetic classification and drug resistance mutation identification. Different, overlapping subsets were used to *i)* compare SNP calls obtained with `KvarQ` with SNP calls of our standard SNP-calling pipeline based on BWA and SAMtools for 206 isolates, *ii)* compare `KvarQ` phylogenetic classification of 321 MTBC isolates with previous phylogenetic information, *iii)* compare `KvarQ` drug resistance mutations with previously identified drug resistance associated mutations in 19 MTBC isolates, and *iv)* obtain additional information, i.e. phylogenetic classification and drug resistance mutations, from a "blind" set of 388 genome sequences from a recent study (Walker *et al.*, 2013b). More information can be found in Additional Information 2 in Appendix A.1.

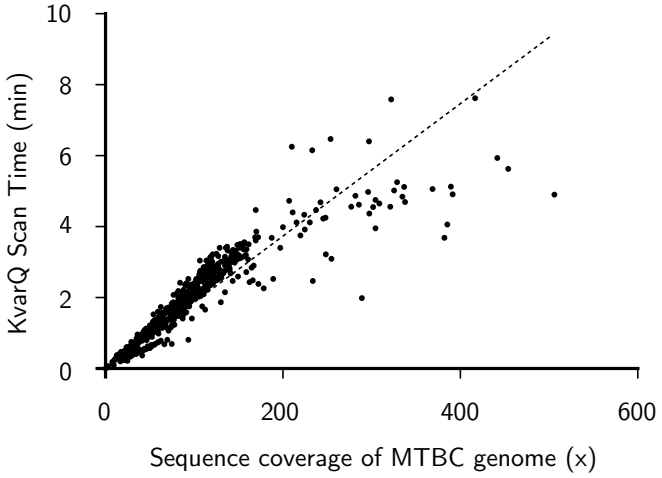


Figure 5.5.: **KvarQ scanning time for all 206 polymorphic positions with respect to average sequence coverage of the MTBC genome per `fastq` file.** Each dot represents one isolate (i.e. `fastq` file). Scanning time was found to be 112 seconds / 100x coverage. Eleven isolates of 880 were excluded due to variable read lengths (difficult calculation of total base output). Three additional isolates isolates were excluded because of file sizes larger than 10 GB.

on the genomes and between `fastq` files, which indicates that the number of hits per SNP depends on the SNP queried (i.e. the particular genomic region being interrogated) and the quality of the base calls in the `fastq` file.

5.4.2. Comparison of SNP-calls with BWA calls

A subset of 206 `fastq` files was used to directly compare SNP-calls obtained with `KvarQ` with SNP-calls generated with our standard pipeline using BWA, SAMtools, BCFtools and custom scripts for filtering (Coscolla *et al.*, 2013). For each of the 206 positions (single nucleotide positions plus regions of interest), the BWA list of variants in the corresponding genome was interrogated for presence of an alternative allele compared to the reference genome (reconstructed hypothetical ancestor (Comas *et al.*, 2010; Comas *et al.*, 2013)). With this, a direct comparison of all 206 positions in 206 strains was performed, resulting in 42,436 total data points (Table 5.1).

A total of 782 mutations were called by both `KvarQ` and the BWA-pipeline (1.8% of all data points). This corresponds to a sensitivity of 99.2% in mutation calling of `KvarQ` (782/788) taking the BWA pipeline as the gold standard. Specificity was found to be 99.9% (41,615/41,648 positions).

Thirty-three mutations (0.08% of all data points) from 20 `fastq` files were called by `KvarQ`, but no mutation was found in the corresponding BWA SNP-list. Upon manual

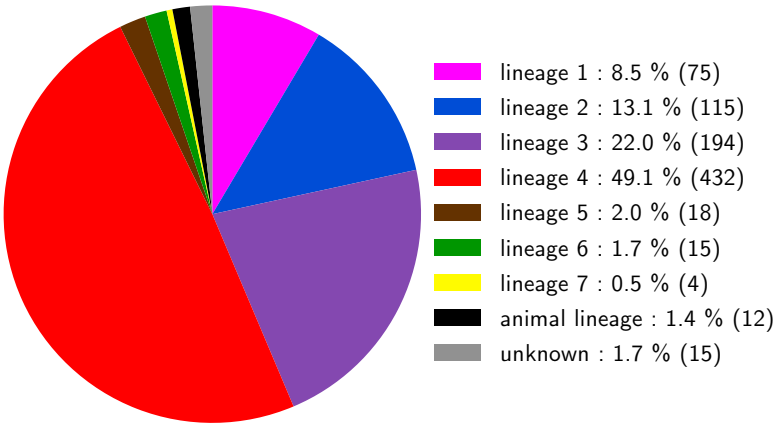


Figure 5.6.: **Phylogenetic classification of all 880 isolates used in this study.** This figure shows the distribution of phylogenetic markers in all of the scanned genomes. For 865 isolates, an MTBC lineage-specific SNP was found. In 15 isolates, no MTBC lineage-specific SNPs were found, either because isolates were non-MTBC (10 were known to be non-MTBC), or because coverage was low.

Table 5.1.: **Comparison of SNP-calls at 206 polymorphic positions in 206 MTBC genome sequence `fastq` files.**

	SNP in BWA mapping pipeline					
	yes		no		Total	
	N	%	N	%	N	%
SNP in KvarQ	782	99.24	33	0.08	815	1.92
no SNP in KvarQ	6	0.76	41,615	99.92	41,621	98.08
Total	788	100	41,648	100	42,436	100

inspection, we found a successful SNP-call in the SAMtools *pileup*-list for each of the 33 mutations, but these mutations were filtered out in the subsequent heuristic filtering step including the SAMtools `varFilter` command for unknown reasons despite good quality and coverage.

We found three `fastq` files where `KvarQ` missed a single SNP and one file where `KvarQ` missed 3 SNPs despite six calls in the BWA-pipeline. The BWA read depths for these six calls (0.01% of all data points) were 5, 6, 21, 38, 64 and 77. Corresponding `KvarQ` number of hits were 0, 1, 2, 3, 4 and 5, respectively. For strain MTB_erdman_SRR017677, when the `KvarQ` quality cutoff was removed (all reads considered), the number of hits increased to 57, 84 and 120, respectively. Subsequent FastQC analysis [1] showed poor `fastq` quality over the whole read length (data not shown).

[1] `http://www.bioinformatics.babraham.ac.uk/projects/fastqc/`

5.4.3. Validation of phylogenetic classification

A subset of 321 isolates was used to compare the phylogenetic classification by `KvarQ` with data obtained previously by various genotyping methods (molecular assays in our laboratory (Stucki *et al.*, 2012)), data obtained from the collaborating laboratory that provided DNA for WGS, or meta-data from the sequence read archive in the case of downloaded files. The comparison included seven main phylogenetic lineages (Comas *et al.*, 2013), the animal-associated MTBC clade (*M. bovis*/*M. caprae*), the Beijing sublineage of Lineage 2, and 10 non-MTBC isolates. In 309/321 (96.3%) isolates, `KvarQ` detected MTBC lineage-specific SNPs that were in agreement with the previous classifications (Additional Information 2 in Appendix A.1). The remaining 12 genome sequences for which no MTBC lineage-specific SNP was found included eight non-MTBC, the *chimpanzee bacillus* and *M. canettii*, for which no MTBC-lineage specific SNP was expected, and two isolates (0.6%) with low `KvarQ` coverage and therefore no lineage-call.

5.4.4. Drug resistance associated mutations

`KvarQ` was applied to a subset of 19 strains harbouring known mutations associated with drug resistance to validate the interpretation (annotation) of the mutations. The 19 strains overlapped with the 206 strains of the direct comparison described above, but the resistance-associated mutations were identified with separate scripts from the annotated SNP-list in the BWA pipeline. The 179 drug resistance associated genomic positions were interrogated with `KvarQ`, including the 81 bp Rifampicin Resistance Determining Region (RRDR) of *rpoB*, the 63 bp Quinolone Resistance Determining Region of *gyrA* (QRDR) and 3 bp in the codon 315 of *katG* (Table A.2 in Appendix A.1).

Among the 19 strains, 16 (84.2%) had perfectly congruent drug resistance mutations identified by `KvarQ` when compared to the BWA pipeline. For one strain (GQ1580), one mutation in *rpoC* was not found by `KvarQ`. However, this mutation was not in the list of target SNPs (Table A.2 in Appendix A.1). For one isolate, MTB_KZN_605, an additional mutation in *gyrA* was found with `KvarQ`, and was not found with the BWA pipeline due to low read depth and the presence of an alternative allele. For another isolate, MTB_russia_ERR015616_1, six additional mutations in five genes were found by `KvarQ`. All mutations in these two isolates were traced back to the filtering problem discussed above, and were found in the *pileup* file.

In total, `KvarQ` detected 314 drug resistance associated mutations in 139/880 files (15.8%) (Figure 5.7).

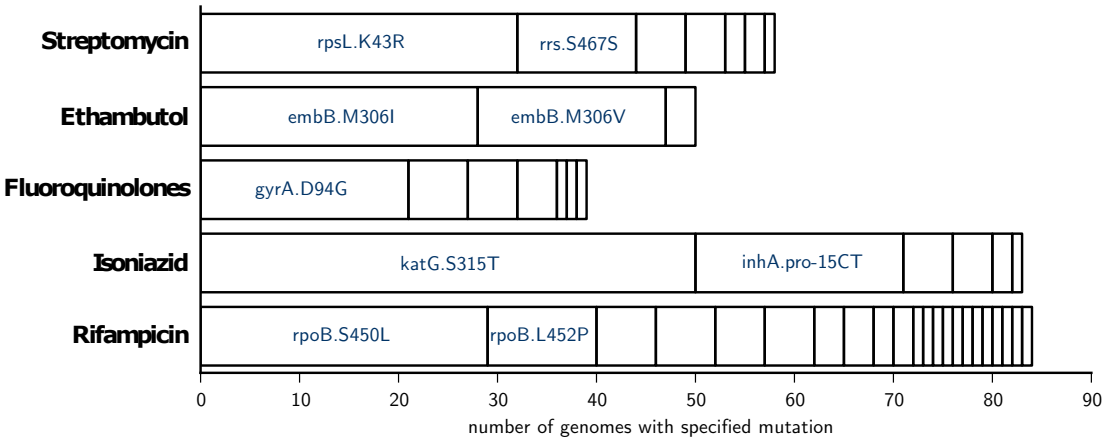


Figure 5.7.: **MTBC drug resistance associated mutations found in all 880 isolates.** This figure shows all 314 mutations that were found with `KvarQ` in 139 isolates with at least one mutation in any of the drug resistance associated genes that were analysed (see Table A.2 in Appendix A.1). Only mutations that were found in at least ten isolates are labeled.

5.4.5. "Blind" set of FastQ files from a public repository

Finally, a "blind" set of whole genome sequences was used to illustrate the added value provided by `KvarQ`. We downloaded available sequences of a recent publication that used whole genome sequencing for molecular epidemiology of MTBC transmission (Walker *et al.*, 2013b). This publication did not report drug resistance data or phylogenetic data. We therefore analyzed all available sequences for phylogenetic SNPs and drug resistance mutations. The original sample set consisted of 390 isolates, but two `fastq` files were not found for download (`http://www.sanger.co.uk`, ERR192249 and ERR192250). Results for 388 isolates were obtained in less than 12 hours. A total of 62 drug resistance mutations were identified in 33 patient isolates from 11 patients (Figure 5.8). All paired isolates (cross-sectional: paired pulmonary and extrapulmonary isolates from the same patients; longitudinal: paired isolates from the same patient separated by at least 6 months; shown as smaller dotted boxes in Figure 5.8) were found to harbour identical drug resistance associated mutations, except for one patient (P000155), where an additional mutation was found in two of three isolates. Isolates of one community transmission cluster (defined by MIRU-VNTR), cluster 9, were found to harbour drug resistance associated mutations (large box in Figure 5.8). Within cluster 9, all patient isolates harboured the same drug resistance mutations, except for one patient (P000179) with three otherwise unidentified mutations. Among the 388 isolates, we found a predominance of Lineage 4 (i.e. the Euro-American lineage) in the sample set (65%), followed by Lineage 3 (24%) and Lineage 1 (6%) (Figure 5.9).

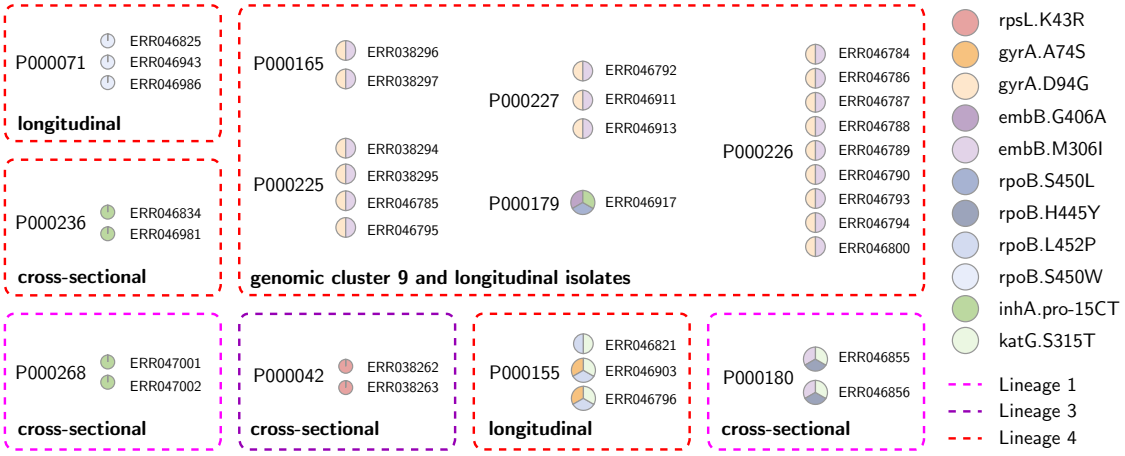


Figure 5.8.: **KvarQ identified 62 drug resistance associated mutations in 33 of 388 patient isolates in a "blind" set of genome sequences from Walker *et al.* (2013b).** Each circle represents one isolate (Patient number given as P-number, isolate given as ERR-number). Only isolates with drug resistance associated mutation identified by KvarQ are shown. The 62 individual drug resistance mutations are shown as colors of pies. Phylogenetic lineages were obtained with KvarQ and are shown as colored and dashed boxes. Group definitions were obtained from the original publication (Walker *et al.*, 2013b): cross-sectional isolates were paired pulmonary and extrapulmonary isolates from the same patient, longitudinal isolates were paired isolates from the same patient separated by at least 6 months, and genomic cluster 9 was a large MIRU-VNTR defined cluster of transmission in the community. Paired isolates (longitudinal and cross-sectional) as well as isolates of cluster 9 were found to harbour the same mutations, except for P000179, where three otherwise not detected mutations were found by KvarQ. For patient P000155, an additionally acquired mutation was found in two of three isolates. Two of the 390 whole genome sequences in the original publication were not found for download.

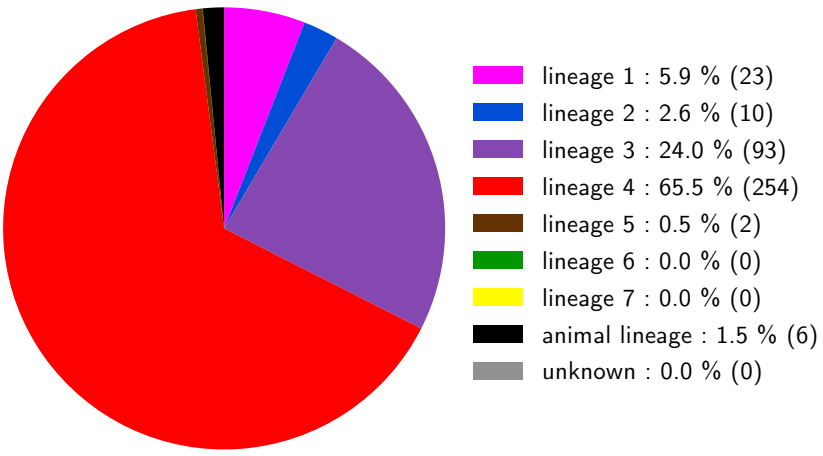


Figure 5.9.: **Main MTBC phylogenetic lineage classification of the fastq set of Walker *et al.* (2013b), consisting of 388 isolates.** Successful lineage assignment was obtained for 388/388 (100 %) isolates. Figure includes all clustered isolates as well as longitudinal, cross-sectional and household isolates.

5.5. Discussion

In this article, we present a new approach to *in silico* SNP-typing from whole genome sequence data. `KvarQ` scans `fastq` short sequence reads directly for known polymorphisms instead of mapping or assembling the complete genome. This results in a massively reduced computational time required compared to established WGS analysis tools. At the same time, `KvarQ` offers a high degree of flexibility and is also very user-friendly. This was achieved by developing an extended command-line interface together with pre-compiled packages with a graphical user interface.

`KvarQ` was designed for bacterial genomes, but is applicable to any `fastq` sequence data. The default parameters work best for the different Illumina machines, but can easily be adapted to data from other manufacturers. For example, the Roche 454 reads in our testset were scanned with a lower quality cutoff. Due to the lack of relevant data from MTBC isolates, we were not able to test reads from Pacific Bioscience devices.

`KvarQ` currently takes a single `fastq` file as input. Paired-end data can be used as well, but reads have to be concatenated before running `KvarQ`. Trimming of files is not necessary, as `KvarQ` will accept only high-quality parts of the reads anyway.

Both the reference sequence and the interrogated polymorphisms can be changed or extended. Changing the SNPs of interest can be achieved by *template* files, which are lists of nucleotide positions screened by `KvarQ`. An extended documentation including instructions for changing the target polymorphisms is distributed with the software and available online[2]. Scanning parameters are adjusted by invoking a number of options in the command-line interface or via the settings dialog in the graphical user interface.

Using our test set of 880 `fastq` files, we obtained `KvarQ` running times of less than 2 minutes on a high-performance computer (average running time 110 seconds) and less than 20 minutes on a standard desktop computer (Figure 5.5). This important decrease in time for analysis (when compared to the conventional pipeline in which the whole genome sequence is first reconstructed) is crucial for the efficient treatment of data as WGS becomes more widely available. The overall performance of `KvarQ` was high, as reflected by 98.3% of genome sequences successfully assigned to a phylogenetic MTBC lineage (Figure 5.6). By including phylogenetically redundant SNPs, the sensitivity of classification was increased. Specifically, classifications were obtained if 2 of 3 redundant SNPs were found mutant. The 96.3% congruence of lineage-assignment of `KvarQ` with *a priori* lineage classification illustrates how WGS combined with a quick tool such as `KvarQ` has the potential to replace previous genotyping methods.

[2] `http://www.swisstph.ch/kvarq`

When comparing KvarQ SNP-calls with calls from our mapping pipeline using BWA and SAMtools, we found a high sensitivity and specificity of 99.2% and 99.9%, respectively. The three mutations that were called by the BWA pipeline but not by KvarQ were resolved when lowering the quality cutoff parameter. FastQC[3] confirmed the low quality of the reads in these files.

Drug resistance mutations obtained with KvarQ and BWA in the 19 compared files were nearly identical, but showed one limitation. Mutations conferring resistance to pyrazinamide are found in any part of *pncA*, which is 561 bp long (Cole *et al.*, 1998). KvarQ was designed to scan short stretches of DNA, and we therefore excluded mutations in *pncA* to avoid an increase in scanning time.

In the total set of 880 genomes, we found many additional drug resistance associated mutations (Figure 5.7) that could not be validated due to the lack of laboratory DST data for many of the isolates and the complex genetic nature of drug resistance in MTBC (Warner *et al.*, 2013).

Successful KvarQ calls depended on the selection of suitable SNPs. Several SNPs had to be replaced in the testing phase. Most SNPs that were replaced occurred in repetitive regions, and these SNPs were also never called when using the BWA-pipeline. We therefore recommend including only SNPs that were previously found to be in core genome regions, in non-repetitive regions and where no deletions have been described.

With a "blind" set of genome sequences obtained from a recent publication (Walker *et al.*, 2013b), we were able to obtain lineage classifications and drug resistance calls that had not been reported in the original publication in less than one day. Using the clustering information of the publication, all isolates identified as drug-resistant by KvarQ and from the same patient (paired isolates) or the same patient cluster shared the same drug resistance mutations (except patient P000179, see Figure 5.8) and were assigned to the same phylogenetic lineage, highlighting the consistency in mutation calling. The fast generation of drug resistance and phylogenetic data illustrates the application of KvarQ for quick targeted scanning of large public datasets.

KvarQ was designed to interrogate positions with known mutations. However, as some regions (e.g. hot spot regions of drug resistance) can harbour several mutations in close proximity, we included the capability to scan regions of interest for the presence of any (new) mutation. These regions can be of any length, but the scanning time increases in square with sequence length. The DST testsuite in the current version checks some short regions known to harbour many different drug resistance associated mutations and reports any new mutation that would result in a amino acid change of the associated gene. In

[3] http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

future releases, KvarQ could be extended to scan for other genomic polymorphisms, such as small insertions and deletions (InDels) as well as large sequence deletions or duplications.

5.6. Conclusion

In conclusion, KvarQ provides a user-friendly and highly flexible platform for rapid and targeted analysis of fastq files. KvarQ will help overcome the hurdles of whole genome sequence analysis in clinical microbiological laboratories and other settings where bioinformatics capacity is limited. The short running times, the user-friendly graphical interface and the high configurability at the command line level allows analysing hundreds of fastq files in a short time.

5.7. Materials and methods

5.7.1. Informative SNPs for MTBC

We used known phylogenetic markers and drug resistance associated mutations of MTBC to validate and benchmark KvarQ when scanning genome sequences of clinical MTBC strains. Phylogenetically informative SNPs were selected from previous publications (Stucki *et al.*, 2012; Comas *et al.*, 2009b). For each of the seven main phylogenetic lineages of human-associated MTBC, plus the *Mycobacterium bovis*/*M. caprae* lineage, we included three redundant canonical SNPs as markers for the corresponding lineage. Previously published SNPs were complemented with SNPs obtained as described before (Stucki *et al.*, 2012) for cases where less than three SNPs per phylogenetic clade were available. The extraction of these additional SNPs was based on 172 genomes described by (Comas *et al.*, 2013). For Lineage 2, we additionally included known polymorphisms to discriminate the so-called "Beijing" lineage from non-Beijing Lineage 2 strains. An overview of the phylogenetically informative SNPs included in KvarQ is shown in Table A.1 in Appendix A.1. We included drug resistance-mutations obtained from the Tuberculosis Drug Resistance Database (TBDReaMDB) (Sandgren *et al.*, 2009), and additionally compensatory mutations from (Comas *et al.*, 2011a). High-confidence mutations for the most important anti-tuberculosis drugs were selected (isoniazid, rifampicin, ethambutol, streptomycin, fluoroquinolones and second-line injectable drugs). Pyrazinamide-resistance conferring mutations were excluded. Mutations in TBDReamDB are listed as codon changes, therefore we generated the corresponding nucleotide changes for inclusion in KvarQ. For rifampicin- and fluoroquinolone-resistance conferring mutations, we included

the *rifampicin-resistance determining region* (RRDR) (Ramaswamy *et al.*, 1998) and the *quinolone resistance determining region* (QRDR) (Takiff *et al.*, 1994), respectively, rather than specific positions. The codon 315 of *katG* was also treated as a region of three base pairs. All included mutations and regions associated with drug resistance mutations are listed in Table A.2 in Appendix A.1. All genomic positions given in this study refer to the genome of the strain H37Rv (NC000962.3 / AL123456.3) (Cole *et al.*, 1998; Camus *et al.*, 2002; Lew *et al.*, 2011).

5.7.2. Short-read datasets

We used a test set of 880 bacterial whole genome sequences in `fastq` format to screen for mutations with `KvarQ`. This set represents a global and diverse collection of clinical isolates of MTBC, and includes drug-resistant strains from patients and from in vitro evolution experiments, as well as non-MTBC bacterial genome sequences. Additionally, the genome sequences were chosen to represent a technically diverse collection (Illumina HiSeq 2000, GAIIx and MiSeq). In addition to the genome sequences generated in-house, the test set contains `fastq` files downloaded from `www.sanger.ac.uk`, the European Nucleotide Archive (ENA; `http://www.ebi.ac.uk/ena/`) and the Sequence Read Archive (SRA; `http://www.ncbi.nlm.nih.gov/sra`). More information on the `fastq` files including accession numbers can be found in Additional Information 2 in Appendix A.1.

5.7.3. Mapping data for comparison with KvarQ

To compare `KvarQ` SNP calls with SNP calls obtained from conventional mapping of `fastq` short sequencing reads to a reference, we used a previously published computational pipeline (Coscolla *et al.*, 2013). In brief, all short reads were mapped to a hypothetical reconstructed ancestor with BWA, reads were piled up with SAMtools and variants called with BCFtools (Li *et al.*, 2009). Filtering for low quality calls was done with customized Perl scripts. The 206 positions that were used in `KvarQ` for allele detection were extracted with a Python script and compared to the `KvarQ` calls in the corresponding sequence. Discrepant results were manually checked with Artemis (Rutherford *et al.*, 2000).

5.7.4. KvarQ parameters used

All `KvarQ` results were generated using default parameters, i.e. a quality cutoff of 13, a minimum read length of 25, a minimum overlap of 25, a minimum coverage of two reads for allele calls and a maximum of two errors per read. Parameters were adjusted for `fastq`

files with low coverage or low quality values, and are reported in Additional Information 2 (Appendix A.1). In case of paired-end sequencing files, only one of two read files was used. Files were manually concatenated in case of low coverage (reported in Additional Information 2 in Appendix A.1).

5.7.5. Hardware specification

`KvarQ` runs on any system that runs Python and has a POSIX threads implementation (this includes notably Windows, OS X and Linux). Analyses in this study were performed on a Red Hat Enterprise Linux Server release 5.9 (Tikanga) with four six-core AMD Opteron CPUs and 132 GB RAM, using 8 threads.

5.8. Availability and requirements

Project name: KvarQ: Targeted and Direct Variant Calling from FastQ Reads of Bacterial Genomes
Project home page: http://www.swisstph.ch/kvarq
Operating systems: platform independent; requires Python 2.7, POSIX thread library; pre-compiled packages for OS X (10.6.8 and later) and Windows (7 and later)
Programming Language: Python/C
Licence: GNU GPL v3

6. Tracking a tuberculosis outbreak over 21 years: strain-specific single nucleotide polymorphism-typing combined with targeted whole genome sequencing

David Stucki [1,2,§], Marie Ballif [3,§], Thomas Bodmer [4,5], Mireia Coscollà [1,2], Anne-Marie Maurer [6], Sara Droz [4], Christa Butz [7], Sonia Borrell [1,2], Christel Längle [4], Julia Feldmann [1,2], Hansjakob Furrer [8], Carlo Mordasini [7], Peter Helbling [9], Hans L. Rieder [10,11], Matthias Egger [3,12], Sebastien Gagneux [1,2,*], Lukas Fenner [1,2,3,*]

[1] Swiss Tropical and Public Health Institute, Basel, Switzerland

[2] University of Basel, Switzerland

[3] Institute of Social and Preventive Medicine, University of Bern, Switzerland

[4] Institute for Infectious Diseases, University of Bern, Switzerland

[5] labormedizinisches zentrum Dr Risch, Bern-Liebefeld, Switzerland

[6] Cantonal Health Authorities, Canton of Bern, Switzerland

[7] Bernese Lung Association, Bern, Switzerland

[8] Department of Infectious Diseases, Bern University Hospital and University of Bern, Switzerland

[9] Federal Office of Public Health, Bern, Switzerland

[10] International Union Against Tuberculosis and Lung Disease

[11] Institute of Social and Preventive Medicine, University of Zürich, Switzerland

[12] School of Public Health and Family Medicine, University of Cape Town, South Africa

[§] Contributed equally

[*] Corresponding authors

This paper was accepted for publication in *The Journal of Infectious Diseases* in 2014.

6.1. Abstract

Background. Whole-genome sequencing (WGS) is increasingly used in molecular-epidemiological investigations of bacterial pathogens, despite cost- and time-intensive analyses. We combined strain-specific single-nucleotide polymorphism (SNP) typing and targeted WGS to investigate a tuberculosis cluster spanning 21 years in Bern, Switzerland.
Methods. On the basis of genome sequences of three historical outbreak *Mycobacterium tuberculosis* isolates, we developed a strain-specific SNP-typing assay to identify further cases. We screened 1,642 patient isolates and performed WGS on all identified cluster isolates. We extracted SNPs to construct genomic networks. Clinical and social data were retrospectively collected.
Results. We identified 68 patients associated with the outbreak strain. Most received a tuberculosis diagnosis in 1991–1995, but cases were observed until 2011. Two thirds were homeless and/or substance abusers. Targeted WGS revealed 133 variable SNP positions among outbreak isolates. Genomic network analyses suggested a single origin of the outbreak, with subsequent division into three subclusters. Isolates from patients with confirmed epidemiological links differed by 0–11 SNPs.
Conclusions. Strain-specific SNP genotyping allowed rapid and inexpensive identification of *M. tuberculosis* outbreak isolates in a population-based strain collection. Subsequent targeted WGS provided detailed insights into transmission dynamics. This combined approach could be applied to track bacterial pathogens in real time and at high resolution.

6.2. Introduction

Tuberculosis transmission has traditionally been investigated using contact tracing and molecular typing (Cook *et al.*, 2012; McElroy *et al.*, 2003). However, social contact data are often hard to obtain retrospectively, especially in high-risk groups such as homeless individuals and substance abusers, who are difficult to trace (Anderson *et al.*, 2014; Asghar *et al.*, 2009; Burki, 2010; Lambregts-van Weezenbeek *et al.*, 2003). Moreover, classical molecular epidemiological techniques such as IS6110 restriction fragment-length polymorphism (RFLP) analysis and mycobacterial interspersed repetitive unit – variable number of tandem repeat (MIRU-VNTR) analysis interrogate only a small proportion of the mycobacterial genome and therefore suffer from limited resolution (Kato-Maeda *et al.*, 2011b).

More recently, whole-genome sequencing (WGS) of *Mycobacterium tuberculosis* has been used to investigate tuberculosis outbreaks (Walker *et al.*, 2013a). Known as "genomic epidemiology" (Gardy, 2013), this emerging field uses WGS to detect unknown transmission events, identify superspreaders, and exclude or confirm epidemiologically suspected transmission links (Bryant *et al.*, 2013b; Gardy *et al.*, 2011; Roetzer *et al.*, 2013; Walker *et al.*, 2013b). Moreover, WGS can also be used to detect drug resistance mutations (Köser *et al.*, 2013). Even though routine WGS has the potential to replace classical genotyping (Walker *et al.*, 2013b; Diep, 2013), analyzing WGS data remains resource-intensive and requires further standardization to meet public health needs, particularly for tracking ongoing outbreaks in real time (Walker *et al.*, 2013a).

In 1993, a tuberculosis outbreak was reported in the Canton of Bern, Switzerland (Genewein *et al.*, 1993). Twenty-two cases were involved, and their *M. tuberculosis* isolates shared identical IS6110 RFLP patterns. As in other affluent countries (Abubakar *et al.*, 2012; Bamrah *et al.*, 2013), this outbreak involved mainly homeless individuals and substance abusers. In 2012, we studied the molecular epidemiology of tuberculosis in Switzerland. We used the classic methods of spoligotyping and MIRU-VNTR to genotype 520 *M. tuberculosis* isolates from patients in whom tuberculosis was diagnosed between 2000 and 2008 (12.3% of all culture-confirmed tuberculosis cases in Switzerland during the study period) (Fenner *et al.*, 2012b). Among 68 isolates from the Canton of Bern, two were identified as belonging to the same outbreak described in 1993, indicating that this particular strain was still circulating in the area.

In the present study, we used a combination of single-nucleotide polymorphism (SNP) typing and targeted WGS to track the spread of the outbreak over two decades. Using representative isolates from the historical outbreak, we first developed a novel strain-specific SNP-typing assay to rapidly and inexpensively identify all tuberculosis cases caused by

this strain in the Canton of Bern between 1991 and 2011. We then applied targeted WGS to all cluster isolates identified by the screening assay to study the outbreak dynamics in relation to social contact information.

6.3. Methods

6.3.1. Study setting and sample set

We subcultured 1,642 patient isolates available from the *M. tuberculosis* strain collection at the Institute for Infectious Diseases (Bern, Switzerland). These isolates were all collected between 1991 and 2011 and correspond to 84.6% of all 1,940 tuberculosis cases (all forms) notified in the Canton of Bern during the same period (Figure 6.1). Subcultures were performed on Löwenstein-Jensen slants according to international laboratory standards. Purified DNA for WGS was obtained using the CTAB extraction method after subculturing a single colony in 7H9 liquid medium (Embden *et al.*, 1993). Bulk extracts (i.e. without single colony selection) were available for four isolates.

The collection included the 22 historical outbreak isolates reported by Genewein *et al.* (1993). One strain isolated in 1987, before systematic collection was started in 1991, was also included (P028, originally termed "patient 1" (Genewein *et al.*, 1993)). Finally, for one outbreak patient from 1991, we identified an additional strain isolated in 1988 (P006A, originally termed "patient 2"; Figure 6.1).

6.3.2. Cluster strain-specific SNP-typing assay and screening of strain collection

We performed WGS on one historical outbreak isolate from 1992 (Genewein *et al.*, 1993) and two isolates from 2005 and 2008 with the same MIRU-VNTR and spoligotyping patterns (Fenner *et al.*, 2012b) associated with the so-called "Bernese cluster", as described below. We also performed WGS on two isolates with the same spoligotyping pattern (isolation years 2001 and 2004) but a different MIRU-VNTR pattern (three and four different loci, compared with the outbreak isolate), one additional Lineage 4 isolate from another study, and the reference strain H37Rv (Figure 6.2). The three outbreak isolates shared 118 SNPs not observed in any of the control isolates (Figure 6.2). We used one of these outbreak-specific SNPs (878,174 GA; position in reference to H37Rv) to develop a real-time polymerase chain reaction (PCR) SNP-typing assay (TaqMan, Life Technologies, Switzerland), as described previously (Stucki *et al.*, 2012). All 1,642 available isolates were screened for the presence of that SNP. For confirmation, we subjected all isolates with

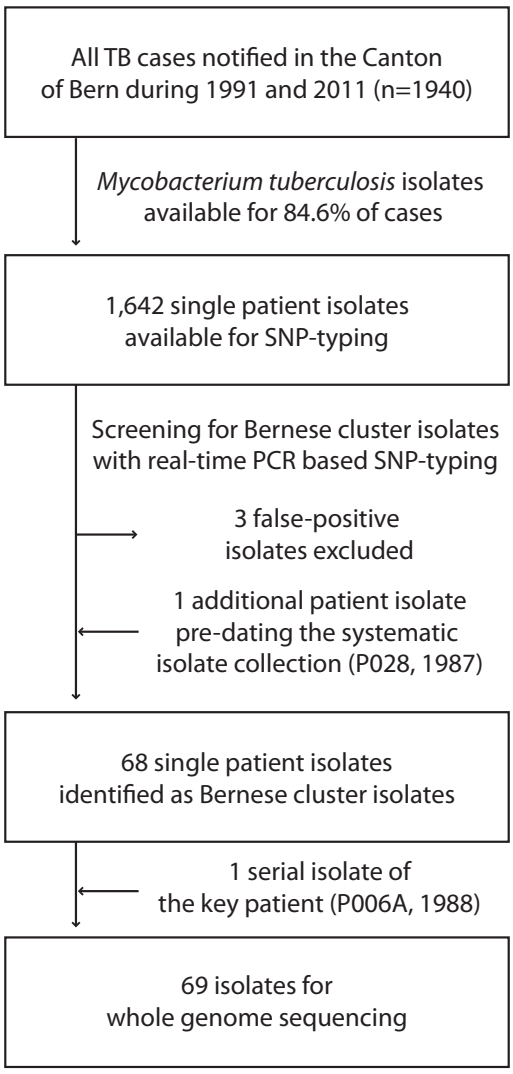


Figure 6.1.: **Overview of patient isolates and whole genome sequences generated.** A total of 1,642 isolates collected between 1991 and 2011 were available for single-nucleotide polymorphism (SNP) genotyping. Three isolates showed ambiguous SNP-typing results and were excluded. One additional patient isolate (P028; isolated in 1987) reported in the original publication (Genewein *et al.*, 1993) and predating the systematic collection of isolates in 1991 was included in the study. For the key patient, a second isolate (P006B; isolated in 1991 (Genewein *et al.*, 1993)) was available and was included in the genomic analyses.

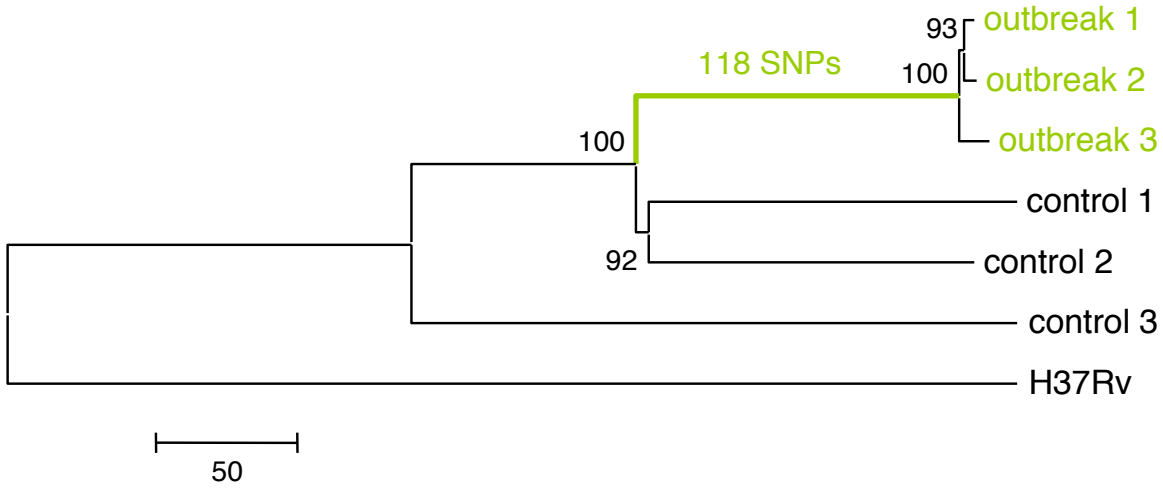


Figure 6.2.: **Initial neighbor joining phylogeny of *Mycobacterium tuberculosis* isolates.** Three whole-genome sequences from the historic outbreak and four control isolates were used to identify single-nucleotide polymorphisms (SNPs) specific to the outbreak genotype. Node support was assessed by bootstrapping over 1000 pseudo-replicates and is indicated as a percentage.

a mutation at this position to a second, phylogenetically redundant SNP-typing assay (981,565 CT). Both SNPs were selected to be synonymous and located in genomic regions suitable for primer and probe design.

6.3.3. WGS and phylogenetic analyses

All isolates identified by the screening assay and the additional serial isolate from patient P006 were subjected to Illumina WGS at GATC Biotech (Konstanz, Germany), with a median nucleotide coverage of 157.3 reads (range, 29.1–8 96.9 reads). Sequence read mapping and SNP calling was done as previously described (Coscolla *et al.*, 2013). We considered SNPs with a coverage of at least 10 sequencing reads and a value of 20 in the Phred-scaled quality score. SNPs in genes annotated as "PE/PPE/PGRS", "maturase", "phage", "insertion sequence", or "13E12 repeat family protein" were removed. Additionally, positions with missing nucleotide calls in at least three isolates were excluded. We used a second short-read alignment tool (SMALT, Wellcome Trust Sanger Institute, United Kingdom) to obtain SNP calls. Only positions called by both methods after filtering for the criteria mentioned above were included for further analysis. A subset of 28 SNPs was confirmed by Sanger sequencing (Supplementary Materials in Appendix A.2).

A genomic network with all variable positions was generated using Fluxus Network Software (available at: `www.fluxus-engineering.com`) and the Median Joining algorithm. Arlequin 3.5.1.12 (Excoffier *et al.*, 2005) was used to calculate genetic distances

and fixation indices (F_{ST}) to estimate population separation between genomic subclusters. Statistical significance was calculated with permutations.

Raw sequencing data are available under accession number PRJEB5925 (European Nucleotide Archive).

6.3.4. Clinical and socio-demographic data collection

We collected clinical and sociodemographic data for all patients identified as belonging to the cluster. Treating physicians, hospital archives, the Bernese Lung Association, and the Cantonal health authorities collected the data, using standardized questionnaires. We also collected contact tracing information for confirmed or presumptive links among cluster patients. The National Tuberculosis Surveillance Registry (Federal Office of Public Health) provided basic demographic data (age, sex, birth place, and disease site) for all tuberculosis cases notified in the Canton of Bern between 1991 and 2011.

6.3.5. Definitions

We categorized patients as having new tuberculosis, recurrent tuberculosis, or an unknown previous treatment status, according to international definitions (Rieder *et al.*, 1996). Links between cases were considered confirmed for contacts named in the contact tracing information. Links between cases were considered presumptive when contacts were not clearly named but were strongly supported by other contact tracing information (i.e. visiting common hotspots of transmission, shared housing, and shared place of work). Alcohol abuse was defined as daily consumption of alcohol, and smoking was defined as past or current smoking. We defined "milieu" as a combined variable capturing high-risk populations (substance abusers and/or homeless individuals) and/or patients frequenting high-risk settings (i.e. drug injection places, methadone distribution places, and homeless shelters).

6.3.6. Statistical analyses

We used χ^2 tests or Fisher exact tests to assess differences between groups in binary variables and the Wilcoxon rank sum test for analysis of continuous variables. We investigated differences between (1) the characteristics of patients in the Bernese cluster and all other notified tuberculosis cases in the Canton of Bern between 1991 and 2011 and (2) patients in the genomic subclusters.

6.3.7. Ethics statement

Ethics approval for this study was obtained from the ethics committee of the Canton of Bern. The treating physicians sought written informed consent from study participants. In most cases, however, informed consent could not be obtained because the patient could not be located or was known to have died. We therefore obtained permission from the Federal Expert Commission on Confidentiality in Medical Research to use the data provided by the treating physicians.

6.3.8. Role of the funding sources

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the article.

6.4. Results

6.4.1. Identification of Bernese cluster isolates by strain-specific SNP typing

Using the strain-specific real-time PCR SNP-typing assay, we screened 1,642 *M. tuberculosis* single patient isolates for the Bernese cluster–specific SNP (878,174 GA) and identified 71 of 1,642 (4.3%) isolates as belonging to this cluster. All isolates but three were confirmed by use of a second, phylogenetically redundant SNP (981,565 CT). These three isolates with ambiguous results were excluded from further analysis, as subsequent WGS revealed mean pair-wise distances of 91, 148, and 174 SNPs to the other cluster isolates. In contrast, all other cluster isolates were separated by ≤19 SNPs (Supplementary Table 1 in Appendix A.2). This corresponds to a specificity of 99.8% (three false-positive results out of 1,574 noncluster isolates), when considering our screening results based on only the first strain-specific SNP. All 22 historical isolates described in 1993 (Genewein *et al.*, 1993) were correctly identified by both SNP-typing assays (sensitivity, 100%) as having Bernese cluster genotype. Hence, we identified a total of 68 patients linked to the cluster strain (Figures 6.1 and 6.3).

For one patient, two isolates (P006A and P006B, isolated in 1988 and 1991, respectively) were available, and we included both for further analyses because of the central role of this patient described in the original study (originally termed "patient 2" (Genewein *et al.*, 1993)). Hence, we included 69 cluster isolates in the WGS analyses (Figure 6.1).

To illustrate the strengths of our novel combined approach, we compared the costs, advantages, and disadvantages of the different methods used in outbreak investigations of tuberculosis (Table 6.1). Our strain-specific SNP assay to screen 1,642 isolates was approximately 15 times less expensive than the current gold standard based on MIRU-VNTR. A combination of the SNP assay and targeted WGS for 69 cluster isolates was approximately 20 times less expensive than performing WGS of the entire collection.

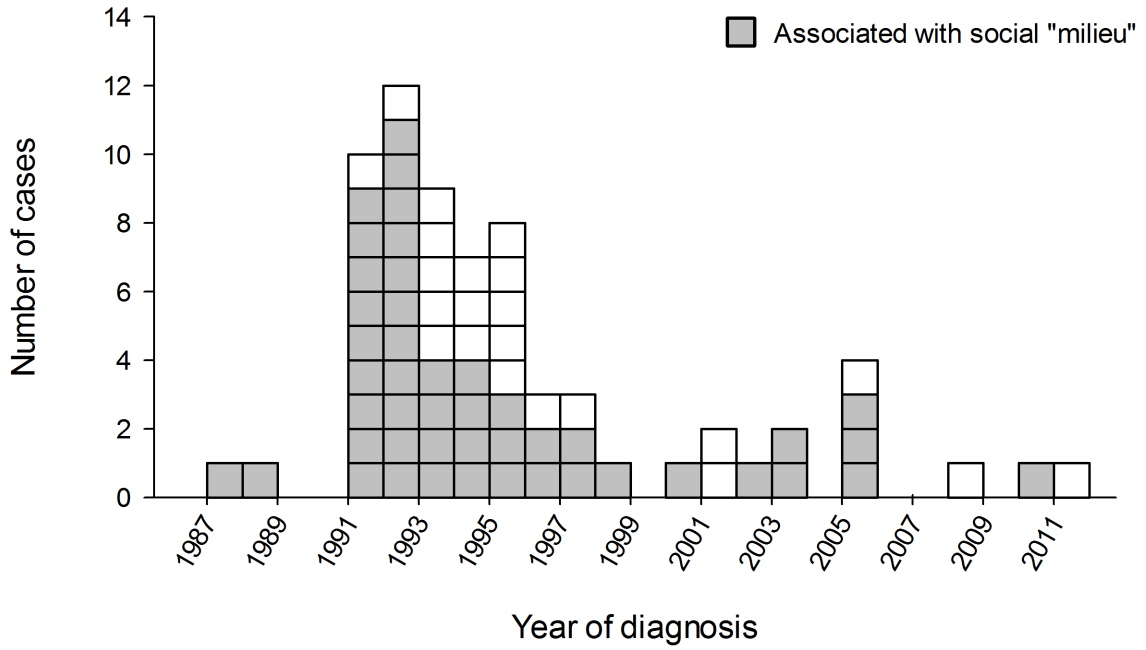


Figure 6.3.: **Epidemic curve of the 68 patients identified as tuberculosis cluster cases.** Gray boxes indicate patients associated with the social milieu (homeless individuals and/or substance abusers). One additional patient isolate ((P028, isolated in 1987) reported in the original publication (Genewein *et al.*, 1993) and pre-dating the systematic collection of isolates in 1991 was included in the study. For one patient from 1991 (Genewein *et al.*, 1993), a second isolate (P006A, isolated in 1988) was available and was therefore backdated.

Table 6.1.: **Comparison of tuberculosis outbreak tracking methods, considering the scenario of the present study (68 outbreak patients of 1642 patient isolates to be screened)**

	This study					
	SNP-assay for identification of cluster isolates (n=1,642)	Targeted WGS of identified cluster isolates (n=69)	WGS of entire collection (n=1,642)	MIRU-VNTR (n=1,642)	Spoligotyping (n=1,642)	Contact tracing
Estimated costs (US$)[1]						
per isolate	3 (SNP assay)	330 (targeted WGS)	330	49	26	High
total	4,926[2] (SNP assay)	22,770 (targeted WGS)	541,860	80,458	42,692	
	Total: 27,696					
Advantages	Rapid identification of cluster isolates; inexpensive	Highest resolution among cluster isolates; additional information (e.g. drug-resistance mutations)	Highest resolution among all isolates (in all clusters); additional information obtained (e.g. drug-resistance mutations)	Current gold-standard for molecular epidemiology; can be semi-automatized	Low technology requirement	Information on transmission hotspots for targeted prevention; information on secondary cases
Disadvantages	No further resolution among cluster isolates; performance of assay depends on selection of initial isolates for WGS (e.g. SNP-selection)	Previous identification of outbreak isolates necessary	Expensive; bioinformatic expertise needed	Limited resolution within outbreak clusters	Low resolution of outbreak cluster analysis when used as single method	Expensive and time-consuming; misses many cases, particularly in high-risk populations
Prospective use	Can be used in real-time once outbreak is identified and an assay established. If used in combination with targeted WGS: highest resolution among outbreak isolates	In combination with SNP-assay	Can be used in real-time once bioinformatics expertise is established	Routine use	Yes	Yes

MIRU-VNTR, mycobacterial interspersed repetitive unit–variable number of tandem repeat; SNP, single-nucleotide polymorphism; WGS, whole genome sequencing.

[1] Cost calculations were based on commercially available services (available at: `www.genoscreen.fr` and `www.gatc-biotech.com`) or estimated according to in-house costs (as of August 2014).

[2] For WGS of six initial isolates and screening using the strain-specific SNP-genotyping assay of 1,642 isolates.

6.4.2. Description of the tuberculosis cluster over 21 years

Of the 68 patients in the Bernese cluster, 55 (80.9%) received their diagnosis during 1998 or earlier, and 13 received their diagnosis between 1999 and 2011 (Figure 6.3). The characteristics of the 68 cluster patients, compared with characteristics of all other tuberculosis cases diagnosed in the same region and period, are presented in Table 6.2. Cluster patients were more likely to be male (79.4% vs 57.3%; $P < 0.001$), born in Switzerland (83.8% vs 47.4%; $P < 0.001$), and to have pulmonary tuberculosis as opposed to extrapulmonary tuberculosis (94.1% vs 75.1%; $P < 0.001$). The median age of the cluster patients was 41 years (interquartile range [IQR], 34–53 years), compared with 44 years (IQR, 29–71 years) for all other patients with tuberculosis ($P = 0.12$). Most (67.6%) of the cluster patients were part of the local injection drug scene and/or homeless milieu. Among cluster patients, 19.1% were infected with human immunodeficiency virus (HIV; HIV information was unavailable for noncluster patients). Four particular hotspots of tuberculosis transmission were identified within the milieu: a short-term homeless shelter, a long-term social integration home, a meeting point for injection drug and methadone supply, and a bar where substance abusers met. The distribution of cluster patients among these four transmission hotspots, the social milieu, and the general population are presented in Figure 6.4.

Contact investigation provided information on potential patient-to-patient links (Figure 6.4). Fourteen of 68 patients (20.6%) had confirmed epidemiological links (Figure 6.4). Confirmed links were more frequent between patients sharing transmission hotspots, indicating the large degree of social interaction in these settings. Only one confirmed link, between a father and his daughter (P040-P041), was identified in the general population.

6.4.3. WGS of Bernese cluster isolates

A total of 133 variable positions were identified among the 69 cluster isolates (Supplementary Table 2 in Appendix A.2). We generated a Median Joining network using these 133 variable positions (Figure 6.5). Despite identical MIRU-VNTR and spoligotyping patterns (Supplementary Table 3 in Appendix A.2), 52 of 69 isolates (75.4%) were discriminated by at least one SNP from their most closely related neighbor. The maximum number of SNPs between any two isolates was 19 (Supplementary Table 1 in Appendix A.2), and the mean pair-wise distance ($\pm$SD) between all isolates was 6.0 ± 2.9 SNPs. Patient isolates with confirmed epidemiological links differed by 0–11 SNPs (Supplementary Table 1 in A.2). No drug resistance–associated mutation was detected among the 69 cluster isolates (Supplementary Materials in Appendix A.2).

Table 6.2.: **Characteristics of cases confirmed to be in the tuberculosis cluster, compared with all other notified tuberculosis cases in the Canton of Bern between 1991 and 2011**

Characteristic	Bernese cluster cases	All other cases	P-value
Total	68 (100)	1,872 (100)	
Age at TB diagnosis (median, IQR)	41 (34-53)	44 (29-71)	0.12
Sex			
Male	54 (79.4)	1072 (57.3)	
Female	14 (20.6)	800 (42.7)	
Birth region			<0.001
Switzerland	57 (83.8)	888 (47.4)	
Europe (without Switzerland)	10 (14.7)	398 (21.3)	
Sub-Saharan Africa	0	227 (12.1)	
Asia	1 (1.5)	284 (15.2)	
Caribbean and Latin America	0	31 (1.7)	
Other regions	0	36 (1.9)	
Unknown	0	8 (0.4)	
Tuberculosis site			<0.001
Pulmonary	64 (94.1)	1,406 (75.1)	
Extra-pulmonary	4 (5.9)	466 (24.9)	
Tuberculosis category			
New case	54 (79.4)		
Recurrent	7 (10.3)		
Unknown	7 (10.3)		
Imprisonment within 2 years of diagnosis	9 (13.2)		
Diabetes	3 (4.4)		
Alcohol abuse	39 (57.4)		
Smoker	41 (60.3)		
Injection drug user	18 (26.5)		
Homeless	21 (30.9)		
HIV positive	13 (19.1)		
Homeless/substance abuser milieu	46 (67.6)		
Residence			
Bern City	37 (54.4)		
Outside Bern City	29 (42.6)		
Unknown	2 (2.9)		

TB, tuberculosis

IQR, interquartile range

Data are no. (%) patients or median (interquartile range).

Data for some characteristics were missing for the other cases notified to the Canton of Bern.

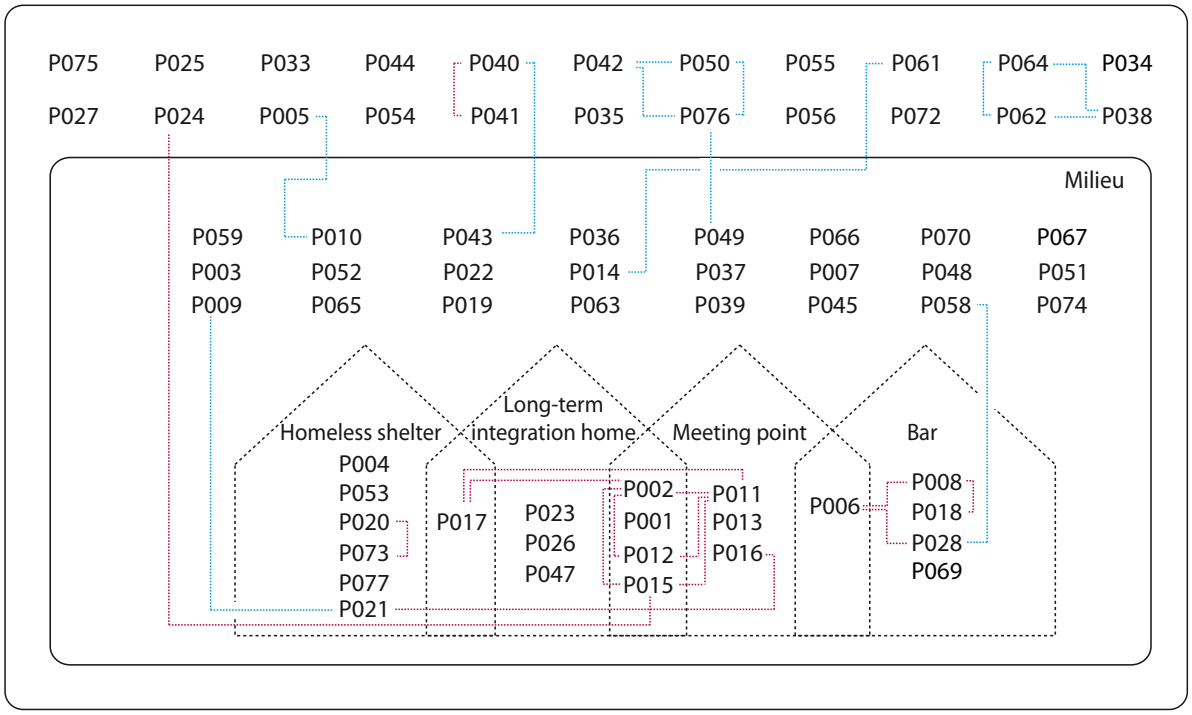


Figure 6.4.: **Distribution of tuberculosis cluster patients in the milieu of substance abusers and homeless people.** The four main hotspots of transmission that were identified by social contact tracing are shown (a short-term homeless shelter, a long-term social integration home, a meeting point for substance abusers and a bar). Milieu patients are associated with a particular social milieu (homeless, substance abuser scene). Red lines indicate confirmed epidemiological links, blue lines indicate suspected social links. Presumptive individual links between milieu patients are not shown because these patients are highly interlinked.

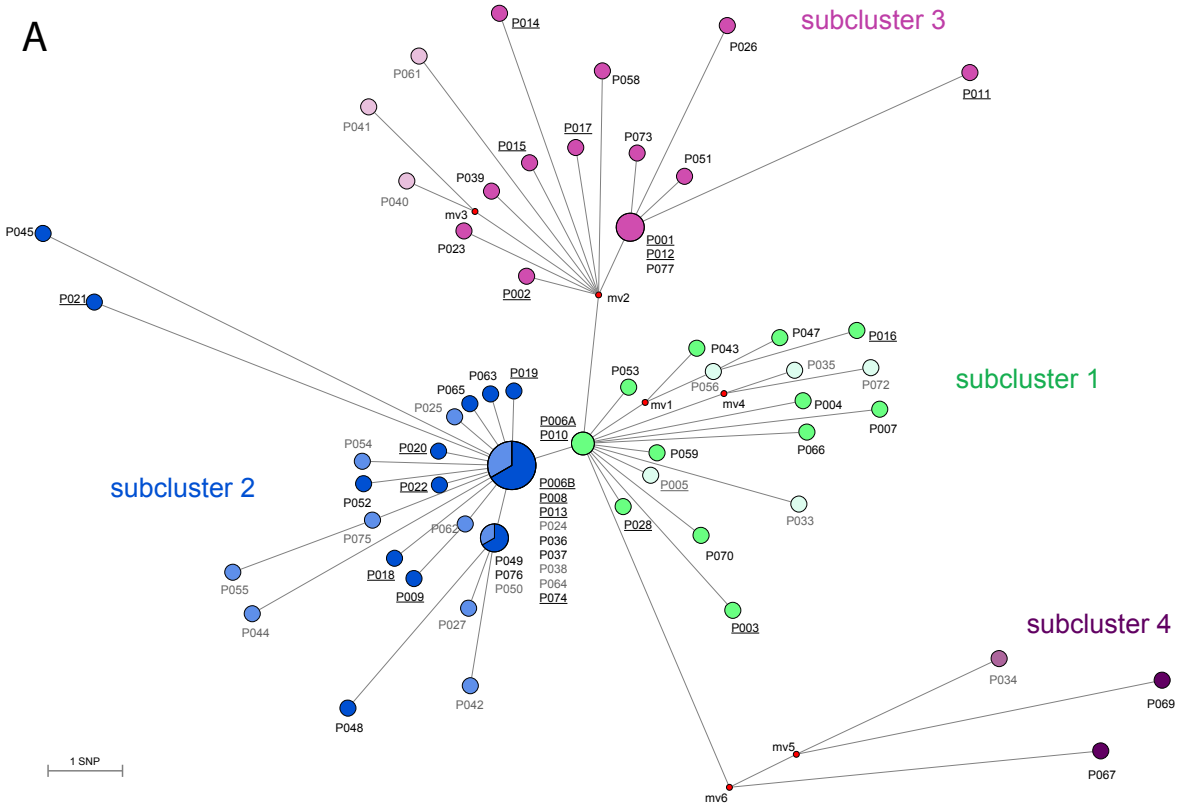


Figure 6.5.: **Median Joining network using 133 variable single nucleotide positions (SNP) among whole genome sequences of *Mycobacterium tuberculosis* cluster isolates of the "Bernese cluster" outbreak**. Branch lengths correspond to number of SNPs. Circle sizes correspond to number of isolates, median vectors (mv) are hypothetical genotypes. Position of "mv6" is the root of the network. **A**. Network showing the four identified subclusters. Darker circles in each subcluster indicate patients that were associated with the particular social milieu (homeless, substance abuser scene), lighter circles are patients in the non-milieu population. Underlined labels represent isolates that were identified in the original publication (Genewein *et al.*, 1993). **B**. Network colored according to time period when the *M. tuberculosis* strains were isolated. **C**. Network showing patients associated with a particular hotspot.

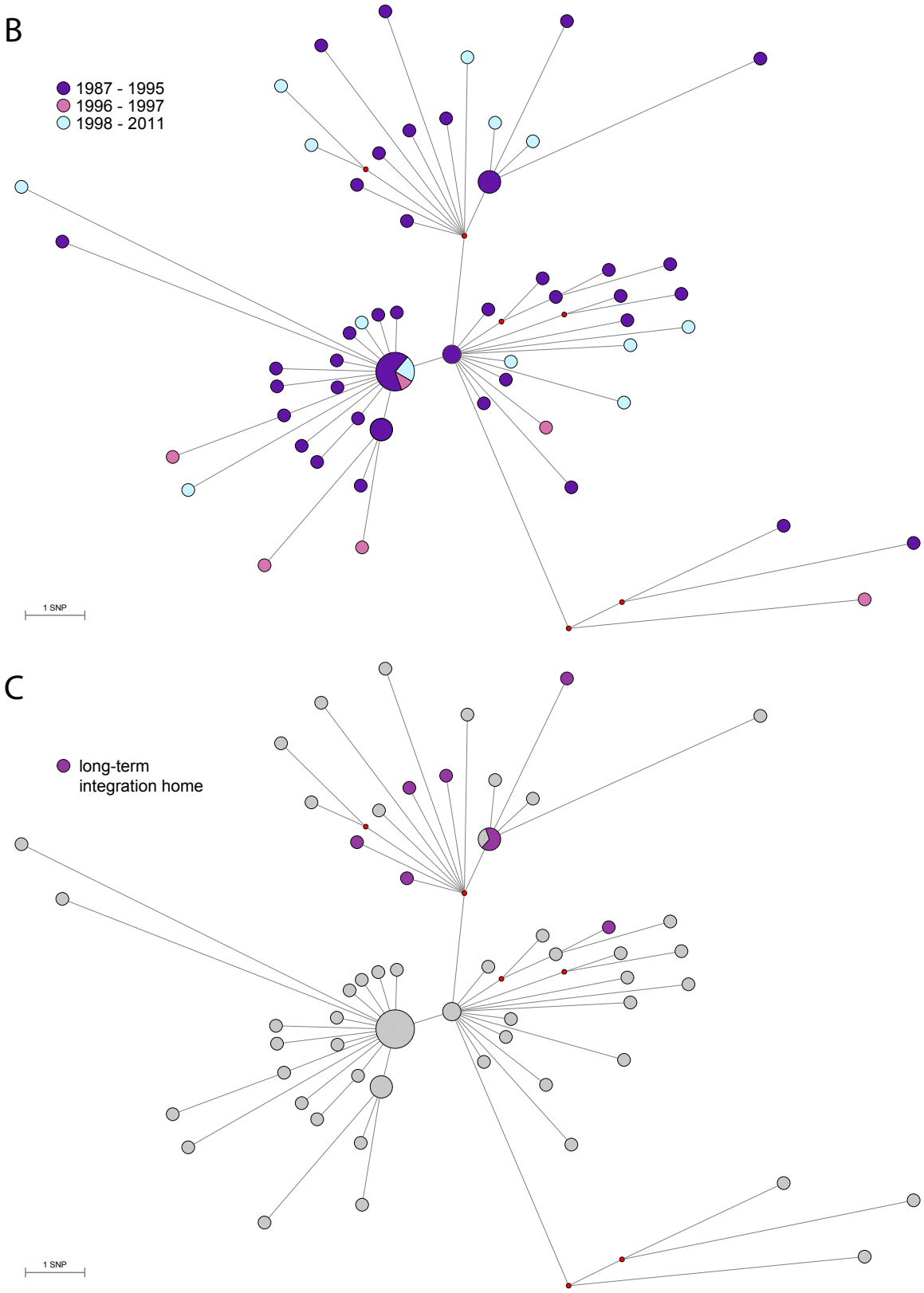


Figure 6.5 continued.

6.4.4. Subclusters and key patient

The genomic network revealed three independent star-like structures, suggesting (1) an early divergence of a shared ancestor strain into three subclusters (Figure 6.5) and (2) the presence of one or several superspreaders. A fourth, more distantly related subcluster was separated by nine SNPs from the nearest isolate (P006A; Figure 6.5). In contrast, subclusters 1 and 2 were separated by one SNP, and subclusters 1 and 3 were separated by two SNPs. The average pairwise distance within each subcluster was 4.2 (subcluster 1), 3.1 (subcluster 2), 5 (subcluster 3), and 9.3 SNPs (subcluster 4). All corresponding SNP distances were larger between subclusters than within subclusters (4.8 SNPs between subclusters 1 and 2, 6.8 SNPs between subclusters 1 and 3, 12.2 SNPs between subclusters 1 and 4, 7.3 SNPs between subclusters 2 and 3, 12.6 SNPs between subclusters 2 and 4, and 14.6 SNPs between subclusters 3 and 4). Pairwise fixation indices (F_{ST}) between subclusters were between 0.24 (subcluster 1 and 2) and 0.67 (subcluster 2 and 4; $P < 0.005$ for all comparisons), further supporting the subcluster distinction.

The central position of subcluster 1 was occupied by the first isolate from the key patient, patient P006 (P006A; isolated in 1988) together with isolate P010. The second isolate from the key patient (P006B; isolated 1991) was found in the central position of subcluster 2. This suggests a second tuberculosis episode in patient P006, generating further secondary cases. Seven other isolates were genomically clustered with P006B and could also be the source of transmission in subcluster 2. However, patient P006 was a homeless substance abuser who was known to have interrupted treatment and had a history of treatment failure, and a key role for this patient in the outbreak was already suspected in 1993.

The central position of subcluster 3 remained unoccupied (i.e. no isolate was available with this hypothetical genotype). This could be explained by an unsampled strain variant from patient P002 (i.e. from a mixed infection of microevolved strains) that had been transmitted to other subcluster 3 patients. Such a variant would have been missed as a result of the single colony isolation step preceding WGS. Manual inspection of the corresponding Sanger sequence trace files generated from a separate DNA preparation without the single colony isolation step revealed a double peak at the SNP separating P002 from the hypothetical node mv2 (genomic position 2,156,041) in the bulk isolate of P002, indicating the presence of a mixed population in P002. Alternatively, an unsampled patient isolate (reported as “A”, “B”, and “C” by Genewein *et al.* (1993)) might correspond to the central position of subcluster 3.

When plotting the period of isolation of *M. tuberculosis* strains in the genomic network (Figure 6.5B), we found no evidence that the different subclusters were associated with

specific periods.

Finally, we compared the patient characteristics between the subclusters, excluding subcluster 4 (genetically distant, epidemiologically unrelated) and patient P006, whose isolates belonged to both subclusters 1 and 2. We found that HIV infection was more frequent in subcluster 1 (7 of 17 [41.2%]) than in the other subclusters (3 of 30 [9.7%] in subcluster 2 and 3 of 17 [17.6%] in subcluster 3; $P = 0.04$). All three subclusters included a majority of individuals from the social milieu (12 of 17 [70.6%] in subcluster 1, 17 of 30 [56.7%] in subcluster 2, and 14 of 17 [82.4%] in subcluster 3; $P = 0.19$). Subcluster 3 was associated with two particular transmission hotspots: a long-term social integration home (1 of 17 [5.9%] in subcluster 1, 0 of 30 in subcluster 2, and 7 of 17 [41.2%] in subcluster 3; $P < 0.001$; Figure 6.5C), and a meeting point for methadone supply (1 of 17 [5.9%] in subcluster 1, 1 of 30 [3.3%] in subcluster 2, and 5 of 17 [29.4%] in subcluster 3; $P = 0.02$).

6.5. Discussion

Using a novel combination of a rapid and inexpensive strain-specific SNP screening assay and targeted WGS, we tracked a tuberculosis cluster spanning 21 years and revealed the transmission dynamics among the outbreak patients.

Our study demonstrated the feasibility and advantages of tracking a tuberculosis outbreak by using a strain-specific SNP-based screening assay in a large population-based collection of *M. tuberculosis* isolates. We subsequently performed targeted WGS on the 69 identified cluster isolates, which, combined with social contact data, enabled us to retrace transmission dynamics at high resolution. The combined cost of six initial whole-genome sequences used to design the SNP-typing assay and the subsequent population-wide screening by real-time PCR was low (approximately US$ 4900), compared with the cost of screening all isolates with any other genotyping method. Additionally, the time required for screening was substantially reduced, making this a powerful approach for identifying and tracking tuberculosis outbreaks in real time. Even though our study was retrospective, our approach could be used to screen isolates prospectively.

Our results indicate that the sensitivity and specificity of this approach are nearly 100%. However, we could only estimate the technical test characteristic, because WGS data were not available for the entire collection. Importantly, the performance of a strain-specific SNP-typing assay depends on the selection of SNPs, and the selection of SNPs depends on the isolates initially sequenced. For the successful design of such an assay, we recommend the following: (1) select at least two known clustered isolates for WGS, (2) include at least two control isolates with genotyping patterns closely related to but different from

those of the clustered isolates (e.g. MIRU-VNTR), (3) identify SNPs specific to all clustered isolates and absent in the control isolates, (4) exclude SNPs in genes known to be associated with drug resistance, (5) select SNPs suitable for probe and primer design, and (6) use at least two SNPs for the screening of isolates, as the specificity of each SNP might vary.

Using our combined approach and linking it to clinical and contact tracing data, we found that the tuberculosis cluster continued to propagate in Bern, mainly in particular transmission hotspots and in the originally described high-risk populations of substance abusers and homeless people (Genewein *et al.*, 1993). This is consistent with previous reports from other low-incidence settings (Anderson *et al.*, 2014; Bamrah *et al.*, 2013; Mitruka *et al.*, 2011; Zenner *et al.*, 2013). The outbreak involved a key patient who caused numerous secondary cases, which corresponds to superspreader behavior. Most outbreak cases occurred between 1991 and 1995, followed by 16 years of sporadic cases, the majority of which were likely cases of reactivation tuberculosis. The cases in the early 1990s coincided with known peaks of heroin abuse in Switzerland. However, 32.4% of all cluster cases involved the nonmilieu population, possibly reflecting transmission from the milieu to the wider community.

In retrospect, more secondary cases could have been identified if our novel screening method had been available in the 1990s. Indeed, strain-specific SNP typing would have provided an inexpensive method to identify outbreak cases more rapidly. Furthermore, targeted WGS would have identified superspreaders in a context where contact tracing is particularly difficult. Such superspreader behavior could have then been targeted with intensified control measures to interrupt transmission.

The targeted WGS analyses of all cluster isolates identified by strain-specific SNP typing shed new light on the transmission dynamics of the outbreak, compared with traditional genotyping methods. Whereas MIRU-VNTR and spoligotyping showed identical genotyping patterns, WGS revealed distinct genotypes for 76.5% of the Bernese cluster isolates. In particular, we identified four genomic subclusters not revealed by classical genotyping, likely reflecting concomitant but independent clusters of a common ancestral strain. However, the genetic distances between subclusters were small (one, two, and three SNPs between subclusters 1, 2, and 3), and therefore the definition of "subcluster" may be debatable. The subclusters were, however, supported by F_{ST} values indicating separation of these populations.

Two sequential isolates from the key patient, isolated in 1988 and 1991, occupied the

central positions of subclusters 1 and 3, respectively. This suggests that two disease episodes of this patient led to two independent star-like patterns in the genomic network, indicating superspreader behavior (Gardy *et al.*, 2011; Roetzer *et al.*, 2013; Walker *et al.*, 2013b). The central role of this patient was already suspected in the original description of the outbreak (Genewein *et al.*, 1993). Hence, WGS analyses indicated that this patient likely caused more secondary cases than previously assumed.

Despite the many advantages of WGS, our results also showed that interpreting WGS data has limitations. For example, nearly 25% of cluster isolates were genomically indistinguishable from at least one other isolate. This emphasizes the need to include repetitive regions of the genome that are currently excluded because of technical limitations (Bryant *et al.*, 2013a; Copin *et al.*, 2014). Furthermore, there is increasing evidence that bacterial populations within patients are heterogeneous as a consequence of ongoing microevolution, further complicating the interpretation of transmission events (Pérez-Lago *et al.*, 2014). In our study, genomes were generated from single colonies for most isolates. Yet, considering potential clonal variants that were randomly excluded from the sequencing process could influence the way transmission events are inferred. With improving sequencing technologies, future studies should sequence bulk isolates rather than single colonies and consider within-host heterogeneity in bacterial populations. Mutations can also arise during laboratory culture; these could be avoided by performing WGS directly from sputum (Blainey, 2013).

In conclusion, our strain-specific SNP-based screening approach offers a rapid and inexpensive way of tracking tuberculosis outbreaks retrospectively and prospectively. This novel screening method, combined with targeted WGS, can be used to guide control interventions by rapid and inexpensive screening, revealing transmission hotspots and missing links in transmission chains. Future studies could use this approach in real time to track ongoing outbreaks of tuberculosis and other infectious diseases in hospital settings, as well as population-wide.

6.6. Acknowledgements

We thank all of the patients with tuberculosis who participated in this study; the treating physicians and hospitals, as well as the Swiss HIV Cohort Study, for providing clinical information; the Institute for Infectious Diseases, University of Bern, Switzerland, for providing the clinical isolates; and the National Tuberculosis Surveillance Registry at the Federal Office of Public Health, the Bernese Lung Association, and the Cantonal health authorities, for supporting the collection of clinical data and contact-tracing information.

6.7. Financial support

This work was supported by the Bernese Lung Association, the Swiss National Science Foundation (grants PP00P3_150750 and 33CS30_134277, to the Swiss HIV Cohort Study), the US National Institutes of Health (grants AI090928 and U01AI069924), and the European Research Council (grant 309540-EVODRTB).

6.8. Potential conflicts of interest

All authors: No reported conflicts. All authors have submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Conflicts that the editors consider relevant to the content of the manuscript have been disclosed.

7. The global spread of the Euro-American lineage of *Mycobacterium tuberculosis*

David Stucki [1,2], Liliana Rutaihwa [1,2], Marie Ballif [3], Tao Luo [4], Rita Macedo [5], Helmi Mardassi [6], Sonia Borrell [1,2], Griselda Tudo Vilanova [7], Janet Fyfe [8], Maria Globan [8], Jackson Thomas [9], Francis Jamieson [10], Jennifer Guthrie [10], Dorothy Yeboah-Manu [11], Moses Joloba [12], Eddie Wampande [12], James Bower [13], Qian Gao [4], Midori Kato-Maeda [14], Iñaki Comas [15], Mireia Coscollà [1,2], Lukas Fenner [1,2], Sebastien Gagneux [1,2]

[1] Swiss Tropical and Public Health Institute, Basel, Switzerland

[2] University of Basel, Switzerland

[3] Institute for Social and Preventive Medicine, Bern, Switzerland

[4] Institutes of Biomedical Sciences and Medical Microbiology, Fudan University, Shanghai, China

[5] Laboratòrio de Saùde Publica, Lisbon, Portugal

[6] Institut Pasteur de Tunis, Tunis-Belvédère, Tunisia

[7] Servei de Microbiologia, Hospital Clinic-CRESIB-FCRB, Barcelona, Spain

[8] Victorian Infectious Diseases Reference Laboratory, Victoria, Australia

[9] Ifakara Health Institute, Bagamoyo, Tanzania

[10] Public Health Ontario, Toronto, Canada

[11] Noguchi Memorial Institute for Medical Research, University of Ghana, Accra, Ghana

[12] Department of Medical Microbiology, Makerere University, Kampala, Uganda

[13] Department of Clinical Microbiology, Auckland City Hospital, Auckland, New Zealand

[14] School of Medicine, University of California, San Francisco, USA

[15] Genomics and Health Unit, Centre for Public Health Research, Valencia, Spain

7.1. Abstract

Tuberculosis has been co-evolving with humans for 70,000 years. Today, seven main phylogenetic lineages of the *Mycobacterium tuberculosis* complex (MTBC) are found among human isolates, each associated with particular human populations. An "out-of-Africa" scenario was proposed to explain the present-day phylogeographic distribution. The "Euro-American" lineage, also known as "Lineage 4", has likely emerged as one of three "modern" lineages of MTBC around 46,000 years ago. Despite the hypothesized expansion in Europe, Lineage 4 has been observed on all continents, and is also the main cause of TB in the Americas. The reasons for this apparent "success" of Lineage 4 are unclear. Lineage 4 comprises various "subtypes", but whether these are homogeneously distributed in the world is unknown. We hypothesized that European explorations (starting with the Age of Discovery in the 15th century) and European colonialism in Africa, Asia and the Americas at least partially account for the prevalence of the "European" genotype on all continents. We aimed at inferring the evolutionary history of Lineage 4.

We first used 72 whole genome sequences of Lineage 4 isolates to define 10 sublineages and to extract sublineage-specific phylogenetic markers. We developed a new single nucleotide polymorphism (SNP)-genotyping assay based on the Luminex platform, which we used to screen 3,366 isolates for sublineage classification. Plotting isolate sublineage against patient place of birth revealed a phylogeographically structured population of Lineage 4 sublineages. Five of the 10 sublineages were not observed among 228 isolates from European patients, and were locally restricted to particular geographical areas (L4.2, L4.4, L4.7, L4.8, L4.9). These included three sublineages only seen in Africa (L4.2, L4.8, L4.9). Three sublineages were found frequently in Europe and, at the same time, particularly widespread globally (L4.3, L4.5, L4.10). Each of these three global sublineages was observed in more than 40 of the 100 countries represented in this study.

Next, we focused on the most frequent of the globally distributed sublineages, L4.5 (also known as the "Latin-American Mediterranean" (LAM) family), using 150 whole genome sequences from MTBC isolates from all continents. Analysis of 12,256 SNPs resulted in a phylogeny with five genetically distinct clades among the LAM sublineage. Maximum Parsimony and Bayesian reconstruction of ancestral characters supported a hypothetical ancestor of LAM isolates in Europe with 100% and 66% probability, respectively. We also found that LAM isolates from TB patients in Europe harboured significantly more genetic diversity than isolates from Africa and the Americas, and that the average nucleotide diversity of these isolates decreased with the geographic distance from Lisbon.

Our data show that Lineage 4 is a globally prevalent lineage with a remarkably structured phylogeography. The fact that five MTB Lineage 4 sublineages do not appear to

occur in Europe points at an early dispersal of Lineage 4 genotypes, preceding the introduction of Lineage 4 into Europe. The three sublineages most frequently seen in Europe and also occurring on all other continents support a (perhaps more recent) dispersal from Europe. In particular, our data are consistent with the notion that the global distribution of the L4.5/LAM sublineage reflects global waves of European exploration and colonization. However, the relatively weak negative correlation of genetic diversity with geographic distance, as well as the lack of phylogenetic clustering by geography, indicates that LAM genotypes have also been circulating globally, perhaps as a consequence of increasing globalization.

7.2. Introduction

Tuberculosis (TB) is estimated to have caused more than one billion human deaths in the last two centuries. Today, TB still causes 8.6 million new cases and 1.3 million deaths a year (Selgelid, 2008; WHO, 2013). Evolutionary studies of the causative agent, *Mycobacterium tuberculosis* (MTB), suggest that MTB has co-evolved with humans for thousands of years and spread globally with human migrations (Hershberg *et al.*, 2008; Wirth *et al.*, 2008; Comas *et al.*, 2013). Seven main human-associated phylogenetic lineages of MTB are found today; these are associated with particular geographic regions and human populations (Gagneux *et al.*, 2006b; Jong *et al.*, 2010; Pareek *et al.*, 2012; Fenner *et al.*, 2013; Firdessa *et al.*, 2013; Comas *et al.*, 2013). Over recent years, there has been increasing evidence that strain variation in MTB influences the outcome of TB (Coscolla *et al.*, 2010; Click *et al.*, 2012; Caws *et al.*, 2008). MTB Lineage 4, also called the "Euro-American" lineage, is one of three phylogenetically "modern" lineages of the *Mycobacterium tuberculosis* complex (MTBC), and has likely evolved from a common ancestor of Lineages 2, 3 and 4 around 46,000 years ago in the Eastern Mediterranean region, spreading to Europe subsequently with human migrations (Wirth *et al.*, 2008; Comas *et al.*, 2013). Today, MTB Lineage 4 is the predominant cause of TB in Europe as well as in North- and South-America (Gagneux *et al.*, 2006b). But Lineage 4 has repeatedly also been described in high proportions in Africa (Ani *et al.*, 2010; Wampande *et al.*, 2013), and with varying prevalence in Asia (Panel A in Figure 7.1) (Nahid *et al.*, 2010; Li *et al.*, 2011). Reasons for the global distribution and broad "success" of MTB Lineage 4 are unknown. The global prevalence of Lineage 4 was proposed to have its origin in the explorations and migrations of Europeans to the world in the past centuries (Hershberg *et al.*, 2008). The huge historic burden of TB in Western Europe (a mortality of up to 1% between the 18th and 19th century (Murray, 2001)), the exponential population growth in the last centuries (Panel B in Figure 7.1) and the subsequent migrations from Europe to other continents may have shaped the global population structure of Lineage 4. However to date, no comprehensive study adressing the evolutionary scenario of MTB Lineage 4 using genomic data has been performed.

Various *sub*groups or *sub*lineages of MTB Lineage 4 (strain/genotype families) have been described. Among them are the well-known families known as "H" (Haarlem), "X", "T" and "LAM" that were originally defined based on *IS*6110-RFLP and spoligotyping (Brudey *et al.*, 2006; Comas *et al.*, 2009b). Using early genomic approaches, sublineage-specific large sequence polymorphisms were identified (Gagneux *et al.*, 2006b), and are now used as sublineage-specific markers (Anderson *et al.*, 2013; Rindi *et al.*, 2012). Later,

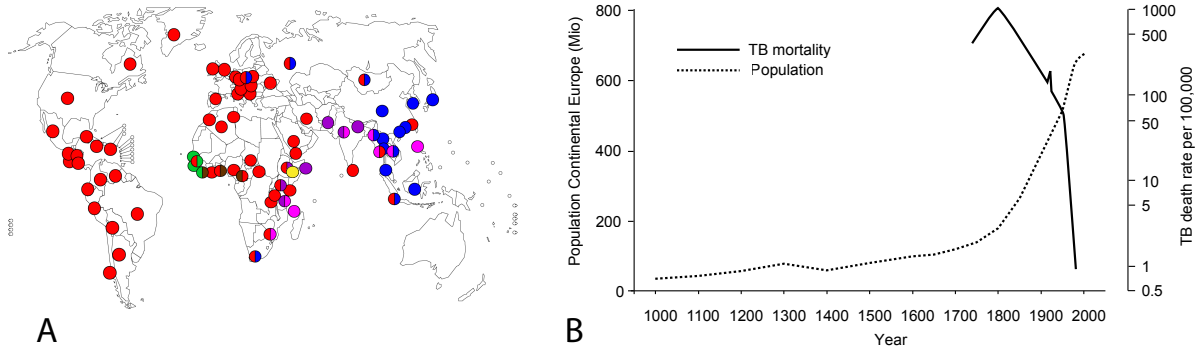


Figure 7.1.: **Today's global phylogeography of the MTBC, and the TB mortality in Europe during the last thousand years. A**. Strains of the "Euro-American" lineage, also called MTB Lineage 4, were reported to be the predominant cause of TB in Europe, North and South America, but are also highly prevalent in Africa, and partly in Asia. Dots indicate predominant phylogenetic lineages per country; colors correspond to lineage colors in 1.4. Modified from Gagneux *et al.* (2006b). **B**. The European population has exponentially grown in the last 1000 years. In parallel, the mortality due to TB in Western countries was up to 1000 times higher between the 17th and 20th century than today. European explorations and colonizations of other continents potentially spread particular "European" genotypes of the MTBC globally. Mortality data from Murray (2001), population data from Feuerwerker (1990). Population data point for 2009 from `www.worldpopulationstatistics.com`.

single nucleotide polymorphism (SNP)-schemes were developed for the classification of clinical isolates (Baker *et al.*, 2004; Filliol *et al.*, 2006; Abadia *et al.*, 2010; Bergval *et al.*, 2012; Homolka *et al.*, 2012). All classification schemes, however, were either based on suboptimal markers (Comas *et al.*, 2009b), or suffer from discovery bias (Achtman, 2012), and therefore provide an incomplete picture. Furthermore, the various molecular markers and genotyping schemes have lead to confusing strain definitions and nomenclatures. In the last years, large-scale whole genome sequencing (WGS) has become widely available and allows generating unambiguous and robust phylogenies of MTBC (Comas *et al.*, 2009b; Achtman, 2012). A robust phylogenetic framework for MTB Lineage 4, based on a large representative collection of isolates and WGS data, is now needed to be able to compare between studies.

One sublineage of Lineage 4, the "Latin American Mediterranean" (LAM) family, has been described as particularly widespread (Lazzarini *et al.*, 2007; Brudey *et al.*, 2006; Gibson *et al.*, 2008; Ignatova *et al.*, 2006; Casali *et al.*, 2012; Lanzas *et al.*, 2013). The genotype "Latin-American Mediterranean" was named in 2001 based on spoligotyping and VNTR data after the regions where it was first observed (Sola *et al.*, 2001). Thousands of LAM isolates have been reported since then, with particular subtypes of LAM dominating in different regions (Gibson *et al.*, 2008; Demay *et al.*, 2012). The evolution and global distribution of LAM strains has been the subject of ongoing discussions (Mokrousov *et*

al., 2014). Interestingly, strains of the LAM family were described to be the predominant genotypes in Southern Europe (Lopes *et al.*, 2013), but also in Sub-Saharan Africa (particularly in the coastal regions) (Pillay *et al.*, 2007; Viegas *et al.*, 2010), in South America (Gibson *et al.*, 2008; Lanzas *et al.*, 2013; Gomes *et al.*, 2012; Lopes *et al.*, 2013), and in some countries of South-East Asia (Demay *et al.*, 2012). We hypothesized that the LAM genotype, potentially representative for the whole MTB Lineage 4, was distributed from (Southern) Europe to other continents by early Spanish and Portuguese explorations starting in the 15th century, colonization of the Americas, and imperial colonialism in Africa and Asia (Lehning, 2013; Barr *et al.*, 2014). We asked the question if the phylogenetic structure seen today supports the origin and dissemination from a European origin, or, if more recent and increasing migrations and trade have lead to a "everything-is-everywhere" situation of the LAM genotype. Specifically, assuming an "export" of the LAM sublineage from Europe, we expected a phylogeny structured by geographic region and a higher genetic diversity of LAM strains in Europe compared to all other continents (due to bottlenecks and clonal expansion in the "receiving" regions) compared to Europe.

To test these hypotheses, we used population genomic methods to infer the evolutionary trajectory of MTB Lineage 4. First, we used a reference collection of 72 MTB Lineage 4 whole genome sequences to define evolutionary robust sublineages of MTB Lineage 4. We then used SNP-typing with sublineage-specific SNPs to classify 3,366 clinical isolates and to estimate the frequencies of Lineage 4 sublineages in different regions of the world. Finally, we applied targeted whole genome sequencing to 150 global LAM isolates, and used population diversity estimates to test the hypothesis of global dissemination of the LAM sublineage from Europe.

7.3. Methods

7.3.1. MTBC isolates and DNA extraction

All MTBC isolates were obtained from existing strain collections, where ethics approval was obtained in the original study.

For the definition of Lineage 4 sublineages, we aimed at covering maximal diversity of MTB genotypes, i.e. selected isolates based on spoligotyping data, large sequence polymorphisms and diverse geographical origins. For the total of 72 whole genome sequences, we included all MTB Lineage 4 genome sequences and representative strains of other main

phylogenetic lineages from Coscolla *et al.* (2013) and Comas *et al.* (2013), and additionally generated WGS of selected isolates from various ongoing studies. For the latter, purified DNA was obtained with the CTAB method (Embden *et al.*, 1993).

For the screening of clinical isolates for sublineage-classification, we used 2,763 MTB Lineage 4 DNA isolates from 100 countries (country of birth of the patient) and various ongoing studies. All isolates were previously identified as MTB Lineage 4 isolates by SNP-typing, RD-typing or spoligotyping. Purified DNA or crude, heat-inactivated extracts were obtained as previously described (Stucki *et al.*, 2012), or with comparable protocols. All 2,763 isolates were randomly selected. Additionally, we obtained sublineage data for 156 isolates from a collection of strains from Uganda, in which the L4.8/"Uganda" genotype had been previously identified (Wampande *et al.*, 2013). To correct for the proportion of L4.8/"Uganda" isolates in the same time period, we estimated the corresponding proportion of L4.8/"Uganda" genotype in Uganda based on reported data (Wampande *et al.*, 2013) and added 447 data points of L4.8 for the sublineage phylogeography. These data were not included for the frequency figures (Figure 7.6). Including isolates from Uganda, we therefore used sublineage classification data for 3,366 isolates with known country of birth of the patient. Approximately one third of these isolates were from patients immigrated to Switzerland or to the US (1,106; 32.9%), where we used country of birth of the patient as a proxy for country of infection with the MTB strain, as previously done (Fenner *et al.*, 2012b; Anderson *et al.*, 2013; Gagneux *et al.*, 2006b). Two thirds of the isolates (2,260; 67.1%) were from countries where both country of isolation and country of birth were identical.

For the set of 150 "LAM" strains used for WGS, we included previously sequenced strains and added selected isolates identified during the sublineage-SNP-typing. Additionally, we downloaded genome sequences from Lanzas *et al.* (2013) and Coll *et al.* (2014). The reference strain H37Rv and one isolate of L4.6 were included as the outgroup.

7.3.2. Whole genome sequencing, SNP calling and diversity analyses

Whole genome sequencing was performed using Illumina chemistry (GAIIx and HiSeq, paired end or single end). We used a pipeline as previously described for the mapping of short sequencing reads to the reference genome (a hypothetical MTBC ancestor strain (Comas *et al.*, 2010)) with BWA (Coscolla *et al.*, 2013)(and Chapter 6). As a difference to the previous pipeline, we used a dynamic minimum reads threshold for SNPs to be kept, corresponding to 10% of the mean coverage over the whole genome. SNPs with a

quality of less than 30 (QUAL field in VCF files), and SNPs with a coverage of more than double the average genome coverage were excluded. Illumina MiSeq-generated sequencing reads were clipped for Illumina adapters with Trimmomatic (Bolger *et al.*, 2014) before mapping. All SNPs were then annotated using H37Rv reference annotation (AL123456.1) with Annovar (Wang *et al.*, 2010) and customized scripts. SNPs in regions annotated as "PE/PPE/PGRS", "maturase", "phage" and "insertion sequence" were removed. The final alignment of polymorphic positions in all strains was imported into MEGA5 (Tamura *et al.*, 2011). Phylogenetic trees were built using Maximum Likelihood. Positions with gaps in more than 90% of sequences were ignored. Pairwise SNP distances were calculated using the "Compute pairwise distance" function in MEGA5. Genetic diversity (nucleotide diversity π, the average pairwise distance between two DNA sequences per site) and fixation indices (F_{ST}; an estimation of population separation) were calculated using Arlequin 3.5 (Excoffier *et al.*, 2005) and DnaSP v5 (Librado *et al.*, 2009). Populations were defined according to country of birth of the patient, either as continents or large geographic regions according to the United Nations definition[1]. The presence of large deletions (large sequence polymorphisms, LSP, according to Gagneux *et al.* (2006b)) was assessed by manually inspecting BAM alignment files from BWA mappings in Artemis (Rutherford *et al.*, 2000) for the presence of reads at the genomic regions with described deletions. For the definition of MTB Lineage 4 sublineages, we performed principal component analysis (PCA) with Jalview[2], using the alignment of all variable positions extracted from 72 whole genome sequences of Lineage 4 isolates.

7.3.3. Identification of sublineage-specific SNPs

The alignment of all SNPs from the initial 72 MTB Lineage 4 strains was converted to NEXUS format and imported into Mesquite (Maddison *et al.*, 2011), in parallel with the phylogenetic tree generated from the same data in MEGA5. We then used the "Trace Character History" module of Mesquite to map polymorphisms to clades. The full dataset of reconstructed positions was exported to an OpenOffice spreadsheet, and sublineage-specific SNPs were extracted as nucleotide differences between internal nodes of the phylogeny.

[1] `https://unstats.un.org/unsd/methods/m49/m49regin.htm`
[2] `http://www.jalview.org`

7.3.4. SNP-typing to screen for MTB Lineage 4 sublineages

We developed a new SNP-typing assay to screen clinical isolates for the defined sublineages. For this, we selected one canonical SNP per sublineage using previously defined methods and criteria. Oligonucleotides were designed for a 10-plex MOL-PCR assay based on the Luminex xTag platform (Luminex, Austin, USA) (Stucki *et al.*, 2012) (Figure 7.2 and Table 7.1). DNA extracts from clinical MTB isolates (previously confirmed as MTB Lineage 4 isolates) were screened with i.) the new MOL-PCR assay, ii.) standard PCR amplification plus subsequent Sanger sequencing of the region up- and downstream of the sublineage-specific SNP (isolates from New Zealand and Canada; PCR primers in Table 7.2), iii.) a real-time PCR melting curve assay using the same SNPs (isolates from China), or iv.) the MassARRAY platform (Sequenom, San Diego CA, USA) using phylogenetically redundant SNPs (samples isolated in San Francisco).

Table 7.1.: **Sublineage-specific SNPs and oligonucleotides used for MOL-PCR.** Fields defined as in Table 4.1 of Chapter 4.

MTBC Sublineage	Alternative Name	SNP	Reference Allele	Mutant Allele	Gene	Ess/Syn	Sense	Oligonucleotide	Bead Region	Sequence
L4.1	X	3798451	C	G	Rv3383c (idsB)	noness/syn	sense	LPO_ancestral	29	GGGTTCCCTAAGGGTTGGAtactacttctataactcacttaaaAATGCTTTCCCAAGGTCAGCAGGGACACTC
								LPO_mutant	28	GGGTTCCCTAAGGGTTGGAcacttaattcattctaaatctatcAATGCTTTCCCAAGGTCAGCAGGGACACTG
								RPO	-	P-CCCAGAAAAGCCGCATCCAGAGTCAATATCTAGATTGGATCTTGCTGGCAC
L4.2	Ghana	4409231	T	G	Rv3921c	ess/nonsyn	sense	LPO_ancestral	27	GGGTTCCCTAAGGGTTGGAtaacttacacttaactatcatcttCCTGCTTTTTGGCCTCCTCCT
								LPO_mutant	12	GGGTTCCCTAAGGGTTGGAcataatcaatttcaactttctactCCTGCTTTTTGGCCTCCTCCG
								RPO	-	P–CCTTTTCGATCATGCCGAAGACGTAATGCTCTAGATTGGATCTTGCTGGCAC
L4.3	Haarlem	3013784	C	G	Rv2697c (dut)	ess/nonsyn	anti-sense	LPO_ancestral	14	GGGTTCCCTAAGGGTTGGAaatttcttctctttctttcacaatAGTTGCTAGTGCAACGGGTTGAGTTGG
								LPO_mutant	30	GGGTTCCCTAAGGGTTGGActtaacatttaacttctataacacAGTTGCTAGTGCAACGGGTTGAGTTGC
								RPO	-	P-TCGAGCTGGTCGAGGTCTCGTCGTTTCTAGATTGGATCTTGCTGGCAC
L4.4	Ural	2181026	G	C	Rv1928c	noness/syn	sense	LPO_ancestral	33	GGGTTCCCTAAGGGTTGGAactacttattctcaaactctaataTGCTGGCTCACATCGCAGCAGACG
								LPO_mutant	15	GGGTTCCCTAAGGGTTGGAtacttctttactacaatttacaacTGCTGGCTCACATCGCAGCAGACC
								RPO	-	P-GGCACGACCTTGCCACCTGATGTTCTAGATTGGATCTTGCTGGCAC
L4.5	LAM	1480024	G	T	Rv1318c	noness/nonsyn	anti-sense	LPO_ancestral	20	GGGTTCCCTAAGGGTTGGActttctcatactttcaactaatttGCTGATCATCTCGATGGTCACATTGGTGTTC
								LPO_mutant	21	GGGTTCCCTAAGGGTTGGAtcaaactctcaattcttacttaatGCTGATCATCTCGATGGTCACATTGGTGTTA
								RPO	-	P-GGGTTCATCCTGATGTGGATCCTGGCCTTCTAGATTGGATCTTGCTGGCAC
L4.6	Vietnam	3966059	G	C	Rv3529c	noness/nonsyn	anti-sense	LPO_ancestral	36	GGGTTCCCTAAGGGTTGGAattaaacaactcttaactacacaaACTTGATTGCCGATCCGCTGGGTAC
								LPO_mutant	37	GGGTTCCCTAAGGGTTGGAtacaacatctcattaacatatacaACTTGATTGCCGATCCGCTGGGTAG
								RPO	-	P-GGTGGCAGATATCTACCGGCACTTCGTCTAGATTGGATCTTGCTGGCAC
L4.7	Iran	2789341	A	C	Rv2483c (plsC)	ess/nonsyn	anti-sense	LPO_ancestral	18	GGGTTCCCTAAGGGTTGGAacacttatctttcaattcaattacATCGCCGAAGGCCAAACCCAGCGAATCTAAGAT
								LPO_mutant	22	GGGTTCCCTAAGGGTTGGAcaaacaaacattcaaatatcaatcATCGCCGAAGGCCAAACCCAGCGAATCTAAGAG
								RPO	-	P-CGCTGGCAAGGATGGTGAGGCCTCCGCATCGCCAAGCTCTATCTAGATTGGATCTTGCTGGCAC
L4.8	Uganda	990626	T	A	Rv0890c	noness/nonsyn	sense	LPO_ancestral	34	GGGTTCCCTAAGGGTTGGAacttatttcttcactactatatcaATCGCATCACCTCCTGCCAGGGCT
								LPO_mutant	35	GGGTTCCCTAAGGGTTGGAcatcttcatatcaattctcttattATCGCATCACCTCCTGCCAGGGCA
								RPO	-	P-AACTGCGCCATCAGGACCTGGTGCATTCTAGATTGGATCTTGCTGGCAC
L4.9	Cameroon	3191099	C	A	Rv2881c (cdsA)	ess/nonsyn	anti-sense	LPO_ancestral	19	GGGTTCCCTAAGGGTTGGAatactttacaaacaaataacacacCGCAATGCTGGTCTACCCGGAAAATG
								LPO_mutant	25	GGGTTCCCTAAGGGTTGGActttcttaatacattacaacatacCGCAATGCTGGTCTACCCGGAAAATT
								RPO	-	P-GCTCGGGATGGGTGTTCTGCATGATGATTCTAGATTGGATCTTGCTGGCAC
L4.10	PGG3	1692141	C	A	Rv1501	noness/syn	sense	LPO_ancestral	13	GGGTTCCCTAAGGGTTGGAcaaatacataatcttacattcactCGACTCATGATGAAGTATGACCCTCATTTC-TTTACCTTTCTTGAAATC
								LPO_mutant	26	GGGTTCCCTAAGGGTTGGAtacattcaacactcttaaatcaaaCGACTCATGATGAAGTATGACCCTCATTTC-TTTACCTTTCTTGAAATA
								RPO	-	P-CCCGAAGTCCTAAGCATCGTTGATCGTGTGCTATCTGAAACTCTAGATTGGATCTTGCTGGCAC

Table 7.2.: **PCR primers for amplification of regions flanking Lineage 4 sublineage specific SNPs.**

Sublineage	SNP	Forward Primer	Reverse Primer	Amplification Fragment Length
L4.1	3798451	TTGTGCACCAACTCCACAGCCG	CGTGTCTTTTCTGTAGTGGATGACC	518bp
L4.3	3013784	GCGTGTCCGGCCTTGCGTTTG	CCGGCGTTGATCTCTACAGC	522bp
L4.2	4409231	AGGATTGTCAACGTTTGCGT	GGATGCGTTCGTCGATTTCAG	563bp
L4.4	2181026	CGCCTTGGAGGCGCAGTAGTGG	GGCATGTGATTCCATCAGGTATC	557bp
L4.6	3966059	CGGGCAAATTGCGTATCTGC	GGTACTAAAGAATCCGAGTCATC	531bp
L4.5	1480024	GCCGGCTGGTCACCAATTGCGTC	GTTCGCCGCCCAGGCGCTCGAG	542bp
L4.7	2789341	TAGAACGGTCCTCGCCAGATTG	GCAACTCCACCACGATCAATC	656bp
L4.8	990626	CGGACACCTTCGGAGTGACTG	GCCCAGGTGCTGGCGTATTGC	500bp
L4.9	3191099	CACGTTTGACCTGCGACTCCAC	CCTTGGTCGCTACCCATGAG	603bp
L4.10	1692141	AGGTGAATAAGCGTAGCATGATTC	GCGCGAGGTAGGTATGGTCC	540bp

7.3.5. Reconstruction of ancestral geographical origin of LAM strains

The software RASP (Yu *et al.*, 2010) was used to reconstruct the hypothetical geographic origin of the MTBC LAM ancestor genotype. As input, the alignment of all 12,256 SNPs among 150 LAM strains plus the two outgroup strains was used. Both Maxiumum Parsimony (S-DIVA) as well as Bayesian Binary Method were used.

7.3.6. Genetic diversity of LAM strains compared to geographical distances

To analyze the association between genetic diversity of LAM strains and distance from a reference point, we used great circle distances between capitals of each country and the reference point (Lisbon, Mexico City or Brasilia) as geographical distances. For each large geographical UN region (see above), we calculated within-population nucleotide diversity (π). In parallel, we calculated for each isolate the geographical distance between capitals of the countries of origin and the reference point. We plotted the mean geographical distance against the nucleotide diversity. One LAM strain was excluded as the patient place of birth was unknown. The single isolate from the Oceania region (a patient from Australia) was also excluded. GraphPad Prism (GraphPad Software, La Jolla, USA) was used for statistical analyses and figures.

7.3.7. Estimation of MTB Lineage 4 prevalence

To estimate the proportion of MTB Lineage 4 among all MTBC isolates, we used the "Online Tools"/"Clade Distribution" function of the SITVITWEB database on `http://www.pasteur-guadeloupe.fr:8081/SITVIT_ONLINE/index.jsp` and extracted data for each country separately. Spoligotyping families with either of the text strings "LAM", "S", "H", "MANU", "T" or "X" were categorized as MTB Lineage 4; spoligotypes with text strings "EAI", "PINI", "AFRI", ,"BOV", "CAS" or "BEIJING" as non-Lineage 4, and "ZERO" and "UNKNOWN" as "unknown" (referring to Coscolla *et al.* (2010) and Comas *et al.* (2009b)). Orphan spoligotypes are not included in the "Clade Distribution" function of the SITVITWEB database. A total of 49,799 data points were used.

7.4. Results

7.4.1. Definition of 10 MTB Lineage 4 sublineages

We first aimed at defining sublineages within MTB Lineage 4. We started with 72 MTB Lineage 4 whole genome sequences. A total of 9,455 genomic single nucleotide positions were variable among these 72 strains, and used to generate a Maximum Likelihood phylogeny. The overall topology of this phylogeny was congruent with previously published topologies (Comas *et al.*, 2013). Based on the Lineage 4 phylogeny (Figure 7.2), we initially defined sublineages by the branching depth of the topology (dashed line in Figure 7.2), indicating in 9 sublineages. We further subdivided one clade into L4.8 and L4.9, as these two sublineages had been previously defined based on specific regions of difference (RDs) (Gagneux *et al.*, 2006b) and spoligotyping data (Niobe-Eyangoh *et al.*, 2003; Niemann *et al.*, 2002), and the definitions have been widely used in various typing schemes ("Uganda" family, RD724-deleted, historically known as *M. africanum* subtype II (Jong *et al.*, 2010), and "Cameroon" family, RD726-deleted). We continued our study based on 10 defined sublineages.

To validate our definitions, we used five different approaches. First, we performed principal component analysis (PCA) using the complete SNP alignment of 9,455 SNPs. PCA showed a clear separation of seven sublineages (L4.1, L4.2, L4.3, L4.4, L4.5, L4.6, L4.10; Figure 7.3). Sublineages L4.7, L4.8 and L4.9, however, were less clearly separated, likely mirroring the (intended) subdivision of the initial sublineage into these two "sub-sublineages" (see above). L4.8 and L4.9 ("Uganda" and "Cameroon") are part of a larger clade including three additional isolates from Ethiopia.
Second, we calculated pairwise SNP distances between and within the defined sublineages. Whereas the mean pairwise SNP distance between any two of all 2,556 pairs (72 isolates) was 564 SNPs, the distance between pairs *within* the defined sublineages was significantly lower than between pairs *between* sublineages (242 SNPs vs. 601 SNPs, $p < 0.0001$, Table 7.3), supporting our sublineage definitions.

Third, we calculated sublineage-pairwise fixation index values (F_{ST}). F_{ST} values are a concept from diploid organisms and indicate the degree of separation between two populations based on nucleotide differences, and range from 0 ("panmixis", i.e.
no population separation) to 1 (completely separated populations). All F_{ST} values (Table 7.4) between our defined sublineages were found to be larger than 0.33, indicating good population separation, i.e. a reasonable definition of sublineages.

Fourth, we compared the genetic diversity among all 72 isolates with the genetic diversity *within* each sublineage. We used nucleotide diversity π, the mean number of

Table 7.3.: **Mean pairwise SNP distances within and between isolates of defined sublineages.**

	Distance within sublineages	Distance to other sublineages
L4.1	325	621
L4.2	26	622
L4.3	162	600
L4.4	323	656
L4.5	274	605
L4.6	330	556
L4.7	321	608
L4.8	220	554
L4.9	136	589
L4.10	307	596
Mean	**242**	**601**

nucleotide differences per site, between any two DNA sequences of a population (in our case the sublineages). The nucleotide diversity was found lower *within* each sublineage than among all 72 isolates, indicating a valid level of sublineage definition (Figure 7.4). However, for L4.4, the difference to the nucleotide diversity among all isolates was not statistically significant (overlapping confidence intervals). Differences *between* sublineages were not significantly different except for sublineages 4.2 and 4.3.

Last, to further support our definitions and to make a comparison with other studies possible, we mapped previously reported markers, i.e. SNPs and RDs, onto our genome-based phylogenetic tree (Sreevatsan *et al.*, 1997; Gagneux *et al.*, 2006b; Filliol *et al.*, 2006; Abadia *et al.*, 2010; Homolka *et al.*, 2012) (Supplementary Figure 7.5). Many of the markers overlapped and therefore supported our definitions. However, the figure also revealed that previous sets of markers were incomplete, and that no comprehensive RD/SNP set had been defined hat would include all our sublineages. In particular, no markers had been reported for sublineages L4.4 and L4.6. So far, no marker has been identified that was not covered by our definitions, indicating the completeness of our set of 10 defined sublineages.

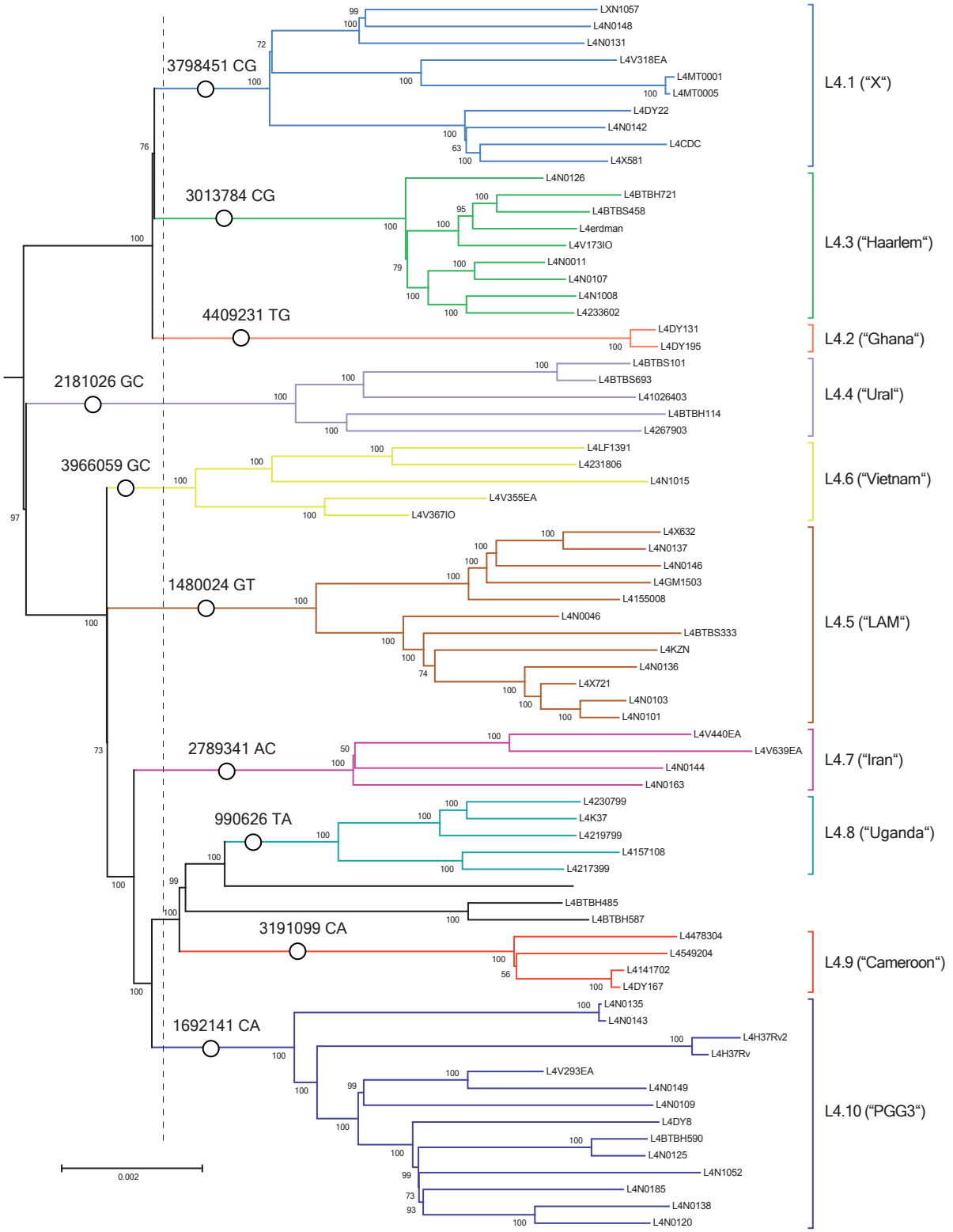


Figure 7.2.: **Maximum likelihood phylogeny of 72 whole genome sequences of Lineage 4 strains and 9455 SNPs.** Bootstrap values are indicated. On the right side of the tree and colored are the newly defined sublineages of Lineage 4. Previously reported markers from other publications are mapped to the same tree in Figure 7.5.

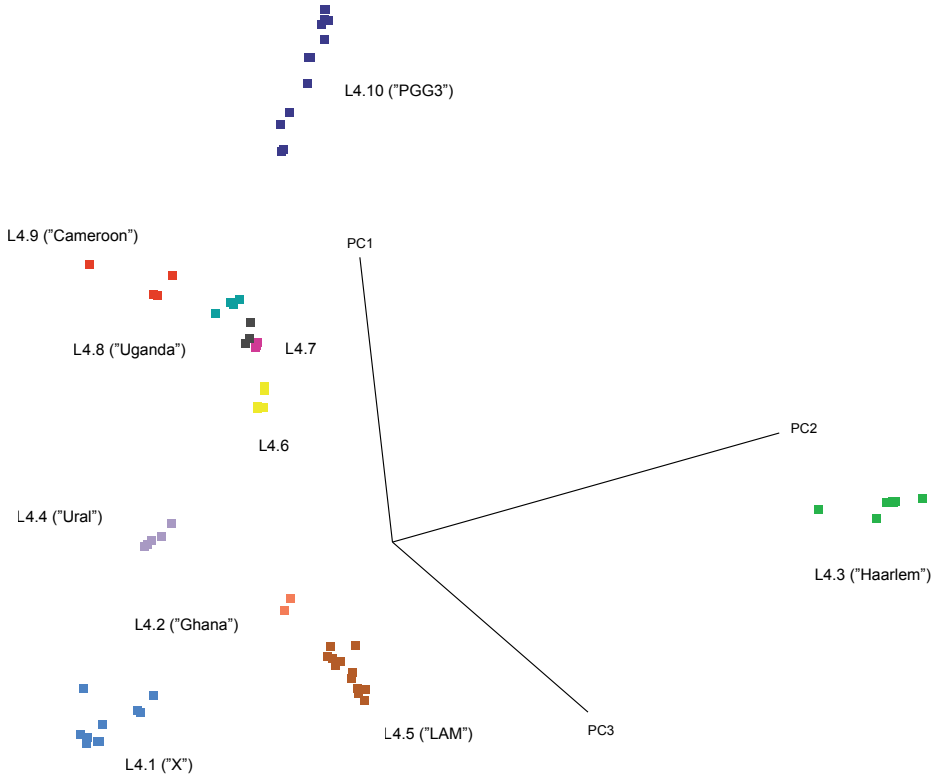


Figure 7.3.: **Principal component analysis of 9455 SNPs in 72 Lineage 4 strains.** Each dot represents one of 72 MTB Lineage 4 strains. Colors correspond to Figure 7.2.

Table 7.4.: **Pairwise F_{ST} values between MTB lineage 4 sublineages.** Numbers in brackets are values with p>0.05. All other values were statistically significant.

	L4.1	L4.2	L4.3	L4.4	L4.5	L4.6	L4.7	L4.8	L4.9	L4.10
L4.1	0									
L4.1	0									
L4.2	0.408	0								
L4.3	0.432	0.683	0							
L4.4	0.387	0.448	0.527	0						
L4.5	0.510	0.631	0.630	0.413	0					
L4.6	0.417	(0.57)	0.590	0.330	0.430	0				
L4.7	0.471	(0.63)	0.649	0.387	0.500	0.389	0			
L4.8	0.486	0.679	0.660	0.414	0.510	0.411	0.471	0		
L4.9	0.523	(0.67)	0.692	0.455	0.563	0.471	0.509	0.508	0	
L4.10	0.487	0.588	0.594	0.429	0.482	0.389	0.437	0.427	0.499	0

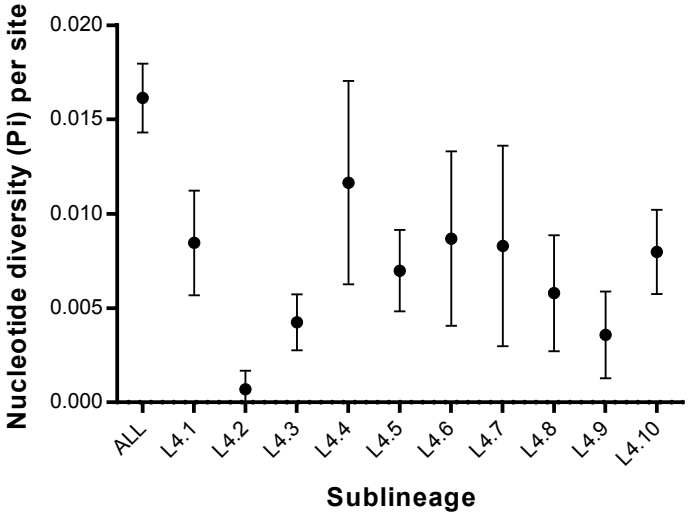


Figure 7.4.: **Nucleotide diversity (per polymorphic site) of MTB Lineage 4 sublineages, calculated with Arlequin software.** A total of 8,872 polymorphic sites were used. Error bars indicate confidence intervals.

7.4.2. Phylogeographic distribution of Lineage 4 sublineages

We then aimed at extracting informative, sublineage-specific SNPs as markers for the newly defined sublineages. Using the 9,455 variable positions among the 72 MTB Lineage 4 strains, we found between 51 and 277 SNPs specific for the corresponding sublineage. Suitable markers were extracted and globally representative clinical isolates screened for sublineage classification. We obtained sublineage-classification data for 3,273 (97.2%) of a total of 3,366 SNP-typed isolates. The remaining 93 (2.8%) isolates failed during SNP-typing. A total of 92/3,273 (2.8%) isolates were found to harbour the reference allele for all 10 sublineages (i.e. not classified into any of the ten sublineages), but were confirmed to be Lineage 4 isolates, and potentially representing additional sublineages. Each of the remaining 3,181 isolates was classifed into one of the ten defined sublineages. Sublineage L4.5 ("LAM") was the most frequent sublineage with a global proportion of 22.9% among all sublineages, followed by L4.10 (14%), L4.3 (11.6%), L4.1 (11%), L4.9 (10.4%) and L4.6 (10.7%) (Panel A in Figure 7.6). However, frequencies were likely biased by the unequal distribution of isolates among countries (Panel B in Figure 7.6). We therefore plotted the relative frequencies of sublineages in each country (country of birth of patient) to a global map. Figure 7.7 shows that MTB Lineage 4 sublineages are phylogeographically structured, i.e. were found in unequal proportions between countries, and different sublineages were found to be predominant in different geographic regions of the world. Strikingly, individual sublineage 'heat maps" (Figure 7.8) revealed sublin-

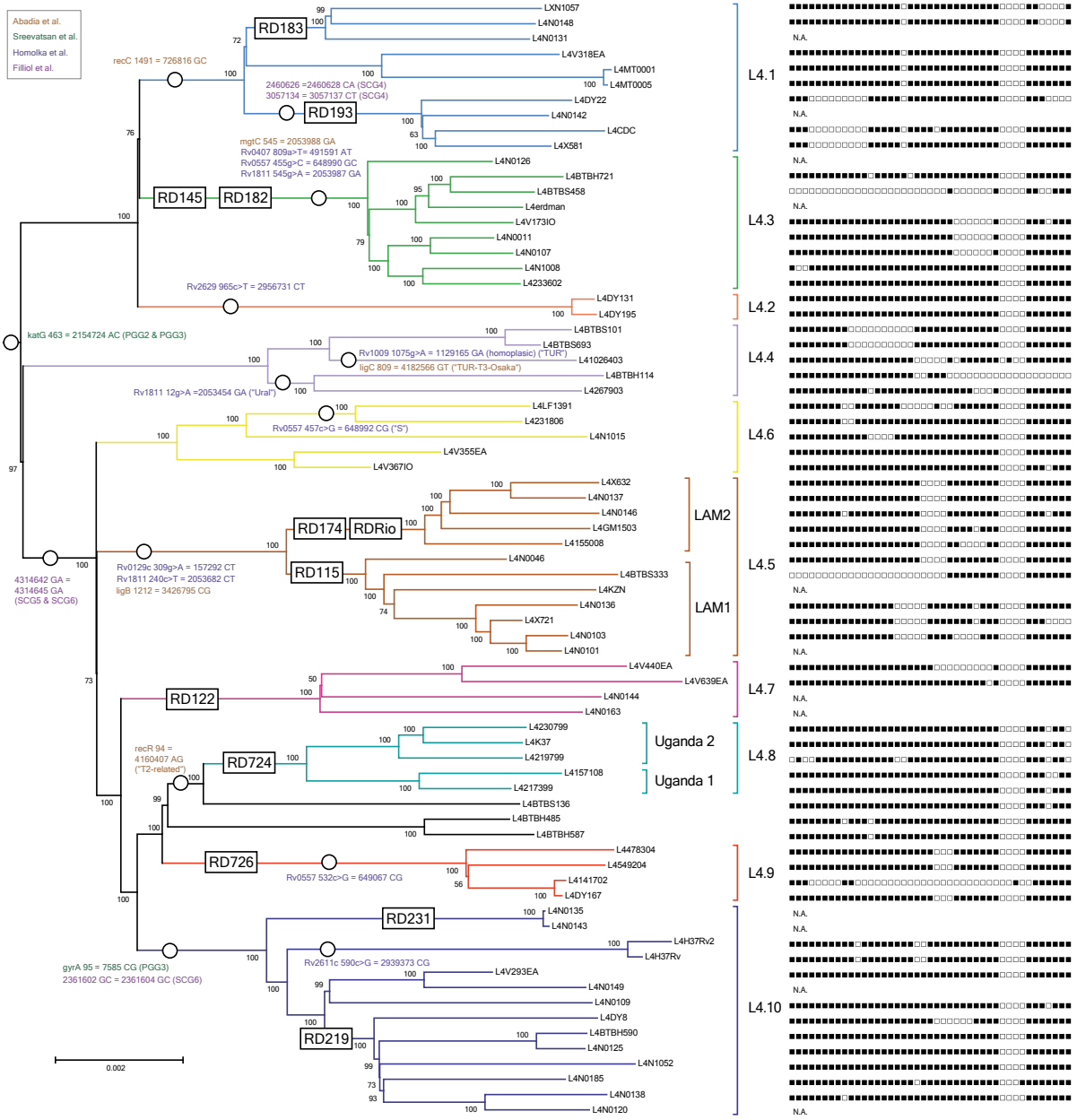


Figure 7.5.: **Phylogenetic tree of 72 whole genome sequences as in Figure 7.2, with previously reported Lineage 4 sublineage specific markers mapped.** Boxes on branches represent large sequence polymorphisms (LSP), also known as regions of difference (RD), as defined by Gagneux *et al.* (2006b) and used by Lazzarini *et al.* (2007). Circles are clade-specific SNPs previously defined by Sreevatsan *et al.* (1997), Filliol *et al.* (2006), Abadia *et al.* (2010) and Homolka *et al.* (2012).

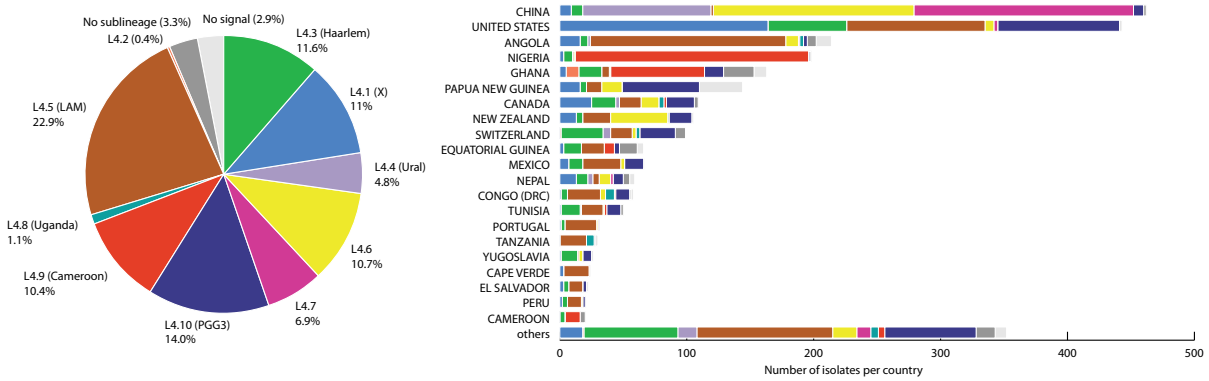


Figure 7.6.: **Sublineage L4.5 of MTB Lineage 4 ("LAM") was the most frequent sublineage in our dataset. A**. Pie chart showing global frequencies of sublineages. Data of 2763 isolates were used (excluding data from Uganda, see above). **B**. Bar chart showing isolates by country of birth of the patients, stratified by sublineages. Category "others" includes countries with less than 20 isolates. Colors correspond to sublineage colors in panel A.

eages that were present globally, and sublineages that were locally restricted. Sublineages L4.3 (Haarlem), L4.5 (LAM) and L4.10 (PGG3) were found globally. Sublineages L4.2 (Ghana), L4.4 (Ural), L4.7, L4.8 and L4.9 were restricted to specific geographical regions. Sublineage L4.1 (X) was observed in the Americas and in lower proportions in few countries of Southern Africa, Asia and Europe. L4.6 was seen in high proportions among isolates from Asia, and in lower proportions among isolates from few countries in Africa, Europe and the Americas.

Each of the three global sublineages (L4.3, L4.5, L4.10) were found in more than 45 countries (Figure 7.9). The L4.5/"LAM" sublineage was found with a proportion of 30-100% in 47 countries in which it was observed (Figure 7.9).

Five of the 10 sublineages were not observed among patients born in Europe. Among the other five sublineages, L4.3 (Haarlem), L4.5 (LAM) and L4.10 (PGG3) were found frequently in Europe and in high proportions. L4.1 and L4.6, on the other hand, were observed only in four and five European countries, respectively, and in low proportions. Importantly, the three sublineages most frequent in Europe were also observed globally.

7.4.3. Discovery of new diversity

We generated whole genome sequences of 17 of the 92 isolates (18.5%) that were not classified into any of the 10 defined sublineages i.e. harbouring the ancestral allele for all 10 SNPs). The resulting phylogenetic tree (Supplementary Figure 7.10) showed the same topology as in Figure 7.2, but revealed that the 17 isolates (plus the previously uncategorized 3 isolates clustering with L4.8 and L4.9) branched off towards to root of tree compared to the branching points of the defined sublineages. Interestingly, these

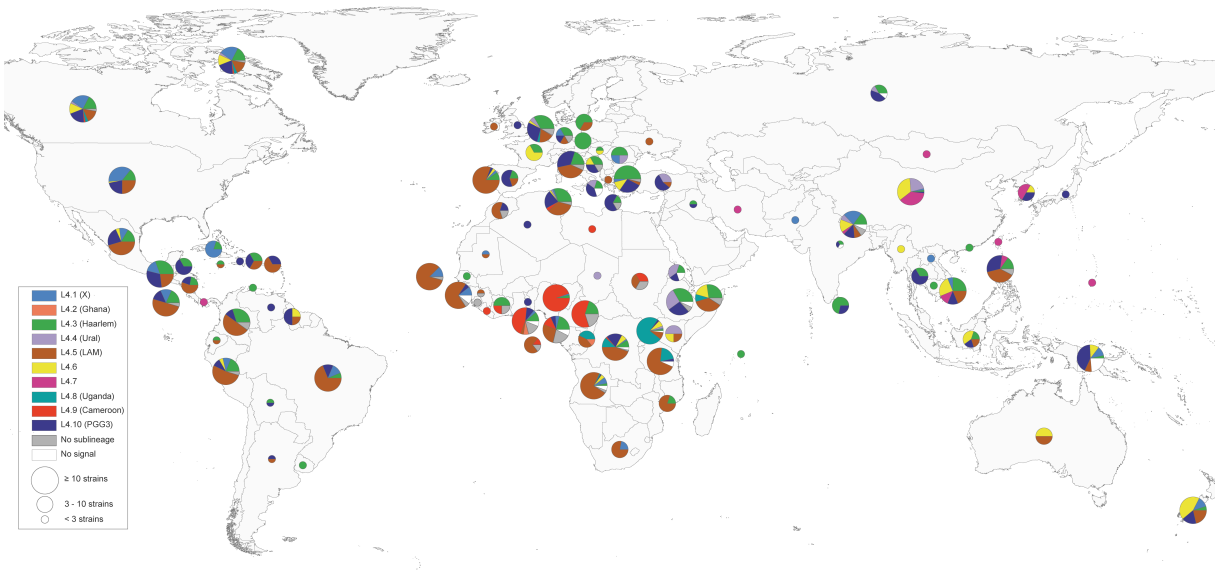


Figure 7.7.: **Lineage 4 sublineages were found globally distributed, but proportions of sublineages varied depending on the country and the geographical region.** Pie charts show proportions of the 10 sublineages among all MTB Lineage 4 isolates per country. Circle sizes correspond to number of isolates per country. Data obtained from screening 3,366 MTB Lineage 4 isolates with sublineage-specific SNPs.

strains did not form a distinct new sublineages; instead they were spread across the phylogeny.

7.4.4. Reconstruction of the global dispersal of LAM strains by whole genome sequencing

To address the evolutionary scenario of MTB Lineage 4, we focused on the presumably most "successful" sublineage, L4.5/"LAM", and generated whole genome sequences of 150 isolates. We found 11,352 SNPs separating the 150 LAM isolates (12,256 SNPs including the outgroup). The mean pairwise distance between two LAM isolates was 387 SNPs (range 6-581; standard deviation 97). The resulting Maximum Likelihood phylogeny showed several sub-groups which we labelled subgroups A to E (Figure 7.11); we also included the previous definitions of RDRio/RD174 (Lazzarini *et al.*, 2007) and RD115 (Gagneux *et al.*, 2006b).

We first tested the hypothesis that the phylogeny reflects a LAM strain population structured by geography, i.e. that particular LAM sub-clades are associated with geographical regions. To that end, we mapped the strain origin (patient place of birth) of each LAM strain to the phylogeny (colored tree tips by continent in Figure 7.5). Overall, the four continents were heterogenously distributed among all subgroups (Figure 7.11),

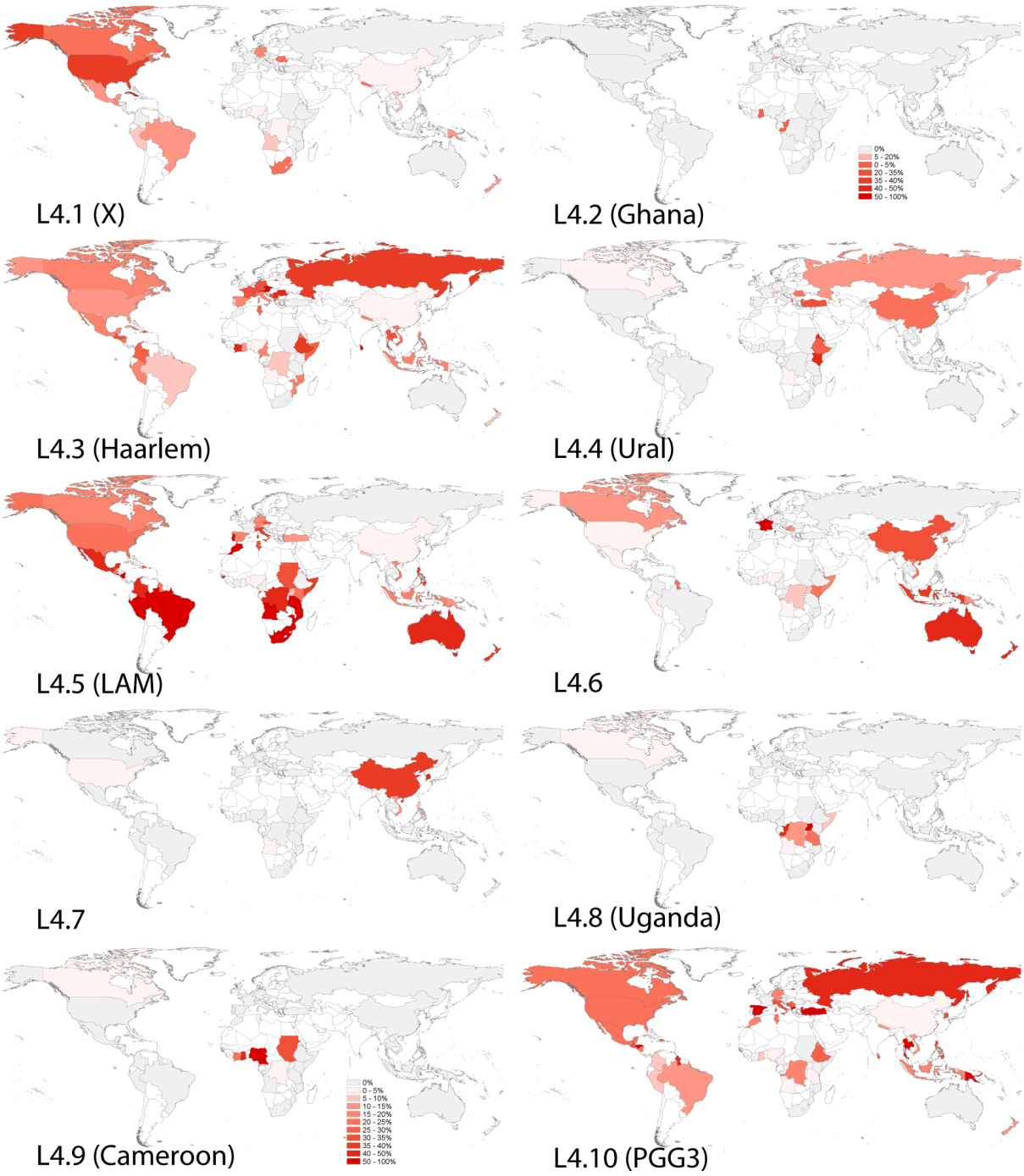


Figure 7.8.: **Heat maps of the proportions of the different sublineages reveal "global" versus "local" sublineages.** Sublineages L4.1 (X), L4.3 (Haarlem), L4.5 (LAM), L4.6 and L4.10 (PGG3) were found in higher proportions globally spread, whereas sublineages L4.2 (Ghana), L4.4 (Ural), L4.7, L4.8 and L4.9 were restricted to certain geographical regions. Intensity of red indicates proportion of the sublineage among all Lineage 4 isolates. Countries with less than three isolates in total are shown as "no data" and are filled white. A total of 3,366 isolates was used. The scale for red intensity is the same for all sublineages (and indicated at L4.9) except for the L4.2 sublineages (separate scale shown).

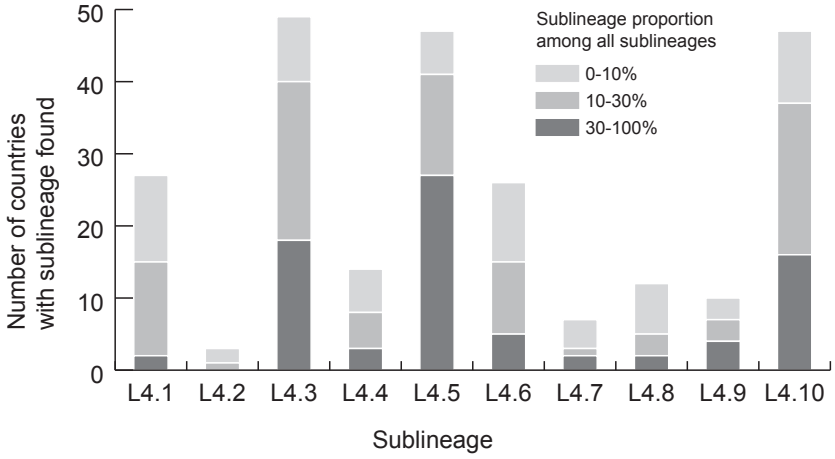


Figure 7.9.: **Isolates of five sublineages of MTB Lineage 4 (L4.1, L4.3, L4.5, L4.6, L4.10) were found in more than twenty countries.** Three sublineages were found in more than 45 countries each (L4.3, L4.5 and L4.10). One of those sublineages, L4.5 ("LAM"), was found with a proportion of 30-100% in more than 50% of the countries of occurrence. The differences in proportions among sublineages were statistically significant (p=0.001, χ^2 test).

but some regions were overrepresented in particular clades (Table 7.5). In subgroup A, a subclade of the well-described RDRio family, 43% and 38% of isolates were from the Americas and Europe, respectively, and less than 10% from Africa or Asia. By contrast in subgroup C, no isolate from the Americas was detected. Most remarkable, 100% (19/19) of the isolates in subgroup B1, also part of RDRio (named after its first description in Brazil), were from Africa.

However, irrespective of the varying proportions of the different subgroups, we did not identify any clade that branched off ancestrally and was represented by isolates from a single continent only.

We continued by reconstructing the most likely geographical origin of the hypothetical LAM ancestor genotype. Both Maximum Parsimony (100% probability) and Bayesian (66% probability) approaches predicted "Europe" as the most likely character state of the LAM ancestor (Figure 7.12). When we split Europe into Southern Europe and the rest of Europe (following the hypothesis that LAM was mainly spread with early migrations from Portugal and Spain), the signal for Europe was lost in the Bayesian approach (and combined Central America and Africa was obtained as most likely origin; data not shown), but remained strong for combined Southern and other Europe in the Maximum Parsimony reconstruction (specific combinations of regions can not be defined in the Bayesian approach, but can be defined in the Maximum Parsimony approach).

Next, we tested the hypothesis that the genetic diversity of LAM strains was highest in Europe if the LAM sublineage originated in Europe. The "export" to other continents

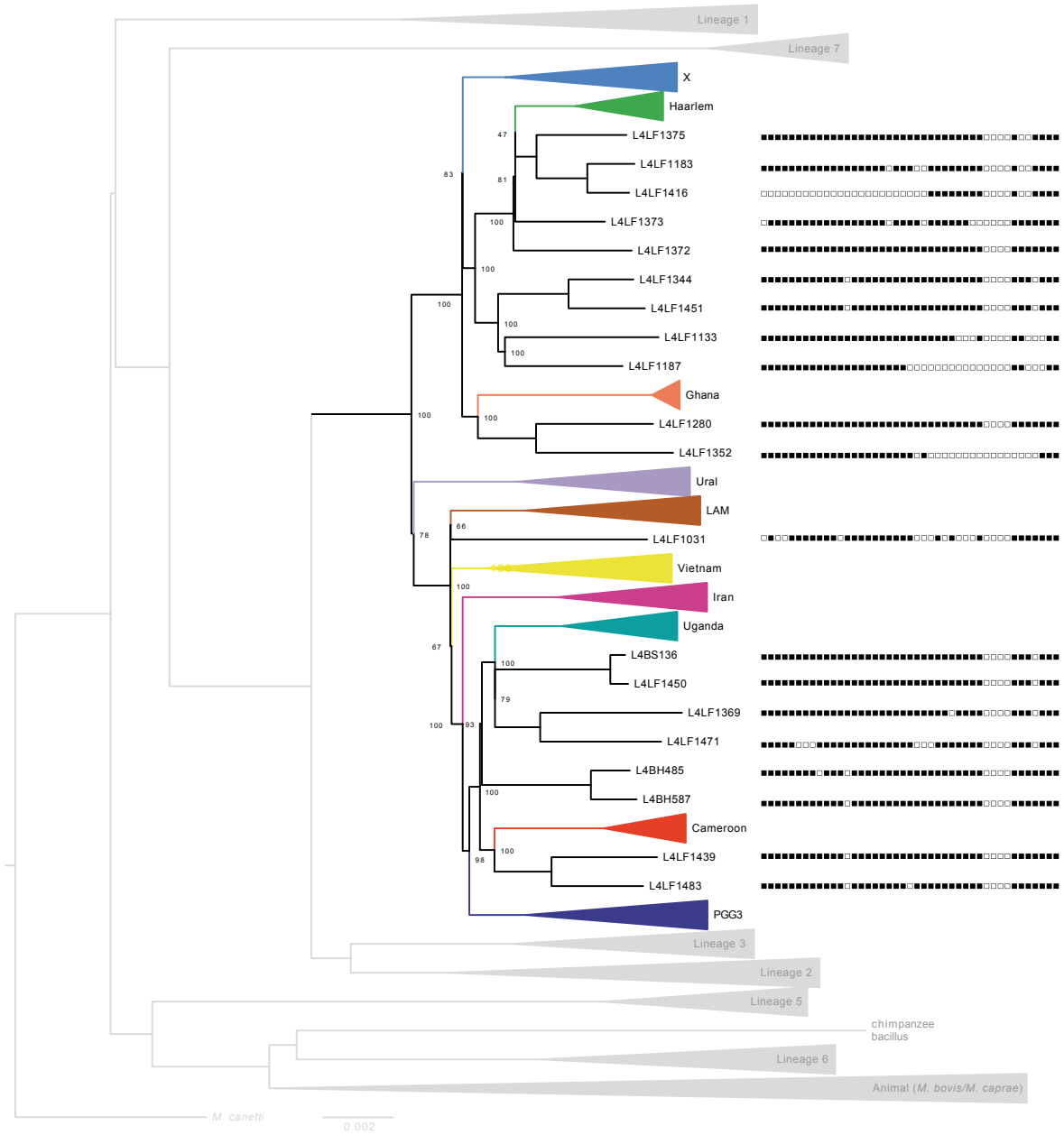


Figure 7.10.: **Lineage 4 isolates that were not classified into any of the 10 defined sublineages branch off more ancestrally compared to the ancestral node of the defined sublineages.** The phylogenetic tree was generated with all Lineage 4 isolates as in Figure 7.2 (collapsed branches), plus 17 newly sequenced isolates that were not classified into any of the 10 sublineages. Colored, collapsed clades represent sublineages as shown in Figure 7.2. Bootstrap values are indicated.

Table 7.5.: **Proportions of isolates from each continent in the subgroups of the LAM-phylogeny as defined in Figure 7.11.**

LAM group	Europe	Africa	Asia	Americas	others
A	8 (38.1%)	2 (9.52%)	1 (4.76%)	9 (42.86%)	1 (4.76%)
B	7 (18.92%)	22 (59.46%)	2 (5.41%)	6 (16.22%)	0 (0%)
B1	0 (0%)	19 (100%)	0 (0%)	0 (0%)	0 (0%)
C	3 (42.86%)	2 (28.57%)	2 (28.57%)	0 (0%)	0 (0%)
D	8 (38.1%)	8 (38.1%)	0 (0%)	5 (23.81%)	0 (0%)
E	11 (17.46%)	17 (26.98%)	3 (4.76%)	27 (42.86%)	5 (7.94%)
E1	2 (9.09%)	7 (31.82%)	2 (9.09%)	11 (50%)	0 (0%)

would have represented population bottlenecks (followed by clonal expansion), in sum reducing the genetic diversity in the non-European regions. Nucleotide diversity analysis was indeed found significantly higher in Europe than in Africa and the Americas (ANOVA $p<0.0001$ in Panel A in Figure 7.13). LAM strains in Asia were found to harbour a similar level of diversity; however, only 12 isolates were included from Asia.

Finally, we assessed if the genetic diversity of LAM strains decreases with the distance from Europe. Indeed, we found a weak correlation ($R=-0.22$, $p=0.142$) between distance from Lisbon and nucleotide diversity (Panel B in Figure 7.13). In contrast, when we used Mexico City or Brasilia as reference points, this correlation was lost ($-0.05<R<0.05$, n.s.).

7.4.5. Global prevalence of MTB Lineage 4

In order to estimate the TB burden caused by MTB Lineage 4 in different geographical regions of the world (in relation to all main phylogenetic lineages), we calculated country-specific proportions of Lineage 4 among all phylogenetic lineages using data from the SITVITWEB database (Demay *et al.*, 2012). Knowing the limitations of phylogenetic classification using spoligotyping, we used data from 49,799 MTB strains from 76 countries in the SITVITWEB database, and found MTB Lineage 4 globally prevalent with proportions between 0% (Singapore) and 100% (Dominican Republic) (Figure 7.14). As expected, markedly different proportions of Lineage 4 were found between continents. In the Americas, all countries were found to harbour more than 45% Lineage 4 isolates; in Africa, all countries except Sudan had more than 30% Lineage 4 isolates, with many countries more than 70%. In South-East Asia, on the other hand, no country was found to harbour more than 35% Lineage 4 isolates. Interestingly, countries in Northern and

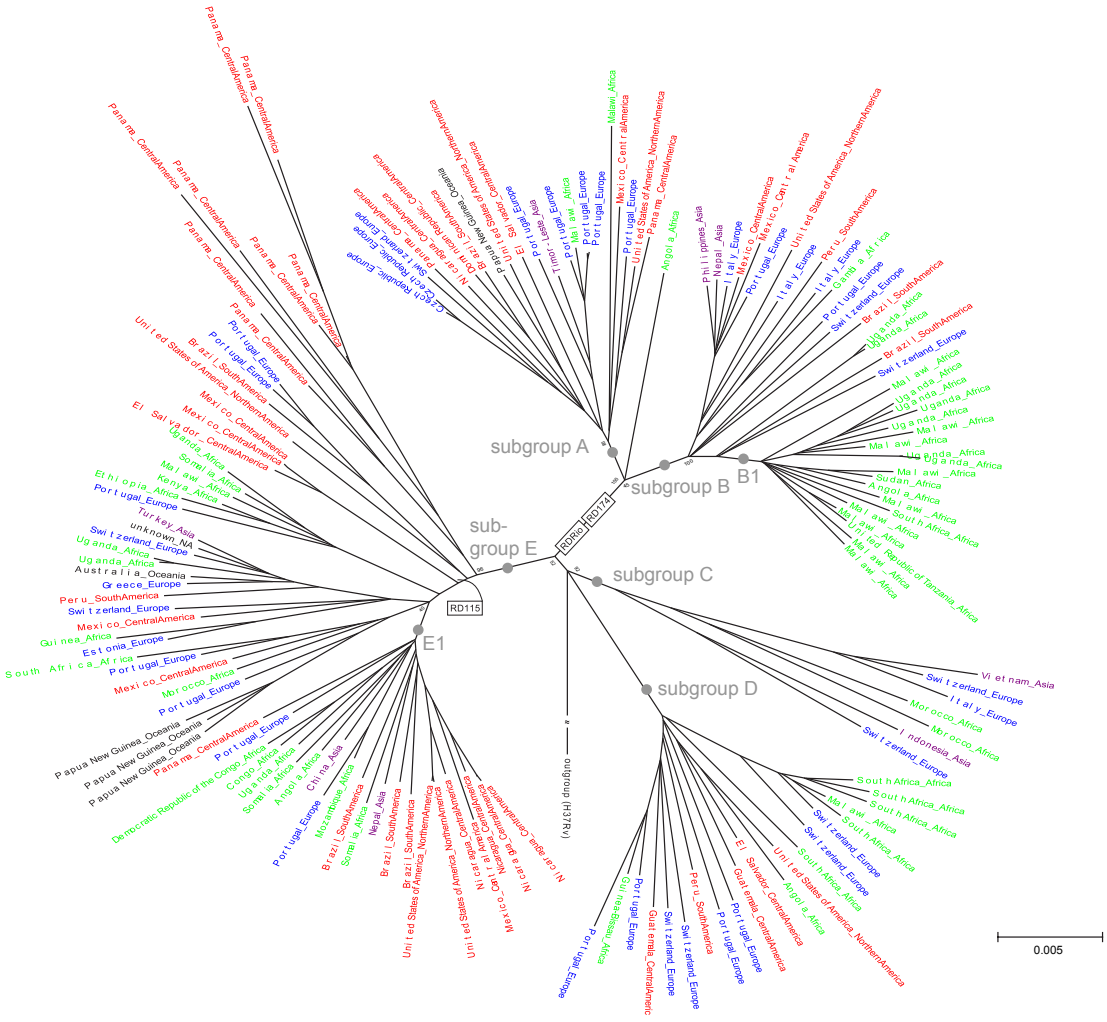


Figure 7.11.: **Maximum likelihood phylogeny of 150 isolates of the L4.5/"LAM" family of MTB Lineage 4.** H37Rv and one isolate of sublineage L4.6 were used to root the tree. A total of 12,256 SNPs separated isolates. Bootstrap values are indicated in small font. Boxes with "RD" designations indicate branches with the particular region deleted, identified by a lack of sequencing reads mapping to the corresponding genomic region. The scale bar shows number of substitutions (SNPs) per site.

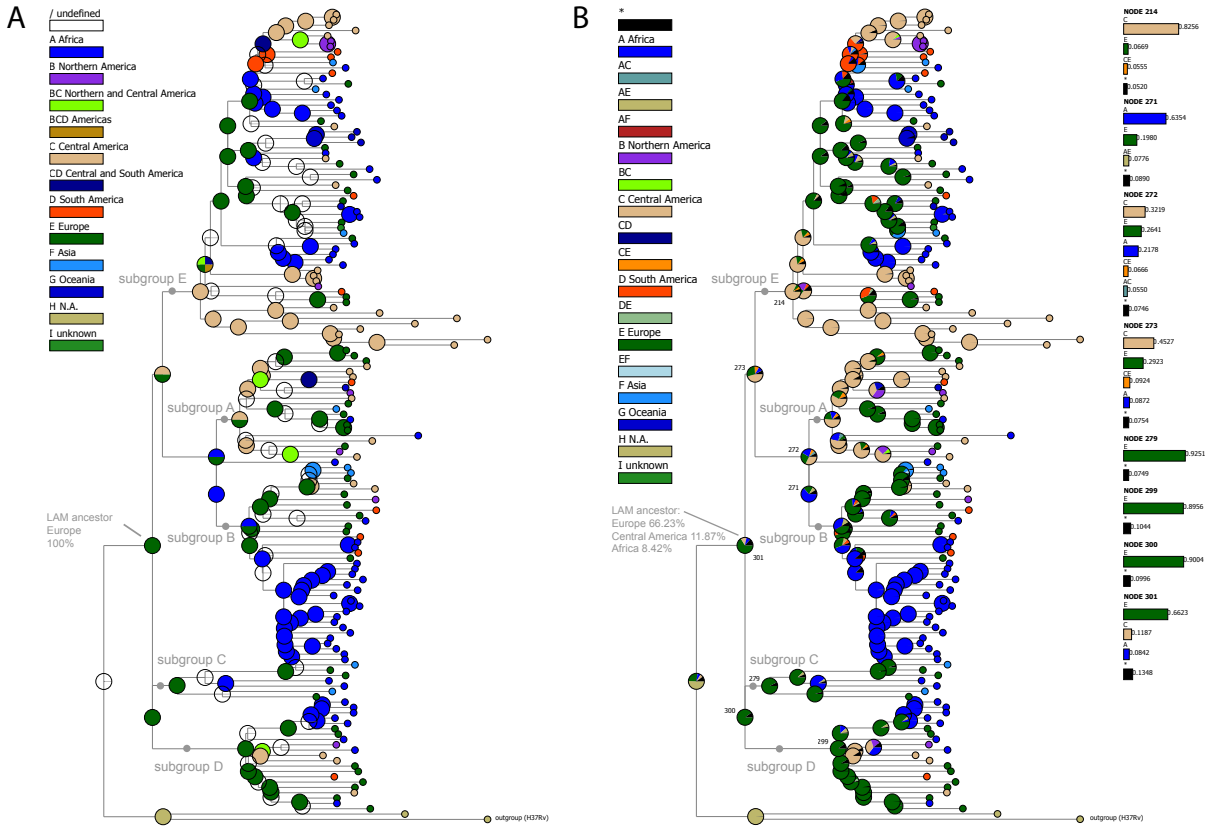


Figure 7.12.: **The most likely geographical origin of the LAM ancestor genotype was found to be *Europe* by both Maximum Parsimony and Bayesian estimations.** Maximum likelihood tree as in Figure 7.11 and known patient place of birth were used as input for RASP software (Yu *et al.*, 2010). Colors of circles show patient place of birth of the corresponding MTB isolate (small circles) or the likely geographical region for the internal nodes (large circles). Colors are indicated in the legend and differ between panels. Transparent circles indicate undetermined character state. **A**. Reconstruction of ancestral character states (continent of origin) using the "Statistical Dispersal Vicariance Analysis" (S-DIVA) algorithm of RASP, a Maximum Parsimony approach. Combinations of continents were allowed for Northern, Central and South America (combinations of B, C and D). The LAM ancestor, indicated with an arrow, had a 100% probability of *Europe* as reconstructed ancestral state. **B**. Reconstruction of ancestral states (continent) using the Bayesian Binary Method, for which combinations of populations can not be given *a priori*, but are defined by the algorithm.

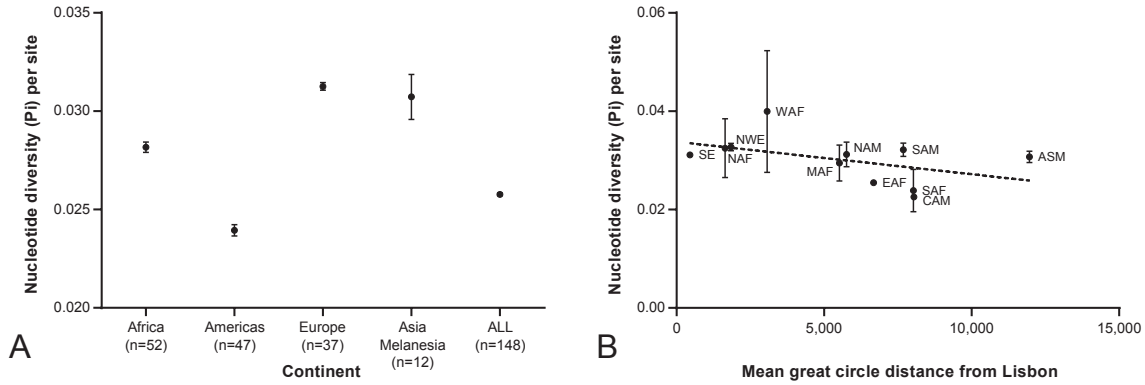


Figure 7.13.: **Genetic diversity of LAM isolates per continent and world region. A**. The genetic diversity of LAM isolates, measured as nucleotide diversity per site (π), was found to be highest in Europe and Asia/Melanesia. Mean nucleotide diversity was significantly different across all four continents (ANOVA, $p<0.0001$). Error bars indicate 95% confidence intervals. **B**. A regression coefficient of $R^2=0.22$ (dashed line, linear regression) indicates that the genetic diversity (π) decreases with geographical distance from Portugal, indicating an origin of the LAM genotype in Europe (largest diversity due to origin) and subsequent introduction to other continents (bottlenecks and clonal expansion leading to lower diversity). Abbreviations are: WAF, Western Africa; SE, Southern Europe; EAF, Eastern Africa; AM, Asia Melanesia; SAF, Southern Africa; CAM, Central America and Carribean; NAM, Northern America; SAM, South America; MAF, Middle Africa; NAF, Northern Africa; NWE, Northern, Western and Eastern Europe. The software DnaSP (Librado *et al.*, 2009) was used for nucleotide diversity calculations, using 10,120 of a total of 12,256 variable sites without gaps. Mean great circle distances were used for geographical distance. For each geographical region, we calculated the distance between Lisbon and the capitals of all countries of origin of the isolates, and used a mean value of the distance to be plotted against genetic diversity. Error bars indicate confidence intervals of the population mean.

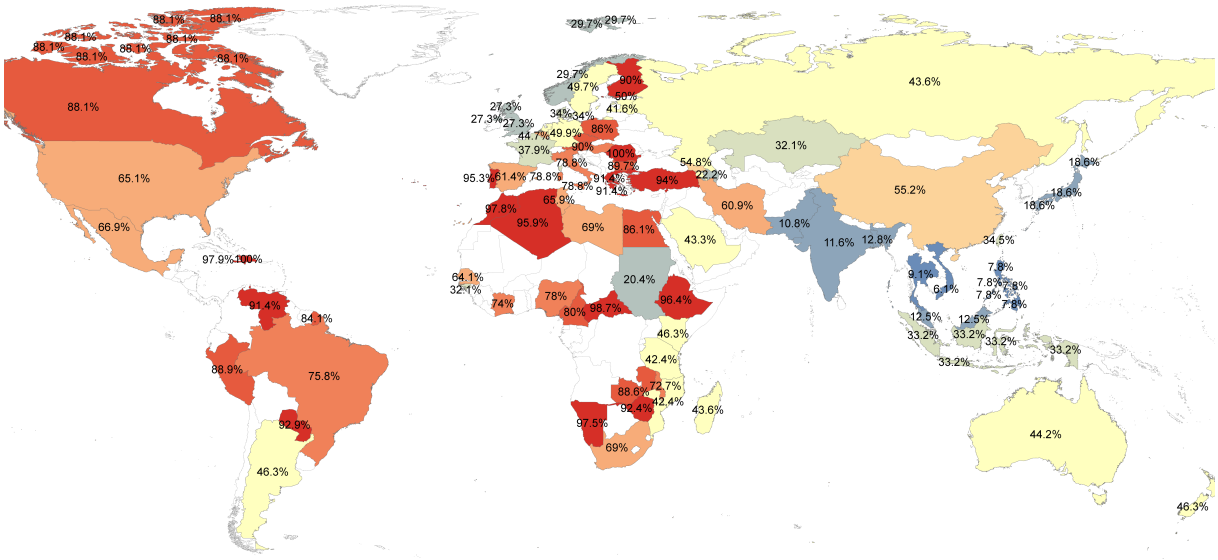


Figure 7.14.: **Proportion of Lineage 4 among all isolates (all lineages) shown by country.** Dark red color indicates a high proportion of MTB Lineage 4 among all isolates, blue color low proportion of Lineage 4. Data was extracted from SITVITWEB (Demay *et al.*, 2012) using 49,799 MTB strains with known country of isolation and known spoligotypes, as accessed on 22 April 2014. Percentages shown are Lineage 4 isolates among all isolates of the corresponding country dataset, including unknown or uninterpretable spoligotyping families. No data or less than 10 isolates is shown as white filled countries.

Western Europe showed a Lineage 4 prevalence of less than 50%, likely reflecting a large proportion of immigrants among TB cases, who were infected with a non-Lineage 4 strain from their country of origin (Gagneux *et al.*, 2006b).

7.5. Discussion

MTB Lineage 4, previously called the "Euro-American" lineage, is geographically the most widely distributed lineage of MTB. To better understand the distribution and evolution of MTB Lineage 4, we defined 10 sublineages and found a spatially structured global population with 5 locally restricted and 5 globally distributed sublineages. Remarkably, the three sublineages observed frequently in Europe were also frequent globally. Zooming into the most frequent and most widespread sublineage, the L4.5/LAM family of strains, we used 150 WGS and found phylogenetic evidence that the origin of the LAM family might have been in Europe. Additionally, the genetic diversity of LAM strains was found highest in Europe and decreasing with distance from Europe.

As a first in the MTB strain genotyping era, we were able to comprehensively define robust sublineages of MTB Lineage 4 based on whole genome data. These sublineages

included all major, previously reported definitions using SNPs and LSPs (Supplementary Figure 7.5), and also the most important spoligotyping families. Our definitions, though, were more complete and included sublineages not defined before. To the best of our knowledge, no SNP markers has been described to include all genotypes of our sublineages L4.4 and L4.5. We recommend that future typing schemes should consider the complete diversity of MTB Lineage 4, as data might otherwise suffer from discovery bias (Achtman, 2008). One objective of our study was to provide a framework for unambiguous MTB Lineage 4 strain classification.

Our SNP-screening of clinical isolates demonstrated that more than 95% of MTB Lineage 4 isolates can be categorized into one of our 10 sublineages. Furthermore, not a single isolate of totally 3,366 was found with two simultaneously mutant sublineage-specific alleles, once more underlining the clonal population structure of MTBC (Comas *et al.*, 2009b). However, we found 3% of isolates with all 10 sublineage-specific SNP position showing the ancestral allele. This indicated the presence of additional sublineages. Sequencing of a subset of these isolates identified them as taxa branching off ancestrally in the phylogeny rather than forming genetically distinct new sublineages. However, we only sequenced 17 of 92 (18.5%) of those isolates, and additional strains might indeed reveal new sublineages. The constantly growing phylogeny shows that sublineage definitions are not static and closing definitions, but working definitions that will change with increasing coverage of the diversity. The reduction of branch lengths by adding taxa questions the definitions of sublineages or "families" in the long run, and will push strain classification from a categorical to a more continuous, dynamic process revealed by, and manageable only using the high resolution of WGS.

We conclude that our iterative process of using WGS-defined clades and WGS-extracted markers for subsequent SNP-screening of large isolate collections, and the subsequent sequencing of uncategorized isolates, is a successful strategy to uncover genetic diversity in MTBC.

The global phylogeography of MTB Lineage 4 sublineages revealed a pronounced structure. Five sublineages were found in less than 20 of 100 countries included in the study, and were locally restricted. Three of these sublineages, L4.2/"Ghana", L4.8/"Uganda" and L4.9/"Cameroon", were found exclusively in sub-saharan Africa, except for anecdotal cases in immigration countries. This finding is particularly interesting in the light of a postulated European origin of MTB Lineage 4, and at least two alternative explanations could account for this geographic restriction. On the one hand, L4.2, L4.8 and L4.9 might have existed in Europe in low frequencies before, and were introduced by European ex-

plorations somewhere along the coasts of Africa likely after 1471 (Diffie *et al.*, 1977). A scenario of clonal and spatially restricted expansion in Africa and the simultaneous extinction in Europe would then explain today's distribution. It is tempting to speculate that these sublineages are less virulent than other Lineage 4 genotypes and were therefore extinct in Europe, but persisted in Africa as they were still more transmissible than *Mycobacterium africanum* (Jong *et al.*, 2010). *M. africanum* consists of two "ancient" lineages of MTBC (Lineage 5 and Lineage 6) and was associated with a longer latency period and a putative lower virulence as an adaption to low host-densities (Jong *et al.*, 2008; Gagneux, 2012). As an alternative explanation, if an ancestral Lineage 4 strain originated in Africa rather than Europe or the Middle East, these three sublineages would represent the "ancient" sublineages that never migrated out of Africa. However, the complete lack of these genotypes in the Americas despite the large population migrations from Africa to the Americas make this scenario unlikely. Furthermore, the three sublineages L4.2, L4.8 and L4.9 do not form a monophyletic group in the phylogeny.
Sublineages L4.4 and L4.7 were found in large proportions in Eastern Europe and Asia. These sublineages potentially represent "ancient" MTB Lineage 4 sublineages that diverged from a common MTB Lineage 4 ancestor preceding the introduction of MTB Lineage 4 in Europe. Consistent with this finding is the recent description of up to 45.9% non-"Beijing" isolates (but mainly MTB Lineage 4) in China (Li *et al.*, 2011), mainly in Western China. However, these two sublineages do not form a monophyletic group, and we therefore conclude that strains from at least two independent branching events of MTB Lineage 4 have migrated from a common ancestor eastwards. Consistently, the branching point of L4.4/"Ural" was found to be the most ancestral of all sublineages.

Sublineages L4.1 and L4.6 were inconclusive. Sublineage L4.1/"X" was found mainly in Northern America, but in smaller proportions also in South America, Africa and Asia. Interestingly, a "sub-sublineage" of L4.1 (RD183 deleted) has recently been found to be associated with genotypic clustering in San Francisco (Anderson *et al.*, 2013). This potential hypertransmissibility could be associated with host genetic factors and/or social factors. Isolates of L4.6 (including the "S" spoligotyping family) were observed in higher proportions in East and South-East Asia, partly in Africa and the Americas and almost absent from Europe. This also indicates a dispersal preceding the introduction into Europe.

Sublineages L4.3/"Haarlem", L4.5/"LAM" and L4.10/"PGG3" were found in more than 40 countries each, and had been described frequently by others (Durmaz *et al.*, 2007; Homolka *et al.*, 2008; Demay *et al.*, 2012; Rindi *et al.*, 2012; Traore *et al.*, 2012; Lopes

et al., 2013). This finding is consistent with a global distribution from *Europe*. The LAM sublineage was additionally present with a frequency of more than 30% among all MTB Lineage 4 strains in more than 50% of the countries in which it occurred, potentially reflecting hypertransmissibility or other factors contributing to its global success.

Concluding on the phylogeography of MTB Lineage 4 sublineage, we provide a solid basis for future studies looking at clinical and biological differences between sublineages of MTB Lineage 4, and for evolutionary studies. Lineage 4 was found globally prevalent despite its original name "Euro-American". However, the absence of at least five of the 10 sublineages from Europe speaks against the hypothesis that the complete MTB Lineage 4 dispersed from (Central/Southern) Europe. A more complex evolutionary scenario is therefore expected and needs further research using increased sample numbers especially from the Eastern Mediterranean and Central Asian regions.

We acknowledge that the proportions we calculated represent proportions among all MTB Lineage 4 isolates, not among all MTBC isolates (all lineages). The global frequencies of MTB Lineage 4 overall, however, indicated that Lineage 4 isolates are frequent throughout the world (Figure 7.14).

To investigate the evolution of MTB Lineage 4 in more detail, we focused on the L4.5/"LAM" sublineage. A total of 150 LAM whole genome sequences revealed again a genetically structured, but geographically less clearly separated population. Overall, we did not observe LAM isolates phylogenetically clustering by continent, except for the subgroup B1 (part of the RDRio group) that harboured exclusively isolates from subsaharan Africa. Such a grouping indicates a single introduction of an ancestral genotype in Africa and a subsequent clonal expansion. Interestingly, the RDRio genotype was otherwise described as a cosmopolitan genotype, with a predominance in Brazil (Dalla Costa *et al.*, 2013) and New York (Weisenberg *et al.*, 2012), but is also known to have caused cases of multi- and extensively drug resistant TB in South Africa in case of the KwaZuluNatal genotype (KZN) (Pillay *et al.*, 2007).

As expected, isolates from Europe were found in all subgroups of the LAM phylogeny, consistent with a global dissemination from Europe. However, we note that also isolates from other continents were found in all subgroups.

As a more powerful method, we used phylogenetic reconstruction of character states (continent of origin) and inferred Europe as the most likely continent for the geographical origin of the LAM-ancestor. When splitting Europe into Southern and Northern Europe,

the signal disappeared and other continents were found as more likely ancestral origin in the Bayesian approach. This might indicate that the spread of LAM strains was not exclusively originating from Southern Europe, but also from other (coastal) regions of North-Western Europe. The sample size of North-Western European isolates, however, was low (n=14) and most of the isolates were from Switzerland, likely not representative of North-Western European LAM isolates. Further studies are required to investigate the diversity of LAM strains, historically as well as contemporary, in Europe. The presence of LAM-related genotypes in Northern UK in the 13th to 14th centry AD was recently shown (Muller *et al.*, 2014). Also, reports of the genetically distinct LAM-RUS subfamily in Eastern Europe, Russia and Central Asia in relevant frequencies today (Mokrousov *et al.*, 2014) indicate that the LAM sublineage has been present in North-Eastern Europe and Central Asia for a long time.

Our data also showed that the genetic diversity of LAM strains is highest in Europe and decreases with the distance from Europe, as would be expected if the LAM genotype diversified for the longest time in Europe and was introduced to other continents going through population bottlenecks. However, the weak signal for diversification by distance stands in contrast with data for all seven main phylogenetic lineages, for which 77% of genetic variation was explained by distance from the likely cradle of the MTBC, Ethiopia (Comas et al., manuscript in preparation). Possibly, the L4.5/LAM genotype might represent a "generalist" that has been distributed in the last centuries as an "emerging" genotype, and as a consequence of globalization is found in a globally mixed population today. We speculate that MTB Lineage 4 was transmitted globally by Europeans before the "Beijing" genotype / MTB Lineage 2 spread with increasing migrations from Asia and a selective advantage associated with drug resistance, and has now become an emerging pathogen (Cowley *et al.*, 2008). Interestingly, MDR-TB cases were recently found at higher rates in both "LAM" and 'Beijing" strains than in other strains (Mokrousov *et al.*, 2012).

We hypothesize that the massive historic and contemporary migrations to and from Europe lead to a "cosmopolitan" population structure, and to MTB Lineage 4 strains also transmitting "allopatrically" on all continents (if "sympatry" was interpreted as MTB Lineage 4 strains infecting primarily Europeans and Americans). These migration patterns will likely continue to impact the MTB population structure. Historic samples from old strain collections, or ancient DNA, will allow inferring the historically prevalent genotypes in different parts of the world.

7.5.1. Limitations

We acknowledge that the definition of the term "sublineage" as well as the number of sublineages is arbitrary to a certain extent. We chose a level of discrimination that resulted in 10 sublineages, but could have defined only 5 sublineages by moving the dashed line in Figure 7.2 towards the root of the tree. We selected the level to be most compatible with existing genotyping schemes and known molecular markers. Therefore, sublineage markers from other studies using LSPs or SNPs are covered by our definitions (Figure 7.5), except for the group defined by SNP 4314642G>A that defines SNP cluster groups (SCG) 5 and 6.

A potential weakness of our study is the sample selection. We used, on the one hand, convenience samples from different countries, that might not be representative for the whole country. Additionally, particular countries were overrepresented (Figure 7.6). On the other hand, many countries were not included in this study. Also, we used immigration collections from Switzerland and the US. Place of birth of TB patients has in several studies proven to be a valid proxy for place of infection (Baker *et al.*, 2004; Gagneux *et al.*, 2006b; Fenner *et al.*, 2013), but we can not exclude that patients were infected with a circulating strain of their country of immigration. To be as conservative as possible, we excluded all strains previously identified as part of a transmission cluster from the LAM analysis. Overall, we are confident that the signal that we observe for the MTB Lineage 4 strains population structure would rather become stronger with a population-based sample. Obtaining globally representative collections of MTB (Lineage 4) isolates from 100 countries, as in our study, would be very difficult, hence our approach. One region that future studies should particularly seek strains from is the Eastern Mediterranean and Central Asia, which were underrepresented in our collection. These regions could be key to study the phylogeography of the sublineages absent from Europe and likely having diverged from a MTB Lineage 4 ancestor in that region.

We used "country" as a geographical origin rather than geographical coordinates, which means, in the case of large countries, we might have lost signal due to the prevalence of different predominant genotypes in different regions (such as Southern versus Northern Brazil). However, obtaining geographical coordinates was nearly impossible for the large and diverse sample collection we used, and the more detailed geographical information should rather increase the signal.

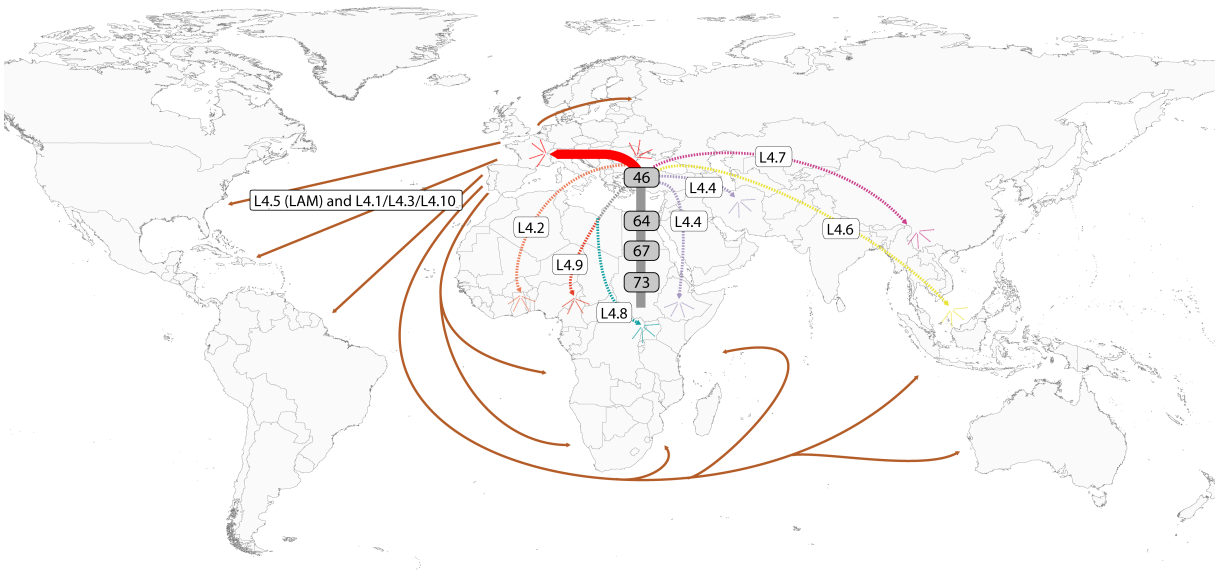


Figure 7.15.: **Our findings support a scenario in which the "LAM" family of MTB Lineage 4 was dispersed to other continents from Europe, most likely starting with the age of discovery in the 15th century.** At least two other sublineages, L4.3 and L4.10, might have followed the same path. Numbers in grey boxes indicate the time of coalescence of MTBC lineages as proposed by Comas *et al.* (2013).

7.5.2. Conclusions

With this study, we provide the first comprehensive MTB Lineage 4 classification scheme based on a large collection of whole genome sequences. We recommend to consider the full genetic diversity of MTB Lineage 4 and to apply robust markers for future studies.

Applying the sublineage-screening to a large clinical collection, our data shows that strains of MTB Lineage 4 are globally present, but with varying proportions of the 10 sublineages between geographical regions. The phylogeographic structure and the absence of five sublineages from Europe indicate that the evolution of MTB Lineage 4 did not follow a simple "out-of-Europe" scenario, but is more complex. A proposed scenario is indicated in Figure 7.15. Several sublineages are restricted to Asian and African regions and might represent "ancient" genotypes that emerged in the Eastern Mediterranean / Central Asian region. Strikingly, the three sublineages observed as frequent in Europe were also observed as globally frequent.

The most frequent MTB Lineage 4 sublineage, the L4.5/"LAM" family, was found to be globally distributed and at a high proportion in many countries. The results of the phylogenetic reconstruction, as well as the elevated genetic diversity in Europe, support our initial hypothesis that the L4.5/"LAM" sublineage (potentially representative also for L4.3 and L4.10) has expanded dramatically with human populations in Europe and was

dispersed to other continents with European migrations. Further studies, including the dating of the phylogeny, ideally with ancient DNA, are necessary to gain more detailed insights in the evolution of the globally important MTB Lineage 4. Clearly, the designation "Euro-American" should to be abandoned.

8. General discussion

The completed genome sequence of H37Rv in 1998 (Cole *et al.*, 1998) has revolutionized TB research. It was not until 2010, however, that 21 whole genome sequences and the first WGS-based phylogeny of the MTBC lead to the appreciation of the extent of genetic diversity of MTBC (Comas *et al.*, 2010). Nearly 10,000 SNPs were discovered with these 21 genomes. In the last four years, more than 3,000 (draft) genome sequences have been generated (Stucki *et al.*, 2013). For the first time, we are now able to obtain complete sets of phylogenetically informative, clade-specific SNPs from comprehensive WGS data. With these molecular markers, we can screen large collections of MTBC isolates.

The aim of this study was to develop and apply combinations of SNP-screening and targeted whole genome sequencing to delineate micro- and macroevolutionary events in a transmission cluster and on the global diversity scale. We have successfully developed two laboratory and one *in silico* SNP-typing assays. With these assays, we provide a robust framework for MTBC strain classification. Using a combination of WGS and SNP-genotyping, we were able track the transmission of a single genotype causing a TB outbreak in Bern, Switzerland, over two decades. This was possible in a time and cost frame that would not have been possible with classical genotyping assays. Expanding our perspective from micro-evolution to macro-evolution, we were able to delineate the historical origin and map the likely spread of MTB Lineage 4 from Europe to the rest of the world. Both the characterization of a TB outbreak as well as the evolutionary reconstruction of the particularly successful MTBC Lineage 4 highlight the potential that the "era of genomics" can bring to TB research and control.

In this chapter, we summarize the main findings and discuss the future applicability of WGS and SNP-typing in molecular epidemiology and TB research.

8.1. The role of SNP-typing and whole genome sequencing

Obtaining genomic information has become a daily task in basic and clinical TB research. In this study, we present different approaches to extract SNP data directly from DNA in

the laboratory, and from WGS data. These approaches show the complementary nature but also the rapid evolution of methods.

In 2010, when this study started, WGS was available only for selected isolates and was still expensive. First studies using WGS for molecular epidemiological purposes were just being published (Schürch *et al.*, 2010; Gardy *et al.*, 2011). Multiplexed, barcoded libraries only became available during the course of the study, and sequencing prices decreased with the roll-out of new DNA sequencing devices. While SNP-genotyping had been expected to be widely used, WGS has gained strong *momentum* and is now in reach for routine application on complete MTBC collections. WGS will likely replace all other genotyping methods in the future, as molecular epidemiological as well as drug resistance and phylogenetic information can be easily extracted. This "ultimate barcode" of a clinical isolate is of great value and will allow comparability and exchange across laboratories. Nevertheless, the challenges of implementing cutting-edge technologies, particularly in resource-limited settings, where control of TB is most urgent, should not be underestimated. Until robust, Xpert-like devices become available (ideally battery-powered) that can be operated by any lab personnel (Lawn *et al.*, 2013), established genotyping methods including SNP-typing will continue to play a major role. Furthermore, the data analysis and storage capacities are often limited, meaning that useful information can not be obtained quickly (although tools are now becoming available for simplified analysis, such as the software `KvarQ` described in Chapter 5 of this thesis.) SNP-genotyping, on the other hand, is - depending on the protocol - very robust, and can be implemented in any laboratory that has access to a PCR machine (Hillemann *et al.*, 2005).

Almost four years after the beginning of this PhD thesis, the price of WGS per MTBC genome has massively decreased (`http://www.genome.gov/images/content/cost_per_megabase.jpg`), and is now almost comparable to standard 24-loci MIRU-VNTR analysis (MIRU commercially available for approximately EUR 50; MiSeq libary and sequencing costs of around EUR 100). First public health authorities, particularly in the U.K., are starting to implement WGS as a close-to-routine method (Walker *et al.*, 2014). Despite the great value added, *targeted* WGS with previous screening of informative SNPs can be more cost-effective and faster, also in high-income countries. Likewise, in our last two chapters, we show that SNP-typing and WGS are perfectly complementary.

8.2. Technical and analytical limitations of WGS

From a technical aspect, there are several limitations in our study. WGS of MTBC isolates was performed with different generations of Illumina devices. We obtained sequencing

read lengths of between 36 and 250 bp. This range of read length is in general too short to extract repetitive regions such as the PE/PPE/PGRS genes or insertion sequences. Therefore, we excluded approximately 10% of the genome, corresponding to potentially 10% of SNPs that remain undetected. Considering a potential excess of diversity in these genes (Copin *et al.*, 2014; McEvoy *et al.*, 2012), the actual number of mutations in these regions could be higher than expected. Ten percent of SNPs might be less relevant for phylogenetic purposes such as the ones described in Chapter 7. But the improved resolution by (potential) additional SNPs might allow discriminating the so far indistinguishable 25% of the outbreak isolates described in Chapter 6. Furthermore, longer sequencing reads would also allow the extraction of an *in silico* MIRU-VNTR pattern with e.g. the `KvarQ` software, allowing the comparison with existing MIRU-VNTR data and databases. A first approach of extracting MIRU-pattern has been made using read depth and paired-end mapping (Ragheb *et al.*, 2013), but needs further validation for clinical isolates (an *in vivo* mutation model with MTBC Erdman strain was used in that study). Longer reads might overcome the need for coverage-based MIRU-extraction, but reads will need to be at least 500 bp long (each MIRU repeat unit is between 51 and 111 bp long (Supply *et al.*, 2006)). Sequencing technologies achieving these lengths are available, albeit not cost-effective for small prokaryotic genomes. On the other hand, future sequencing platforms such as Nanopore sequencing promise to deliver longer reads (Loman *et al.*, 2012).

Another source of potentially missed diversity, i.e. molecular resolution, are insertions and deletions (InDels). These were so far ignored due to technical difficulties in analysing such genomic regions (Albers *et al.*, 2011). Although 90% of the diversity in MTBC is likely represented by SNPs (Comas *et al.*, 2010), InDels are another important source of genomic plasticity.

The third analytical difficulty are SNPs that had to be excluded due to "heterozygous" SNP calls, i.e. positions with more than two alleles (e.g. reference and at least two alternative alleles). These calls can occur due to the presence of two or more alleles in the genomic DNA (either a microevolutionary event or a "true" mixed infection with different strains; sequencing errors or alleles occurring with a low frequency of only few reads are ignored by the SNP-calling software). On the other hand, calls with ambiguous allele frequencies ("AF1" field between 0 and 1 in the variant call format (VCF) file), were excluded in our pipeline if a polymorphisms at that genomic position occurred only in one strain. Again, this is likely negligible for phylogenetic purposes as described in Chapter 7, but potentially harbouring additional information for transmission detection. In the "Bern outbreak" dataset, one ambiguous position in one isolate was detected (as another isolate

harboured the same mutation, otherwise it would have been missed) and confirmed by subsequent Sanger sequencing to be an additional allele rather than a sequencing artefact.

These limitations also highlight the need for standardized and validated WGS data analysis pipelines, as well as guidelines on how to generate WGS data. First guidelines to ensure a high WGS data quality have been published by the Centre of Disease Control (CDC) (Gargis *et al.*, 2012), but these now need to be implemented. This is particularly important with regards to future clinical assays using WGS, where software packages and variant calling/filtering criteria must be clearly defined. Criteria to filter-out false-positive SNP-calls lead to a trade-off between excluding a portion of true-positive calls (by keeping only high-confidence calls) and keeping false-positive calls, and thresholds should be validated before analysis. In the "Bern outbreak" dataset, 37 polymorphic positions were excluded due to low quality or gaps, but these could potentially harbour additional information. Manual confirmation of such positions as well as further research regarding analysis software is therefore necessary.

An upcoming discussion for the near future will be the comparability of WGS data between laboratories. How can a particular strain (e.g. the outbreak strain described in Chapter 6 be identified by other researchers in their database (the key patient was known to have travelled across Europe)? Current data compatibility would likely require MIRU-VNTR or spoligotyping data, but WGS data as of today can often not easily be reconciled. Recently, WGS-adapted multi-locus sequence typing schemes have been developed for bacteria in general (Jolley *et al.*, 2010; Jolley *et al.*, 2012), and as a "core genome multi-locus sequence type" (cgMLST) approach for MTBC (Kohl *et al.*, 2014), which extracts allelic information for 3,041 genes and transfers the SNP diversity into an allele numbering system, which can easily be exchanged.

8.3. Future improvements of WGS

In addition to all analytical challenges, innovations are needed on the sequencing device side, if WGS is to be used widely, particularly in countries with a large TB burden. On the one hand, devices need to be smaller and more robust, thereby avoiding the need for highly equipped laboratories and highly skilled personnel. The Nanopore sequencing platform is a promising platform (Gut, 2013; Ying *et al.*, 2013; Liang *et al.*, 2014). On the other hand, library preparation must be simplified (Tan *et al.*, 2013), which could be implemented in a lab-on-a-chip device (Sackmann *et al.*, 2014). More down-to-earth, a decrease in sequencing costs is still necessary. Indeed, sequencing prices for a mega-basepair (Mbp) of (human) DNA have fallen to around 0.1 $ (around 1000 $ per human

genome; `http://www.genome.gov/images/content/cost_per_megabase.jpg`), but due to the large data redundancy and the low levels of multiplexing currently achieved for bacteria, the corresponding costs for MTBC sequencing per Mbp are still orders of magnitude higher. Only recently, smaller benchtop devices have become available that allow more flexible and cost-effective sequencing of small genomes.

Last, for future studies it will be desirable to perform WGS directly from sputum samples to investigate the within-host diversity, ideally in the form of single-cell sequencing (Blainey, 2013; Gole *et al.*, 2013; Lasken, 2012).

8.4. Micro-evolutionary aspects

The retrospective identification of a TB outbreak in a large, population-based study with more than 1600 isolates collected over two decades was for the first time possible in such a short time and at such a low cost (around 3 hours for 96 isolates; reagent costs of $ 4800 including the genome sequencing; Chapter 6). In the current era of transition between classical genotyping methods and routine WGS, we have shown that a combination of both (SNP-typing as the "classical method") can be most appropriate and successfully applied to identify outbreak isolates. Realizing that the outbreak continued to propagate mainly in the same population of substance abusers and homeless, we emphasize the need for continuous intervention in that population, particularly to identify the patients most likely to cause secondary cases, i.e. the so-called "super spreaders". In the Canton of Bern, further outbreaks have been described already in 1993 (Genewein *et al.*, 1993), but never investigated. We note that the only study so far that has *compared* TB clusters by WGS (Walker *et al.*, 2013b) has identifified different topologies of genomic networks, depending on the setting and the population. These varied from a star-like topology in cluster 5 to a more linear topology in cluster 7. WGS of all clustered isolates in the "Bernese outbreak" between 1991 and 2011 might reveal similar differences. By comparing patient characteristics between genomic clusters of different topologies, the question could be addressed if only social factors or also biological differences between MTBC strains contribute to the different transmission patterns. From a public health perspective, the topology could potentially inform public health staff where an targeted intervention is most urgent, if e.g. if a "super spreader" can be identified in a star-like topology.

It is interesting to note that most of the previous reports studying TB outbreaks with WGS have reported "super-spreader" topologies (Roetzer *et al.*, 2013) (Walker *et al.*, 2013b) (Gardy *et al.*, 2011). Super-spreader behaviour has been identified by social contact tracing (McElroy *et al.*, 2003), but has not been possible with classical molecular

epidemiological methods, urging for the need to implement prospective WGS-screening of TB cases (Walker *et al.*, 2013b).

Also, the described oubreak needs to be followed up, as new cases might still appear many years later (the last case associated with the Bern outbreak described in Chapter 6 was identified in 2011). We hypothesized that the isolates post-dating 1998 resulted from re-activation of infection during 1991 and 1995. Assuming that the 13 cases after 1998 (one child excluded) represent only the 10% of infected persons ever progressing to active TB in their lifetime (Barry *et al.*, 2009) (but potentially more in this highly susceptible population), we would except a higher number of *latently* infected persons in the population, and TB reactivation cases could potentially develop in the years to come. From a research point of view, these re-activation cases might also be a unique opportunity to study micro-evolutionary events in the dormant stage of MTBC. It has recently been shown that MTBC mutation rates in latently infected macaques were comparable to MTBC during active disease (Ford *et al.*, 2011), but no study was so far able to investigate micro-evolutionary events in humans rather than in animal models.
Interestingly, looking at the "Bernese outbreak cluster" from a macro-evolutionary perspective, on a side note, the causative MTB strain belongs to the L4.6 sublineage of MTB Lineage 4 ("S" spoligotype). In Chapter 7, this sublineage was observed mainly in patients from Asian countries. An early transmission event from an Asian immigrant to the first patient in the outbreak (pre-dating 1987) would explain the "allopatric" sublineage.

8.5. Macro-evolutionary aspects

Using a total of 239 whole genome sequences and SNP-genotyping data of 3,366 clinical MTBC isolates, we showed for the first time that the previously called "Euro-American" Lineage 4 is globally dispersed and more diverse than previously acknowledged, and that the geographic origin of at least 3 sublineages with Lineage 4 was most likely Europe. In a broader context, the distribution from Europe would be compatible with an origin of MTBC in Africa (Wirth *et al.*, 2008; Comas *et al.*, 2013) and a subsequent migration via Levant into Europe. The detailed migratory paths of humans and the MTBC to Europe, however, are complex and need further investigation. Presumably, MTB Lineage 4 could have paralleled the evolutionary pathway described for the Lineage 2 ("East Asian" lineage, including the "Beijing" sublineage), with a population increase during the Neolithic transition, which started around 10,000 years ago in Central China as well as in the Fertile Crescent (Diamond *et al.*, 2003).

Despite the focus on the "Beijing" family and its association with drug resistance, HIV co-infection and hypervirulence (e.g. De Beer *et al.*, 2014; Duong *et al.*, 2009; Fenner *et al.*, 2012a; Middelkoop *et al.*, 2009; Klopper *et al.*, 2013), future studies should also explore the phenotypic characteristics of MTB Lineage 4 and its sublineages. This was previously hindered by a confusion of genotyping families, which were based on suboptimal markers. Indeed, with this study, we do not only provide an evolutionary scenario of MTB Lineage 4, but also a robust classification scheme (including SNPs as ideal markers) for unambiguous classification of MTB Lineage 4 isolates. An important study by Rindi *et al.* (2012) had already shown that spoligotyping families are not always concordant with RD-defined sublineages due to the large degree of homoplasy, and recommended therefore not to use spoligotyping data for MTB Lineage 4 classifications. And although RD definitions are phylogenetically robust, we have now shown that the RD-defined sublineages do not represent the complete phylogenetic diversity of Lineage 4.

8.6. Evolutionary dating

For both micro- and macro-evolutionary aspects (Chapters 6 and 7), the dating of the phylogeny might reveal the time of the coalescent points of subclusters and sublineages. The feasibility of dating the origin of an outbreak has recently been shown using tip-dates (isolation dates) (Roetzer *et al.*, 2013). However, we would like to note that in the "Bernese outbreak cluster" (chapter 6), the numbers of SNPs between isolates was very low. Single SNPs might skew the estimated mutation rates significantly, and thus, the complete *within-patient* diversity, as discussed above, needs to be taken into account. Furthermore, culture-derived mutations also need to be excluded. Ideally, sequencing should be applied to the bulk MTBC population isolated from the patient sputum sample.

Dating the MTB Lineage 4 phylogeny could be done with an estimated date of an internal node, as recently shown with the complete MTBC phylogeny (Comas *et al.*, 2013) and the coalescent time of the human L3 mitochondrial haplogroup 70,000 years ago. Three alternative dates were used and scenarios compared. The scenario of 70,000 years was consistent with an expansion of the effective population size during the Neolithic transition starting 10,000 years ago. Such an approach might be difficult to apply to a single MTBC lineage, as the time of emergence of MTB Lineage 4 might be more difficult to link to a particular event in human evolutionary history. More promising would be to use whole genome data from ancient DNA (aDNA). Two studies have reported WGS data from MTBC aDNA recently, but both date back only to the late 19th century (St.

George's Church Crypt, England) and 1797 (Vàc, Hungary) (Bouwman *et al.*, 2012; Chan *et al.*, 2013). Assuming 46,000 years since the emergence of MTB Lineage 4, 100 to 200 years will likely not lead to an accurate estimation of mutation rates. We are confident, however, that in the future more studies using aDNA will reveal sequencing data from older specimens. Clearly, the combination of WGS data from contemporary isolates with data from historical isolates has a great potential to unravel the evolutionary history of MTBC.

8.7. Public health relevance

With this study, we provide molecular markers and assays for the classification of MTBC isolates. These low-cost assays can be implemented in any laboratory in high-burden as well as low-burden TB settings to study phenotypic and epidemiological differences between MTBC lineages and sublineages. Such studies can potentially reveal MTBC strain differences that will be relevant for TB product design and treatment.

With `KvarQ`, we reduce the gap between the massive amounts of WGS data that are being generated, and the lack of analysis capacity. This is today mainly of interest for research applications, but if WGS becomes a routine "genotyping" method in the near future, any lab personnel will be able to extract e.g. genotypic drug resistance information with `KvarQ`.

WGS in general will likely change our understanding of TB transmission in general. Our retrospective analysis of a TB outbreak in Switzerland has shown that an epidemiologically suspected "super spreader" can be confirmed with WGS data. This was not possible with previous genotyping methods. The identification of super-spreaders is most crucial for public health responsibles in high-income countries, as these persons are potentially causing a large number of clustered cases. Furthermore, if a "super spreader" person shall be isolated (as a public health intervention), strong molecular evidence will help supporting the action.

As long as routine WGS of *all* isolates is not practicable and not affordable, the SNP-screening can inform public health staff which isolates should be subjected to WGS, as shown in the case of an outbreak. Indeed, in most settings, routine (classical) molecular typing of all MTBC isolates is not performed so far, and WGS will likely also not be an option. There, epidemiologically suspected outbreaks could be tracked by an approach as we have described it in Chapter 6.

The role of WGS in high-burden TB countries is so far unclear. On the one hand, the costs of WGS are high, and the analytical capacities limited. On the other hand, it is also

unclear if - with a higher force of infection - WGS will still provide enough discriminatory power to resolve transmission chains. Studies are therefore needed to evaluate the benefit of WGS in these countries, where the burden of TB is highest and most cases could be prevented. In any case, only a selected number of isolates can probably be sequenced.

Our study contributes a SNP-screening approach to rationally narrow down the number of isolates to be sequenced (e.g. only clusters), and that approach is equally suited for high-burden as well as low-burden settings.

8.8. Conclusions

TB remains a major public health concern. The diversity of the pathogen and its diversity is one aspect of TB control. The first part of this thesis has provided new markers and tools for robust strain classification. SNP-typing allows low-cost genotypic classification in virtually any laboratory and can nicely complement targeted WGS, which is so far only available in high-income countries.

The second part of this thesis indicated the need for screening efforts and targeted interventions for TB control in particular populations, as shown with the example of a low TB-burden country such as Switzerland. The approach of combined SNP-screening and WGS applied here can be implemented at larger scales to identify outbreaks in real-time. Compared to classical molecular markers that suffered from a lack of discrimination power, WGS was able to resolve a majority of isolates.

The third part of this thesis has found an "out-of-Europe" evolutionary scenario of Lineage 4 of the MTBC. We found substantial genetic information within Lineage 4 and also within one sublineage, the "LAM" family. However, Lineage 4 is only one lineage of the MTBC, and the focus on other lineages might reveal similarly interesting evolutionary pathways of this important pathogen.

9. Bibliography

Abadia E, Zhang J, Vultos T dos, Ritacco V, Kremer K, Aktas E, Matsumoto T, Refregier G, Soolingen D van, Gicquel B, Sola C (2010) Resolving lineage assignation on *Mycobacterium tuberculosis* clinical isolates classified by spoligotyping with a new high-throughput 3R SNPs based method. *Infection, Genetics and Evolution*, **10**(7): 1066–1074.

Abubakar I, Stagg HR, Cohen T, Mangtani P, Rodrigues LC, Pimpin L, Watson JM, Squire SB, Zumla A (2012) Controversies and unresolved issues in tuberculosis prevention and control: a low-burden-country perspective. *The Journal of Infectious Diseases*, **205** S293–300.

Achtman M (2008) Evolution, population structure, and phylogeography of genetically monomorphic bacterial pathogens. *Annual Review of Microbiology*, **62** 53–70.

Achtman M (2012) Insights from Genomic Comparisons of Genetically Monomorphic Bacterial Pathogens. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **367**(1590): 860–867.

Aderem A, Adkins JN, Ansong C, Galagan J, Kaiser S, Korth MJ, Law GL, McDermott JG, Proll SC, Rosenberger C, Schoolnik G, Katze MG (2011) A systems biology approach to infectious disease research: innovating the pathogen-host research paradigm. *mBio*, **2**(1): e00325–00310.

Albers CA, Lunter G, MacArthur DG, McVean G, Ouwehand WH, Durbin R (2011) Dindel: accurate indel calls from short-read data. *Genome Research*, **21**(6): 961–973.

Alexander KA, Laver PN, Michel AL, Williams M, Helden PD van, Warren RM, Gey van Pittius NC (2010) Novel *Mycobacterium tuberculosis* Complex Pathogen, *M. mungi*. *Emerging Infectious Diseases*, **16**(8): 1296–1299.

Alland D, Whittam TS, Murray MB, Cave MD, Hazbon MH, Dix K, Kokoris M, Duesterhoeft A, Eisen JA, Fraser CM, Fleischmann RD (2003) Modeling bacterial evolution with comparative-genome-based marker systems: application to *Mycobacterium tuberculosis* evolution and pathogenesis. *Journal of Bacteriology*, **185**(11): 3392–3399.

Allix-Béguec C, Harmsen D, Weniger T, Supply P, Niemann S (2008) Evaluation and strategy for use of MIRU-VNTRplus, a multifunctional database for online analysis of

genotyping data and phylogenetic identification of *Mycobacterium tuberculosis* complex isolates. *Journal of Clinical Microbiology*, **46**(8): 2692–2699.

Anderson J, Jarlsberg LG, Grindsdale J, Osmond D, Kawamura M, Hopewell PC, Kato-Maeda M (2013) Sublineages of lineage 4 (Euro-American) *Mycobacterium tuberculosis* differ in genotypic clustering. *The International Journal of Tuberculosis and Lung Disease*, **17**(7): 885–891.

Anderson LF, Tamne S, Brown T, Watson JP, Mullarkey C, Zenner D, Abubakar I (2014) Transmission of multidrug-resistant tuberculosis in the UK: a cross-sectional molecular and epidemiological study of clustering and contact tracing. *The Lancet Infectious Diseases*, **14**(5): 406–415.

Ani A, Bruvik T, Okoh Y, Agaba P, Agbaji O, Idoko J, Dahle U (2010) Genetic diversity of *Mycobacterium tuberculosis* Complex in Jos, Nigeria. *BMC Infectious Diseases*, **10**(1): 189.

Ansorge WJ (2009) Next-generation DNA sequencing techniques. *New Biotechnology*, **25**(4): 195–203.

Arnvig KB, Comas I, Thomson NR, Houghton J, Boshoff HI, Croucher NJ, Rose G, Perkins TT, Parkhill J, Dougan G, Young DB (2011) Sequence-Based Analysis Uncovers an Abundance of Non-Coding RNA in the Total Transcriptome of *Mycobacterium tuberculosis*. *PLoS Pathogens*, **7**(11): e1002342.

Arriaza BT, Salo W, Aufderheide AC, Holcomb TA (1995) Pre-Columbian tuberculosis in Northern Chile: Molecular and skeletal evidence. *American Journal of Physical Anthropology*, **98**(1): 37–45.

Asghar RJ, Patlan DE, Miner MC, Rhodes HD, Solages A, Katz DJ, Beall DS, Ijaz K, Oeltmann JE (2009) Limited utility of name-based tuberculosis contact investigations among persons using illicit drugs: results of an outbreak investigation. *Journal of Urban Health*, **86**(5): 776–780.

Baker L, Brown T, Maiden MC, Drobniewski F (2004) Silent nucleotide polymorphisms and a phylogeny for *Mycobacterium tuberculosis*. *Emerging Infectious Diseases*, **10**(9): 1568–1577.

Bamrah S, Yelk Woodruff RS, Powell K, Ghosh S, Kammerer JS, Haddad MB (2013) Tuberculosis among the homeless, United States, 1994-2010. *The International Journal of Tuberculosis and Lung Disease*, **17**(11): 1414–1419.

Barnes PF, Cave MD (2003) Molecular epidemiology of tuberculosis. *New England Journal of Medicine*, **349**(12): 1149–1156.

Barr J, Countryman E (2014) *Contested Spaces of Early America*. Philadelphia: University of Pennsylvania Press, 426.

Barry CE, Boshoff HI, Dartois V, Dick T, Ehrt S, Flynn J, Schnappinger D, Wilkinson RJ, Young D (2009) The spectrum of latent tuberculosis: rethinking the biology and intervention strategies. *Nature Reviews Microbiology*, **7**(12): 845–855.

Behr MA (2001) Comparative genomics of BCG vaccines. *Tuberculosis (Edinburgh, Scotland)*, **81**(1-2): 165–168.

Ben Shabat M, Mikula I, Gerchman I, Lysnyansky I (2010) Development and evaluation of a novel single-nucleotide-polymorphism real-time PCR assay for rapid detection of fluoroquinolone-resistant Mycoplasma bovis. *Journal of Clinical Microbiology*, **48**(8): 2909–2915.

Bentley SD, Comas I, Bryant JM, Walker D, Smith NH, Harris SR, Thurston S, Gagneux S, Wood J, Antonio M, Quail MA, Gehre F, Adegbola RA, Parkhill J, Jong BC de (2012) The genome of Mycobacterium Africanum West African 2 reveals a lineage-specific locus and genome erosion common to the *M. tuberculosis* complex. *PLoS Neglected Tropical Diseases*, **6**(2): e1552.

Bergval IL, Vijzelaar RNCP, Dalla Costa ER, Schuitema ARJ, Oskam L, Kritski AL, Klatser PR, Anthony RM (2008) Development of multiplex assay for rapid characterization of *Mycobacterium tuberculosis*. *Journal of Clinical Microbiology*, **46**(2): 689–699.

Bergval I, Sengstake S, Brankova N, Levterova V, Abadía E, Tadumaze N, Bablishvili N, Akhalaia M, Tuin K, Schuitema A, Panaiotov S, Bachiyska E, Kantardjiev T, Zwaan R de, Schürch A, Soolingen D van, van 't Hoog A, Cobelens F, Aspindzelashvili R, Sola C, Klatser P, Anthony R (2012) Combined Species Identification, Genotyping, and Drug Resistance Detection of *Mycobacterium tuberculosis* Cultures by MLPA on a Bead-Based Array. *PLoS ONE*, **7**(8): e43240.

Bertelli C, Greub G (2013) Rapid bacterial genome sequencing: methods and applications in clinical microbiology. *Clinical Microbiology and Infection*, **19**(9): 803–813.

Bharti R, Das R, Sharma P, Katoch K, Bhattacharya A (2012) MTCID: a database of genetic polymorphisms in clinical isolates of *Mycobacterium tuberculosis*. *Tuberculosis (Edinburgh, Scotland)*, **92**(2): 166–172.

Blainey PC (2013) The future is now: single-cell genomics of bacteria and archaea. *FEMS Microbiology Reviews*, **37**(3):

Boehme CC, Nabeta P, Hillemann D, Nicol MP, Shenai S, Krapp F, Allen J, Tahirli R, Blakemore R, Rustomjee R, Milovic A, Jones M, O'Brien SM, Persing DH, Ruesch-Gerdes S, Gotuzzo E, Rodrigues C, Alland D, Perkins MD (2010) Rapid Molecular Detection of Tuberculosis and Rifampin Resistance. *New England Journal of Medicine*, **363**(11): 1005–1015.

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)*, **btu170** 1–7.

Borrell S, Gagneux S (2009) Infectiousness, reproductive fitness and evolution of drug-resistant *Mycobacterium tuberculosis*. *The International Journal of Tuberculosis and Lung Disease*, **13**(12): 1456–1466.

Borrell S, Gagneux S (2011) Strain diversity, epistasis and the evolution of drug resistance in *Mycobacterium tuberculosis*. *Clinical Microbiology and Infection*, **17**(6): 815–820.

Bouakaze C, Keyser C, Gonzalez A, Sougakoff W, Veziris N, Dabernat H, Jaulhac B, Ludes B (2011) Matrix-assisted laser desorption ionization-time of flight mass spectrometry-based single nucleotide polymorphism genotyping assay using iPLEX gold technology for identification of *Mycobacterium tuberculosis* complex species and lineages. *Journal of Clinical Microbiology*, **49**(9): 3292–3299.

Bouakaze C, Keyser C, Martino SJ de, Sougakoff W, Veziris N, Dabernat H, Ludes B (2010) Identification and genotyping of *Mycobacterium tuberculosis* complex species by use of a SNaPshot Minisequencing-based assay. *Journal of Clinical Microbiology*, **48**(5): 1758–1766.

Bouwman AS, Kennedy SL, Müller R, Stephens RH, Holst M, Caffell AC, Roberts CA, Brown TA (2012) Genotype of a historic strain of *Mycobacterium tuberculosis*. *Proceedings of the National Academy of Sciences*, **109**(45): 18511–18516.

Branco M (2013) Bridging genomics technology and biology. *Genome Biology*, **14**(10): 312.

Breitling R (2010) What is Systems Biology? *Frontiers in Physiology*, **1** 9.

Brosch R, Gordon SV, Marmiesse M, Brodin P, Buchrieser C, Eiglmeier K, Garnier T, Gutierrez C, Hewinson G, Kremer K, Parsons LM, Pym AS, Samper S, Soolingen D van, Cole ST (2002) A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. *Proceedings of the National Academy of Sciences*, **99**(6): 3684–3689.

Brudey K, Driscoll J, Rigouts L, Prodinger W, Gori A, Al-Hajoj S, Allix C, Aristimuno L, Arora J, Baumanis V, Binder L, Cafrune P, Cataldi A, Cheong S, Diel R, Ellermeier C, Evans J, Fauville-Dufaux M, Ferdinand S, Viedma D de, Garzelli C, Gazzola L, Gomes H, Guttierez MC, Hawkey P, Helden P van, Kadival G, Kreiswirth B, Kremer K, Kubin M, Kulkarni S, Liens B, Lillebaek T, Ly H, Martin C, Martin C, Mokrousov I, Narvskaia O, Ngeow Y, Naumann L, Niemann S, Parwati I, Rahim Z, Rasolofo-Razanamparany V, Rasolonavalona T, Rossetti ML, Rusch-Gerdes S, Sajduda A, Samper S, Shemyakin I, Singh U, Somoskovi A, Skuce R, Soolingen D van, Streicher E, Suffys P, Tortoli E, Tracevska T, Vincent V, Victor T, Warren R, Yap S, Zaman K, Portaels F, Rastogi N, Sola C (2006) *Mycobacterium tuberculosis* complex genetic diversity: mining the fourth

international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology. *BMC Microbiology*, **6**(1): 23.

Bryant JM, Harris SR, Parkhill J, Dawson R, Diacon AH, Helden P van, Pym A, Mahayiddin AA, Chuchottaworn C, Sanne IM, Louw C, Boeree MJ, Hoelscher M, McHugh TD, Bateson ALC, Hunt RD, Mwaigwisya S, Wright L, Gillespie SH, Bentley SD (2013a) Whole-genome sequencing to establish relapse or re-infection with *Mycobacterium tuberculosis*: a retrospective observational study. *The Lancet Respiratory Medicine*, **1**(10): 786–792.

Bryant JM, Schürch AC, Deutekom H van, Harris SR, Beer JL de, Jager V de, Kremer K, Hijum SA van, Siezen RJ, Borgdorff M, Bentley SD, Parkhill J, Soolingen D van (2013b) Inferring patient to patient transmission of *Mycobacterium tuberculosis* from whole genome sequencing data. *BMC Infectious Diseases*, **13**(1): 110.

Burki T (2010) Tackling tuberculosis in London's homeless population. *The Lancet*, **376**(9758): 2055–2056.

Camus JC, Pryor MJ, Médigue C, Cole ST (2002) Re-annotation of the genome sequence of *Mycobacterium tuberculosis* H37Rv. *Microbiology (Reading, England)*, **148**(Pt 10): 2967–2973.

Casali N, Nikolayevskyy V, Balabanova Y, Harris SR, Ignatyeva O, Kontsevaya I, Corander J, Bryant J, Parkhill J, Nejentsev S, Horstmann RD, Brown T, Drobniewski F (2014) Evolution and transmission of drug-resistant tuberculosis in a Russian population. *Nature Genetics*, **46**(3): 279–286.

Casali N, Nikolayevskyy V, Balabanova Y, Ignatyeva O, Kontsevaya I, Harris SR, Bentley SD, Parkhill J, Nejentsev S, Hoffner SE, Horstmann RD, Brown T, Drobniewski F (2012) Microevolution of Extensively Drug-Resistant Tuberculosis in Russia. *Genome Research*, **22**(4): 735–745.

Castets M, Sarrat H (1969) [Experimental study of the virulence of *Mycobacterium africanum* (preliminary note)]. *Bulletin De La Société Médicale d'Afrique Noire De Langue Française*, **14**(4): 693–696.

Castillo-Ramirez S, Corander J, Marttinen P, Aldeljawi M, Hanage W, Westh H, Boye K, Gulay Z, Bentley S, Parkhill J, Holden M, Feil E (2012) Phylogeographic variation in recombination rates within a global clone of methicillin-resistant *Staphylococcus aureus*. *Genome Biology*, **13**(12): R126.

Caws M, Thwaites G, Dunstan S, Hawn TR, Lan NTN, Thuong NTT, Stepniewska K, Huyen MNT, Bang ND, Loc TH, Gagneux S, Soolingen D van, Kremer K, Sande M van der, Small P, Anh PTH, Chinh NT, Quy HT, Duyen NTH, Tho DQ, Hieu NT, Torok E, Hien TT, Dung NH, Nhu NTQ, Duy PM, van Vinh Chau N, Farrar J (2008) The

influence of host and bacterial genotype on the development of disseminated disease with *Mycobacterium tuberculosis*. *PLoS Pathogens*, **4**(3): e1000034.

Caws M, Thwaites G, Stepniewska K, Nguyen TNL, Nguyen THD, Nguyen TP, Mai NTH, Phan MD, Tran HL, Tran THC, Soolingen D van, Kremer K, Nguyen VVC, Nguyen TC, Farrar J (2006) Beijing genotype of *Mycobacterium tuberculosis* is significantly associated with human immunodeficiency virus infection and multidrug resistance in cases of tuberculous meningitis. *Journal of Clinical Microbiology*, **44**(11): 3934–3939.

Chan JZM, Sergeant MJ, Lee OYC, Minnikin DE, Besra GS, Pap I, Spigelman M, Donoghue HD, Pallen MJ (2013) Metagenomic Analysis of Tuberculosis in a Mummy. *New England Journal of Medicine*, **369**(3): 289–290.

Chen Y, Mukherjee S, Hoffmann M, Kotewicz ML, Young S, Abbott J, Luo Y, Davidson MK, Allard M, McDermott P, Zhao S (2013) Whole-genome sequencing of gentamicin-resistant *Campylobacter coli* isolated from U.S. retail meats reveals novel plasmid-mediated aminoglycoside resistance genes. *Antimicrobial Agents and Chemotherapy*, **57**(11): 5398–5405.

Click ES, Moonan PK, Winston CA, Cowan LS, Oeltmann JE (2012) Relationship between *Mycobacterium tuberculosis* phylogenetic lineage and clinical site of tuberculosis. *Clinical Infectious Diseases*, **54**(2): 211–219.

Clotilde LM, Bernard C4, Hartman GL, Lau DK, Carter JM (2011) Microbead-based immunoassay for simultaneous detection of Shiga toxins and isolation of Escherichia coli O157 in foods. *Journal of Food Protection*, **74**(3): 373–379.

Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry CE, Tekaia F, Badcock K, Basham D, Brown D, Chillingworth T, Connor R, Davies R, Devlin K, Feltwell T, Gentles S, Hamlin N, Holroyd S, Hornsby T, Jagels K, Krogh A, McLean J, Moule S, Murphy L, Oliver K, Osborne J, Quail MA, Rajandream MA, Rogers J, Rutter S, Seeger K, Skelton J, Squares R, Squares S, Sulston JE, Taylor K, Whitehead S, Barrell BG (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature*, **393**(6685): 537–544.

Coll F, Mallard K, Preston MD, Bentley S, Parkhill J, McNerney R, Martin N, Clark TG (2012) SpolPred: Rapid and accurate prediction of *Mycobacterium tuberculosis* spoligotypes from short genomic sequences. *Bioinformatics (Oxford, England)*, **28**(22): 2991–2993.

Coll F, Preston M, Guerra-Assunção JA, Hill-Cawthorn G, Harris D, Perdigão J, Viveiros M, Portugal I, Drobniewski F, Gagneux S, Glynn JR, Pain A, Parkhill J, McNerney

R, Martin N, Clark TG (2014) PolyTB: A genomic variation map for *Mycobacterium tuberculosis*. *Tuberculosis (Edinburgh, Scotland)*,

Collins CH, Yates MD, Grange JM (1982) Subdivision of *Mycobacterium tuberculosis* into five variants for epidemiological purposes: methods and nomenclature. *The Journal of Hygiene*, **89**(2): 235–242.

Comas I, Borrell S, Roetzer A, Rose G, Malla B, Kato-Maeda M, Galagan J, Niemann S, Gagneux S (2011a) Whole-genome sequencing of rifampicin-resistant *Mycobacterium tuberculosis* strains identifies compensatory mutations in RNA polymerase genes. *Nature Genetics*, **44**(1): 106–110.

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

Comas I, Coscolla M, Luo T, Borrell S, Holt KE, Kato-Maeda M, Parkhill J, Malla B, Berg S, Thwaites G, Yeboah-Manu D, Bothamley G, Mei J, Wei L, Bentley S, Harris SR, Niemann S, Diel R, Aseffa A, Gao Q, Young D, Gagneux S (2013) Out-of-Africa migration and Neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans. *Nature Genetics*, **45**(10): 1176–1182.

Comas I, Gagneux S (2009a) The past and future of tuberculosis research. *PLoS Pathogens*, **5**(10): e1000600.

Comas I, Gagneux S (2011b) A role for systems epidemiology in tuberculosis research. *Trends in Microbiology*, **19**(10): 492–500.

Comas I, Homolka S, Niemann S, Gagneux S (2009b) Genotyping of genetically monomorphic bacteria: DNA sequencing in *Mycobacterium tuberculosis* highlights the limitations of current methodologies. *PloS ONE*, **4**(11): e7815.

Cook VJ, Shah L, Gardy J, Bourgeois AC (2012) Recommendations on modern contact investigation methods for enhancing tuberculosis control. *The International Journal of Tuberculosis and Lung Disease*, **16**(3): 297–305.

Copin R, Coscollá M, Seiffert SN, Bothamley G, Sutherland J, Mbayo G, Gagneux S, Ernst JD (2014) Sequence diversity in the pe_pgrs genes of *Mycobacterium tuberculosis* is independent of human T cell recognition. *mBio*, **5**(1): e00960–00913.

Coscolla M, Lewin A, Metzger S, Maetz-Rennsing K, Calvignac-Spencer S, Nitsche A, Dabrowski PW, Radonic A, Niemann S, Parkhill J, Couacy-Hymann E, Feldman J, Comas I, Boesch C, Gagneux S, Leendertz FH (2013) Novel *Mycobacterium tuberculosis* Complex Isolate from a Wild Chimpanzee. *Emerging Infectious Diseases*, **19**(6): 969–976.

Coscolla M, Gagneux S (2010) Does *M. tuberculosis* genomic diversity explain disease diversity? *Drug Discovery Today. Disease Mechanisms*, **7**(1): e43–e59.

Cowan LS, Diem L, Brake MC, Crawford JT (2004) Transfer of a *Mycobacterium tuberculosis* genotyping method, Spoligotyping, from a reverse line-blot hybridization, membrane-based assay to the Luminex multianalyte profiling system. *Journal of Clinical Microbiology*, **42**(1): 474–477.

Cowley D, Govender D, February B, Wolfe M, Steyn L, Evans J, Wilkinson RJ, Nicol MP (2008) Recent and rapid emergence of W-Beijing strains of *Mycobacterium tuberculosis* in Cape Town, South Africa. *Clinical Infectious Diseases*, **47**(10): 1252–1259.

Croucher NJ, Harris SR, Grad YH, Hanage WP (2013) Bacterial genomes in epidemiology–present and future. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **368**(1614): 20120202.

Dalla Costa ER, Lazzarini LCO, Perizzolo PF, Díaz CA, Spies FS, Costa LL, Ribeiro AW, Barroco C, Schuh SJ, da Silva Pereira MA, Dias CF, Gomes HM, Unis G, Zaha A, Almeida da Silva PE, Suffys PN, Rossetti MLR (2013) *Mycobacterium tuberculosis* of the RDRio genotype is the predominant cause of tuberculosis and associated with multidrug resistance in Porto Alegre City, South Brazil. *Journal of Clinical Microbiology*, **51**(4): 1071–1077.

Daniel TM (2006) The history of tuberculosis. *Respiratory Medicine*, **100**(11): 1862–1870.

De Beer JL, Kodmon C, Werf MJ van der, Ingen J van, Soolingen D van, ECDC MDR-TB Molecular Surveillance Project Participants (2014) Molecular surveillance of multi- and extensively drug-resistant tuberculosis transmission in the European Union from 2003 to 2011. *Euro surveillance: bulletin Européen sur les maladies transmissibles = European communicable disease bulletin*, **19**(11):

Demay C, Liens B, Burguière T, Hill V, Couvin D, Millet J, Mokrousov I, Sola C, Zozio T, Rastogi N (2012) SITVITWEB – A publicly available international multimarker database for studying *Mycobacterium tuberculosis* genetic diversity and molecular epidemiology. *Infection, Genetics and Evolution*, **12**(4): 755–766.

Deshpande A, Gans J, Graves SW, Green L, Taylor L, Kim HB, Kunde YA, Leonard PM, Li PE, Mark J, Song J, Vuyisich M, White PS (2010) A rapid multiplex assay for nucleic acid-based diagnostics. *Journal of Microbiological Methods*, **80**(2): 155–163.

Diamond J, Bellwood P (2003) Farmers and Their Languages: The First Expansions. *Science*, **300**(5619): 597–603.

Didelot X, Eyre D, Cule M, Ip C, Ansari M, Griffiths D, Vaughan A, O'Connor L, Golubchik T, Batty E, Piazza P, Wilson D, Bowden R, Donnelly P, Dingle K, Wilcox M,

Walker A, Crook D, A Peto T, Harding R (2012) Microevolutionary analysis of *Clostridium difficile* genomes to investigate transmission. *Genome Biology*, **13**(12): R118.

Diep BA (2013) Use of whole-genome sequencing for outbreak investigations. *The Lancet Infectious Diseases*, **13**(2): 99–101.

Diffie BW, Winius GD (1977) *Foundations of the Portuguese Empire: 1415 - 1580*. University of Minnesota Press, 590.

Donoghue HD (2011) Insights gained from palaeomicrobiology into ancient and modern tuberculosis. *Clinical Microbiology and Infection*, **17**(6): 821–829.

Donoghue HD (2009) Human tuberculosis – an ancient disease, as elucidated by ancient microbial biomolecules. *Microbes and Infection*, Forum on Chikungunya **11**(14–15): 1156–1162.

Dos Vultos T, Mestre O, Rauzier J, Golec M, Rastogi N, Rasolofo V, Tonjum T, Sola C, Matic I, Gicquel B (2008) Evolution and Diversity of Clonal Bacteria: The Paradigm of *Mycobacterium tuberculosis*. *PLoS ONE*, **3**(2):

Drobniewski F, Nikolayevskyy V, Maxeiner H, Balabanova Y, Casali N, Kontsevaya I, Ignatyeva O (2013) Rapid diagnostics of tuberculosis and drug resistance in the industrialized world: clinical and public health benefits and barriers to implementation. *BMC Medicine*, **11** 190.

Dunbar SA (2006) Applications of Luminex xMAP technology for rapid, high-throughput multiplexed nucleic acid detection. *Clinica Chimica Acta*, **363**(1-2): 71–82.

Duong DA, Nguyen THD, Nguyen TNL, Dai VH, Dang TMH, Vo SK, Do DAT, Nguyen VVC, Nguyen HD, Dinh NS, Farrar J, Caws M (2009) Beijing genotype of *Mycobacterium tuberculosis* is significantly associated with high-level fluoroquinolone resistance in Vietnam. *Antimicrobial Agents and Chemotherapy*, **53**(11): 4835–4839.

Durmaz R, Zozio T, Gunal S, Allix C, Fauville-Dufaux M, Rastogi N (2007) Population-based molecular epidemiological study of tuberculosis in Malatya, Turkey. *Journal of Clinical Microbiology*, **45**(12): 4027–4035.

Dye C, Scheele S, Dolin P, Pathania V, Raviglione MC, for the WHO Global Surveillance and Monitoring Project (1999) Global burden of tuberculosis: Estimated incidence, prevalence, and mortality by country. *JAMA*, **282**(7): 677–686.

Dye C, Williams BG (2010) The population dynamics and control of tuberculosis. *Science*, **328**(5980): 856–861.

Embden JD van, Cave MD, Crawford JT, Dale JW, Eisenach KD, Gicquel B, Hermans P, Martin C, McAdam R, Shinnick TM (1993) Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology. *Journal of Clinical Microbiology*, **31**(2): 406–409.

Espinosa de los Monteros LE, Galán JC, Gutiérrez M, Samper S, García Marín JF, Martín C, Domínguez L, Rafael L de, Baquero F, Gómez-Mampaso E, Blázquez J (1998) Allele-specific PCR method based on pncA and oxyR sequences for distinguishing *Mycobacterium bovis* from *Mycobacterium tuberculosis*: intraspecific M. bovis pncA sequence polymorphism. *Journal of Clinical Microbiology*, **36**(1): 239–242.

Excoffier L, Laval G, Schneider S (2005) Arlequin (version 3.0): An integrated software package for population genetics data analysis. *Evolutionary Bioinformatics Online*, **1** 47–50.

Farhat MR, Shapiro BJ, Kieser KJ, Sultana R, Jacobson KR, Victor TC, Warren RM, Streicher EM, Calver A, Sloutsky A, Kaur D, Posey JE, Plikaytis B, Oggioni MR, Gardy JL, Johnston JC, Rodrigues M, Tang PKC, Kato-Maeda M, Borowsky ML, Muddukrishna B, Kreiswirth BN, Kurepina N, Galagan J, Gagneux S, Birren B, Rubin EJ, Lander ES, Sabeti PC, Murray M (2013) Genomic analysis identifies targets of convergent positive selection in drug-resistant *Mycobacterium tuberculosis*. *Nature Genetics*, **45**(10): 1183–1189.

Fenner L, Egger M, Bodmer T, Altpeter E, Zwahlen M, Jaton K, Pfyffer GE, Borrell S, Dubuis O, Bruderer T, Siegrist HH, Furrer H, Calmy A, Fehr J, Stalder JM, Ninet B, Böttger EC, Gagneux S (2012a) Effect of mutation and genetic background on drug resistance in *Mycobacterium tuberculosis*. *Antimicrobial Agents and Chemotherapy*, **56**(6): 3047–53.

Fenner L, Egger M, Bodmer T, Furrer H, Ballif M, Battegay M, Helbling P, Fehr J, Gsponer T, Rieder HL, Zwahlen M, Hoffmann M, Bernasconi E, Cavassini M, Calmy A, Dolina M, Frei R, Janssens JP, Borrell S, Stucki D, Schrenzel J, Böttger EC, Gagneux S, for the Swiss HIV Cohort and Molecular Epidemiology of Tuberculosis Study Groups (2013) HIV Infection Disrupts the Sympatric Host–Pathogen Relationship in Human Tuberculosis. *PLoS Genetics*, **9**(3): e1003318.

Fenner L, Gagneux S, Helbling P, Battegay M, Rieder HL, Pfyffer GE, Zwahlen M, Furrer H, Siegrist HH, Fehr J, Dolina M, Calmy A, Stucki D, Jaton K, Janssens JP, Stalder JM, Bodmer T, Ninet B, Böttger EC, Egger M (2012b) *Mycobacterium tuberculosis* transmission in a country with low tuberculosis incidence: role of immigration and HIV infection. *Journal of Clinical Microbiology*, **50**(2): 388–395.

Fenner L, Malla B, Ninet B, Dubuis O, Stucki D, Borrell S, Huna T, Bodmer T, Egger M, Gagneux S (2011) "Pseudo-Beijing": Evidence for convergent evolution in the direct repeat region of *Mycobacterium tuberculosis*. *PloS ONE*, **6**(9): e24737.

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

Feuerwerker A (1990) "Chinese Economic History in Comparative Perspective". In: *Ropp PS, Heritage of China: Contemporary Perspectives on Chinese Civilization*. Berkeley: University of California Press, pp. 224–241.

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

Firdessa R, Berg S, Hailu E, Schelling E, Gumi B, Erenso G, Gadisa E, Kiros T, Habtamu M, Hussein J, Zinsstag J, Robertson BD, Ameni G, Lohan AJ, Loftus B, Comas I, Gagneux S, Tschopp R, Yamuah L, Hewinson G, Gordon SV, Young DB, Aseffa A (2013) Mycobacterial lineages causing pulmonary and extrapulmonary tuberculosis, ethiopia. *Emerging Infectious Diseases*, **19**(3): 460–463.

Fleischmann RD, Alland D, Eisen JA, Carpenter L, White O, Peterson J, DeBoy R, Dodson R, Gwinn M, Haft D, Hickey E, Kolonay JF, Nelson WC, Umayam LA, Ermolaeva M, Salzberg SL, Delcher A, Utterback T, Weidman J, Khouri H, Gill J, Mikula A, Bishai W, Jacobs, Jr. WR, Venter JC, Fraser CM (2002) Whole-Genome Comparison of *Mycobacterium tuberculosis* Clinical and Laboratory Strains. *Journal of Bacteriology*, **184**(19): 5479–5490.

Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S, Gil L, Gordon L, Hendrix M, Hourlier T, Johnson N, Kahari AK, Keefe D, Keenan S, Kinsella R, Komorowska M, Koscielny G, Kulesha E, Larsson P, Longden I, McLaren W, Muffato M, Overduin B, Pignatelli M, Pritchard B, Riat HS, Ritchie GRS, Ruffier M, Schuster M, Sobral D, Tang YA, Taylor K, Trevanion S, Vandrovcova J, White S, Wilson M, Wilder SP, Aken BL, Birney E, Cunningham F, Dunham I, Durbin R, Fernandez-Suarez XM, Harrow J, Herrero J, Hubbard TJP, Parker A, Proctor G, Spudich G, Vogel J, Yates A, Zadissa A, Searle SMJ (2011) Ensembl 2012. *Nucleic Acids Research*, **40**(D1): D84–D90.

Flores L, Van T, Narayanan S, DeRiemer K, Kato-Maeda M, Gagneux S (2007) Large sequence polymorphisms classify *Mycobacterium tuberculosis* strains with ancestral spoligotyping patterns. *Journal of Clinical Microbiology*, **45**(10): 3393–3395.

Ford CB, Lin PL, Chase MR, Shah RR, Iartchouk O, Galagan J, Mohaideen N, Ioerger TR, Sacchettini JC, Lipsitch M, Flynn JL, Fortune SM (2011) Use of whole genome sequencing to estimate the mutation rate of *Mycobacterium tuberculosis* during latent infection. *Nat Genet*, **43**(5): 482–6.

Ford CB, Shah RR, Maeda MK, Gagneux S, Murray MB, Cohen T, Johnston JC, Gardy J, Lipsitch M, Fortune SM (2013) *Mycobacterium tuberculosis* mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis. *Nature Genetics*, **45**(7): 784–790.

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

Gagneux S, Burgos MV, DeRiemer K, Encisco A, Muñoz S, Hopewell PC, Small PM, Pym AS (2006a) Impact of bacterial genetics on the transmission of isoniazid-resistant *Mycobacterium tuberculosis*. *PLoS Pathogens*, **2**(6): e61.

Gagneux S (2012) Host-pathogen coevolution in human tuberculosis. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **367**(1590): 850–859.

Gagneux S, DeRiemer K, Van T, Kato-Maeda M, Jong BC de, Narayanan S, Nicol M, Niemann S, Kremer K, Gutierrez MC, Hilty M, Hopewell PC, Small PM (2006b) Variable host-pathogen compatibility in *Mycobacterium tuberculosis*. *Proceedings of the National Academy of Sciences*, **103**(8): 2869–2873.

Gardy JL (2013) Investigation of disease outbreaks with genome sequencing. *The Lancet Infectious Diseases*, **13**(2): 101–102.

Gardy JL, Johnston JC, Sui SJH, Cook VJ, Shah L, Brodkin E, Rempel S, Moore R, Zhao Y, Holt R, Varhol R, Birol I, Lem M, Sharma MK, Elwood K, Jones SJM, Brinkman FSL, Brunham RC, Tang P, Ho Sui SJ (2011) Whole-Genome Sequencing and Social-Network Analysis of a Tuberculosis Outbreak. *New England Journal of Medicine*, **364**(8): 730–739.

Gargis AS, Kalman L, Berry MW, Bick DP, Dimmock DP, Hambuch T, Lu F, Lyon E, Voelkerding KV, Zehnbauer BA, Agarwala R, Bennett SF, Chen B, Chin ELH, Compton JG, Das S, Farkas DH, Ferber MJ, Funke BH, Furtado MR, Ganova-Raeva LM, Geigenmüller U, Gunselman SJ, Hegde MR, Johnson PLF, Kasarskis A, Kulkarni S, Lenk T, Liu CSJ, Manion M, Manolio TA, Mardis ER, Merker JD, Rajeevan MS, Reese

MG, Rehm HL, Simen BB, Yeakley JM, Zook JM, Lubin IM (2012) Assuring the quality of next-generation sequencing in clinical laboratory practice. *Nature Biotechnology*, **30**(11): 1033–1036.

Garnier T, Eiglmeier K, Camus JC, Medina N, Mansoor H, Pryor M, Duthoy S, Grondin S, Lacroix C, Monsempe C, Simon S, Harris B, Atkin R, Doggett J, Mayes R, Keating L, Wheeler PR, Parkhill J, Barrell BG, Cole ST, Gordon SV, Hewinson RG (2003) The complete genome sequence of *Mycobacterium bovis*. *Proceedings of the National Academy of Sciences*, **100**(13): 7877–7882.

Genewein A, Telenti A, Bernasconi C, Schopfer K, Bodmer T, Mordasini C, Weiss S, Maurer AM, Rieder H (1993) Molecular approach to identifying route of transmission of tuberculosis in the community. *The Lancet*, **342**(8875): 841–844.

Gey van Pittius NC, Perrett KD, Michel AL, Keet DF, Hlokwe T, Streicher EM, Warren RM, Helden PD van (2012) Infection of african buffalo (syncerus caffer) by oryx bacillus, a rare member of the antelope clade of the mycobacterium tuberculosis complex. *Journal of Wildlife Diseases*, **48**(4): 849–857.

Gibson AL, Huard RC, Pittius NCG van, Lazzarini LCO, Driscoll J, Kurepina N, Zozio T, Sola C, Spindola SM, Kritski AL, Fitzgerald D, Kremer K, Mardassi H, Chitale P, Brinkworth J, Viedma DG de, Gicquel B, Pape JW, Soolingen D van, Kreiswirth BN, Warren RM, Helden PD van, Rastogi N, Suffys PN, Silva JLe, Ho JL (2008) Application of Sensitive and Specific Molecular Methods To Uncover Global Dissemination of the Major RDRio Sublineage of the Latin American-Mediterranean *Mycobacterium tuberculosis* Spoligotype Family. *Journal of Clinical Microbiology*, **46**(4): 1259–1267.

Gillespie JJ, Wattam AR, Cammer SA, Gabbard JL, Shukla MP, Dalay O, Driscoll T, Hix D, Mane SP, Mao C, Nordberg EK, Scott M, Schulman JR, Snyder EE, Sullivan DE, Wang C, Warren A, Williams KP, Xue T, Yoo HS, Zhang C, Zhang Y, Will R, Kenyon RW, Sobral BW (2011) PATRIC: the Comprehensive Bacterial Bioinformatics Resource with a Focus on Human Pathogenic Species. *Infection and Immunity*, **79**(11): 4286–4298.

Gilmour MW, Graham M, Reimer A, Van Domselaar G (2013) Public health genomics and the new molecular epidemiology of bacterial pathogens. *Public Health Genomics*, **16**(1-2): 25–30.

Goecks J, Nekrutenko A, Taylor J (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology*, **11**(8): R86.

Gole J, Gore A, Richards A, Chiu YJ, Fung HL, Bushman D, Chiang HI, Chun J, Lo YH, Zhang K (2013) Massively parallel polymerase cloning and genome sequencing of single cells using nanoliter microwells. *Nature Biotechnology*, **31**(12): 1126–1132.

Gomes HM, Elias AR, Oelemann MAC, Pereira MAdS, Montes FFO, Marsico AG, Kritski AL, Filho LdA, Caldas PC, Possuelo LG, Cafrune P, Rossetti ML, Lucena N, Saad MHF, Cavalcanti HR, Leite CQF, Brito RC de, Lopes ML, Lima K, Souza M, Trindade RdC, Zozio T, Sola C, Rastogi N, Suffys PN (2012) Spoligotypes of *Mycobacterium tuberculosis* complex isolates from patients residents of 11 states of Brazil. *Infection, Genetics and Evolution*, **12**(4): 649–656.

Groenen PM, Bunschoten AE, Soolingen D van, Embden JD van (1993) Nature of DNA polymorphism in the direct repeat cluster of *Mycobacterium tuberculosis*; application for strain differentiation by a novel typing method. *Molecular Microbiology*, **10**(5): 1057–1065.

Gut IG (2013) New sequencing technologies. *Clinical and Translational Oncology*, **15**(11): 879–881.

Gutacker MM, Mathema B, Soini H, Shashkina E, Kreiswirth BN, Graviss EA, Musser JM (2006) Single-nucleotide polymorphism-based population genetic analysis of *Mycobacterium tuberculosis* strains from 4 geographic sites. *The Journal of Infectious Diseases*, **193**(1): 121–128.

Gutacker MM, Smoot JC, Migliaccio CAL, Ricklefs SM, Hua S, Cousins DV, Graviss EA, Shashkina E, Kreiswirth BN, Musser JM (2002) Genome-wide analysis of synonymous single nucleotide polymorphisms in *Mycobacterium tuberculosis* complex organisms: resolution of genetic relationships among closely related microbial strains. *Genetics*, **162**(4): 1533–1543.

Gutierrez MC, Brisse S, Brosch R, Fabre M, Omaïs B, Marmiesse M, Supply P, Vincent V (2005) Ancient origin and gene mosaicism of the progenitor of *Mycobacterium tuberculosis*. *PLoS Pathogens*, **1**(1): e5.

Hain Lifescience - Mycobacteria (2012).

Halse TA, Escuyer VE, Musser KA (2011) Evaluation of a single-tube multiplex real-time PCR for differentiation of members of the *Mycobacterium tuberculosis* complex in clinical specimens. *Journal of Clinical Microbiology*, **49**(7): 2562–2567.

Hanekom M, Gey van Pittius NC, McEvoy C, Victor TC, Van Helden PD, Warren RM (2011) *Mycobacterium tuberculosis* Beijing genotype: A template for success. *Tuberculosis (Edinburgh, Scotland)*, **91**(6): 510–523.

Hanekom M, Spuy GD van der, Streicher E, Ndabambi SL, McEvoy CRE, Kidd M, Beyers N, Victor TC, Helden PD van, Warren RM (2007) A recently evolved sublineage of the

Mycobacterium tuberculosis Beijing strain family is associated with an increased ability to spread and cause disease. *Journal of Clinical Microbiology*, **45**(5): 1483–1490.

Hartl DL, Clark AG (2007) *Principles of Population Genetics*. 4th ed. Sinauer Associates, 565.

Hershberg R, Lipatov M, Small PM, Sheffer H, Niemann S, Homolka S, Roach JC, Kremer K, Petrov DA, Feldman MW, Gagneux S (2008) High functional diversity in {M}ycobacterium tuberculosis driven by genetic drift and human demography. *PLoS biology*, **6**(12): e311.

Hershberg R, Petrov DA (2010) Evidence that mutation is universally biased towards AT in bacteria. *PLoS Genetics*, **6**(12): e311.

Hershkovitz I, Donoghue HD, Minnikin DE, Besra GS, Lee OYC, Gernaey AM, Galili E, Eshed V, Greenblatt CL, Lemma E, Bar-Gal GK, Spigelman M (2008) Detection and Molecular Characterization of 9000-Year-Old *Mycobacterium tuberculosis* from a Neolithic Settlement in the Eastern Mediterranean. *PLoS ONE*, **3**(10): e3426.

Hillemann D, Weizenegger M, Kubica T, Richter E, Niemann S (2005) Use of the genotype MTBDR assay for rapid detection of rifampin and isoniazid resistance in {M}ycobacterium tuberculosis complex isolates. *Journal of clinical microbiology*, **43**(8): 3699–3703.

Hirakawa M, Tanaka T, Hashimoto Y, Kuroda M, Takagi T, Nakamura Y (2002) JSNP: a database of common gene variations in the Japanese population. *Nucleic Acids Research*, **30**(1): 158–162.

Hirsh AE, Tsolaki AG, DeRiemer K, Feldman MW, Small PM (2004) Stable association between strains of *Mycobacterium tuberculosis* and their human host populations. *Proceedings of the National Academy of Sciences*, **101**(14): 4871–4876.

Holland PM, Abramson RD, Watson R, Gelfand DH (1991) Detection of specific polymerase chain reaction product by utilizing the 5'—-3' exonuclease activity of Thermus aquaticus DNA polymerase. *Proceedings of the National Academy of Sciences*, **88**(16): 7276–7280.

Homolka S, Niemann S, Russell DG, Rohde KH (2010) Functional genetic diversity among *Mycobacterium tuberculosis* complex clinical isolates: delineation of conserved core and lineage-specific transcriptomes during intracellular survival. *PLoS Pathogens*, **6**(7): e1000988.

Homolka S, Post E, Oberhauser B, George AG, Westman L, Dafae F, Rüsch-Gerdes S, Niemann S (2008) High genetic diversity among *Mycobacterium tuberculosis* complex strains from Sierra Leone. *BMC Microbiology*, **8** 103.

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

Huard RC, Fabre M, Haas P de, Lazzarini LCO, Soolingen D van, Cousins D, Ho JL, Claudio Oliveira Lazzarini L (2006) Novel Genetic Polymorphisms That Further Delineate the Phylogeny of the *Mycobacterium tuberculosis* Complex. *Journal of Bacteriology*, **188**(12): 4271–4287.

Ignatova A, Dubiley S, Stepanshina V, Shemyakin I (2006) Predominance of multi-drug-resistant LAM and Beijing family strains among *Mycobacterium tuberculosis* isolates recovered from prison inmates in Tula Region, Russia. *Journal of Medical Microbiology*, **55**(10): 1413–1418.

Inouye M, Conway TC, Zobel J, Holt KE (2012) Short read sequence typing (SRST): multi-locus sequence types from short reads. *BMC Genomics*, **13**(1): 338.

Integrating common and rare genetic variation in diverse human populations (2010). *Nature*, **467**(7311): 52–58.

Jolley KA, Bliss CM, Bennett JS, Bratcher HB, Brehony C, Colles FM, Wimalarathna H, Harrison OB, Sheppard SK, Cody AJ, Maiden MCJ (2012) Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. *Microbiology*, **158**(Pt 4): 1005–1015.

Jolley K, Maiden M (2010) BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics*, **11**(1): 595.

Jong BC de, Antonio M, Gagneux S (2010) *Mycobacterium africanum*–review of an important cause of human tuberculosis in West Africa. *PLoS Neglected Tropical Diseases*, **4**(9): e744.

Jong BC de, Hill PC, Aiken A, Awine T, Antonio M, Adetifa IM, Jackson-Sillah DJ, Fox A, Deriemer K, Gagneux S, Borgdorff MW, McAdam KPWJ, Corrah T, Small PM, Adegbola RA (2008) Progression to active tuberculosis, but not transmission, varies by *Mycobacterium tuberculosis* lineage in The Gambia. *The Journal of Infectious Diseases*, **198**(7): 1037–1043.

Kaderali L, Deshpande A, Nolan JP, White PS (2003) Primer-design for multiplexed genotyping. *Nucleic Acids Research*, **31**(6): 1796–1802.

Kamerbeek J, Schouls L, Kolk A, Agterveld M van, Soolingen D van, Kuijper S, Bunschoten A, Molhuizen H, Shaw R, Goyal M, Embden J van (1997) Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. *Journal of Clinical Microbiology*, **35**(4): 907–914.

Kappelman J, Alçiçek MC, Kazancı N, Schultz M, Özkul M, Şen Ş (2008) First Homo erectus from Turkey and implications for migrations into temperate Eurasia. *American Journal of Physical Anthropology*, **135**(1): 110–116.

Kato-Maeda M, Gagneux S, Flores LL, Kim EY, Small PM, Desmond EP, Hopewell PC (2011a) Strain classification of *Mycobacterium tuberculosis*: congruence between large sequence polymorphisms and spoligotypes. *The International Journal of Tuberculosis and Lung Disease*, **15**(1): 131–133.

Kato-Maeda M, Metcalfe JZ, Flores L (2011b) Genotyping of *Mycobacterium tuberculosis*: application in epidemiologic studies. *Future Microbiology*, **6** 203–216.

Kato-Maeda M, Shanley CA, Ackart D, Jarlsberg LG, Shang S, Obregon-Henao A, Harton M, Basaraba RJ, Henao-Tamayo M, Barrozo JC, Rose J, Kawamura LM, Coscolla M, Fofanov VY, Koshinsky H, Gagneux S, Hopewell PC, Ordway DJ, Orme IM (2012) Beijing Sublineages of *Mycobacterium tuberculosis* Differ in Pathogenicity in the Guinea Pig. *Clinical and Vaccine Immunology*, **19**(8): 1227–1237.

Keim P, Van Ert MN, Pearson T, Vogler AJ, Huynh LY, Wagner DM (2004) Anthrax molecular epidemiology and forensics: using the appropriate marker for different evolutionary scales. *Infection, Genetics and Evolution*, **4**(3): 205–213.

Kim S, Misra A (2007) SNP genotyping: technologies and biomedical applications. *Annual Review of Biomedical Engineering*, **9**(1): 289–320.

Kirschner DE, Young D, Flynn JL (2010) Tuberculosis: Global Approaches to a Global Disease. *Current Opinion in Biotechnology*, **21**(4): 524–531.

Kitts A, Sherry S (2011) "The Single Nucleotide Polymorphism Database (dbSNP) of Nucleotide Sequence Variation." In: *2nd edition.* Vol. Chapter 5. McEntyre J, Ostell J editors.

Klopper M, Warren RM, Hayes C, Gey van Pittius NC, Streicher EM, Müller B, Sirgel FA, Chabula-Nxiweni M, Hoosain E, Coetzee G, David van Helden P, Victor TC, Trollip AP (2013) Emergence and spread of extensively and totally drug-resistant tuberculosis, South Africa. *Emerging Infectious Diseases*, **19**(3): 449–455.

Koch R (1932) Die Aetiologie der Tuberkulose. *Klinische Wochenschrift*, **11**(12): 490–492.

Kohl TA, Diel R, Harmsen D, Rothgänger J, Walter KM, Merker M, Weniger T, Niemann S (2014) Whole genome based *Mycobacterium tuberculosis* surveillance: A standardized, portable and expandable approach. *Journal of Clinical Microbiology*, JCM.00567–14.

Köser CU, Bryant JM, Becq J, Török ME, Ellington MJ, Marti-Renom MA, Carmichael AJ, Parkhill J, Smith GP, Peacock SJ (2013) Whole-genome sequencing for rapid susceptibility testing of *M. tuberculosis*. *The New England Journal of Medicine*, **369**(3): 290–292.

Köser CU, Feuerriegel S, Summers DK, Archer JAC, Niemann S (2012a) Importance of the Genetic Diversity within the *Mycobacterium tuberculosis* Complex for the Development of Novel Antibiotics and Diagnostic Tests of Drug Resistance. *Antimicrobial Agents and Chemotherapy*, **56**(12): 6080–6087.

Köser CCUC, Holden MMTG, Ellington MJM, Cartwright EJPE, Brown NMN, Ogilvy-Stuart ALA, Hsu LLY, Chewapreecha C, Croucher NJN, Harris SR, Sanders M, Enright MC, Dougan G, Bentley SD, Parkhill J, Fraser LJ, Betley JR, Schulz-Trieglaff OB, Smith GP, Peacock SJ (2012b) Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. *New England Journal of Medicine*, **366**(24): 2267–2275.

Kruczkiewicz P, Mutschall S, Barker D, Thomas J, Van Domselaar G, Gannon VP, Carrillo CD, Taboada EN (2012) MIST: a tool for rapid in silico generation of molecular data from bacterial genome sequences.

Kumar P, Henikoff S, Ng PC (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols*, **4**(7): 1073–1081.

Lambregts-van Weezenbeek CSB, Sebek MMGG, Gerven PJHJ van, Vries G de, Verver S, Kalisvaart NA, Soolingen D van (2003) Tuberculosis contact investigation and DNA fingerprint surveillance in The Netherlands: 6 years' experience with nation-wide cluster feedback and cluster monitoring. *The International Journal of Tuberculosis and Lung Disease*, **7**(12): S463–S470.

Lamrabet O, Drancourt M (2012) Genetic engineering of *Mycobacterium tuberculosis*: A review. *Tuberculosis (Edinburgh, Scotland)*, **92**(5): 365–376.

Lan NTN, Lien HTK, Tung LB, Borgdorff MW, Kremer K, Soolingen D van (2003) *Mycobacterium tuberculosis* Beijing genotype and risk for treatment failure and relapse, Vietnam. *Emerging Infectious Diseases*, **9**(12): 1633–1635.

Lanzas F, Karakousis PC, Sacchettini JC, Ioerger TR (2013) Multidrug-resistant tuberculosis in panama is driven by clonal expansion of a multidrug-resistant *Mycobacterium tuberculosis* strain related to the KZN extensively drug-resistant *M. tuberculosis* strain from South Africa. *Journal of Clinical Microbiology*, **51**(10): 3277–3285.

Lasken RS (2012) Genomic sequencing of uncultured microorganisms from single cells. *Nature Reviews Microbiology*, **10**(9): 631–640.

Lavezzo E, Toppo S, Franchin E, Di Camillo B, Finotello F, Falda M, Manganelli R, Palu G, Barzon L (2013) Genomic comparative analysis and gene function prediction in infectious diseases: application to the investigation of a meningitis outbreak. *BMC infectious Diseases*, **13** 554.

Lawn SD, Mwaba P, Bates M, Piatek A, Alexander H, Marais BJ, Cuevas LE, McHugh TD, Zijenah L, Kapata N, Abubakar I, McNerney R, Hoelscher M, Memish ZA, Migliori GB, Kim P, Maeurer M, Schito M, Zumla A (2013) Advances in tuberculosis diagnostics: the Xpert MTB/RIF assay and future prospects for a point-of-care test. *The Lancet Infectious Diseases*, **13**(4): 349–361.

Lazzarini LCO, Huard RC, Boechat NL, Gomes HM, Oelemann MC, Kurepina N, Shashkina E, Mello FCQ, Gibson AL, Virginio MJ, Marsico AG, Butler WR, Kreiswirth BN, Suffys PN, Silva JRLe, Ho JL (2007) Discovery of a Novel *Mycobacterium tuberculosis* Lineage That Is a Major Cause of Tuberculosis in Rio de Janeiro, Brazil. *Journal of Clinical Microbiology*, **45**(12): 3891–3902.

Le VTM, Diep BA (2013) Selected insights from application of whole-genome sequencing for outbreak investigations. *Current Opinion in Critical Care*,

Lee H, Tang H (2012) Next-generation sequencing technologies and fragment assembly algorithms. *Methods in Molecular Biology*, **855** 155–174.

Lee SH, Walker DR, Cregan PB, Boerma HR (2004) Comparison of four flow cytometric SNP detection assays and their use in plant improvement. *Theoretical and Applied Genetics*, **110**(1): 167–174.

Lehmann J (1946) Para-aminosalicylic acid in the treatment of tuberculosis. *The Lancet*, **1**(6384): 15.

Lehmann K, Neumann R (1896) *Atlas und Grundriss der Bakteriologie und Lehrbuch der speciellen bacteriologischen Diagnostik.*

Lehning J (2013) *European Colonialism since 1700*. Cambridge ; New York: Cambridge University Press, 322.

Lew JM, Kapopoulou A, Jones LM, Cole ST (2011) TubercuList–10 years after. *Tuberculosis (Edinburgh, Scotland)*, **91**(1): 1–7.

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, **25**(14): 1754–1760.

Li X, Xu P, Shen X, Qi L, DeRiemer K, Mei J, Gao Q (2011) Non-Beijing strains of *Mycobacterium tuberculosis* in China. *Journal of Clinical Microbiology*, **49**(1): 392–395.

Liang L, Wang Q, Agren H, Tu Y (2014) Computational studies of DNA sequencing with solid-state nanopores: key issues and future prospects. *Frontiers in Chemistry*, **2** 5.

Librado P, Rozas J (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics*, **25**(11): 1451–1452.

Liu X, Gutacker MM, Musser JM, Fu YX (2006) Evidence for recombination in *Mycobacterium tuberculosis*. *Journal of Bacteriology*, **188**(23): 8169–8177.

Loman NJ, Constantinidou C, Chan JZM, Halachev M, Sergeant M, Penn CW, Robinson ER, Pallen MJ (2012) High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. *Nature Reviews Microbiology*, **10**(9): 599–606.

Lopes JS, Marques I, Soares P, Nebenzahl-Guimaraes H, Costa J, Miranda A, Duarte R, Alves A, Macedo R, Duarte TA, Barbosa T, Oliveira M, Nery JS, Boechat N, Pereira SM, Barreto ML, Pereira-Leal J, Gabriela M Gomes M, Penha-Goncalves C (2013) SNP typing reveals similarity in *Mycobacterium tuberculosis* genetic diversity between Portugal and Northeast Brazil. *Infection, Genetics and Evolution*, **238** 46.

Maddison W, Maddison D (2011) *Mesquite: a modular system for evolutionary analysis. Version 2.73*.

Maiden MC, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, Zhang Q, Zhou J, Zurth K, Caugant DA, Feavers IM, Achtman M, Spratt BG (1998) Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences*, **95**(6): 3140–3145.

Malla B, Stucki D, Borrell S, Feldmann J, Maharjan B, Shrestha B, Fenner L, Gagneux S (2012) First Insights into the Phylogenetic Diversity of *Mycobacterium tuberculosis* in Nepal. *PLoS ONE*, **7**(12): e52297.

Maq (2012).

Mather AE, Reid SWJ, Maskell DJ, Parkhill J, Fookes MC, Harris SR, Brown DJ, Coia JE, Mulvey MR, Gilmour MW, Petrovska L, Pinna E de, Kuroda M, Akiba M, Izumiya H, Connor TR, Suchard MA, Lemey P, Mellor DJ, Haydon DT, Thomson NR (2013) Distinguishable epidemics of multidrug-resistant *Salmonella Typhimurium* DT104 in different hosts. *Science*, **341**(6153): 1514–1517.

McElroy PD, Rothenberg RB, Varghese R, Woodruff R, Minns GO, Muth SQ, Lambert LA, Ridzon R (2003) A network-informed approach to investigating a tuberculosis outbreak: implications for enhancing contact investigations. *The International Journal of Tuberculosis and Lung Disease*, **7**(12 Suppl 3): S486–493.

McEvoy CRE, Cloete R, Müller B, Schürch AC, Helden PD van, Gagneux S, Warren RM, Gey van Pittius NC (2012) Comparative Analysis of *Mycobacterium tuberculosis* pe and ppe Genes Reveals High Sequence Variation and an Apparent Absence of Selective Constraints. *PloS ONE*, **7**(3): e30593.

McEvoy CR, Falmer AA, Pittius NCG van, Victor TC, Helden PD van, Warren RM (2007) The role of IS6110 in the evolution of *Mycobacterium tuberculosis*. *Tuberculosis (Edinburgh, Scotland)*, **87**(5): 393–404.

Mestre O, Luo T, Dos Vultos T, Kremer K, Murray A, Namouchi A, Jackson C, Rauzier J, Bifani P, Warren R, Rasolofo V, Mei J, Gao Q, Gicquel B (2011) Phylogeny of

Mycobacterium tuberculosis Beijing strains constructed from polymorphisms in genes involved in DNA replication, recombination and repair. *PloS ONE*, **6**(1): e16020.

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

Middelkoop K, Bekker LG, Mathema B, Shashkina E, Kurepina N, Whitelaw A, Fallows D, Morrow C, Kreiswirth B, Kaplan G, Wood R (2009) Molecular epidemiology of *Mycobacterium tuberculosis* in a South African community with high HIV prevalence. *The Journal of Infectious Diseases*, **200**(8): 1207–1211.

Migliori GB, De Iaco G, Besozzi G, Centis R, Cirillo DM (2007) First tuberculosis cases in Italy resistant to all tested drugs. *Euro Surveillance*, **12**(5): E070517.1.

Milner DAJ, Vareta J, Valim C, Montgomery J, Daniels RF, Volkman SK, Neafsey DE, Park DJ, Schaffner SF, Mahesh NC, Barnes KG, Rosen DM, Lukens AK, Van Tyne D, Wiegand RC, Sabeti PC, Seydel KB, Glover SJ, Kamiza S, Molyneux ME, Taylor TE, Wirth DF (2012) Human cerebral malaria and *Plasmodium falciparum* genotypes in Malawi. *Malaria Journal*, **11**(1): 35.

Mitchison DA, Wallace JG, Bhatia AL, Selkon JB, Subbaiah TV, Lancaster MC (1960) A comparison of the virulence in guinea-pigs of South Indian and British tubercle bacilli. *Tubercle*, **41** 1–22.

Mitruka K, Oeltmann JE, Ijaz K, Haddad MB (2011) Tuberculosis outbreak investigations in the United States, 2002-2008. *Emerging Infectious Diseases*, **17**(3): 425–431.

Mokrousov I, Vyazovaya A, Narvskaya O (2014) *Mycobacterium tuberculosis* Latin-American Mediterranean family and its sublineages: in the light of evolutionary robust markers. *Journal of Bacteriology*, **196**(10): 1833–41.

Mokrousov I, Vyazovaya A, Otten T, Zhuravlev V, Pavlova E, Tarashkevich L, Krishevich V, Vishnevsky B, Narvskaya O (2012) *Mycobacterium tuberculosis* Population in Northwestern Russia: An Update from Russian-EU/Latvian Border Region. *PLoS ONE*, **7**(7):

Mostowy S, Cousins D, Behr MA (2004) Genomic interrogation of the dassie bacillus reveals it as a unique RD1 mutant within the *Mycobacterium tuberculosis* complex. *Journal of bacteriology*, **186**(1): 104–109.

Mostowy S, Cousins D, Brinkman J, Aranaz A, Behr MA (2002) Genomic deletions suggest a phylogeny for the *Mycobacterium tuberculosis* complex. *The Journal of Infectious Diseases*, **186**(1): 74–80.

Müller B, Dürr S, Alonso S, Hattendorf J, Laisse CJ, Parsons SD, Helden PD van, Zinsstag J (2013) Zoonotic *Mycobacterium bovis* –induced Tuberculosis in Humans. *Emerging Infectious Diseases*, **19**(6): 899–908.

Muller R, Roberts CA, Brown TA (2014) Genotyping of ancient *Mycobacterium tuberculosis* strains reveals historic genetic diversity. *Proceedings of the Royal Society B: Biological Sciences*, **281**(1781): 20133236.

Murray JF (2001) A thousand years of pulmonary medicine: good news and bad. *European Respiratory Journal*, **17**(3): 558–565.

Musser JM (1995) Antimicrobial agent resistance in mycobacteria: molecular genetic insights. *Clinical Microbiology Reviews*, **8**(4): 496–514.

Nahid P, Bliven EE, Kim EY, Mac Kenzie WR, Stout JE, Diem L, Johnson JL, Gagneux S, Hopewell PC, Kato-Maeda M, the Tuberculosis Trials Consortium (2010) Influence of *M. tuberculosis* Lineage Variability within a Clinical Trial for Pulmonary Tuberculosis. *PLoS ONE*, **5**(5): e10753.

Namouchi A, Didelot X, Schöck U, Gicquel B, Rocha EP (2012) After the bottleneck: Genome-wide diversification of the *Mycobacterium tuberculosis* complex by mutation, recombination, and natural selection. *Genome Research*, **22**(4): 721–734.

Ngamphiw C, Assawamakin A, Xu S, Shaw PJ, Yang JO, Ghang H, Bhak J, Liu E, Tongsima S, and the HUGO Pan-Asian SNP Consortium (2011) PanSNPdb: The Pan-Asian SNP Genotyping Database. *PLoS ONE*, **6**(6): e21451.

Nicol MP, Wilkinson RJ (2008) The clinical consequences of strain diversity in *Mycobacterium tuberculosis*. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, **102**(10): 955–965.

Niemann S, Rusch-Gerdes S, Joloba ML, Whalen CC, Guwatudde D, Ellner JJ, Eisenach K, Fumokong N, Johnson JL, Aisu T, Mugerwa RD, Okwera A, Schwander SK (2002) *Mycobacterium africanum* Subtype II Is Associated with Two Distinct Genotypes and Is a Major Cause of Human Tuberculosis in Kampala, Uganda. *Journal of Clinical Microbiology*, **40**(9): 3398–3405.

Niemann S, Köser CU, Gagneux S, Plinke C, Homolka S, Bignell H, Carter RJ, Cheetham RK, Cox A, Gormley NA, Kokko-Gonzales P, Murray LJ, Rigatti R, Smith VP, Arends FPM, Cox HS, Smith G, Archer JAC (2009) Genomic diversity among drug sensitive and multidrug resistant isolates of *Mycobacterium tuberculosis* with identical DNA fingerprints. *PloS ONE*, **4**(10): e7407.

Niobe-Eyangoh SN, Kuaban C, Sorlin P, Cunin P, Thonnon J, Sola C, Rastogi N, Vincent V, Gutierrez MC (2003) Genetic biodiversity of *Mycobacterium tuberculosis* complex strains from patients with pulmonary tuberculosis in Cameroon. *Journal of Clinical Microbiology*, **41**(6): 2547–2553.

Nübel U, Nitsche A, Layer F, Strommenger B, Witte W (2012) Single-nucleotide polymorphism genotyping identifies a locally endemic clone of methicillin-resistant *Staphylococcus aureus*. *PLoS ONE*, **7**(3): e32698.

Okonechnikov K, Golosova O, Fursov M (2012) Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics (Oxford, England)*, **28**(8): 1166–1167.

Olsen RJ, Long SW, Musser JM (2012) Bacterial genomics in infectious disease and the clinical pathology laboratory. *Archives of Pathology & Laboratory Medicine*, **136**(11): 1414–1422.

Pallen MJ, Loman NJ, Penn CW (2010) High-throughput sequencing and clinical microbiology: progress, opportunities and challenges. *Current Opinion in Microbiology*, **13**(5): 625–631.

Palomino JC, Leão SC, Ritacco V (2007) *Tuberculosis 2007 | A Medical Textbook, 700 pages*.

Pareek M, Evans J, Innes J, Smith G, Hingley-Wilson S, Lougheed KE, Sridhar S, Dedicoat M, Hawkey P, Lalvani A (2012) Ethnicity and mycobacterial lineage as determinants of tuberculosis disease phenotype. *Thorax*, **68**(3): 221–9.

Parkhill J, Wren B (2011) Bacterial epidemiology and biology - lessons from genome sequencing. *Genome Biology*, **12**(10): 230.

Parwati I, Crevel R van, Soolingen D van (2010) Possible underlying mechanisms for successful emergence of the *Mycobacterium tuberculosis* Beijing genotype strains. *The Lancet Infectious Diseases*, **10**(2): 103–111.

Pasricha R, Chandolia A, Ponnan P, Saini N, Sharma S, Chopra M, Basil M, Brahmachari V, Bose M (2011) Single nucleotide polymorphism in the genes of mce1 and mce4 operons of *Mycobacterium tuberculosis*: analysis of clinical isolates and standard reference strains. *BMC Microbiology*, **11**(1): 41.

Paterson GK, Morgan FJE, Harrison EM, Cartwright EJP, Torok ME, Zadoks RN, Parkhill J, Peacock SJ, Holmes MA (2013) Prevalence and characterization of human mecC methicillin-resistant *Staphylococcus aureus* isolates in England. *Journal of Antimicrobial Chemotherapy*, **69**(4): 907–10.

Pearson T, Busch JD, Ravel J, Read TD, Rhoton SD, U'Ren JM, Simonson TS, Kachur SM, Leadem RR, Cardon ML, Van Ert MN, Huynh LY, Fraser CM, Keim P (2004) Phylogenetic discovery bias in Bacillus anthracis using single-nucleotide polymorphisms from whole-genome sequencing. *Proceedings of the National Academy of Sciences*, **101**(37): 13536–13541.

Pepperell CS, Casto AM, Kitchen A, Granka JM, Cornejo OE, Holmes EC, Birren B, Galagan J, Feldman MW (2013) The Role of Selection in Shaping Diversity of Natural *M. tuberculosis* Populations. *PLoS Pathogens*, **9**(8): e1003543.

Pepperell C, Hoeppner VH, Lipatov M, Wobeser W, Schoolnik GK, Feldman MW (2010) Bacterial Genetic Signatures of Human Social Phenomena among *M. tuberculosis* from an Aboriginal Canadian Population. *Molecular Biology and Evolution*, **27**(2): 427–440.

Pérez-Lago L, Comas I, Navarro Y, González-Candelas F, Herranz M, Bouza E, García-de-Viedma D (2014) Whole Genome Sequencing Analysis of Intrapatient Microevolution in *Mycobacterium tuberculosis*: Potential Impact on the Inference of Tuberculosis Transmission. *Journal of Infectious Diseases*, **209**(1): 98–108.

Pillay M, Sturm AW (2007) Evolution of the Extensively Drug-Resistant F15/LAM4/KZN Strain of *Mycobacterium tuberculosis* in KwaZulu-Natal, South Africa. *Clinical Infectious Diseases*, **45**(11): 1409–1414.

Portevin D, Gagneux S, Comas I, Young D (2011) Human macrophage responses to clinical isolates from the *Mycobacterium tuberculosis* complex discriminate between ancient and modern lineages. *PLoS Pathogens*, **7**(3): e1001307.

Ragheb MN, Ford CB, Chase MR, Lin PL, Flynn JL, Fortune SM (2013) The mutation rate of mycobacterial repetitive unit loci in strains of M. Tuberculosis from cynomolgus macaque infection. *BMC Genomics*, **14**(1): 145.

Ramaswamy S, Musser JM (1998) Molecular genetic basis of antimicrobial agent resistance in *Mycobacterium tuberculosis*: 1998 update. *Tubercle and Lung Disease*, **79**(1): 3–29.

Reddy TBK, Riley R, Wymore F, Montgomery P, DeCaprio D, Engels R, Gellesch M, Hubble J, Jen D, Jin H, Koehrsen M, Larson L, Mao M, Nitzberg M, Sisk P, Stolte C, Weiner B, White J, Zachariah ZK, Sherlock G, Galagan JE, Ball CA, Schoolnik GK (2009) TB database: an integrated platform for tuberculosis research. *Nucleic Acids Research*, **37**(Database issue): D499–508.

Reiling N, Homolka S, Walter K, Brandenburg J, Niwinski L, Ernst M, Herzmann C, Lange C, Diel R, Ehlers S, Niemann S (2013) Clade-specific virulence patterns of *Mycobacterium tuberculosis* complex strains in human primary macrophages and aerogenically infected mice. *mBio*, **4**(4): e00250–13.

Rieder HL, Watson JM, Raviglione MC, Forssbohm M, Migliori GB, Schwoebel V, Leitch AG, Zellweger JP (1996) Surveillance of tuberculosis in Europe. Working Group of the World Health Organization ({WHO)} and the European Region of the International Union Against Tuberculosis and Lung Disease ({IUATLD)} for uniform reporting on tuberculosis cases. *European Respiratory Journal*, **9**(5): 1097–1104.

Rindi L, Lari N, Garzelli C (2012) Large Sequence Polymorphisms of the Euro-American lineage of *Mycobacterium tuberculosis*: a phylogenetic reconstruction and evidence for convergent evolution in the DR locus. *Infection, Genetics and Evolution*, **12**(7): 1551–7.

Riska PF, Jacobs WRJ, Alland D (2000) Molecular determinants of drug resistance in tuberculosis. *The International Journal of Tuberculosis and Lung Disease*, **4**(2): S4–10.

Robinson ER, Walker TM, Pallen MJ (2013) Genomics and outbreak investigation: from sequence to consequence. *Genome Medicine*, **5**(4): 36.

Rocha EPC, Smith JM, Hurst LD, Holden MTG, Cooper JE, Smith NH, Feil EJ (2006) Comparisons of dN/dS are time dependent for closely related bacterial genomes. *Journal of Theoretical Biology*, **239**(2): 226–235.

Rodwell TC, Valafar F, Douglas J, Qian L, Garfein RS, Chawla A, Torres J, Zadorozhny V, Soo Kim M, Hoshide M, Catanzaro D, Jackson L, Lin G, Desmond E, Rodrigues C, Eisenach K, Victor TC, Ismail N, Crudu V, Gle MT, Catanzaro A (2013) Predicting Extensively Drug-resistant Tuberculosis (XDR-TB) Phenotypes with Genetic Mutations. *Journal Clinical Microbiology*, **52.**(3): 781–9.

Roetzer A, Diel R, Kohl TA, Rückert C, Nübel U, Blom J, Wirth T, Jaenicke S, Schuback S, Rüsch-Gerdes S, Supply P, Kalinowski J, Niemann S (2013) Whole Genome Sequencing versus Traditional Genotyping for Investigation of a *Mycobacterium tuberculosis* Outbreak: A Longitudinal Molecular Epidemiological Study. *PLoS Medicine*, **10**(2): e1001387.

Rothschild BM, Martin LD, Lev G, Bercovier H, Bar-Gal GK, Greenblatt C, Donoghue H, Spigelman M, Brittain D (2001) *Mycobacterium tuberculosis* Complex DNA from an Extinct Bison Dated 17,000 Years before the Present. *Clinical Infectious Diseases*, **33**(3): 305–311.

Rustad TR, Sherrid AM, Minch KJ, Sherman DR (2009) Hypoxia: a window into *Mycobacterium tuberculosis* latency. *Cellular Microbiology*, **11**(8): 1151–1159.

Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MAA, Barrell B (2000) Artemis: sequence visualization and annotation. *Bioinformatics*, **16**(10): 944–945.

Sackmann EK, Fulton AL, Beebe DJ (2014) The present and future role of microfluidics in biomedical research. *Nature*, **507**(7491): 181–189.

Salo WL, Aufderheide AC, Buikstra J, Holcomb TA (1994) Identification of *Mycobacterium tuberculosis* DNA in a pre-Columbian Peruvian mummy. *Proceedings of the National Academy of Sciences*, **91**(6): 2091–2094.

Samtools (2012).

Sandgren A, Strong M, Muthukrishnan P, Weiner BK, Church GM, Murray MB (2009) Tuberculosis Drug Resistance Mutation Database. *PLoS Medicine*, **6**(2): e1000002.

Sanger Institute (2012a) *Mycobacterium - Wellcome Trust Sanger Institute.*

Sanger Institute (2012b) *SMALT - Wellcome Trust Sanger Institute.*

Sassetti CM, Boyd DH, Rubin EJ (2003a) Genes required for mycobacterial growth defined by high density mutagenesis. *Molecular Microbiology*, **48**(1): 77–84.

Sassetti CM, Rubin EJ (2003b) Genetic requirements for mycobacterial survival during infection. *Proceedings of the National Academy of Sciences*, **100**(22): 12989–12994.

Schatz A, Bugle E, Waksman SA (1944) Streptomycin, a Substance Exhibiting Antibiotic Activity Against Gram-Positive and Gram-Negative Bacteria. *Experimental Biology and Medicine*, **55**(1): 66–69.

Schouten JP, McElgunn CJ, Waaijer R, Zwijnenburg D, Diepvens F, Pals G (2002) Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. *Nucleic Acids Research*, **30**(12): e57.

Schürch AC, Kremer K, Daviena O, Kiers A, Boeree MJ, Siezen RJ, Soolingen D van (2010) High-resolution typing by integration of genome sequencing data in a large tuberculosis cluster. *Journal of Clinical Microbiology*, **48**(9): 3403–3406.

Schürch AC, Kremer K, Hendriks ACA, Freyee B, McEvoy CRE, Crevel R van, Boeree MJ, Helden P van, Warren RM, Siezen RJ, Soolingen D van (2011) SNP/RD Typing of *Mycobacterium tuberculosis* Beijing Strains Reveals Local and Worldwide Disseminated Clonal Complexes. *PLoS ONE*, **6**(12): e28365.

Schürch AC, Soolingen D van (2012) DNA fingerprinting of *Mycobacterium tuberculosis*: From phage typing to whole-genome sequencing. *Infection, Genetics and Evolution*, **12**(4): 602–609.

Selgelid MJ (2008) Ethics, Tuberculosis and Globalization. *Public Health Ethics*, **1**(1): 10–20.

Sharma D, Surolia A (2011) Computational tools to study and understand the intricate biology of mycobacteria. *Tuberculosis (Edinburgh, Scotland)*, **91**(3): 273–276.

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, **29**(1): 308–311.

Small PM, Hopewell PC, Singh SP, Paz A, Parsonnet J, Ruston DC, Schecter GF, Daley CL, Schoolnik GK (1994) The epidemiology of tuberculosis in San Francisco. A population-based study using conventional and molecular methods. *New England Journal of Medicine*, **330**(24): 1703–1709.

Smith NH, Hewinson RG, Kremer K, Brosch R, Gordon SV (2009) Myths and misconceptions: the origin and evolution of *Mycobacterium tuberculosis*. *Nature Reviews. Microbiology*, **7**(7): 537–544.

Smith NH, Kremer K, Inwald J, Dale J, Driscoll JR, Gordon SV, Soolingen D van, Hewinson RG, Smith JM (2006) Ecotypes of the *Mycobacterium tuberculosis* complex. *Journal of Theoretical Biology*, **239**(2): 220–225.

Snitkin ES, Zelazny AM, Thomas PJ, Stock F, Henderson DK, Palmore TN, Segre JA (2012) Tracking a hospital outbreak of carbapenem-resistant {K}lebsiella pneumoniae with whole-genome sequencing. *Science Translational Medicine*, **4**(148): 148ra116.

Snyder LA, Loman NJ, Faraj LA, Levi K, Weinstock G, Boswell TC, Pallen MJ, Ala'Aldeen DA (2013) Epidemiological investigation of *Pseudomonas aeruginosa* isolates from a six-year-long hospital outbreak using high-throughput whole genome sequencing. *Euro Surveillance*, **18**(42): pii: 20611.

Sola C, Filliol I, Legrand E, Mokrousov I, Rastogi N (2001) *Mycobacterium tuberculosis* Phylogeny Reconstruction Based on Combined Numerical Analysis with IS1081, IS6110, VNTR, and DR-Based Spoligotyping Suggests the Existence of Two New Phylogeographical Clades. *Journal of Molecular Evolution*, **53**(6): 680–689.

Song J, Li PE, Gans J, Vuyisich M, Deshpande A, Wolinsky M, White PS (2010) Simultaneous pathogen detection and antibiotic resistance characterization using SNP-based multiplexed oligonucleotide ligation-PCR (MOL-PCR). *Advances in Experimental Medicine and Biology*, **680** 455–464.

Spuy GD van der, Kremer K, Ndabambi SL, Beyers N, Dunbar R, Marais BJ, Helden PD van, Warren RM (2009) Changing *Mycobacterium tuberculosis* population highlights clade-specific pathogenic characteristics. *Tuberculosis (Edinburgh, Scotland)*, **89**(2): 120–125.

Sreevatsan S, Pan X, Stockbauer KE, Connell ND, Kreiswirth BN, Whittam TS, Musser JM (1997) Restricted structural gene polymorphism in the *Mycobacterium tuberculosis* complex indicates evolutionarily recent global dissemination. *Proceedings of the National Academy of Sciences*, **94**(18): 9869–9874.

Stein LD (2010) The case for cloud computing in genome informatics. *Genome Biology*, **11**(5): 207.

Stitziel NO, Binkowski TA, Tseng YY, Kasif S, Liang J (2004) topoSNP: a topographic database of non-synonymous single nucleotide polymorphisms with and without known disease association. *Nucleic Acids Research*, **32**(Database issue): D520–D522.

Stucki D, Gagneux S (2013) Single nucleotide polymorphisms in *Mycobacterium tuberculosis* and the need for a curated database. *Tuberculosis (Edinburgh, Scotland)*, Featuring

Reports From The Tuberculosis Community Annotation Project Jamboree held at Virginia Tech, USA, March 7-8, 2012 **93**(1): 30–39.

Stucki D, Malla B, Hostettler S, Huna T, Feldmann J, Yeboah-Manu D, Borrell S, Fenner L, Comas I, Coscollà M, Gagneux S (2012) Two New Rapid SNP-Typing Methods for Classifying *Mycobacterium tuberculosis* Complex into the Main Phylogenetic Lineages. *PLoS ONE*, **7**(7): e41253.

Supply P, Mazars E, Lesjean S, Vincent V, Gicquel B, Locht C (2000) Variable human minisatellite-like regions in the *Mycobacterium tuberculosis* genome. *Molecular Microbiology*, **36**(3): 762–771.

Supply P, Allix C, Lesjean S, Cardoso-Oelemann M, Rüsch-Gerdes S, Willery E, Savine E, Haas P de, Deutekom H van, Roring S, Bifani P, Kurepina N, Kreiswirth B, Sola C, Rastogi N, Vatin V, Gutierrez MC, Fauville M, Niemann S, Skuce R, Kremer K, Locht C, Soolingen D van (2006) Proposal for standardization of optimized mycobacterial interspersed repetitive unit-variable-number tandem repeat typing of *Mycobacterium tuberculosis*. *Journal of Clinical Microbiology*, **44**(12): 4498–4510.

Supply P, Marceau M, Mangenot S, Roche D, Rouanet C, Khanna V, Majlessi L, Criscuolo A, Tap J, Pawlik A, Fiette L, Orgeur M, Fabre M, Parmentier C, Frigui W, Simeone R, Boritsch EC, Debrie AS, Willery E, Walker D, Quail MA, Ma L, Bouchier C, Salvignol G, Sayes F, Cascioferro A, Seemann T, Barbe V, Locht C, Gutierrez MC, Leclerc C, Bentley SD, Stinear TP, Brisse S, Médigue C, Parkhill J, Cruveiller S, Brosch R (2013) Genomic analysis of smooth tubercle bacilli provides insights into ancestry and pathoadaptation of *Mycobacterium tuberculosis*. *Nature Genetics*, **45**(2): 172–179.

Supply P, Warren RM, Bañuls AL, Lesjean S, Van Der Spuy GD, Lewis LA, Tibayrenc M, Van Helden PD, Locht C (2003) Linkage disequilibrium between minisatellite loci supports clonal evolution of *Mycobacterium tuberculosis* in a high tuberculosis incidence area. *Molecular Microbiology*, **47**(2): 529–538.

Takiff HE, Salazar L, Guerrero C, Philipp W, Huang WM, Kreiswirth B, Cole ST, Jacobs WR, Telenti A (1994) Cloning and nucleotide sequence of *Mycobacterium tuberculosis* gyrA and gyrB genes and detection of quinolone resistance mutations. *Antimicrobial Agents and Chemotherapy*, **38**(4): 773–780.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution*, **28**(10): 2731–2739.

Tan SJ, Phan H, Gerry BM, Kuhn A, Hong LZ, Min Ong Y, Poon PSY, Unger MA, Jones RC, Quake SR, Burkholder WF (2013) A Microfluidic Device for Preparing Next

Generation DNA Sequencing Libraries and for Automating Other Laboratory Protocols That Require One or More Column Chromatography Steps. *PLoS ONE*, **8**(7): e64084.

Telenti A (1997) Genetics of drug resistance in tuberculosis. *Clinics in Chest Medicine*, **18**(1): 55–64.

Thorisson GA, Lancaster O, Free RC, Hastings RK, Sarmah P, Dash D, Brahmachari SK, Brookes AJ (2009) HGVbaseG2P: a central genetic association database. *Nucleic Acids Research*, **37**(Database issue): D797–802.

Thwaites G, Caws M, Chau TTH, D'Sa A, Lan NTN, Huyen MNT, Gagneux S, Anh PTH, Tho DQ, Torok E, Nhu NTQ, Duyen NTH, Duy PM, Richenberg J, Simmons C, Hien TT, Farrar J (2008) Relationship between *Mycobacterium tuberculosis* genotype and the clinical phenotype of pulmonary and meningeal tuberculosis. *Journal of Clinical Microbiology*, **46**(4): 1363–1368.

Török ME, Peacock SJ (2012) Rapid whole-genome sequencing of bacterial pathogens in the clinical microbiology laboratory—pipe dream or reality? *Journal of Antimicrobial Chemotherapy*, **67**(10): 2307–2308.

Traore B, Diarra B, Dembele BPP, Somboro AM, Hammond AS, Siddiqui S, Maiga M, Kone B, Sarro YS, Washington J, Parta M, Coulibaly N, M'baye O, Diallo S, Koita O, Tounkara A, Polis MA (2012) Molecular strain typing of *Mycobacterium tuberculosis* complex in Bamako, Mali. *The International Journal of Tuberculosis and Lung Disease*, **16**(7): 911–916.

Tsolaki AG, Hirsh AE, DeRiemer K, Enciso JA, Wong MZ, Hannan M, Goguet de la Salmoniere YOL, Aman K, Kato-Maeda M, Small PM (2004) Functional and evolutionary genomics of *Mycobacterium tuberculosis*: insights from genomic deletions in 100 strains. *Proceedings of the National Academy of Sciences*, **101**(14): 4865–4870.

Tsukamura M (1976) Numerical Classification of Slowly Growing Mycobacteria. *International Journal of Systematic Bacteriology*, **26**(4): 409–420.

Vallone PM, Butler JM (2004) AutoDimer: a screening tool for primer-dimer and hairpin structures. *BioTechniques*, **37**(2): 226–231.

Viegas SO, Machado A, Groenheit R, Ghebremichael S, Pennhag A, Gudo PS, Cuna Z, Miotto P, Hill V, Marrufo T, Cirillo DM, Rastogi N, Källenius G, Koivula T (2010) Molecular diversity of *Mycobacterium tuberculosis* isolates from patients with pulmonary tuberculosis in Mozambique. *BMC Microbiology*, **10** 195.

Vishnoi A, Srivastava A, Roy R, Bhattacharya A (2008) MGDD: *Mycobacterium tuberculosis* genome divergence database. *BMC genomics*, **9** 373.

Wada T, Iwamoto T, Hase A, Maeda S (2012) Scanning of genetic diversity of evolutionarily sequential *Mycobacterium tuberculosis* Beijing family strains based on genome wide analysis. *Infection, Genetics and Evolution*, **12**(7): 1392–1396.

Walker TM, Monk P, Smith EG, Peto TEA (2013a) Contact investigations for outbreaks of *Mycobacterium tuberculosis*: advances through whole genome sequencing. *Clinical Microbiology and Infection*, **19**(9): 796–802.

Walker TM, Ip CLC, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, Eyre DW, Wilson DJ, Hawkey PM, Crook DW, Parkhill J, Harris D, Walker AS, Bowden R, Monk P, Smith EG, Peto TEA (2013b) Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study. *The Lancet Infectious Diseases*, **13**(2): 137–146.

Walker TM, Lalor MK, Broda A, Ortega LS, Morgan M, Parker L, Churchill S, Bennett K, Golubchik T, Giess AP, Del Ojo Elias C, Jeffery KJ, Bowler ICJW, Laurenson IF, Barrett A, Drobniewski F, McCarthy ND, Anderson LF, Abubakar I, Thomas HL, Monk P, Smith EG, Walker AS, Crook DW, Peto TEA, Conlon CP (2014) Assessment of *Mycobacterium tuberculosis* transmission in Oxfordshire, UK, 2007–12, with whole pathogen genome sequences: an observational study. *The Lancet Respiratory Medicine*, **2**(4): 285–292.

Wampande EM, Mupere E, Debanne SM, Asiimwe BB, Nsereko M, Mayanja H, Eisenach K, Kaplan G, Boom HW, Sebastien G, Joloba ML (2013) Long-term dominance of *Mycobacterium tuberculosis* Uganda family in peri-urban Kampala-Uganda is not associated with cavitary disease. *BMC Infectious Diseases*, **13**(1): 484.

Wang K, Li M, Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, **38**(16): e164–e164.

Wang L, Luhm R, Lei M (2007) SNP and mutation analysis. *Advances in Experimental Medicine and Biology*, **593** 105–116.

Warner DF, Mizrahi V (2013) Complex genetics of drug resistance in *Mycobacterium tuberculosis*. *Nature Genetics*, **45**(10): 1107–1108.

Weisenberg SA, Gibson AL, Huard RC, Kurepina N, Bang H, Lazzarini LCO, Chiu Y, Li J, Ahuja S, Driscoll J, Kreiswirth BN, Ho JL (2012) Distinct clinical and epidemiological features of tuberculosis in New York City caused by the RD(Rio) *Mycobacterium tuberculosis* sublineage. *Infection, Genetics and Evolution*, **12**(4): 664–670.

Weniger T, Krawczyk J, Supply P, Harmsen D, Niemann S (2012) Online tools for polyphasic analysis of *Mycobacterium tuberculosis* complex genotyping data: now and next. *Infection, Genetics and Evolution*, **12**(4): 748–754.

Weniger T, Krawczyk J, Supply P, Niemann S, Harmsen D (2010) MIRU-VNTRplus: a web tool for polyphasic genotyping of *Mycobacterium tuberculosis* complex bacteria. *Nucleic Acids Research*, **38**(W): 326–331.

Wetterstrand KA (2012) *DNA Sequencing Costs: Data from the NHGRI Large-Scale Genome Sequencing Program.*

Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT, Gomes X, Tartaro K, Niazi F, Turcotte CL, Irzyk GP, Lupski JR, Chinault C, Song Xz, Liu Y, Yuan Y, Nazareth L, Qin X, Muzny DM, Margulies M, Weinstock GM, Gibbs RA, Rothberg JM (2008) The complete genome of an individual by massively parallel DNA sequencing. *Nature*, **452**(7189): 872–876.

WHO (2010) *Guidelines for treatment of tuberculosis, fourth edition.*

WHO (2011) *Global tuberculosis report 2011.*

WHO (2013) *Global tuberculosis report 2013.*

Wirth T, Hildebrand F, Allix-Béguec C, Wölbeling F, Kubica T, Kremer K, Soolingen D van, Rüsch-Gerdes S, Locht C, Brisse S, Meyer A, Supply P, Niemann S (2008) Origin, Spread and Demography of the *Mycobacterium tuberculosis* Complex. *PLoS Pathogens*, **4**(9):

Yeboah-Manu D, Asante-Poku A, Bodmer T, Stucki D, Koram K, Bonsu F, Pluschke G, Gagneux S (2011) Genotypic diversity and drug susceptibility patterns among *M. tuberculosis* complex isolates from South-Western Ghana. *PLoS ONE*, **6**(7): e21906.

Yesilkaya H, Meacci F, Niemann S, Hillemann D, Rüsch-Gerdes S, Barer MR, Andrew PW, Oggioni MR (2006) Evaluation of molecular-Beacon, TaqMan, and fluorescence resonance energy transfer probes for detection of antibiotic resistance-conferring single nucleotide polymorphisms in mixed *Mycobacterium tuberculosis* DNA extracts. *Journal of Clinical Microbiology*, **44**(10): 3826–3829.

Ying YL, Zhang J, Gao R, Long YT (2013) Nanopore-Based Sequencing and Detection of Nucleic Acids. *Angewandte Chemie International Edition*, **52**(50): 13154–13161.

Yu Y, Harris AJ, He X (2010) S-DIVA (Statistical Dispersal-Vicariance Analysis): A tool for inferring biogeographic histories. *Molecular Phylogenetics and Evolution*, **56**(2): 848–850.

Yuan L, Huang Y, Mi LG, Li YX, Liu PZ, Zhang J, Liang HY, Li F, Li H, Zhang SQ, Li WJ (2014) There is no correlation between sublineages and drug resistance of *Mycobacterium tuberculosis* Beijing/W lineage clinical isolates in Xinjiang, China. *Epidemiology & Infection*, 1–9.

Zankari E, Hasman H, Kaas RS, Seyfarth AM, Agersø Y, Lund O, Larsen MV, Aarestrup FM (2013) Genotyping using whole-genome sequencing is a realistic alternative

to surveillance based on phenotypic antimicrobial susceptibility testing. *The Journal of Antimicrobial Chemotherapy*, **68**(4): 771–777.

Zenner D, Southern J, Hest R van, DeVries G, Stagg HR, Antoine D, Abubakar I (2013) Active case finding for tuberculosis among high-risk groups in low-incidence countries. *The International Journal of Tuberculosis and Lung Disease*, **17**(5): 573–582.

Zhang Y, Yew WW (2009) Mechanisms of drug resistance in *Mycobacterium tuberculosis*. *The International Journal of Tuberculosis and Lung Disease*, **13**(11): 1320–1330.

Zhi J (2010) MAPD: a probe design suite for multiplex ligation-dependent probe amplification assays. *BMC Research Notes*, **3**(1): 137.

Zhi J, Hatchwell E (2008) Human MLPA Probe Design (H-MAPD): a probe design tool for both electrophoresis-based and bead-coupled human multiplex ligation-dependent probe amplification assays. *BMC Genomics*, **9**(1): 407.

Zink AR, Sola C, Reischl U, Grabner W, Rastogi N, Wolf H, Nerlich AG (2003) Characterization of *Mycobacterium tuberculosis* complex DNAs from Egyptian mummies by spoligotyping. *Journal of Clinical Microbiology*, **41**(1): 359–367.

Abbreviations

AF	Allele frequency
AR	Allelic ratio
BCG	Bacille de Calmette et Guérin
CAS	Central Asian
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
DOT(S)	Directly Observed Treatment (Short-Course)
DR	Direct Repeat
EAI	East-African-Indian
HIV	Human Immunodeficiency Virus
InDel	Insertion Deletion
IQR	Interquartile Range
IS	Insertion Sequence
LAM	Latin American Mediterranean
LHS	Left Hybridising Sequence
LPO	Left Probe Oligonucleotide
LSP	Large Sequence Polymorphism
MIRU	Mycobacterial Interspersed Repetitive Units
ML	Maximum Likelihood
MOL-PCR	Multiplexed Oligonucleotide Ligation PCR
MTB	Mycobacterium tuberculosis
MTBC	Mycobacterium tuberculosis complex
NGS	Next Generation Sequencing
PCA	Principal Component Analysis
PE	Proline Glutamine
PPE	Proline Proline Glutamine
PGRS	Polymorphic GC-rich Repetitive Sequences
PGG	Principal Genetic Group
QRDR	Quinolone Resistance Determining Region
RD	Region of Difference
RFLP	Restriction Fragment Length Polymorphism
RPO	Right Probe Oligonucleotide
RHS	Right Hybridising Sequence
RRDR	Rifampicin Resistance Determining Region
SCG	SNP Cluster Group
SNP	Single Nucleotide Polymorphism
Spoligotyping	Spacer Oligonucleotide Typing
SRA	Sequence Read Archive
TB	Tuberculosis
VCF	Variant Call Format
VNTR	Variable Number of Tandem Repeats
WGS	Whole Genome Sequencing
WHO	World Health Organization

List of Figures

1.1. TB incidence rates per country in 2012 as estimated by WHO 3
1.2. The global burden of TB as estimated by WHO 3
1.3. Phylogeny of the genus *Mycobacterium* using 16S rRNA sequences 7
1.4. The current phylogeny of 420 whole genome sequences of the MTBC 7
1.5. Proposed evolutionary pathway of the tubercle bacilli 8

3.1. Number of MTBC samples with genome sequences in the SRA/NCBI . . . 21
3.2. Phylogenetic tree of 22 whole genome sequences of MTBC 25
3.3. Example of a genome (re-) sequencing analysis pipeline for MTBC 31
3.4. A simplified scheme of a future SNP-database for MTBC SNPs 37

4.1. Phylogenetic tree of 22 whole genome sequences of MTBC 46
4.2. Schematic illustration of MOL-PCR 50

5.1. Simplified overview of scanning process and preparation 73
5.2. Interactive inspection of `KvarQ` "coverage" 75
5.3. Interactive inspection of `json` file 75
5.4. FastQ dataset used to validate KvarQ 77
5.5. KvarQ scanning time 78
5.6. Phylogenetic classification of all 880 isolates used in `KvarQ` 79
5.7. MTBC drug resistance associated mutations found in all 880 isolates 81
5.8. Drug resistance associated mutations identified by `KvarQ` 82
5.9. Main MTBC phylogenetic lineage classification by `KvarQ` 82

6.1. Overview of patient isolates and whole genome sequences 94
6.2. Initial neighbor joining phylogeny of *M. tuberculosis* isolates 95
6.3. Epidemic curve of the 68 patients 99
6.4. Distribution of tuberculosis cluster patients in the milieu 103
6.5. Median Joining genomic network of the "Bernese outbreak" 104

7.1. Today's global phylogeography of the MTBC 116

7.2. Phylogeny of 72 whole genome sequences of Lineage 4 strains 126
7.3. Principal component analysis of 9455 SNPs in 72 Lineage 4 strains 127
7.4. Nucleotide diversity of MTB Lineage 4 sublineages 128
7.5. Phylogenetic tree of 72 whole genome sequences, plus additional markers . 129
7.6. MTB Lineage 4 sublineage frequencies 130
7.7. Lineage 4 sublineages were found globally distributed 131
7.8. Heat maps of the proportions of the different sublineages 132
7.9. Country frequencies of Lineage 4 sublineages 133
7.10. New MTB Lineage 4 diversity 134
7.11. Maximum likelihood phylogeny of 150 isolates 136
7.12. Reconstruction of the geographical origin of the LAM ancestor 137
7.13. Genetic diversity of LAM isolates per continent and world region 138
7.14. Proportion of Lineage 4 among all isolates shown by country 139
7.15. MTB Lineage 4 dispersal scenario 145

List of Tables

3.1. Relevant SNP databases for MTBC 28

4.1. Oligonucleotides used for Luminex MOL-PCR 53
4.2. Primer and probe sequences for TaqMan assays 56
4.3. Comparison of MOL-PCR and TaqMan 59
4.4. Lineage assignments of MOL-PCR, TaqMan and spoligotyping 61

5.1. Comparison of SNP-calls by `KvarQ` and BWA 79

6.1. Comparison of tuberculosis outbreak tracking methods 100
6.2. Characteristics of cases confirmed to be in the tuberculosis cluster 102

7.1. Sublineage-specific SNPs and oligonucleotides used for MOL-PCR 121
7.2. PCR primers for Lineage 4 sublineage specific SNPs 122
7.3. Mean pairwise SNP distances of defined sublineages 125
7.4. Pairwise F_{ST} values between MTB lineage 4 sublineages 127
7.5. Proportions of isolates from each continent in the subgroups 135

A.1. SNP markers used to identify main MTBC phylogenetic lineages in `KvarQ` 196
A.2. Drug resistance markers used in KvarQ 197

A. Appendix

A.1. Appendix to Chapter 5 (KvarQ)

Additional Information 1—List of SNPs used as markers for phylogenetic classification of MTBC, and drug resistance markers used in this study. Tables A.1 and A.2 are shown below, and the file "Additional_File_1.xlsx" including these two tables is available from the authors upon request.

Additional Information 2—List of 880 `fastq` files used in this study, additional information and `KvarQ` results for all files. Data not shown here due to format limitations, but file "Additional_File_2.xlsx" is available from the authors upon request.

Table A.1.: **SNP markers used to identify main MTBC phylogenetic lineages in `KvarQ`.**

Phylogenetic Lineage	Position Genome H37Rv	Ancestral Allele	Mutant Allele
Lineage 1	3920109	G	T
Lineage 1	3597682	C	T
Lineage 1	1590555	C	T
Lineage 2	1834177	A	C
Lineage 2	3304966	G	A
Lineage 2	2711722	T	G
Lineage 3	301341	C	A
Lineage 3	4266647	A	G
Lineage 3	157129	C	T
Lineage 4	3326554	C	A
Lineage 4	2154724	A	C
Lineage 4	648856	C	T
Lineage 5	1377185	C	G
Lineage 5	801959	C	T
Lineage 5	2859147	C	T
Lineage 6	2427828	G	C
Lineage 6	378404	G	A
Lineage 6	4269522	G	A
Lineage 7	14806	T	C
Lineage 7	1663221	T	G
Lineage 7	497126	G	A
animal associated MTBC (*M. bovis* / *M. caprae*)	3480645	T	G
animal associated MTBC (*M. bovis* / *M. caprae*)	1427476	C	T
animal associated MTBC (*M. bovis* / *M. caprae*)	3624593	C	T
"Beijing" sublineage of Lineage 2	2112832	A	C
"Beijing" sublineage of Lineage 2	3587446	G	A
"Beijing" sublineage of Lineage 2	1849051	C	T

Table A.2.: **Drug resistance markers used in KvarQ.**

Resistance to Drug	Description	Position H37Rv	Ancestral Allele	Mutant Allele	Codon or Base Change	Reference
Isoniazid	katG mutation codon 315	2155167	G	T	S315 = GCT on + strand (S=AGC on − strand)	Ramaswamy *et al.* 1998 / TBDReamDB
Isoniazid	katG mutation codon 315	2155167	G	C	S315 = GCT on + strand (S=AGC on − strand)	Ramaswamy *et al.* 1998 / TBDReamDB
Isoniazid	katG mutation codon 315	2155168	C	A	S315 = GCT on + strand (S=AGC on − strand)	Ramaswamy *et al.* 1998 / TBDReamDB
Isoniazid	katG mutation codon 315	2155168	C	T	S315 = GCT on + strand (S=AGC on − strand)	Ramaswamy *et al.* 1998 / TBDReamDB
Isoniazid	katG mutation codon 315	2155168	C	G	S315 = GCT on + strand (S=AGC on − strand)	Ramaswamy *et al.* 1998 / TBDReamDB
Isoniazid	katG mutation codon 315	2155169	T	A	S315 = GCT on + strand (S=AGC on − strand)	Ramaswamy *et al.* 1998 / TBDReamDB
Isoniazid	katG mutation codon 315	2155169	T	C	S315 = GCT on + strand (S=AGC on − strand)	Ramaswamy *et al.* 1998 / TBDReamDB
Isoniazid	katG mutation codon 315	2155169	T	G	S315 = GCT on + strand (S=AGC on − strand)	Ramaswamy *et al.* 1998 / TBDReamDB
Isoniazid	InhA promoter region -8	1673432	T	A		TBDReamDB
Isoniazid	InhA promoter region -8	1673432	T	C		TBDReamDB
Isoniazid	InhA promoter region -15	1673425	C	T		TBDReamDB
Rifampicin	non-synonymous mutation in rpoB "RRDR" region	761082 – 761162	ggcaccagccagctgagccaattcatggac-cagaacaacccgctgtcggggttgacccacaagcgccgactgtcggcgctg		426 – 452 (E. coli 507 - 533)	Ramaswamy *et al.*, Tuber Lung Dis 1998 / TBDReamDB
Rifampicin / comp. mut.	rpoA comp. mut. for RIFr	3877949	T	C	T187A	Comas *et al.*, Nat Gen 2012
Rifampicin / comp. mut.	rpoA comp. mut. for RIFr	3877949	T	G	T187P	Comas *et al.*, Nat Gen 2012
Rifampicin / comp. mut.	rpoA comp. mut. for RIFr	3877960	A	G	V183A	Comas *et al.*, Nat Gen 2012
Rifampicin / comp. mut.	rpoA comp. mut. for RIFr	3877960	A	C	V183G	Comas *et al.*, Nat Gen 2012
Rifampicin / comp. mut.	rpoC comp. mut. for RIFr	764669	C	G	P434A	Comas *et al.*, Nat Gen 2012
Rifampicin / comp. mut.	rpoC comp. mut. for RIFr	764670	C	G	P434R	Comas *et al.*, Nat Gen 2012
Rifampicin / comp. mut.	rpoC comp. mut. for RIFr	764817	T	C	V483A	Comas *et al.*, Nat Gen 2012
Rifampicin / comp. mut.	rpoC comp. mut. for RIFr	764817	T	G	V483G	Comas *et al.*, Nat Gen 2012
Rifampicin / comp. mut.	rpoC comp. mut. for RIFr	764819	T	G	W484G	Comas *et al.*, Nat Gen 2012
Rifampicin / comp. mut.	rpoC comp. mut. for RIFr	764822	T	A	D485N	Comas *et al.*, Nat Gen 2012
Rifampicin / comp. mut.	rpoC comp. mut. for RIFr	764822	T	C	D485H	Comas *et al.*, Nat Gen 2012
Rifampicin / comp. mut.	rpoC comp. mut. for RIFr	764840	A	G	I491V	Comas *et al.*, Nat Gen 2012
Rifampicin / comp. mut.	rpoC comp. mut. for RIFr	764841	T	C	I491T	Comas *et al.*, Nat Gen 2012
Rifampicin / comp. mut.	rpoC comp. mut. for RIFr	764918	G	C	V517L	Comas *et al.*, Nat Gen 2012
Rifampicin / comp. mut.	rpoC comp. mut. for RIFr	765462	A	G	N698S	Comas *et al.*, Nat Gen 2012
Rifampicin / comp. mut.	rpoC comp. mut. for RIFr	765463	C	G	N698K	Comas *et al.*, Nat Gen 2012
Fluoroquinolone	non-synonymous mutation in gyrA "QRDR" region	7521 – 7583	gcccggtcggttgccgagaccatgggcaacta-ccacccgcacggcgacgcgtcgatctacgac		74 - 94	Sun *et al.*, Antimicr Agents 2008 / TBDReamDB
Fluoroquinolone	non-QRDR gyrA mutation	7606	C	A	102 P>	TBDReamDB
Fluoroquinolone	non-QRDR gyrA mutation	7677	G	A	126 A> R	TBDReamDB
Fluoroquinolone	non-QRDR gyrA mutation	7678	C	G	126 A> R	TBDReamDB
Fluoroquinolone	gyrB codon 510 mutation	6767	G	A	510 AAC>GAC	TBDReamDB
Fluoroquinolone	gyrB codon 510 mutation	6768	G	A	510 AAC>GAC	TBDReamDB
Streptomycin	rpsL mutation	781687	A	G	K43R	TBDReamDB
Streptomycin	rpsL mutation	781822	A	C	K88T	TBDReamDB
Streptomycin	rpsL mutation	781822	A	T	K88R	TBDReamDB
Streptomycin	rpsL mutation	781822	A	G	K88?	TBDReamDB
Streptomycin	rrs mutation	1472337	C	A	491	TBDReamDB
Streptomycin	rrs mutation	1472337	C	G	491	TBDReamDB
Streptomycin	rrs mutation	1472337	C	T	491	TBDReamDB
Streptomycin	rrs mutation	1472358	C	A	512	TBDReamDB
Streptomycin	rrs mutation	1472358	C	G	512	TBDReamDB
Streptomycin	rrs mutation	1472358	C	T	512	TBDReamDB
Streptomycin	rrs mutation	1472359	A	C	513	TBDReamDB
Streptomycin	rrs mutation	1472359	A	G	513	TBDReamDB
Streptomycin	rrs mutation	1472359	A	T	513	TBDReamDB
Streptomycin	rrs mutation	1472362	C	A	516	TBDReamDB
Streptomycin	rrs mutation	1472362	C	G	516	TBDReamDB
Streptomycin	rrs mutation	1472362	C	T	516	TBDReamDB
Streptomycin	rrs mutation	1472752	A	C	906	TBDReamDB
Streptomycin	rrs mutation	1472752	A	G	906	TBDReamDB
Streptomycin	rrs mutation	1472752	A	T	906	TBDReamDB
Kanamycin and Amikacin	rrs mutation	1473246	A	C	1401	TBDReamDB
Kanamycin and Amikacin	rrs mutation	1473246	A	G	1401	TBDReamDB
Kanamycin and Amikacin	rrs mutation	1473246	A	T	1401	TBDReamDB
Kanamycin and Amikacin	rrs mutation	1473247	C	A	1402	TBDReamDB
Kanamycin and Amikacin	rrs mutation	1473247	C	G	1402	TBDReamDB
Kanamycin and Amikacin	rrs mutation	1473247	C	T	1402	TBDReamDB
Ethambutol	embB mutation	4247429	A	G	M306V	TBDReamDB
Ethambutol	embB mutation	4247431	G	A	M306I	TBDReamDB
Ethambutol	embB mutation	4247431	G	T	M306I	TBDReamDB
Ethambutol	embB mutation	4247431	G	C	M306I	TBDReamDB
Ethambutol	embB mutation	4247429	A	C	M306L	TBDReamDB
Ethambutol	embB mutation	4247730	G	C	G406A	TBDReamDB
Ethambutol	embB mutation	4248003	A	G	Q497R	TBDReamDB

A.2. Appendix to Chapter 6 (Transmission of MTBC in an outbreak in Bern)

Supplementary Materials in the submission format (Supplementary Information, Supplementary Tables and Supplementary Figures) is shown below.

SUPPLEMENTARY INFORMATION

MIRU-VNTR and Spoligotyping

MIRU-VNTR analysis for 24 loci was performed according to standard protocols (Genoscreen, Lille, France). Spoligotyping patterns were obtained from WGS with SpolPred [1] and confirmed with bead-based spoligotyping [2].

Sixty-four of 69 (92.8 %) isolates (68 single patient isolates plus the additional isolate of the key patient P006) were found identical by MIRU-VNTR pattern. Three isolates showed a double-band, which likely reflects technical issues, and two isolates had a different number of repeats at one locus each.

Spoligotyping patterns were identical for all isolates except P073, for which two additional spacers were found absent. The “consensus” spoligotyping pattern was identified as “S” family, a sublineage of the Euro-American lineage of *M. tuberculosis* (Supplementary Table 3, see below).

Excluded SNP calls

Thirty-seven additional SNPs were excluded with gaps in more than three strains (i.e. no base call) or that had ambiguous SNP-calls, identified upon manual verification.

Sequencing replicates

As controls, we included four biological replicates for WGS (independent DNA extraction, sequencing library preparation and WGS) and did not detect any discrepant SNPs between replicates. To control for sequencing errors, we sequenced twice the same genomic library of three isolates and obtained identical SNP-calls.

Sanger sequencing for SNP confirmation

To validate SNP-calls from WGS, we performed PCR and Sanger sequencing on a subset of 28 mutations identified by WGS in the corresponding strain with the mutation identified by WGS. All WGS-identified mutations were confirmed by Sanger sequencing.

Drug resistance results

We used the Bactec MGIT 960 system (Becton Dickinson Diagnostic Systems, Sparks, MD, USA) for semi-quantitative drug-susceptibility testing (DST) to first-line drugs as previously described [3]. No phenotypic drug resistance was detected in the 68 single patient isolates and the second isolate of the key patient P006 that were included in the genomic analyses. However, three additional isolates were available from recurrent tuberculosis episodes of patient P006, including the isolate P006C ("2R" in the original figure, isolated 1992, described as isoniazid mono-resistant in the article of Genewein et al. [4]), and the isolates P006D (patient "2", isolated 1992, described as resistant to isoniazid and rifampicin), and P006E (isolated 1993, also resistant to isoniazid and rifampicin). For all isolates, we confirmed the resistance patterns described in 1993 [4]. Of note, despite the multiple tuberculosis episodes of the key patient over four years, none of the isolates with the evolved drug-resistant mutations was transmitted.

We found mutations associated with resistance to isoniazid (*katG* S315N, 2155168CT), rifampicin (*rpoB* D435Y, 761109GT) and ethambutol (*embB* Q497P, 4248003AC) in the last two of the serial isolates of patient P006 (P006D and P006E) [5]. Phenotypic isoniazid mono-resistance was identified in the third isolate of patient P006 (P006C), as originally described by Genewein et al. [4]. The corresponding *katG* S315N mutation was only found in the DNA extracted from the DST culture directly, but not in the DNA extracted from the routine subculture and used for WGS. This indicates a mixed population of sensitive and resistant variants in the original isolate, and the selection for the resistant isolate in the DST.

No other genotypic or phenotypic drug resistance was found in any other isolate. All drug resistance associated mutations were excluded for analyses of the genomic network.

Supplementary Table 1. Pairwise comparison of Single Nucleotide Polymorphism (SNPs) positions across whole genome sequences of 69 *Mycobacterium tuberculosis* “Bernese cluster” isolates.

	P034	P069A	P067	P006A	P010	P053	P043	P056	P047	P016	P072	P035	P004	P007	P066	P059	P005	P033	P070	P003	P028A	P011	P001	P012	P077	P051	P026	P073	P058	P017	P015	P061	P014	P039	P041	P040	P023	P002	P0B6	P008	P013	P024	P036	P074	P037	P064	P038	P019	P063	P065	P025	P045	P021	P020	P054	P052	P022A	P062	P076	P049	P050	P027	P042	P009	P018	P048	P044	P055	P075
P034																																																																					
P069A	8																																																																				
P067	9	11																																																																			
P006A	9	11	10																																																																		
P010	9	11	10	0																																																																	
P053	10	12	11	1	1																																																																
P043	11	13	12	2	2	3																																																															
P056	11	13	12	2	2	3	2																																																														
P047	12	14	13	3	3	4	3	1																																																													
P016	12	14	13	3	3	4	3	1	2																																																												
P072	13	15	14	4	4	5	6	6	7	7																																																											
P035	12	14	13	3	3	4	5	5	6	6	3																																																										
P004	12	14	13	3	3	4	5	5	6	6	7	6																																																									
P007	13	15	14	4	4	5	6	6	7	7	8	7	7																																																								
P066	12	14	13	3	3	4	5	5	6	6	7	6	6	7																																																							
P059	10	12	11	1	1	2	3	3	4	4	5	4	4	5	4																																																					
P005	10	12	11	1	1	2	3	3	4	4	5	4	4	5	4	2																																																				
P033	12	14	13	3	3	4	5	5	6	6	7	6	6	7	6	4	4																																																			
P070	11	13	12	2	2	3	4	4	5	5	6	5	5	6	5	3	3	5																																																		
P003	12	14	13	3	3	4	5	5	6	6	7	6	6	7	6	4	4	6	5																																																	
P028A	10	12	11	1	1	2	3	3	4	4	5	4	4	5	4	2	2	4	3	4																																																
P011	17	19	18	8	8	9	10	10	11	11	12	11	11	12	11	9	9	11	10	11	9																																														
P001	12	14	13	3	3	4	5	5	6	6	7	6	6	7	6	4	4	6	5	6	4	5																																														
P012	12	14	13	3	3	4	5	5	6	6	7	6	6	7	6	4	4	6	5	6	4	5	0																																													
P077	12	14	13	3	3	4	5	5	6	6	7	6	6	7	6	4	4	6	5	6	4	5	0	0																																												
P051	13	15	14	4	4	5	6	6	7	7	8	7	7	8	7	5	5	7	6	7	5	6	1	1	1																																											
P026	15	17	16	6	6	7	8	8	9	9	10	9	9	10	9	7	7	9	8	9	7	8	3	3	3	4																																										
P073	13	15	14	4	4	5	6	6	7	7	8	7	7	8	7	5	5	7	6	7	5	6	1	1	1	2	4																																									
P058	14	16	15	5	5	6	7	7	8	8	9	8	8	9	8	6	6	8	7	8	6	9	4	4	4	5	7	5																																									
P017	13	15	14	4	4	5	6	6	7	7	8	7	7	8	7	5	5	7	6	7	5	8	3	3	3	4	6	4	5																																								
P015	13	15	14	4	4	5	6	6	7	7	8	7	7	8	7	5	5	7	6	7	5	8	3	3	3	4	6	4	5	4																																						
P061	15	17	16	6	6	7	8	8	9	9	10	9	9	10	9	7	7	9	8	9	7	10	5	5	5	6	8	6	7	6	6																																						
P014	15	17	16	6	6	7	8	8	9	9	10	9	9	10	9	7	7	9	8	9	7	10	5	5	5	6	8	6	7	6	6	8																																					
P039	14	16	15	5	5	6	7	7	8	8	9	8	8	9	8	6	6	8	7	8	6	9	4	4	4	5	7	5	6	5	5	7	7																																				
P041	15	17	16	6	6	7	8	8	9	9	10	9	9	10	9	7	7	9	8	9	7	10	5	5	5	6	8	6	7	6	6	8	8	7																																		
P040	14	16	15	5	5	6	7	7	8	8	9	8	8	9	8	6	6	8	7	8	6	9	4	4	4	5	7	5	6	5	5	7	7	6	3																																		
P023	13	15	14	4	4	5	6	6	7	7	8	7	7	8	7	5	5	7	6	7	5	8	3	3	3	4	6	4	5	4	4	6	6	5	6	5																																	
P002	12	14	13	3	3	4	5	5	6	6	7	6	6	7	6	4	4	6	5	6	4	7	2	2	2	3	5	3	4	3	3	5	5	4	5	4	3																																
P0B6	10	12	11	1	1	2	3	3	4	4	5	4	4	5	4	2	2	4	3	4	2	9	4	4	4	5	7	5	6	5	5	7	7	6	7	6	5	4																															
P008	10	12	11	1	1	2	3	3	4	4	5	4	4	5	4	2	2	4	3	4	2	9	4	4	4	5	7	5	6	5	5	7	7	6	7	6	5	4	0																														
P013	10	12	11	1	1	2	3	3	4	4	5	4	4	5	4	2	2	4	3	4	2	9	4	4	4	5	7	5	6	5	5	7	7	6	7	6	5	4	0	0																													
P024	10	12	11	1	1	2	3	3	4	4	5	4	4	5	4	2	2	4	3	4	2	9	4	4	4	5	7	5	6	5	5	7	7	6	7	6	5	4	0	0	0																												
P036	10	12	11	1	1	2	3	3	4	4	5	4	4	5	4	2	2	4	3	4	2	9	4	4	4	5	7	5	6	5	5	7	7	6	7	6	5	4	0	0	0	0																											
P074	10	12	11	1	1	2	3	3	4	4	5	4	4	5	4	2	2	4	3	4	2	9	4	4	4	5	7	5	6	5	5	7	7	6	7	6	5	4	0	0	0	0	0																										
P037	10	12	11	1	1	2	3	3	4	4	5	4	4	5	4	2	2	4	3	4	2	9	4	4	4	5	7	5	6	5	5	7	7	6	7	6	5	4	0	0	0	0	0	0																									
P064	10	12	11	1	1	2	3	3	4	4	5	4	4	5	4	2	2	4	3	4	2	9	4	4	4	5	7	5	6	5	5	7	7	6	7	6	5	4	0	0	0	0	0	0	0																								
P038	10	12	11	1	1	2	3	3	4	4	5	4	4	5	4	2	2	4	3	4	2	9	4	4	4	5	7	5	6	5	5	7	7	6	7	6	5	4	0	0	0	0	0	0	0	0																							
P019	11	13	12	2	2	3	4	4	5	5	6	5	5	6	5	3	3	5	4	5	3	10	5	5	5	6	8	6	7	6	6	8	8	7	8	7	6	5	1	1	1	1	1	1	1	1	1																						
P063	11	13	12	2	2	3	4	4	5	5	6	5	5	6	5	3	3	5	4	5	3	10	5	5	5	6	8	6	7	6	6	8	8	7	8	7	6	5	1	1	1	1	1	1	1	1	1	2																					
P065	11	13	12	2	2	3	4	4	5	5	6	5	5	6	5	3	3	5	4	5	3	10	5	5	5	6	8	6	7	6	6	8	8	7	8	7	6	5	1	1	1	1	1	1	1	1	1	2	2																				
P025	11	13	12	2	2	3	4	4	5	5	6	5	5	6	5	3	3	5	4	5	3	10	5	5	5	6	8	6	7	6	6	8	8	7	8	7	6	5	1	1	1	1	1	1	1	1	1	2	2	2																			
P045	17	19	18	8	8	9	10	10	11	11	12	11	11	12	11	9	9	11	10	11	9	16	11	11	11	12	14	12	13	12	12	14	14	13	14	13	12	11	7	7	7	7	7	7	7	7	7	8	8	8	8																		
P021	16	18	17	7	7	8	9	9	10	10	11	10	10	11	10	8	8	10	9	10	8	15	10	10	10	11	13	11	12	11	11	13	13	12	13	12	11	10	6	6	6	6	6	6	6	6	6	7	7	7	7	13																	
P020	11	13	12	2	2	3	4	4	5	5	6	5	5	6	5	3	3	5	4	5	3	10	5	5	5	6	8	6	7	6	6	8	8	7	8	7	6	5	1	1	1	1	1	1	1	1	1	2	2	2	2	8	7																
P054	12	14	13	3	3	4	5	5	6	6	7	6	6	7	6	4	4	6	5	6	4	11	6	6	6	7	9	7	8	7	7	9	9	8	9	8	7	6	2	2	2	2	2	2	2	2	2	3	3	3	3	9	8	3															
P052	12	14	13	3	3	4	5	5	6	6	7	6	6	7	6	4	4	6	5	6	4	11	6	6	6	7	9	7	8	7	7	9	9	8	9	8	7	6	2	2	2	2	2	2	2	2	2	3	3	3	3	9	8	3	4														
P022A	11	13	12	2	2	3	4	4	5	5	6	5	5	6	5	3	3	5	4	5	3	10	5	5	5	6	8	6	7	6	6	8	8	7	8	7	6	5	1	1	1	1	1	1	1	1	1	2	2	2	2	8	7	2	3	3													
P062	11	13	12	2	2	3	4	4	5	5	6	5	5	6	5	3	3	5	4	5	3	10	5	5	5	6	8	6	7	6	6	8	8	7	8	7	6	5	1	1	1	1	1	1	1	1	1	2	2	2	2	8	7	2	3	3	2												
P076	11	13	12	2	2	3	4	4	5	5	6	5	5	6	5	3	3	5	4	5	3	10	5	5	5	6	8	6	7	6	6	8	8	7	8	7	6	5	1	1	1	1	1	1	1	1	1	2	2	2	2	8	7	2	3	3	2	2											
P049	11	13	12	2	2	3	4	4	5	5	6	5	5	6	5	3	3	5	4	5	3	10	5	5	5	6	8	6	7	6	6	8	8	7	8	7	6	5	1	1	1	1	1	1	1	1	1	2	2	2	2	8	7	2	3	3	2	2	0										
P050	11	13	12	2	2	3	4	4	5	5	6	5	5	6	5	3	3	5	4	5	3	10	5	5	5	6	8	6	7	6	6	8	8	7	8	7	6	5	1	1	1	1	1	1	1	1	1	2	2	2	2	8	7	2	3	3	2	2	0	0									
P027	12	14	13	3	3	4	5	5	6	6	7	6	6	7	6	4	4	6	5	6	4	11	6	6	6	7	9	7	8	7	7	9	9	8	9	8	7	6	2	2	2	2	2	2	2	2	2	3	3	3	3	9	8	3	4	4	3	3	1	1	1								
P042	13	15	14	4	4	5	6	6	7	7	8	7	7	8	7	5	5	7	6	7	5	12	7	7	7	8	10	8	9	8	8	10	10	9	10	9	8	7	3	3	3	3	3	3	3	3	3	4	4	4	4	10	9	4	5	5	4	4	2	2	2	3							
P009	12	14	13	3	3	4	5	5	6	6	7	6	6	7	6	4	4	6	5	6	4	11	6	6	6	7	9	7	8	7	7	9	9	8	9	8	7	6	2	2	2	2	2	2	2	2	2	3	3	3	3	9	8	3	4	4	3	1	3	3	3	4	5						
P018	12	14	13	3	3	4	5	5	6	6	7	6	6	7	6	4	4	6	5	6	4	11	6	6	6	7	9	7	8	7	7	9	9	8	9	8	7	6	2	2	2	2	2	2	2	2	2	3	3	3	3	9	8	3	4	4	3	3	3	3	3	4	5	4					
P048	14	16	15	5	5	6	7	7	8	8	9	8	8	9	8	6	6	8	7	8	6	13	8	8	8	9	11	9	10	9	9	11	11	10	11	10	9	8	4	4	4	4	4	4	4	4	4	5	5	5	5	11	10	5	6	6	5	5	3	3	3	4	5	6	6				
P044	14	16	15	5	5	6	7	7	8	8	9	8	8	9	8	6	6	8	7	8	6	13	8	8	8	9	11	9	10	9	9	11	11	10	11	10	9	8	4	4	4	4	4	4	4	4	4	5	5	5	5	11	10	5	6	6	5	5	5	5	5	6	7	6	6	8			
P055	14	16	15	5	5	6	7	7	8	8	9	8	8	9	8	6	6	8	7	8	6	13	8	8	8	9	11	9	10	9	9	11	11	10	11	10	9	8	4	4	4	4	4	4	4	4	4	5	5	5	5	11	10	5	6	6	5	5	5	5	5	6	7	6	6	8	8		
P075	12	14	13	3	3	4	5	5	6	6	7	6	6	7	6	4	4	6	5	6	4	11	6	6	6	7	9	7	8	7	7	9	9	8	9	8	7	6	2	2	2	2	2	2	2	2	2	3	3	3	3	9	8	3	4	4	3	3	3	3	3	4	5	4	4	6	6	2	

Supplementary Table 2. List of 133 variable positions (SNPs) among whole genome sequences of 69 *Mycobacterium tuberculosis* "Bernese cluster" isolates.

POSITION (H37Rv)	REF	ALT	SYN	GENE	CODONCHANGE	GENENAME	ESSENTIALITY (Sassetti 2003)	ANNOTATION
9416	C	A	nonsynonymous	Rv0006	F705L	gyrA	nonessential	DNA gyrase subunit A
34119	C	A	I	IG31_Rv0031-Rv0032	I			
86009	A	G	nonsynonymous	Rv0077c	V153A	-	nonessential	oxidoreductase
184048	G	A	nonsynonymous	Rv0155	G143R	pntAa	nonessential	NAD(P) transhydrogenase subunit alpha
213643	A	C	synonymous	Rv0182c	A166A	sigG	nonessential	RNA polymerase factor sigma-70
300141	G	A	synonymous	Rv0248c	N221N	sdhA	nonessential	succinate dehydrogenase flavoprotein subunit
322501	G	A	nonsynonymous	Rv0267	V391M	narU	nonessential	integral membrane nitrite extrusion protein NarU
341683	G	T	nonsynonymous	Rv0281	R229L	-	nonessential	hypothetical protein
396452	G	T	synonymous	Rv0331	A84A	-	nonessential	dehydrogenase/reductase
492691	A	G	synonymous	Rv0408	L302L	pta	nonessential	phosphate acetyltransferase
503771	A	G	synonymous	Rv0418	T92T	lpqL	nonessential	lipoprotein aminopeptidase LpqL
523068	G	A	nonsynonymous	Rv0435c	T489I	-	nonessential	putative ATPase
586502	G	C	nonsynonymous	Rv0496	A37P	-	nonessential	hypothetical protein
621578	C	G	synonymous	Rv0530	R224R	-	nonessential	hypothetical protein
632075	A	G	synonymous	Rv0539	P111P	-	nonessential	dolichyl-phosphate sugar synthase
653155	C	A	nonsynonymous	Rv0562	L129M	grcC1	essential	polyprenyl-diphosphate synthase
653176	G	A	nonsynonymous	Rv0562	V136M	grcC1	essential	polyprenyl-diphosphate synthase
653221	T	C	nonsynonymous	Rv0562	F151L	grcC1	essential	polyprenyl-diphosphate synthase
655237	T	C	nonsynonymous	Rv0564c	E238G	gpsA	nonessential	NAD(P)H-dependent glycerol-3-phosphate dehydrogenase
706709	C	T	nonsynonymous	Rv0612	A129V	-	nonessential	hypothetical protein
720785	G	T	synonymous	Rv0629c	G316G	recD	nonessential	exonuclease V alpha chain
772908	C	G	nonsynonymous	Rv0672	D475E	fadE8	nonessential	acyl-CoA dehydrogenase FADE8
811425	A	C	nonsynonymous	Rv0714	E18A	rplN	essential	50S ribosomal protein L14
830636	C	T	I	IG751_Rv0738-	I			
852634	G	C	nonsynonymous	Rv0758	R80T	phoR	nonessential	two component system response sensor kinase membrane associated
867288	G	C	synonymous	Rv0773c	A34A	ggtA	nonessential	bifunctional cephalosporin acylase/gamma-glutamyltranspeptidase
877118	C	A	synonymous	Rv0783c	S441S	emrB	nonessential	multidrug resistance integral membrane efflux protein EmrB
913473	G	A	nonsynonymous	Rv0820	E250K	phoT	nonessential	phosphate ABC transporter ATP-binding protein
992230	T	G	nonsynonymous	Rv0890c	M123L	-	nonessential	LuxR family transcriptional regulator
1082937	C	G	nonsynonymous	Rv0972c	V272L	fadE12	nonessential	acyl-CoA dehydrogenase FADE12
1117590	A	G	nonsynonymous	Rv1001	T136A	arcA	nonessential	arginine deiminase
1120720	G	A	synonymous	Rv1003	S233S	-	nonessential	hypothetical protein
1123098	G	C	synonymous	Rv1005c	T167T	pabB	essential	para-aminobenzoate synthase component I
1166099	C	T	synonymous	Rv1043c	P236P	-	nonessential	hypothetical protein
1167058	T	C	synonymous	Rv1044	C2C	-	nonessential	hypothetical protein
1223015	T	C	nonsynonymous	Rv1095	Y7H	phoH2	nonessential	PhoH-like protein PhoH2 (phosphate starvation-inducible protein PsiH)
1258427	C	T	nonsynonymous	Rv1132	P368L	-	nonessential	hypothetical protein
1312287	C	T	stopgain	Rv1179c	W338X	-	nonessential	hypothetical protein
1323414	A	G	nonsynonymous	Rv1183	D632G	mmpL10	essential	transmembrane transport protein MmpL10
1333799	T	C	nonsynonymous	Rv1190	C274R	-	nonessential	hypothetical protein
1342946	A	G	nonsynonymous	Rv1200	K2R	-	nonessential	integral membrane transport protein
1344507	C	T	nonsynonymous	Rv1201c	M221I	-	essential	transferase
1382073	C	T	nonsynonymous	Rv1239c	G324S	corA	nonessential	magnesium/cobalt transporter CorA
1517741	T	G	nonsynonymous	Rv1350	L84R	fabG	essential	3-ketoacyl-(acyl-carrier-protein) reductase

POSITION (H37Rv)	REF	ALT	SYN	GENE	CODONCHANGE	GENENAME	ESSENTIALITY (Sassetti 2003)	ANNOTATION
1586867	C	T	synonymous	Rv1410c	L300L	-	essential	aminoglycosides/tetracycline-transport integral membrane protein
1590534	G	A	synonymous	Rv1415	V46V	ribA2	essential	bifunctional 3,4-dihydroxy-2-butanone 4-phosphate synthase/GTP
1652328	G	A	nonsynonymous	Rv1464	A271T	csd	essential	cysteine desulfurase
1672983	A	C	nonsynonymous	Rv1482c	V106G	-	nonessential	hypothetical protein
1725795	C	T	nonsynonymous	Rv1527c	G872D	pks5	nonessential	polyketide synthase pks5
1758281	T	A	nonsynonymous	Rv1552	F201I	frdA	nonessential	fumarate reductase flavoprotein subunit
1768164	G	A	synonymous	Rv1563c	C423C	treY	nonessential	maltooligosyltrehalose synthase TreY
1822872	G	A	synonymous	Rv1621c	I134I	cydD	nonessential	vytochrome' transport transmembrane ATP-binding protein ABC
1862001	G	C	synonymous	Rv1650	L748L	pheT	essential	phenylalanyl-tRNA synthetase subunit beta
1914885	A	G	I	IG1718_Rv1689-	I			
1951689	C	A	nonsynonymous	Rv1725c	W21C	-	nonessential	hypothetical protein
2014604	C	G	I	IG1810_Rv1779c-	I			
2023473	T	C	nonsynonymous	Rv1785c	K386E	cyp143	nonessential	cytochrome P450 143
2043368	A	G	I	IG1831_Rv1801-	I			
2075129	G	T	nonsynonymous	Rv1830	A97S	-	nonessential	hypothetical protein
2075223	A	C	nonsynonymous	Rv1830	H128P	-	nonessential	hypothetical protein
2123145	C	T	nonsynonymous	Rv1872c	V3I	lldD2	nonessential	L-lactate dehydrogenase (cytochrome) LldD2
2123146	C	T	synonymous	Rv1872c	A2A	lldD2	nonessential	L-lactate dehydrogenase (cytochrome) LldD2
2148602	C	A	nonsynonymous	Rv1901	A314E	cinA	nonessential	competence damage-inducible protein A
2148603	A	C	synonymous	Rv1901	A314A	cinA	nonessential	competence damage-inducible protein A
2148815	C	T	nonsynonymous	Rv1901	T385I	cinA	nonessential	competence damage-inducible protein A
2156041	C	A	nonsynonymous	Rv1908c	G24V	katG	nonessential	catalase-peroxidase-peroxynitritase T KATG
2190276	C	T	nonsynonymous	Rv1937	T594I	-	nonessential	oxygenase
2257871	G	A	synonymous	Rv2008c	D24D	-	nonessential	hypothetical protein
2330476	T	C	nonsynonymous	Rv2073c	D163G	-	nonessential	shortchain dehydrogenase
2380637	G	A	synonymous	Rv2121c	N8N	hisG	essential	ATP phosphoribosyltransferase
2383349	G	A	nonsynonymous	Rv2124c	R907C	metH	nonessential	5-methyltetrahydrofolate--homocystein methyltransferase
2390561	C	T	stopgain	Rv2129c	W210X	-	nonessential	short chain dehydrogenase
2401673	T	C	nonsynonymous	Rv2141c	Q17R	-	nonessential	hypothetical protein
2417268	G	C	nonsynonymous	Rv2156c	L70V	mraY	essential	phospho-N-acetylmuramoyl-pentapeptide-transferase
2421292	C	T	nonsynonymous	Rv2159c	G125E	-	nonessential	hypothetical protein
2424840	C	T	I	IG2194_Rv2162c-	I			
2439204	A	G	I	IG2208_Rv2176-	I			
2450840	G	A	nonsynonymous	Rv2188c	P104L	-	essential	hypothetical protein
2556606	C	A	nonsynonymous	Rv2284	D154E	lipM	nonessential	esterase LipM
2559251	A	C	nonsynonymous	Rv2286c	C107G	-	nonessential	hypothetical protein
2569115	G	T	nonsynonymous	Rv2298	V12F	-	nonessential	hypothetical protein
2647261	A	G	synonymous	Rv2366c	L369L	-	nonessential	transmembrane protein
2654168	T	C	nonsynonymous	Rv2374c	Y309C	hrcA	essential	heat-inducible transcription repressor
2682465	G	A	synonymous	Rv2388c	Y226Y	hemN	essential	coproporphyrinogen III oxidase
2697944	C	T	stopgain	Rv2401	Q73X	-	nonessential	hypothetical protein
2743299	G	C	nonsynonymous	Rv2444c	D562E	rne	essential	ribonuclease E
2755176	T	C	nonsynonymous	Rv2455c	D510G	-	nonessential	oxidoreductase alpha subunit
2812089	T	G	nonsynonymous	Rv2497c	E3A	pdhA	nonessential	pyruvate dehydrogenase E1 component alpha subunit PdhA
2817672	C	T	stopgain	Rv2502c	W268X	accD1	essential	acetyl-/propionyl-CoA carboxylase subunit beta
2914473	T	A	nonsynonymous	Rv2587c	M422L	secD	nonessential	preprotein translocase subunit SecD
2932369	C	T	synonymous	Rv2605c	G258G	tesB2	nonessential	acyl-CoA thioesterase II

POSITION (H37Rv)	REF	ALT	SYN	GENE	CODONCHANGE	GENENAME	ESSENTIALITY (Sassetti 2003)	ANNOTATION
2952637	G	T	synonymous	Rv2626c	I119I	-	nonessential	hypothetical protein
2962458	C	A	I	IG2678_Rv2634c-	I			
2976575	G	A	I	IG2696_Rv2652c-	I			
3002991	G	A	nonsynonymous	Rv2685	V337I	arsB1	nonessential	arsenic-transport integral membrane protein ArsB1
3028569	G	A	nonsynonymous	Rv2715	A158T	-	nonessential	hydrolase
3037478	A	G	synonymous	Rv2725c	R479R	hflX	nonessential	GTP-binding protein HflX
3053479	G	C	nonsynonymous	Rv2740	D83H	-	nonessential	hypothetical protein
3063788	C	A	synonymous	Rv2751	R51R	-	nonessential	hypothetical protein
3177505	G	C	I	IG2909_Rv2864c-	I			
3185732	G	T	nonsynonymous	Rv2874	A296S	dipZ	nonessential	integral membrane C-type cytochrome biogenesis protein DipZ
3191396	G	A	nonsynonymous	Rv2881c	R76W	cdsA	essential	integral membrane phosphatidate cytidylyltransferase CdsA
3270029	A	G	nonsynonymous	Rv2935	T765A	ppsE	nonessential	phenolpthiocerol synthesis type-I polyketide synthase PPSE
3281845	C	T	nonsynonymous	Rv2940c	V291M	mas	nonessential	multifunctional mycocerosic acid synthase membrane-associated MAS
3303903	G	A	nonsynonymous	Rv2951c	L116F	-	nonessential	oxidoreductase
3343922	C	T	I	IG3034_Rv2986c-	I			
3347492	C	T	synonymous	Rv2990c	T76T	-	nonessential	hypothetical protein
3352210	G	A	synonymous	Rv2994	S314S	-	nonessential	integral membrane protein
3487979	G	A	I	IG3172_Rv3121-	I			
3522352	G	C	nonsynonymous	Rv3156	C40S	nuoL	nonessential	NADH dehydrogenase subunit L
3525549	C	A	nonsynonymous	Rv3157	T473N	nuoM	nonessential	NADH dehydrogenase subunit M
3565708	G	A	nonsynonymous	Rv3195	A449T	-	nonessential	hypothetical protein
3574285	C	A	nonsynonymous	Rv3201c	A918S	-	essential	ATP-dependent DNA helicase
3580551	G	T	I	IG3255_Rv3202c-	I			
3586144	G	C	I	IG3262_Rv3208A-	I			
3641687	G	T	nonsynonymous	Rv3262	K51N	fbiB	nonessential	F420-0--gamma-glutamyl ligase
3661899	G	A	synonymous	Rv3279c	S38S	birA	nonessential	biotin--protein ligase
3778485	C	G	I	IG3424_Rv3366-	I			
3805209	A	C	nonsynonymous	Rv3390	Q115H	lpqD	nonessential	lipoprotein LpqD
3844644	T	G	I	IG3486_Rv3427c-	I			
3906811	T	C	nonsynonymous	Rv3487c	D66G	lipF	nonessential	esterase/lipase LipF
3919936	G	A	nonsynonymous	Rv3500c	R43C	yrbE4B	nonessential	integral membrane protein YrbE4b
3920879	A	C	I	IG3560_Rv3501c-	I			
3968952	A	T	I	IG3590_Rv3531c-	I			
4056268	T	C	synonymous	Rv3616c	L36L	-	essential	hypothetical protein
4070697	C	T	nonsynonymous	Rv3631	H62Y	-	essential	transferase
4092030	C	A	nonsynonymous	Rv3651	H64N	-	essential	hypothetical protein
4115367	C	A	nonsynonymous	Rv3674c	E176D	nth	nonessential	endonuclease III
4135342	A	C	nonsynonymous	Rv3693	D206A	-	nonessential	hypothetical protein
4173545	G	A	synonymous	Rv3727	Q197Q	-	nonessential	oxidoreductase
4285823	C	A	synonymous	Rv3820c	V1V	papA2	nonessential	polyketide synthase associated protein PapA2
4355597	C	G	nonsynonymous	Rv3877	S197R	-	essential	transmembrane protein
4363705	G	A	synonymous	Rv3883c	D351D	mycP1	nonessential	membrane-anchored mycosin MYCP1 (serine protease)

Supplementary Table 3. Mycobacterial interspersed repetitive unit-variable number of tandem repeat (MIRU-VNTR) and spoligotyping data of the 68 *Mycobacterium tuberculosis* "Bernese cluster" isolates identified plus the serial isolate of the key patient (P006). Two alleles are indicated where mixed signals were detected (grey boxes).

ID	154	424	577	580	802	960	1644	1955	2059	2163b	2165	2347	2401	2461	2531	2687	2996	3007	3171	3192	3690	4052	4156	4348	Spoligotype
P001	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	3	3	7	2	2	
P002	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	3	3	7	2	2	
P003	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	3	3	7	2	2	
P004	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	3	3	7	2	2	
P005	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	3	3	7	2	2	
P006A	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	3	3	4 + 7	2	2	
P006B	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	3	3	7	2	2	
P007	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	3	3	7	2	2	
P008	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	3	3	7	2	2	
P009	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	3	3	7	2	2	
P010	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	3	3	7	2	2	
P011	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	3	3	7	2	2	
P012	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	3	3	7	2	2	
P013	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	3	3	7	2	2	
P014	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	3	3	7	2	2	
P015	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	3	3	7	2	2	
P016	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	3	3	7	2	2	
P017	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	3	3	7	2	2	
P018	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	3	3	7	2	2	
P019	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	3	3	7	2	2	
P020	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	3	3	7	2	2	
P021	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	3	3	7	2	2	
P022	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	3	3	7	2	2	
P023	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	3	3	7	2	2	
P024	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	3	3	7	2	2	
P025	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	3	3	7	2	2	
P026	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	3	3	7	2	2	
P027	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	3	3	7	2	2	
P028	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	3	3	7	2	2	
P033	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	3	3	7	2	2	
P034	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	3	3	7	2	2	

P035	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	3	3	7	2	2	
P036	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	3	3	7	2	2	
P037	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	3	3	7	2	2	
P038	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	3	3	7	2	2	
P039	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	3	3	7	2	2	
P040	2	3	4	3	3 + 4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	3	3	7	2	2	
P041	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	3	3	7	2	2	
P042	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	3	3	7	2	2	
P043	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	3	3	7	2	2	
P044	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	3	3	7	2	2	
P045	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	3	3	7	2	2	
P047	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	3	3	7	2	2	
P048	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	3	3	7	2	2	
P049	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	3	3	7	2	2	
P050	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	3	3	7	2	2	
P051	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	3	3	7	2	2	
P052	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	3	3	7	2	2	
P053	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	3	3	7	2	2	
P054	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	3	3	7	2	2	
P055	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	3	3	7	2	2	
P056	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	3	3	7	2	2	
P058	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	3	3	7	2	2	
P059	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	3	3	7	2	2	
P061	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	4 + 5	3	3	3	3	7	2	2	
P062	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	3	3	7	2	2	
P063	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	3	3	7	2	2	
P064	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	3	3	7	2	2	
P065	n.a.																								
P066	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	3	3	7	2	2	
P067	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	2	3	7	2	2	
P069A	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	3	3	7	2	2	
P070	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	3	3	7	2	2	
P072	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	1	3	7	2	2	
P073	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	3	3	7	2	2	
P074	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	3	3	7	2	2	
P075	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	3	3	7	2	2	
P076	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	3	3	7	2	2	
P077	2	3	4	3	4	4	3	1	2	2	4	4	2	2	5	1	5	3	3	3	3	7	2	2	

Supplementary Information References

1. Coll F, Mallard K, Preston MD, et al. SpolPred: rapid and accurate prediction of *Mycobacterium tuberculosis* spoligotypes from short genomic sequences. Bioinformatics **2012**; 28:2991–2993.

2. Cowan LS, Diem L, Brake MC, Crawford JT. Transfer of a *Mycobacterium tuberculosis* genotyping method, Spoligotyping, from a reverse line-blot hybridization, membrane-based assay to the Luminex multianalyte profiling system. J Clin Microbiol **2004**; 42:474–477.

3. Fenner L, Egger M, Bodmer T, et al. Effect of mutation and genetic background on drug resistance in *Mycobacterium tuberculosis*. Antimicrob Agents Chemother **2012**; 56:3047–3053.

4. Genewein A, Telenti A, Bernasconi C, et al. Molecular approach to identifying route of transmission of tuberculosis in the community. Lancet **1993**; 342:841–844.

5. Sandgren A, Strong M, Muthukrishnan P, Weiner BK, Church GM, Murray MB. Tuberculosis Drug Resistance Mutation Database. PLoS Med **2009**; 6:e1000002.

B. List of publications

Fenner L, Malla B, Ninet B, Dubuis O, **Stucki D**, Borrell S, Huna T, Bodmer T, Egger M, Gagneux S. "Pseudo-Beijing": Evidence for convergent evolution in the direct repeat region of *Mycobacterium tuberculosis*. *PLoS ONE* 2011, 6(9):e24737.
doi:10.1371/journal.pone.0024737

Yeboah-Manu D, Asante-Poku A, Bodmer T, **Stucki D**, Koram K, Bonsu F, Pluschke G, Gagneux S. Genotypic diversity and drug susceptibility patterns among *M. tuberculosis* complex isolates from South-Western Ghana. *PLoS ONE* 2011, 6(7):e21906.
doi:10.1371/journal.pone.0021906

Fenner L, Gagneux S, Helbling P, Battegay M, Rieder HL, Pfyffer GE, Zwahlen M, Furrer HJ, Siegrist H, Fehr H, Dolina M, Calmy A, **Stucki D**, Jaton K, Janssens JP, Mazza Stalder J, Bodmer T, Ninet B, Böttger E, Egger M. *Mycobacterium tuberculosis* transmission in a country with low tuberculosis incidence: role of immigration and HIV infection. *Journal of Clinical Microbiology 2012*, 50(2):388–95.
doi:10.1128/JCM.05392-11

Stucki D, Malla B, Hostettler S, Huna T, Feldmann J, Yeboah-Manu D, Borrell S, Fenner L, Comas I, Coscollà M, Gagneux S. Two New Rapid SNP-Typing Methods for Classifying *Mycobacterium tuberculosis* Complex into the Main Phylogenetic Lineages. *PLoS ONE* 2012, 7(7):e41253.
doi:10.1371/journal.pone.0041253

Malla B, **Stucki D**, Borrell S, Feldmann J, Maharjan B, Shrestha B, Fenner L, Gagneux S. First Insights into the Phylogenetic Diversity of *Mycobacterium tuberculosis* in Nepal. *PLoS ONE* 2012, 7(12):e52297.
doi:10.1371/journal.pone.0052297

Bratschi MW, Njih Tabah E, Bolz M, **Stucki D**, Borrell S, Gagneux S, Noumen-Djeunga B, Junghanss T, Um Boock A, Pluschke G. A case of cutaneous tuberculosis in a Buruli ulcer-endemic area. *PLoS Neglected Tropical Diseases* 2012, 6(8):e1751.
doi:10.1371/journal.pntd.0001751

Stucki D, Gagneux S. Single nucleotide polymorphisms in *Mycobacterium tuberculosis* and the need for a curated database. *Tuberculosis (Edinburgh, Scotland)* 2013, 93(1):30–9.
doi:10.1016/j.tube.2012.11.002

Fenner L, Egger M, Bodmer T, Furrer HJ, Ballif M, Battegay M, Helbling P, Fehr J, Gsponer T, Rieder HL, Zwahlen M, Hoffmann M, Bernasconi E, Cavassini M, Calmy A, Dolina M, Frei R, Janssens JP, Borrell S, **Stucki D**, Schrenzel J, Böttger E, Gagneux S, for the Swiss HIV Cohort and Molecular Epidemiology of Tuberculosis Study Groups. HIV Infection Disrupts the Sympatric Host–Pathogen Relationship in Human Tuberculosis. *PLoS Genetics* 2013, 9(3):e1003318.
doi:10.1371/journal.pgen.1003318

Steiner A, **Stucki D**, Borrell S, Coscollà M, Gagneux S. KvarQ: Targeted and direct variant calling from fastq reads of bacterial genomes. *BMC Genomics* 2014, 15:881.
doi:10.1186/1471-2164-15-881

Stucki D, Ballif M, Bodmer T, Coscollà M, Maurer A, Droz S, Butz C, Borrell S, Längle C, Feldmann J, Furrer HJ, Mordasini C, Helbling P, Rieder HL, Egger M, Gagneux S, Fenner L. Tracking a Tuberculosis Outbreak Over 21 Years: Strain-Specific Single-Nucleotide Polymorphism Typing Combined With Targeted Whole-Genome Sequencing. *The Journal of Infectious Diseases*, first published online 2014.
doi:10.1093/infdis/jiu601

Asante-Poku A, Yeboah-Manu D, Otchere ID, Aboagye SY, **Stucki D**, Hattendorf J, Borrell S, Feldmann J, Danso E, Gagneux S. *Mycobacterium africanum* Is Associated with Patient Ethnicity in Ghana. *PLoS Neglected Tropical Diseases* 2015, 9(1):e3370.
doi:10.1371/journal.pntd.0003370