Universidade do Minho
Escola de Engenharia

Bárbara Andreia Andrade Barbosa

Inferring epidemiology and microevolution of *Mycobacterium tuberculosis* strains from deep-sequencing data of patient samples

Dissertação de Mestrado
Mestrado em Bioinformática

Trabalho efetuado sob a orientação de
Professor Doutor Douwe Molenaar
Professora Doutora Isabel Rocha

Janeiro de 2017

DECLARAÇÃO

Nome: Bárbara Andreia Andrade Barbosa

Endereço eletrónico: barbara.barbosabrt@gmail.com

Telefone: +351917878136

Bilhete de Identidade/Cartão do Cidadão: 14312165

Título da dissertação: *Inferring epidemiology and microevolution of Mycobacterium tuberculosis strains from deep-sequencing data of patient samples*

Orientadores:

Professor Doutor Douwe Molenaar, Vrije Universiteit Amsterdam

Professora Doutora Isabel Rocha, Universidade do Minho

Ano de conclusão: 2017

Mestrado em Bioinformática

DE ACORDO COM A LEGISLAÇÃO EM VIGOR, NÃO É PERMITIDA A REPRODUÇÃO DE QUALQUER PARTE DESTA TESE/TRABALHO.

Universidade do Minho, _____/_____/_____

Assinatura:

ACKNOWLEDGEMENTS

Upon completing this project, first I would like to thank my supervisors, Dowue Molenaar and Isabel Rocha, for allowing me to join this fantastic project, for all patience, sympathy, help, guidance and advice towards this work and my future.

I would like to thank Indra Bergval and Sarah Sengstake for all help, ideas, instruction and for being two amazing persons to work with. I wish you all the best!

To my friends for their love and caring, for all the good moments, for all the craziness, for always being there for me, for everything.

To mom and dad, for supporting me in all decisions, for making me who I am today, thank you for being the best parents in the world!

To my brothers, for never disappointing me, for being my brothers and friends, for all the jokes, the annoyances, for always being by my side. I'm proud of you!

To my grandpa, for being the sweetest person in the world.

A sweet special thanks to Jorge, for being always there for me, for being the amazing person you are, for all the help, patient, love, everything. You make me a better person.

To the Netherlands for making me a super active person by riding a bike every day. Dank u wel. Gezellig!

Finally, to everyone else that turned this two years into two fantastic year.

"How wonderful it is that nobody need wait a single moment before starting to improve the world." – Anne Frank

---

RESUMO

A Tuberculose provocada pelo agente patogénico intracelular *Mycobacterium tuberculosis* é uma doença infeciosa que continua a ser um dos maiores problemas de saúde global, estimando-se que aproximadamente um terço da população tenha estado em contacto e esteja infetada de forma latente.

Whole genome sequencing surgiu como um método revolucionário da investigação de genomas de micobactérias. A sua aplicação têm proporcionado conhecimentos inovadores relativamente à evolução da *Mycobacterium tuberculosis* devido a estudos recentes que reportam resultados contraditórios sobre a sua estabilidade genómica, particularmente durante a evolução da sua resistência a antibióticos em linhagens consideradas modernas.

Para abordar esta questão, focámo-nos na análise e compreensão dos fatores genotípicos e epidemiológicos que influenciam a capacidade de disseminação e o *fitness* desta bactéria através da análise de dados *deep-sequencing* provenientes de amostras de 85 pacientes provenientes da Ásia Central. As amostras pertencem a um estudo maior composto por 399 isolados clínicos de pacientes recentemente diagnosticados com tuberculose pulmonar recolhidas entre 2012 e 2013 no *National Center of Tuberculosis and Lung Diseases* (NCTLD) em Tbilisi, Geórgia.

Todas as amostras foram mapeadas contra a estirpe H37Rv. Para a reconstrução de modelos de evolução molecular, focámo-nos apenas em *single-nucleotide polymorphisms* e utilizámos dois métodos distintos, *Maximum Likelihood* e *Bayesian Inference*.

Cerca de 84% da nossa população pertence à linhagem *Beijing*, associada com a propagação em massa de estirpes resistentes a múltiplos antibióticos. Além disso, as mutações no *rpoB* e *rpoC* foram associadas à resistência a rifampicina e mutações na região *pncA* também demonstraram estar relacionadas com a resistência à pirazinamida.

Verificou-se ainda que a quantidade de variabilidade genética acumulada dentro de um paciente pode ser tão alta quanto a observada entre pacientes ao longo, do que supomos ser, uma cadeia de transmissão. Todos os pacientes que foram acompanhados durante tratamento apresentaram variabilidade genética.

O nosso estudo acrescenta novos dados relativamente à variabilidade entre diferentes estirpes de *Mycobacterium tuberculosis* tendo em conta um panorama de microevolução intra e inter paciente.

Palavras-Chave: Tuberculose; *Whole genome sequencing*; Microevolução; Filogenia; Multirresistência a antibióticos.

ABSTRACT

Tuberculosis, caused by the intracellular pathogen Mycobacterium tuberculosis is an infectious disease that remains a global public health problem where approximately one-third of the world population have been at least in contact and is latently infected with.

Whole genome sequencing has revolutionized the investigation of mycobacterial genomes. The application of this technology has provided innovative understandings into the evolution of the Mycobacterium tuberculosis due to recent studies reporting conflicting findings on its genomic stability, particularly during the evolution of drug resistance in modern lineages.

To address this question we focused on understanding the genotypic and epidemiological factors that influence the spread and fitness of this bacterium by analyzing deep –sequencing data of 85 patient samples from Central Asia. Samples were part of a larger study of 399 clinical isolates of newly diagnosed patients with pulmonary TB collected between 2012 and 2013 at the NCTLD in Tbilisi, Georgia.

All the samples were mapped against H37Rv strain. We focused on single-nucleotide polymorphisms to reconstruct models for molecular evolution, using Maximum Likelihood and Bayesian Inference methods. 84% of our population belongs to the Beijing lineage, associated with the massive spread of multidrug-resistant strains. Relationship between mutations on rpoB and rpoC were associated with drug resistance to rifampicin and mutations on pncA region also demonstrated to be related with drug resistance to pyrazinamide.

Furthermore we found that the amount of variation accumulated within a patient can be as high as that observed between patients along, what we assume to be, a chain of transmission. Intrapatient diversity was found in all of the follow up patients.

Our study adds new data to the understandings of the variability among *Mycobacterium tuberculosis* strains in an intra and interpatient microevolution scenario.

KEYWORDS: Tuberculosis; Whole genome sequencing; Microevolution; Phylogeny; Multidrug resistance.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

aLRT - Approximate Likelihood-Ratio Tests

B cells - B lymphocytes

BCG - Bacillus Calmette-Guérin

BER - Base Excision Repair

BI - Bayesian Inference

bp - Base Pair

DC - Dendritic cells

DNA - Deoxyribonucleic Acid

GD - Genome Diff

HIV - Human Immunodeficiency Virus

HPC - High Performance Computing

HTML - Hyper Text Markup Language

LM - Lipomannan

ManLAM - Mannose capped lipoarabinomannan

MCMC - Markov chain Monte Carlo

MDR-TB - Multidrug-resistant tuberculosis

MMR - Mismatch repair

MTB - *Mycobacterium tuberculosis*

MUSCLE - Multiple sequence comparison by log-expectation

NCTLD - National Center of Tuberculosis and Lung Diseases

NER - Nucleotide Excision Repair

NGS - Next Generation Sequencing

NNI - Nearest Neighbor Interchange

PGNs - Peptidoglycans

PIMs - Phosphalidyl-myo-inositol mannosides

PolII - DNA polymerase II

RNA - Ribonucleic Acid

SD - Standard Deviation

SNP - Single Nucleotide Polymorphism

SPR - Subtree Pruning And Regrafting

T cells - T lymphocytes

TB - Tuberculosis

TNF - Tumor Necrosis Factor

UV - Ultraviolet

VCF - Variant Call Format

VM - Virtual Machine

WGS - Whole genome sequencing

WHO - World Health Organization

1. INTRODUCTION

1.1 Context and Motivation

Tuberculosis, caused by the intracellular pathogen *Mycobacterium tuberculosis* (MTB) is an infectious disease that remains a global public health problem (Torrado & Cooper 2010; Lewandowski et al. 2015; da Costa et al. 2014). Indeed, the World Health Organization (WHO) estimates that approximately one-third of the world population have been in contact and is latently infected with MTB (Bozzano et al. 2014; Seo et al. 2014).

Being one of the biggest menaces to human health in a global scale, there is a crucial need to increase our knowledge on the molecular and systemic mechanisms behind the pathological success of this bacterium. Here we will use a combination of genomics, bioinformatics and systems biology approaches to understand the factors that influence this success. The improved understanding of key success factors in growth and transmission of MTB will provide additional or improved therapeutic approaches.

1.2 Main aims

Understanding the genotypic and epidemiological factors that influence the spread and fitness of MTB by analyzing data consisting of raw Illumina and Roche 454 genome sequences of 85 strains of MTB isolated from patients in Central Asia. These will be mapped to existing reference whole-genome sequences and the following aims will be pursued:

- Characterization of the genomic variants and comparison to known variations in MTB;
- Reconstruction of lineages from Central Asia and comparison to known lineages;
- Reconstruction of mutation and selection o4f strains in the host;
- Reconstruction of transmission history of lineages;
- Setting up and testing hypotheses concerning spread and fitness.

2. STATE OF THE ART

2.1 Tuberculosis: A problem with global proportions

Tuberculosis (TB), caused by the intracellular pathogen *Mycobacterium tuberculosis*, is an infectious disease that remains a global public health problem (Torrado & Cooper 2010; Lewandowski et al. 2015; da Costa et al. 2014) . Indeed, the World Health Organization (WHO) estimates that roughly one-third of the world population have been in contact and is latently infected with MTB (Bozzano et al. 2014; Seo et al. 2014). Besides that, in 2015 the WHO reported 9.6 million new cases of TB and 1.5 million deaths (Lewandowski et al. 2015).

Due to the fact that the bacteria is transmitted via aerosol droplets that are suspended in the air, TB is extremely contagious. Infection by MTB can cause a primary TB infection, where the disease is active within two years of the initial infection, or a latent infection, which consists in an asymptomatic condition in a purified protein derivative-positive person. Although latently infected individuals control the initial infection, 5 to 10% of these individuals progress to active TB during their life-time (Lin et al. 2009). The reactivation rates are significantly increased when the immune system is compromised, such as in individuals infected with the Human Immunodeficiency Virus (HIV), old age, during tumor necrosis and other chronic diseases, such as diabetes and alcoholic liver disease (Lin et al. 2009; Flynn, JoAnne L., Chan 2001; Prezzemolo et al. 2014).

The lack of an efficient vaccine has disadvantaged the control of this disease, and although effective drug treatment exists, the procedures are extensive and involve multiple drugs, some of them with considerable toxicity (Raja 2004). Currently, the only vaccine available is an attenuated strain of *Mycobacterium bovis*, known as *Bacillus Calmette-Guérin* (BCG) (Cooper 2009; Reyes et al. 2013; da Costa et al. 2014). Nonetheless, scientific advances have also empowered the search for more sophisticated methodologies to vaccine design. The global pipeline of TB vaccine candidates in clinical trials is stronger than at any previous period in history, now including recombinant BCGs, attenuated MTB strains, recombinant viral-vectored platforms, protein/adjuvant combinations and mycobacterial extracts (Lewandowski et al. 2015).

2.2 *Mycobacterium tuberculosis*: The pathogen

MTB is a facultative intracellular pathogen that has a slow growth rate due to its thick layer of hydrophobic mycolic acid in the cell wall that reduce the entry of nutrients (Flynn & Chan 2001; Kleinnijenhuis et al. 2011; North & Jung 2004). However, this layer is the key for the success of MTB as a pathogen, as it contributes to its resistance to degradation by the lysosomal enzymes of the macrophage's intracellular compartment. Indeed, the cell envelope is a unique characteristic of MTB, with a cell wall comprised of a layer with mostly mycolic acid at the external portion, and arabinogalactan, phosphalidyl-myo-inositol mannosides (PIMs), and peptidoglycans (PGNs) in the internal layers (Kleinnijenhuis et al. 2011). At the surface, are found mannose-containing biomolecules, such as mannose capped lipoarabinomannan (ManLAM), the related lipomannan (LM), PIMs, arabinomannan, mannan and man-noglycoproteins (Kleinnijenhuis et al. 2011; Torrelles & Schlesinger 2011); where ManLAM, LM and PIMs are incorporated into the plasma membrane (Torrelles & Schlesinger 2011). These components act as ligands for host cell receptors and for that reason are responsible for the initiation of the immune response (Torrelles & Schlesinger 2011).

MTB survives and proliferates inside the host macrophages (North & Jung 2004), after being phagocytized, which is induced through the binding of several of the above described molecules to the receptors present in macrophages (Torrelles & Schlesinger 2011). MTB can also be phagocytized by dendritic cells (DCs) through the binding of ManLAM to DC-specific intercellular adhesion molecule-3 (ICAM-3)-grabbing nonintegrin (DC-SIGN) (Maeda et al. 2003; Torrelles & Schlesinger 2011). Therefore, ManLAM is an important participant in the recognition of MTB by the host cell and, for that reason, it is an important virulence factor (Torrelles & Schlesinger 2011). Overall, the initial recognition of MTB is critical for the initiation of the innate immune response, which provides the host's first line of defense.

2.3 *Mycobacterium tuberculosis*: DNA repair system

Pathogenic bacteria are frequently exposed to several hostile conditions, with the host immune system and antibiotic treatments constantly changing their environments. Specifically, intracellular pathogens, like MTB, are challenged by a set of potentially DNA-damaging attacks in vivo, through the host-generated antimicrobial reactive oxygen and nitrogen intermediates (MacMicking et al. 1997; Akaki et al. 2000; Rich et al. 1997; Warner & Mizrahi 2006). Therefore is extremely important for bacteria to have a DNA repair system and also reversal mechanisms that can "counter-attack" powerfully the injurious effects of these encounters (Dos Vultos et al. 2009).

Genome sequencing shown that genes encoding proteins that are required for nucleotide excision repair (NER), base excision repair (BER), recombination and SOS repair and mutagenesis are present in MTB. In particular, a full complement of genes known to be directly involved in the repair of oxidative damage are present in MTB (Dos Vultos et al. 2009). However, MTB lacks both the normally highly conserved mismatch repair (MMR) system, including *dam*, *dcm*, *mutH*, *L* and *H*, and *vsr* genes, and the BER protein mug, DNA glycosylases (Saunders et al. 2011). The nonexistence of the MMR system is parallel with high rates of divergence in other species, such as *Helicobacter pylori*, and hyper mutable lineages and sub-clones of *Pseudomonas aeruginosa* (Mena et al. 2008). Furthermore, the lack of MMR is thought to enable evolution through gene duplication and divergence, and to rise the relative rate of frame-shift mutations relative to point mutations. This assumed "slack in the fidelity of genome maintenance" in MTB is the basis for hypotheses of hyper mutability sustaining rapid drug resistance evolution (Dos Vultos et al. 2008; Dos Vultos et al. 2009). Nonetheless, the amount of point mutations in MTB is not prominent in vitro, a phenomenon attributed to high polymerase fidelity (Springer et al. 2004).

MTB does not possess DNA polymerase II (PolII), but possesses two DnaE proteins, which are apparently functionally redundant. It has been proposed that the main functions of DnaE1 or DnaE2 may be error-prone DNA repair (Dos Vultos et al. 2009). DnaE2 has been shown to be induced in some rifampicin resistance associated *rpoB* mutants and deletion has been shown to reduce UV resistance, to reduce virulence in animal models of MTB, and has been suggested to be "a primary mediator of survival through inducible mutagenesis that can contribute directly to the emergence of drug resistance in vivo" (Bergval et al. 2007; Boshoff, H., Lun 2011).

2.4 The evolution of *Mycobacterium tuberculosis*

2.4.1 Control of infection: the role of adaptive immune system

The adaptive immune response is characterized by being highly specific and by generating a 'memory' response against a specific pathogen, enhancing the adaptive response through each interaction with the same agent. This memory response allows the immune system to react more quickly at the second encounter with the pathogen, being more efficient in neutralizing and clearing the infection (Roitt et al., 1998). This type of immune response is triggered when specific cells, called lymphocytes, recognize a specific molecule characteristic for a microorganism, known as antigen. There are two types of lymphocytes: B lymphocytes (B cells) and T lymphocytes (T cells), both deriving from bone-marrow stem cells (Medzhitov 2007).

Therefore, with the progress of MTB infection, cells that participate in the immune response are recruited to the site of infection and assemble in an organized aggregate consisting of a mass of infected macrophages, mature macrophages, epithelioid cells and neutrophils surrounded by B cells, DCs, CD4+ and CD8+ T cells and fibroblasts, forming a granuloma (Garra et al. 2013; Bozzano et al. 2014). The formation of this structure is assumed to be regulated by IL-10, an anti-inflammatory cytokine produced mainly by macrophages, and tumor necrosis factor (TNF), also produced by macrophages, CD4+ and CD8+ T cells and DCs (Flynn & Chan 2001). The granuloma, on the one hand has an important role in host protection against mycobacterial infection avoiding dissemination of the bacteria, on the other hand, supports the maintenance of mycobacteria in the latent form until its reactivation due to decline in host immunity (Torrado & Copper 2013).

MTB is classified into seven main phylogenetic lineages which are associated with different geographical regions, and are capable of inducing variable inflammatory responses (Krishnan et al. 2011; Portevin & Gagneux 2011; Carmona et al. 2013). Evidences indicate that some strains of MTB are more virulent than others, which can be observed by comparisons between lineages: strains from the ancient lineage (Indo-Oceanic and West African) induce a considerably stronger immune responses, as opposed modern lineage strains (Euro-American/Beijing and Indian/East African) (Krishnan et al. 2011; Portevin &

Gagneux 2011). These differential immune responses have been associated with the differential recognition of MTB strains (Carmona et al. 2013).

2.4.2 A new era: next generation sequencing

Next generation sequencing (NGS) technologies were first introduced in the market around 2005 and since then had a remarkable impact on the genomic field of research. This sort of technologies have been used fundamentally for performing genome sequencing and resequencing and also for innovative applications that were not possible before using the Sanger sequencing method (Sanger et al. 1977).

Whole genome sequencing (WGS) has revolutionized the examination of mycobacterial genomes providing the most comprehensive collection of genetic variations. Recent studies have reported conflicting findings on the genomic stability of MTB during the evolution of drug resistance (Black et al. 2015; Gardy et al. 2011; Comas et al. 2011; Pérez-Lago et al. 2014). Furthermore scientific advances in the molecular area has led to the consent that the infection by MTB can be more heterogeneous than conventionally considered (Pérez-Lago et al. 2014).

Understanding the emergence and spread of multidrug-resistant tuberculosis (MDR-TB) is crucial for its control. MDR-TB in previously treated patients is generally attributed to the selection of drug resistant mutants during inadequate therapy rather than transmission of a resistant strain. Traditional genotyping methods are not sufficient to distinguish strains in populations with a higher burden of tuberculosis and it has previously been difficult to assess the degree of transmission in these settings (Black et al. 2015; Gardy et al. 2011; Comas et al. 2011; Pérez-Lago et al. 2014).

2.4.3 Evolution of antibiotic resistance

To assure its survival and following transmissions, the pathogen has to adapt in the host originally infected. However surviving may be challenging. From physical barriers preventing colonization and infection, struggle with the innate microbiome, repression by the immune system to basic health care, every case of infection is different. Nevertheless,

due to bad and abusive prescription of antibiotics we now see an alarming increase regarding antibiotic resistance and a rise in the facility with which a bacteria can adapt by those selective pressures (WHO 2014).

Evolution of antibiotic resistance within hosts can be described as an evolutionary principle of a selective sweep, i.e., mutations acquired in order to confer antibiotic resistance have a higher frequency and fixation among the host population (Didelot et al. 2016). However in most cases, these mutations only arise in one or more lineages in the within patient population after numerous months of treatment with the respective antibiotic.

Up until now, experimental studies were the possible methodology to characterize bacterial microevolution (Elena & Lenski 2003). Whole genome sequencing has revolutionized the examination of mycobacterial genomes, capturing every detail and providing us now with insightful information about within-host evolution of antibiotic resistance, discovery of new mutations and mechanisms behind resistance. Although it is not possible for all variables to be controlled, witnessing natural evolution in within host populations has the advantage of integrating complexities such as strain differences, the host environment and representative fitness trade-offs, which has highlighted even more the intimidating adaptive potential of bacterial pathogens. Therefore details such as occurrence and spread of individual point mutations can lead to a better and more accurate treatment besides aiding in the development of new drugs (Elena & Lenski 2003).

Common sites of adaptive evolution have been identified as genes with independently arising mutations often hitting different sites inside the same gene (Table 1). Additionally to "traditional" drug-resistance genes, i.e., encoding the protein target of the drug or a drug-metabolizing enzyme, in the presence of drugs there are other three types of mutations in genes that might as well confer a selective advantage by:

1. Reducing the permeability of the cell wall or increasing the activity of drug efflux pumps (Jarher & Nikaido 1994);
2. Enhancing the fitness cost of other resistance mutations and consequently be selected as compensatory mutations (Schrag et al. 1997);
3. Increasing the rate at which rare beneficial mutations occur in specific or mutated phenotypes (Denamur & Matic 2006).

In 2013, Farhat and his colleagues investigated the evolution of resistance to a wide-ranging of drugs in 123 strains of MTB representing transmission clusters and epidemiologically unrelated cases (Farhat et al. 2013). According to their results, resistance evolved independently up to 20 times to the so called “first line” drugs isoniazid, pyrazinamide, ethambutol and rifampicin through substitutions in *katG* and NADH-dependent enoyl-acyl carrier protein reductase (*inhA*; conferring resistance to isoniazid), *pncA* (conferring resistance to pyrazinamide), *embB* (conferring resistance to ethambutol) and *rpoB* (conferring resistance to rifampicin) (Farhat et al. 2013). The adaptive potential and repeatability of bacterial evolution is emphasized by the fact that resistance to these antibiotics was evolving several times in parallel. These results show convergent evolution in independent patients (the occurrence of mutations resulting in the same phenotype in two or more independently evolving lineages; these often arise in the same gene and may even occur at the same site), who were exposed to the same drugs providing the strongest evidence for adaptation (Farhat et al. 2013).

Table 1 - Common targets of antibiotic adaptive evolution in MTB.

Target	Antibiotic
rpoB	Rifampicin
pncA	Pyrazinamide
embB	Ethambutol
rpsL	Streptomycin
katG	Isoniazid

Since bacteria can quickly respond to selective pressures by within host evolution, at first sight it may look unexpected that antibiotic resistance has not spread even more quickly. It has been suggested both by Comas et al. and Sun et al., that this discrepancy is due to the fitness costs associated with resistance because resistance-conferring substitutions in key enzymes may reduce the efficiency of replication and transcription and also resistance-conferring proteins may be costly to produce (Comas et al. 2011; Sun et al. 2012). Nonetheless, compensatory mutations may arise in order to stabilize the fitness costs

associated with antibiotic resistance (Figure 1). Even when fitness costs are expected, bacteria can go towards adaptability.

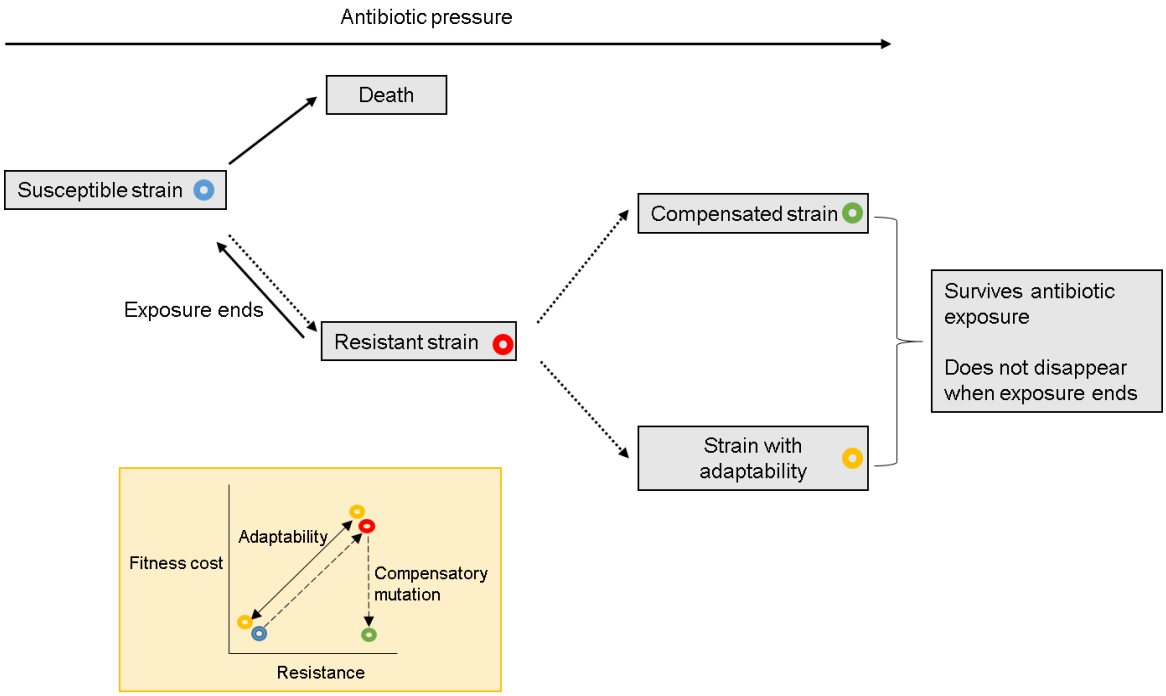

***Figure 1 - Within-host adaptive potential during exposure to antibiotics**. When exposed to an antibiotic, a susceptible bacterial strain (blue) is highly likely to be killed, but may occasionally survive by evolving into a resistant strain (red). Resistance usually has a high fitness cost, so that resistant strains usually disappear when not exposed to the antibiotic. However, resistant strains can evolve compensatory mutations (green) so that they remain resistant without the associated fitness cost. Such compensated strains pose a serious danger to public health, because they do not disappear simply as a result of antibiotic disuse. Alternatively, strains may evolve adaptability (yellow), enabling them to quickly switch resistance on or off and therefore avoid the associated fitness cost, presenting a similar risk to public health as that presented by compensated strains.*

Since the potential for adaptability appears to be different between mycobacterial strains, Ford et al. proposed that the ability for within host evolution might even be able to explain dissimilarities in the prevalence of MTB global lineages (Ford et al. 2014)

Therewithal within host adaptation depends on a series of factors, particularly the rate at which mutations that confer a potential benefit occur, the effective population size and the fitness advantage of mutants (Whitlock 2003). Therefore the larger these factors are, the faster the rate of adaptation in the population as a whole.

3. MATERIALS AND METHODS

3.1 Patient samples

Samples were part of a larger study of 399 clinical isolates of newly diagnosed patients with pulmonary TB collected between 2012 and 2013 at the NCTLD in Tbilisi, Georgia (Tukvadze et al. 2016).

3.2 Whole Genome sequencing

Whole genome sequences were prepared at GATC Biotech (GATC, Konstanz, Germany) on an Illumina HiSeq 2500 device using paired-end reads of 2x150 base pair (bp) and minimum coverage of 400 reads in the core genome.

3.3 Read alignment and variant calling

Sequence reads were aligned to the H37Rv reference genome using *breseq* (Barrick et al. 2014). *breseq* is an open-source computational pipeline designed to analyze short-read re-sequencing data and it's optimized for haploid microbial-sized genomes (Barrick et al. 2014). Specifically it uses *Bowtie2* to map reads against the reference genome (Langmead & Salzberg 2012).

It was used to predict all the single-nucleotide mutations, point insertions and deletions and large deletions. Further details are provided in the Supplementary Appendix.

All the reads located within highly repetitive regions (e.g. PE/PPE family) were excluded because they frequently offer a severe challenge to WGS and posterior data analysis (Lee & Behr 2016). Simultaneous events and events occurring less than 12 base pairs (bp) distance were also excluded from all the samples. All the exclusions from the samples were made using R scripts (Figure 2). Further details are provided in the Supplementary Appendix.

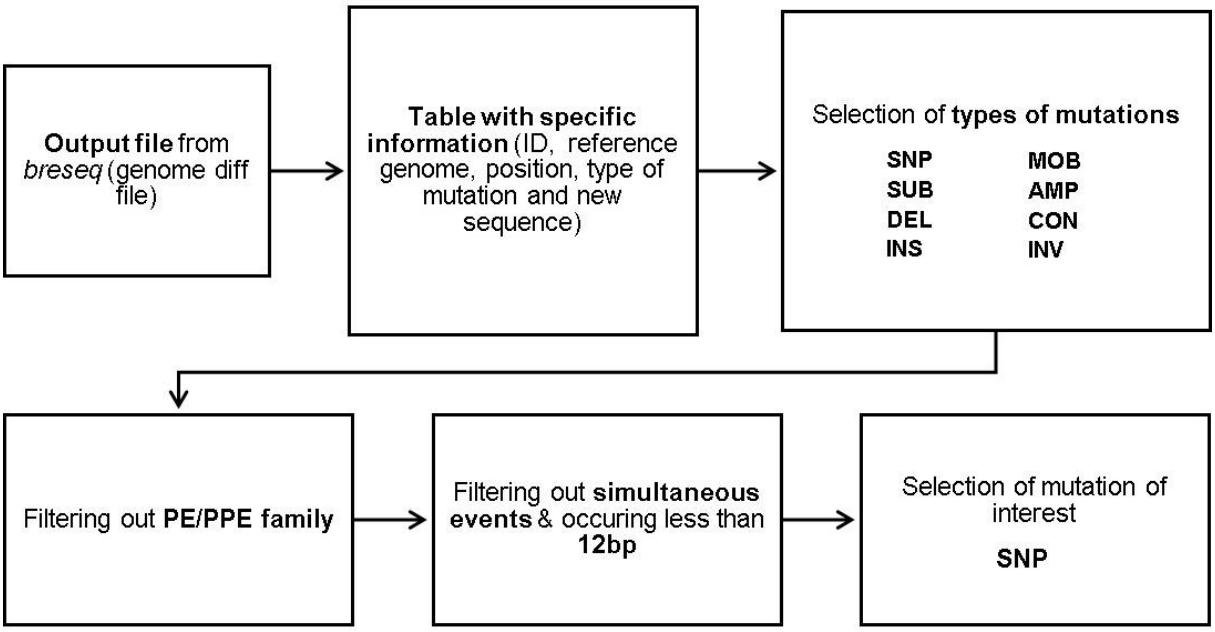

Figure 2 - Fltering steps.

3.4 Comparative Genome Analysis

To check the genetic distance between the samples a differences matrix was implemented in R.

All the Variant Call Format (VCF) files from the samples were analyzed using *PhyTB* to check for specific lineage SNP and drug resistance SNP matches. *PhyTB* is a web-based tool mainly used to support phylogenetic tree visualization with sequence data from 1601 isolates from all over the world and representation of all seven major lineages and sub-lineages in which 91k SNPs have been identified comparing to the H37Rv reference genome (Benavente et al. 2015). It contextualizes MTB genomic variation in epidemiological, geographical and phylogenic backgrounds (Benavente et al. 2015).

To perform a posterior phylogenetic analysis, "artificial" sequences were built using R scripts (Table 2). The artificial sequences contain all the SNP and the H37Rv was used as the reference. The multiple alignment of this sequences was performed using MUSCLE (multiple sequence comparison by log-expectation) with a maximum of 10 iterations. MUSCLE is a computer software used to perform multiple sequence alignments with high accuracy of protein and nucleotide sequences (Edgar et al. 2004).

Table 2 - Example of how to build "artificial" sequences. *All the SNP's were used, if the sample did not contain a specific position then that position would be filled by the nucleotide in the reference genome, in the same position.*

Reference Genome	*A G C G A G C A C T G C G A C C G G C T*
Sample A	**G G C G A G A A C T G C G A C C G G C C**
Sample B	**A G T G A G A A C T G C G A C G G G C T**
Sample C	**A T C G A G C A C T G T G A C C G G C C**
Sample D	**A T C G A A C A C T G C G A C C C G C C**

The first phylogenetic tree was constructed as a Maximum Likelihood (ML) tree using *PhyML* v3.0. PhyML is a software used to estimate maximum likelihood phylogenies from alignment of nucleotides and amino acids sequences (Guindon et al. 2010). The nucleotide substitution model implemented was HKY85. HKY85 is a model developed by Hasegawa, Kishino and Yano that allows uneven base frequencies and differentiates between transitions and transversions (Hasegawa et al. 1985). The internal branch support approach was approximate likelihood-ratio tests (aLRT), proposed as an alternative to the conventional bootstrap re-sampling and even to Bayesian estimation methods by Anisimova and Gascuel (Anisimova & Gascuel 2006). It's a modification to the previous likelihood-ratio tests proposed in 1999 by Alan Stuart and Keith Ord, fundamentally comparing the likelihoods of the best and the second best alternative arrangements around the branch of interest.

The tree searching algorithm was a combination of nearest neighbor interchange (NNI) and subtree pruning and regrafting (SPR). NNI is one of the best known metrics to calculate distances between phylogenies first introduced in 1971 by Robinson. Basically, it improves the likelihood of a given tree by performing a series of exchanges in the internal branches, i.e., for an unrooted tree there are three possibilities of connecting four subtrees and one is the original one, so there are only two possible interchanges that lead to new unrooted trees (Figure 3). This

process is repeated for each internal branch, until the maximum likelihood is obtained. It's considered an exhaustive and slow way of performing this type of search so as a less time consuming and more wide-ranging alternative of search, the SPR (Figure 4) reduces the number of topologies searched by selecting and removing a subtree from the main tree and reinserting it somewhere else on the main tree in order to create a new node. We considered the use of both strategies in order to achieve a better optimization of the whole process.

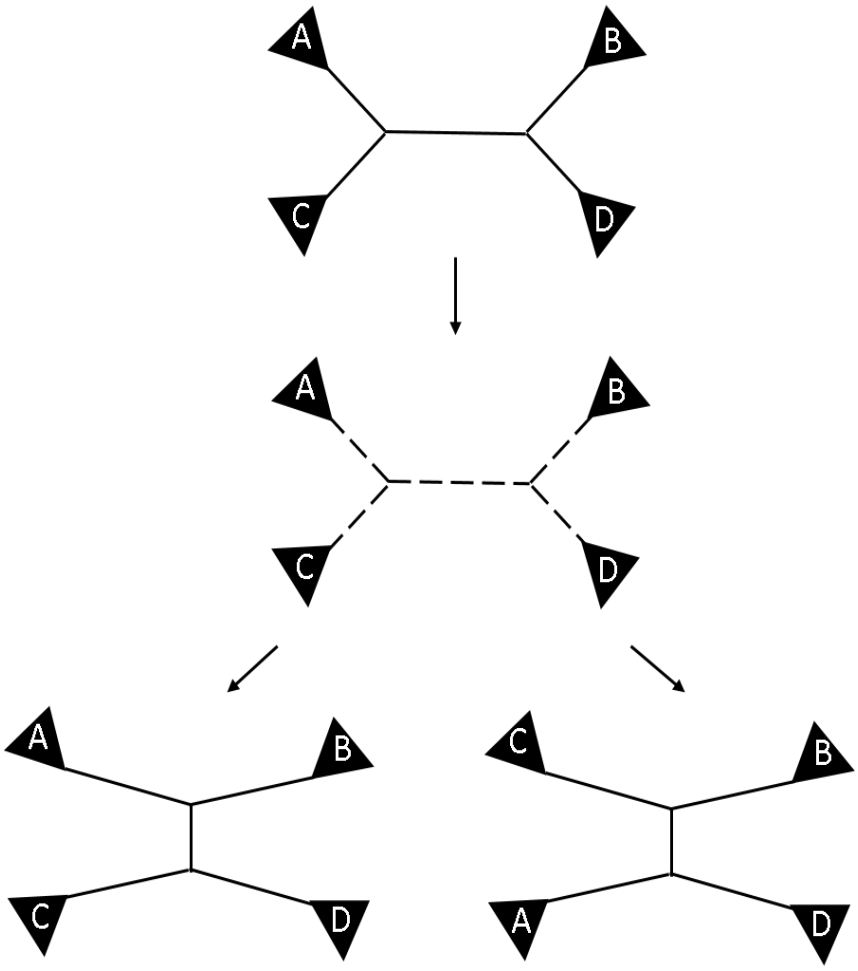

Figure 3 - Stages of nearest-neighbor interchange. *An internal branch is dissolved and rearranged to create two different topologies.*

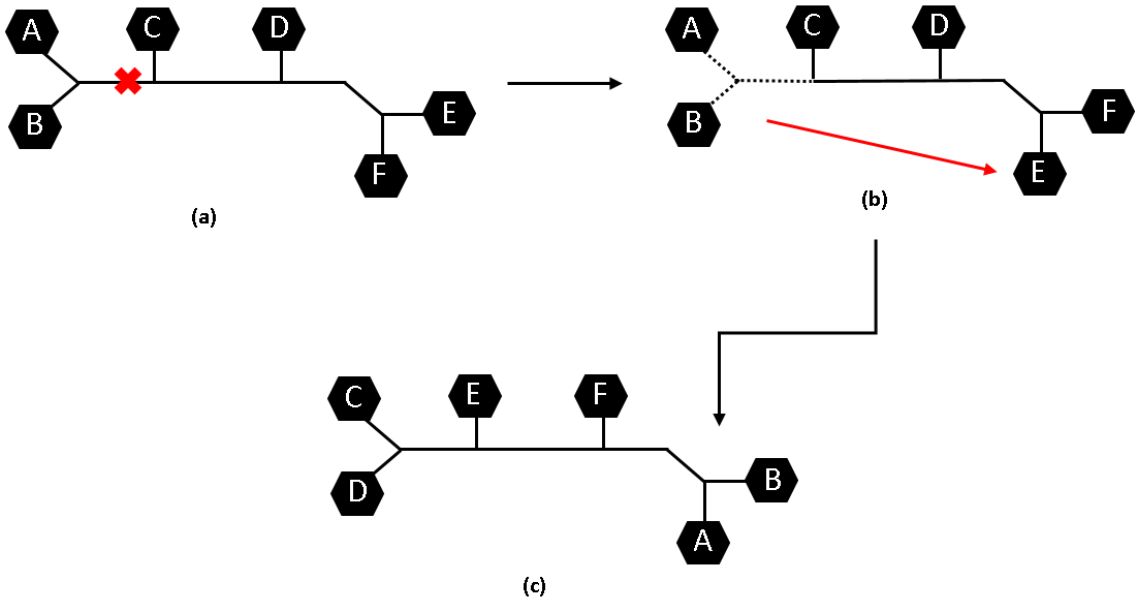

Figure 4 - Stages involved in subtree pruning and regrafting. *A branch of the tree (a) is selected and inserted on another branch (b) in order to create an alternative topology (c).*

A Bayesian evolutionary analysis was made using the program MrBayes v3.2 (Ronquist et al. 2012). MrBayes uses Markov chain Monte Carlo (MCMC) methods to estimate the posterior distribution of model parameters (Ronquist et al. 2012). MCMC is a powerful technique for performing integration by simulation (Metropolis et al. 1953; Hastings 1970).

The nucleotide substitution model implemented was HKY85, mentioned above. The chosen outgroup was the reference genome H37Rv.

4. RESULTS AND DISCUSSION

4.1 Mapping and Variant Calling

A total of 85 isolates collected from 63 patients were analyzed. 82 samples came from patients in Georgia and 3 samples belonged to a European cluster. 34 isolates are follow-ups from 12 different patients (Table 3). The treatment that the patients were subjected to is currently unknown. After sequencing, all the samples were associated with a unique ID number (e.g. 12_13700).

Table 3 – Summary of follow up samples from each patient with information about the time when each sample was collected.

Patient code	Sample ID	Month after start of treatment	Patient code	Sample ID	Month after start of treatment
1	12_16119	baseline	**7**	12_16359	baseline
	13_2210	3rd		13_7366	6th
2	12_16269	baseline		13_6517	7th
	13_1934	3rd	**8**	12_16180	baseline
3	13_1189	baseline		13_6478	3rd
	13_5512	3rd		12_15175	baseline
	13_9242	6th	**9**	13_9017	3rd
4	13_1130	baseline		13_5139	6th
	13_10762	zero		13_819	9th
	13_5974	2nd	**10**	12_17993	baseline
	13_2601	3rd		13_8431	6th
	13_2937	6th	**11**	12_16409	baseline
	13_8557	7th		13_6728	5th
5	13_381	baseline	**12**	13_1972	baseline
	13_8969	3rd		13_8615	3rd
	13_5146	6th		13_6273	5th
6	12_15902	baseline			
	13_2995	3rd			

The MTB strain H37Rv first isolated in 1905, is considered the most studied strain of tuberculosis and has remained pathogenic since. The complete genome sequence and annotation of this strain was first published in 1998 (Cole et al. 1998). In the most recent

annotation there's 4,411,532 bp of DNA sequence representing the whole *Mycobacterium tuberculosis* chromosome (Lew et al. 2011). The reads mapped to more than 95% of the H37Rv genome using *breseq* as shown in Table 4.

Table 4 - Summary of mapping reads results using breseq

ID	*reads*	*bases*	*average*	*longest*	*%mapped*
13_9242	3,2E+06	4,6E+08	144,6 bases	151,0 bases	89,4
13_6273	3,6E+06	5,4E+08	149,1 bases	151,0 bases	94,9
13_6478	3,8E+06	5,7E+08	149,2 bases	151,0 bases	95,0
13_2937	3,4E+06	5,0E+08	149,3 bases	151,0 bases	95,1
12_16850	3,7E+06	5,5E+08	150,0 bases	151,0 bases	95,2
12_18360	3,8E+06	5,8E+08	149,9 bases	151,0 bases	95,2
13_5146	4,0E+06	5,9E+08	148,9 bases	151,0 bases	95,2
13_6517	4,7E+06	7,1E+08	149,1 bases	151,0 bases	95,2
12_15737	5,8E+06	8,6E+08	149,6 bases	151,0 bases	95,3
13_2210	3,2E+06	4,8E+08	149,5 bases	151,0 bases	95,3
13_5139	4,2E+06	6,2E+08	149,1 bases	151,0 bases	95,3
13_8431	4,6E+06	6,9E+08	149,0 bases	151,0 bases	95,3
12_17795	3,5E+06	5,2E+08	149,8 bases	151,0 bases	95,4
13_5974	4,0E+06	5,9E+08	148,8 bases	151,0 bases	95,4
13_7366	5,4E+06	8,1E+08	148,7 bases	151,0 bases	95,4
12_16196	5,5E+06	8,2E+08	149,8 bases	151,0 bases	95,5
12_17736	4,4E+06	6,5E+08	149,6 bases	151,0 bases	95,5
13_2072	5,3E+06	8,0E+08	149,8 bases	151,0 bases	95,5
13_2601	5,2E+06	7,8E+08	149,0 bases	151,0 bases	95,5
13_8557	4,6E+06	6,9E+08	149,3 bases	151,0 bases	95,5
13_8969	4,4E+06	6,5E+08	148,9 bases	151,0 bases	95,5
12_14551	3,7E+06	5,5E+08	149,8 bases	151,0 bases	95,6
12_16734	4,0E+06	5,9E+08	149,7 bases	151,0 bases	95,6
13_421	4,6E+06	6,9E+08	149,6 bases	151,0 bases	95,6
13_8615	5,1E+06	7,6E+08	149,0 bases	151,0 bases	95,6
12_14129	5,5E+06	8,3E+08	149,8 bases	151,0 bases	95,7
12_15155	4,1E+06	6,0E+08	147,8 bases	151,0 bases	95,7
12_17975	5,1E+06	8,0E+08	149,5 bases	151,0 bases	95,7
12_18493	5,3E+06	7,9E+08	149,8 bases	151,0 bases	95,7
12_18942	5,9E+06	8,9E+08	150,0 bases	151,0 bases	95,7
13_10762	3,7E+06	5,2E+08	141,3 bases	151,0 bases	95,7
13_183	3,4E+06	5,0E+08	149,1 bases	151,0 bases	95,7
13_1934	3,7E+06	5,5E+08	149,2 bases	151,0 bases	95,7
13_6728	4,9E+06	7,2E+08	149,1 bases	151,0 bases	95,7
12_16409	3,5E+06	4,4E+08	127,2 bases	151,0 bases	95,8

12_19128	4,7E+06	7,0E+08	149,6 bases	151,0 bases	95,8
13_9017	6,4E+06	8,0E+08	148,8 bases	151,0 bases	95,8
12_13700	5,5E+06	8,2E+08	149.7 bases	151,0 bases	95,9
13_2219	3,5E+06	5,3E+08	149,9 bases	151,0 bases	95,9
13_381	3,6E+06	5,0E+08	138,6 bases	151,0 bases	95,9
13_819	3,5E+06	5,2E+08	149,2 bases	151,0 bases	95,9
12_14180	4,3E+06	6,5E+08	149,8 bases	151,0 bases	96,0
12_15156	5,9E+06	8,9E+08	149,4 bases	151,0 bases	96,0
12_18057	6,0E+06	8,7E+08	149,3 bases	151,0 bases	96,0
13_2995	5,0E+06	7,4E+08	149,8 bases	151,0 bases	96,0
13_5512	5,1E+06	7,7E+08	149,3 bases	151,0 bases	96,0
12_15239	6,7E+06	1,0E+09	149,5 bases	151,0 bases	96,1
12_17047	5,5E+06	8,2E+08	149,9 bases	151,0 bases	96,1
12_18490	7,2E+06	1,1E+09	149,5 bases	151,0 bases	96,1
12_19069	6,7E+06	1,0E+09	149,4 bases	151,0 bases	96,1
13_1	6,0E+06	9,0E+08	149,4 bases	151,0 bases	96,1
13_1130	6,1E+06	7,0E+08	115,2 bases	151,0 bases	96,1
13_774	3,9E+06	5,8E+08	148,7 bases	151,0 bases	96,1
12_19131	5,2E+06	7,7E+08	149,7 bases	151,0 bases	96,2
eu_2	8,9E+06	6,7E+08	75,0 bases	75,0 bases	96,2
13_56	6,7E+06	1,0E+09	149,8 bases	151,0 bases	96,3
13_1972	5,9E+06	8,5E+08	143,3 bases	151,0 bases	96,6
12_16119	4,0E+06	5,8E+08	147,3 bases	151,0 bases	96,9
eu_1	1,4E+07	1,4E+09	100,0 bases	100,0 bases	97,2
eu_3	1,1E+07	1,1E+09	100,0 bases	100,0 bases	97,2
12_17231	3,9E+07	3,9E+09	101,0 bases	101,0 bases	98,1
12_15175	2,2E+07	2,2E+09	101,0 bases	101,0 bases	98,3
12_13963	2,3E+07	2,3E+09	101,0 bases	101,0 bases	98,4
12_14879	2,5E+07	2,5E+09	101,0 bases	101,0 bases	98,4
12_15251	2,4E+07	2,4E+09	101,0 bases	101,0 bases	98,5
12_16180	2,4E+07	2,4E+09	101,0 bases	101,0 bases	98,5
12_17995	2,6E+07	2,6E+09	101,0 bases	101,0 bases	98,5
12_16295	2,7E+07	2,7E+09	101,0 bases	101,0 bases	98,6
12_17593	2,5E+07	2,5E+09	101,0 bases	101,0 bases	98,6
12_15893	2,3E+07	2,3E+09	101,0 bases	101,0 bases	98,7
12_16505	2,1E+07	2,1E+09	101,0 bases	101,0 bases	98,7
12_19027	2,4E+07	2,4E+09	101,0 bases	101,0 bases	98,7
12_18166	2,0E+07	2,1E+09	101,0 bases	101,0 bases	98,8
12_18248	2,3E+07	2,4E+09	101,0 bases	101,0 bases	98,8
12_15460	3,4E+07	3,4E+09	101,0 bases	101,0 bases	98,9
12_16706	2,4E+07	2,5E+09	101,0 bases	101,0 bases	98,9
12_17704	4,9E+07	5,0E+09	101,0 bases	101,0 bases	98,9
12_16496	2,4E+07	2,4E+09	101,0 bases	101,0 bases	99,0

12_17889	2,9E+07	3,0E+09	101,0 bases	101,0 bases	99,0
13_1786	2,6E+07	2,2E+09	101,0 bases	101,0 bases	99,0
12_16269	2,5E+07	2,6E+09	101,0 bases	101,0 bases	99,1
12_16359	2,1E+07	2,1E+09	101,0 bases	101,0 bases	99,1
12_17993	2,3E+07	2,3E+09	101,0 bases	101,0 bases	99,1
12_18055	2,6E+07	2,6E+09	101,0 bases	101,0 bases	99,1
12_18893	1,7E+07	1,7E+09	101,0 bases	101,0 bases	99,1

The difference in the percentage of reads mapped can be explained by the fact that our samples were sequenced using two different methods: Illumina and Roche 454. The two major differences between this two technologies is the read length with Illumina longest read being 101 bases and Roche 454 being 151 bases, also including the sequencing protocol. According to our results the percentage of reads mapped is slightly higher when using the Illumina sequencing technology rather than the Roche 454, as expected by Luo et al (Luo et al. 2012).

breseq is an open-source computational pipeline designed to analyze short-read re-sequencing data and it's optimized for haploid microbial-sized genomes (>10mb) and re-sequenced samples in which the diverging rate is less than 1 mutation per 1000 bp (Barrick et al. 2014). *breseq* predicts mutations in a sample relative to a reference genome using reference-based alignment approaches (Barrick et al. 2014).

breseq is a command line tool implemented in C++ and R. It will compile and function on a variety of UNIX platforms, including MacOSX, Linux and Cygwin. In order to function breseq requires two external dependencies to be installed in your system, *Bowtie2* and *R*. Bowtie2 is an "ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences" (Langmead & Salzberg 2012). R is an "integrated suite of software facilities for data manipulation, calculation and graphical display" (R Development Core Team 2008). *breseq* uses *Bowtie2* to map reads to the reference genome. For more information regarding the algorithms used by *breseq* please look up the manual (http://barricklab.org/twiki/pub/Lab/ToolsBacterialGenomeResequencing/documentation/methods.html).

In order to run *breseq* on a data set like ours, a High Performance Computing (HPC) Cloud, offered by SURFsara was used. "*SURFsara is a Dutch foundation that provides supercomputers, colocation, networks and high-end visualization mainly to*

academic institutions." HPC Cloud is presented as an Infrastructure as a Service (IaaS) platform where we can built our own virtual environment according to our needs.

One Virtual Machine (VM) was used to run 3 processes at the same time, individually performing on average for 5 hours. Each process can be performed by using a command where you input the reference file(s) and the read files.

Example command:

```
>>>breseq –r NC_000962.gbk NG-7755_12_14879_lib58473_3367_8_1.fastq NG-7755_12_14879_lib58473_3367_8_2.fastq
```

The first argument (-r) on the command line corresponds to the reference genome. It's possible to input multiple reference genomes in the same run. The unspecified arguments are the read files, there's also the option to input as many as you need and likewise use FASTQ files from diverse sequencing technologies.

The pipeline used by breseq is a set of 13 steps that are described above in figure 1 and the primary advantage of *breseq* comparatively to other software is essentially based on the ability to predict new sequence junctions in an accurate way, even those associated with mobile element insertions. Additionally it integrates multiple sources of evidence for genetic changes into mutation predictions and it's able to actually produce annotated output describing biologically relevant mutational events. However *breseq* it's still not able to find some types of mutations, like entirely novel sequences that don't exist in the reference sequence. These reads are discarded to an output file suitable for *de novo* assembly in order to be examined by other software programs. Also mutations inserted in repeat regions and chromosomal inversions and rearrangements through repeat sequences are not detected (Barrick et al. 2014).

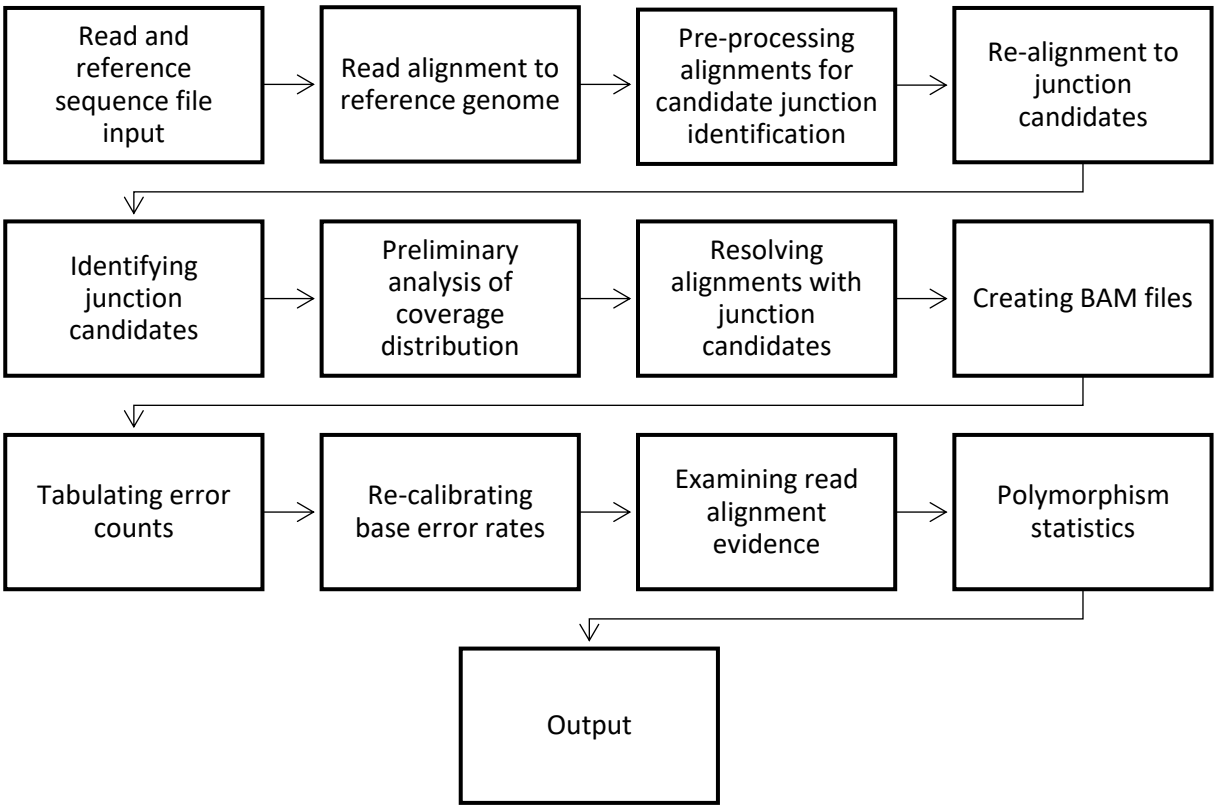

Figure 5 - breseq pipeline
(http://barricklab.org/twiki/pub/Lab/ToolsBacterialGenomeResequencing/documentation/methods.html).

The two main output files provided by *breseq* are a Hyper Text Markup Language (HTML) file and a Genome Diff (.gd) file.

The output.html file consists of an upper table screening predicted mutational events and others tables showing high-quality "orphan" evidence that *breseq* was unable to assign to mutational events. An example is presented on Table 5. A complete output file is provided in the Supplementary Appendix.

Table 5 - Example of HTML output from breseq.

Predicted mutations					
evidence	position	mutation	annotation	gene	description
RA	1,977	A→G	intergenic (+453/-75)	*dnaA* → / → *dnaN*	chromosomal replication initiator protein DnaA/DNA polymerase III subunit beta
RA	4,013	T→C	I245T (ATC→ACC)	*recF* →	DNA replication/repair protein RecF
RA	7,362	G→C	E21Q (GAG→CAG)	*gyrA* →	DNA gyrase subunit A
RA	7,585	G→C	S95T (AGC→ACC)	*gyrA* →	DNA gyrase subunit A
RA	9,304	G→A	G668D (GGC→GAC)	*gyrA* →	DNA gyrase subunit A
RA	11,37	C→T	intergenic (+186/+504)	*alaT* → / ← *Rv0008c*	tRNA-Ala/cell wall synthesis protein CwsA
RA	11,879	A→G	S145P (TCC→CCC)	*Rv0008c* ←	cell wall synthesis protein CwsA
RA	14,785	T→C	C233R (TGC→CGC)	*Rv0012* →	membrane protein
RA	16,601	A→G	D290D (GAT→GAC)	*pknB* ←	serine/threonine-protein kinase PknB
RA	21,795	G→A	P463S (CCG→TCG)	*pstP* ←	phosphoserine/threonine phosphatase PstP
RA	26,959	C→G	intergenic (-78/+64)	*Rv0021c* ← / ← *whiB5*	hypothetical protein/transcriptional regulator WhiB5
RA	29,485	Δ1 bp	coding (241/363 nt)	*Rv0025* →	hypothetical protein
RA	30,943	C→T	P408S (CCC→TCC)	*Rv0026* →	hypothetical protein
RA	31,077	C→T	intergenic (+9/-112)	*Rv0026* → / → *Rv0027*	hypothetical protein/hypothetical protein
RA	34,044	T→C	intergenic (+491/-251)	*Rv0030* → / → *bioF2*	hypothetical protein/8-amino-7-oxononanoate synthase
RA	36,477	$(C)_{7\to6}$	coding (2183/2316 nt)	*bioF2* →	8-amino-7-oxononanoate synthase

The output.gd (figure 6) describes all the mutational differences between the H37Rv and each sample also including evidence from computational analysis or even experiments that supports mutations. For a complete output.gd file please see the Supplementary Appendix.

```
#=GENOME_DIFF   1.0
#=AUTHOR        breseq 0.27.1 revision 87c22d663cc3
#=CREATED       05:18:44 01 May 2016
#=COMMAND       breseq -r NC_000962.gbk 12-13700_lib6779_nextseq_n0066_151bp_R1.fastq 12-13700_lib6779_nextseq_n0066_151bp_R2.fastq
#=REFSEQ        NC_000962.gbk
#=READSEQ       12-13700_lib6779_nextseq_n0066_151bp_R1.fastq
#=READSEQ       12-13700_lib6779_nextseq_n0066_151bp_R2.fastq
SNP     1       1060    NC_000962       1977    G
SNP     2       1074    NC_000962       4013    C
SNP     3       1075    NC_000962       7362    C
SNP     4       1076    NC_000962       7585    C
SNP     5       1077    NC_000962       9304    A
SNP     6       1078    NC_000962       11370   T
SNP     7       1079    NC_000962       11879   G
SNP     8       1080    NC_000962       14785   C
SNP     9       1081    NC_000962       16601   G
SNP     10      1082    NC_000962       21795   A
SNP     11      1097    NC_000962       26959   G
DEL     12      1098    NC_000962       29485   1
SNP     13      1099    NC_000962       30943   T
SNP     14      1100    NC_000962       31077   T
SNP     15      1101    NC_000962       34044   C
DEL     16      1102    NC_000962       36477   1       repeat_length=1 repeat_new_copies=6     repeat_ref_copies=7     repeat_seq=C
SNP     17      1103    NC_000962       37031   G
SNP     18      1104    NC_000962       42967   C
SNP     19      1105    NC_000962       50557   C
SNP     20      1106    NC_000962       51949   G
SNP     21      1107    NC_000962       51954   A
SNP     22      1108    NC_000962       54394   G
SNP     23      1109    NC_000962       55553   T
SNP     24      1110    NC_000962       57393   T
SNP     25      1111    NC_000962       62049   G
SNP     26      1112    NC_000962       62657   A
SNP     27      1113    NC_000962       69834   T
SNP     28      1114    NC_000962       69871   T
SNP     29      1115    NC_000962       69989   A
SNP     30      1116    NC_000962       70026   T
SNP     31      1117    NC_000962       70095   A
SNP     32      1118    NC_000962       70816   G
SNP     33      1119    NC_000962       70924   T
SNP     34      1120    NC_000962       71336   C
INS     35      2858    NC_000962       71586   AGCGCTGTTCTGGCGCTAATCTGACGCTAGAATAGCG
```

Figure 6 - Example of output.gd

Since the outputs provided by breseq had a lot of information that was not needed in this particular work and the data needed to be filtered, a filter was implemented in R as described in the previous Chapter.

The file used for filtering information was the output.gd. The script is provided in the Supplementary Appendix.

In this filter we started by selecting only the mutations with 3-letter codes as shown in Table 6. The reason for this is that the GD file besides containing these mutations also contains many pieces of evidence (RA, JC, MC and UN)[1] that were rejected as a basis for predicting a mutation and are considered marginal predictions. The mean of all mutations, with and without evidence, in all samples was 5103.31 mutations per sample with a standard deviation (SD) of 578.64. After this step the mean of mutations with evidence was 1547.95 mutations per sample with a SD of 242.09. Afterwards, we filtered only the crucial information out of the output.gd to a table in order to perform the further

[1] RA: Read alignment; JC: New Junction; MC: Missing Coverage; UN: Unknown

analysis, such as the sample ID, reference genome, position, type of mutation (Table 6) and new sequence.

Table 6 - Type of mutations with a 3-letter code used by breseq.

Type of mutation	Brief description
SNP	Single-nucleotide polymorphism
SUB	Substitution
DEL	Deletion
INS	Insertion
MOB	Mobile elements
AMP	Sequence Amplification
INV	Inversion

Once that was done, we still had some complications before we could pursue to the samples comparison. One of the challenges of mapping reads is covering repetitive elements which provides an additional issue that may disturb the quality of the MTB sequencing, since it's a known "flaw" of the NGS technology. The PE/PPE gene family with 168 members represents one of the most confusing yet interesting aspects of MTB genome. Even though they were discovered over 15 ago, their function remains uncertain. They are characterized by their high content of GC and repetition throughout the genome, making them very difficult to analyze. Therefore we decided to filter all the members of PE/PPE family out of our samples to make sure that our variant calls are based on rigorous evidence (Lee & Behr 2016). Moreover all the simultaneous events, i.e., mutations and the ones occurring less than 12 bp of distance were filtered out in order to not compromise the analysis.

The last step of the filtering process was selecting only the SNPs to perform a comparative genome analysis. We ended up with an average 1110.87 (±165.41) SNPs per sample. The application of the filter lead to an exclusion of 78.23% of the mutations. More information is provide in the Supplementary Appendix.

4.2 Comparative Genome Analysis

In order to check the genetic distance between all the samples and to do a raw inference of the epidemic clusters in our population a differences matrix with a heat map was made. The matrix was prepared using R. A small example of the matrix is shown in Table 7. The complete matrix and the script are provided in the Supplementary Appendix.

Table 7 - Example of the differences matrix. The genetic distance is based only on SNPs.

	12_14129	**12_16119**	**13_2210**	**12_13700**	**13_421**	**12_15373**
13_2072	756	760	758	757	775	1126
12_15155	725	723	727	726	738	1091
12_19069	740	742	740	741	759	1110
12_15239	536	542	540	539	555	1076
12_14129		212	210	213	229	1068
12_16119	212		6	167	173	1068
13_2210	210	6		165	177	1072
12_13700	213	167	165		150	1071
13_421	229	173	177	150		1081
12_15737	1068	1068	1072	1071	1081	
12_16196	1061	1063	1061	1060	1078	1007

According to the results, eleven epidemic clusters could be inferred taking into account the small genetic distance between each isolate, equal or less than 30 SNP (Table 8). As shown in Table 8 only Clusters I and J don't present any follow up samples. Cluster A is composed by two samples belonging to Patient 1; Cluster B is composed by two samples belonging to Patient 3; Cluster C is composed by three samples, two belonging to Patient 8; Cluster D is composed by three samples belonging to Patient 5; Cluster E is composed by six samples belonging to Patient 4; Cluster F is composed by fifteen samples, three of them belonging to Patient 7 and one belonging to Patient 10 ; Cluster G is composed by three samples, two belonging to Patient 12; Cluster H is composed by two samples, one belonging to Patient 10; Cluster I is composed by ten samples, two of them belonging to Patient 2, four to patient 9 and one to Patient 11. From this results we can by now assume that there's definitely diversity inside clusters and it is even possible to have

isolates from the same patient within different clusters, for example patient 10. Patient 6 and patient 8 don't belong to any cluster.

Table 8 - Association between the clusters and the patients. Eleven clusters with 29 follow up samples from 10 patients, excluding patient 6 and 8.

Patient code	Cluster	Sample ID	Patient code	Cluster	Sample ID
1	A	12_16119		F	12_18893
1		13_2210			12_16295
3	B	13_5512			12_18055
3		13_9242			12_17593
8	C	12_16180		G	12_15893
		12_18166	12		13_6273
8		13_6478	12		13_8615
5	D	13_381		H	12_17889
5		13_5146	10		13_8431
5		13_8969	2	I	12_16269
4	E	13_1130			12_17995
4		13_10762			12_18248
4		13_8557	9		12_15175
4		13_2601	9		13_819
4		13_2937	9		13_5139
4		13_5974	9		13_9017
7	F	12_16359	2		13_1934
7		13_6517			13_774
7		13_7366	11		13_6728
		12_15251		J	12_19131
		12_18057			12_14879
		12_17736			12_19027
		12_19128			12_17704
		13_183		K	12_16505
		12_16850			12_16496
		12_17047			12_17231
10		12_17993			

4.3 Models for Molecular Evolution

4.3.1 Phylogenetic Analysis

What is the best choice between Maximum Likelihood and Bayesian Inference (BI) for inferring phylogenetic relationships? Considered one of the major questions when choosing which method suits the data the best, i.e., which one is the most close to the "true evolution", the truth is that there is no right answer to this question as evolution is a complex subject. On the one hand some assume that in practice, Maximum Likelihood and Bayesian Inference analysis using the same models of evolution frequently produce identical approximations of molecular phylogenies but on the other hand there are some conflicts regarding this assumption (Brooks et al. 2007).

Taking into account what has been said previously, two different phylogenetic analyses were performed, one using the ML method and another using the BI method. In order to perform this analysis, we built "artificial" sequences containing all the SNPs. All the sequences were built using R. The total of SNPs used for this analysis was 5249. Further information of the script is provided in the Supplementary Appendix.

In this next section a comparison between those two methods is made with subsets of every tree due to the high size of each tree. This should be considered a mild analysis since we are only going to consider positions of samples and how there are grouped in each tree, not taking into account other parameters. Therewithal, three main points should be to taken into account:

1. On the right side it is always the BI tree and on the left it is the ML tree;
2. Each subset has its own color code and not related between them neither defining epidemic clusters;
3. The subsets are in order, i.e., starting from the top of each tree and making its way towards the end of it.

Subset I: First 13 samples of each tree. Since is required a outgroup for the BI we decided to use the H37Rv MTB strain ("Original"). In this subset the main difference relies on Red group being "organized" in a different order. This may be due to the existence of the Original sample in the BI since 12_18490 is the one closest to it, in terms of SNP differences. Regarding the Blue group it is equal in both trees.

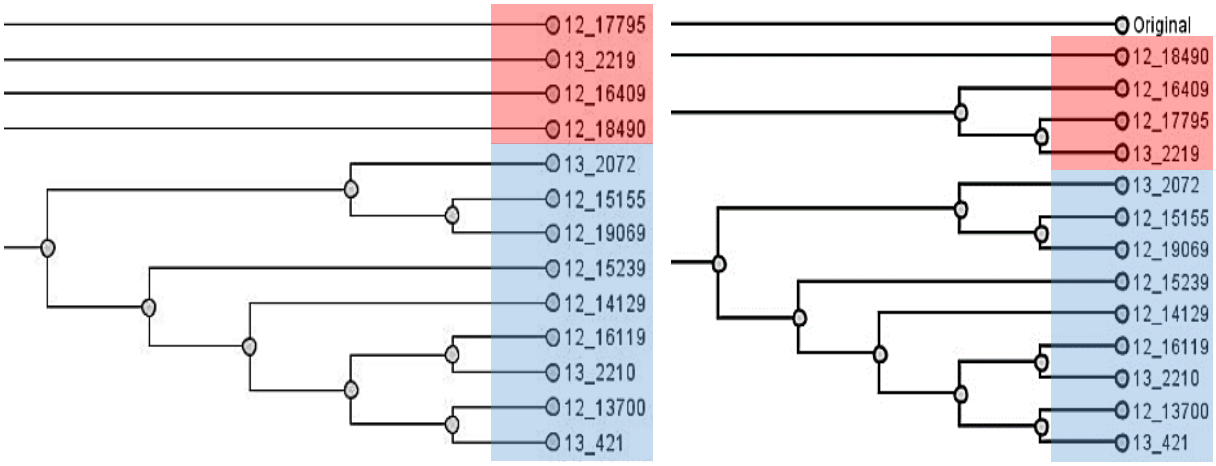

Figure 7 - Subset I: *First 13 samples of each tree. Since the BI requires a outgroup to perform the analysis, in this case the "Original" corresponds to the H37Rv MTB strain. The main difference relies on the first 4 samples being organized in a different order. No differences regarding the blue group.*

Subset II: 21 samples. Five groups of samples, each group containing the same samples. There is a slightly difference in the order of the groups, for example the red group is the last one in the ML and the second one in the BI. Besides that dissimilarity, the green, yellow and orange group are stick together in both analysis. We assume no major differences should be considered.

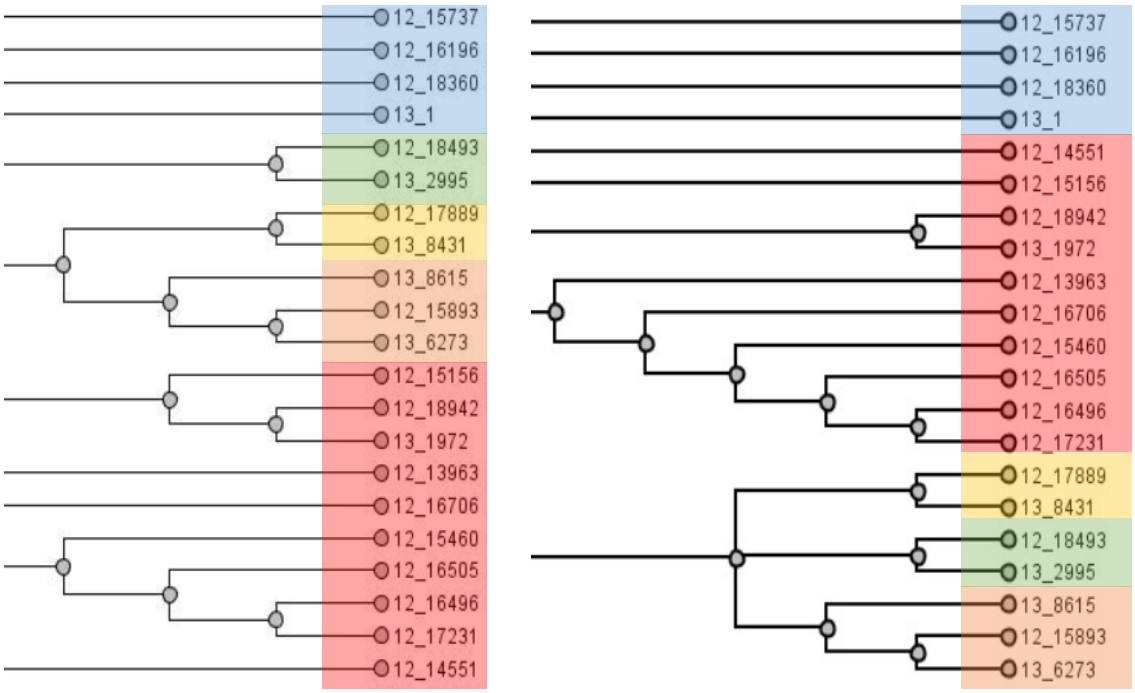

Figure 8 - Subset II: *21 samples. Five groups of samples were considered. It's visible that each group contains the same samples. There is a slightly difference regarding the order of groups, for example the red group is the last one in the ML and the second one in the BI. In addition to that divergence, the green, yellow and orange group are in pair with each other in both analysis. We assume no major differences should be considered.*

Subset III: 15 samples. Three groups were considered. All the groups are in the same order in both trees. Inside the yellow group there is only a minor difference between the order of 13_6728 that appears first in ML and last in BI. Also in the green group there is a minor change between 12_15175 and 12_16269. We assume no major differences should be considered.

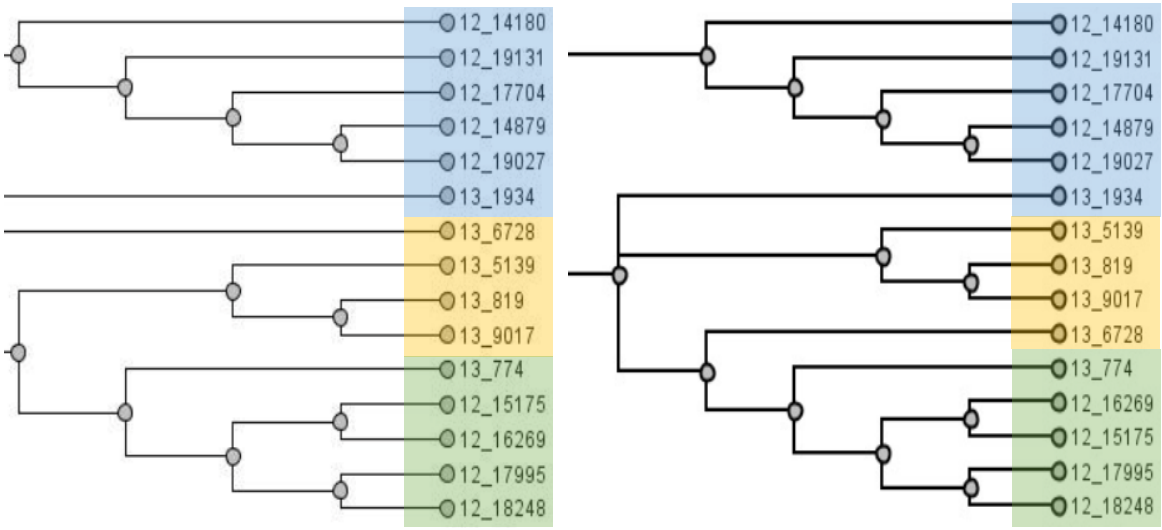

Figure 9 - Subset III: *15 samples. Three groups of samples were considered. It's visible that each group contains the same samples. Inside the yellow group there is only a minor difference between the order of 13_6728 that appears first in ML and last in BI. Also in the green group there is a minor change between 12_15175 and 12_16269. We assume no major differences should be considered.*

Subset IV: 23 samples. Three groups were considered. The samples inside each group are the same in both cases. Inside each group there are some minor differences between the orders of samples. Both subsets start with the blue group but the order of the purple and the red group is the opposite. Despite this variances we assume that no major divergences should be considered since the relationship between samples is the same.

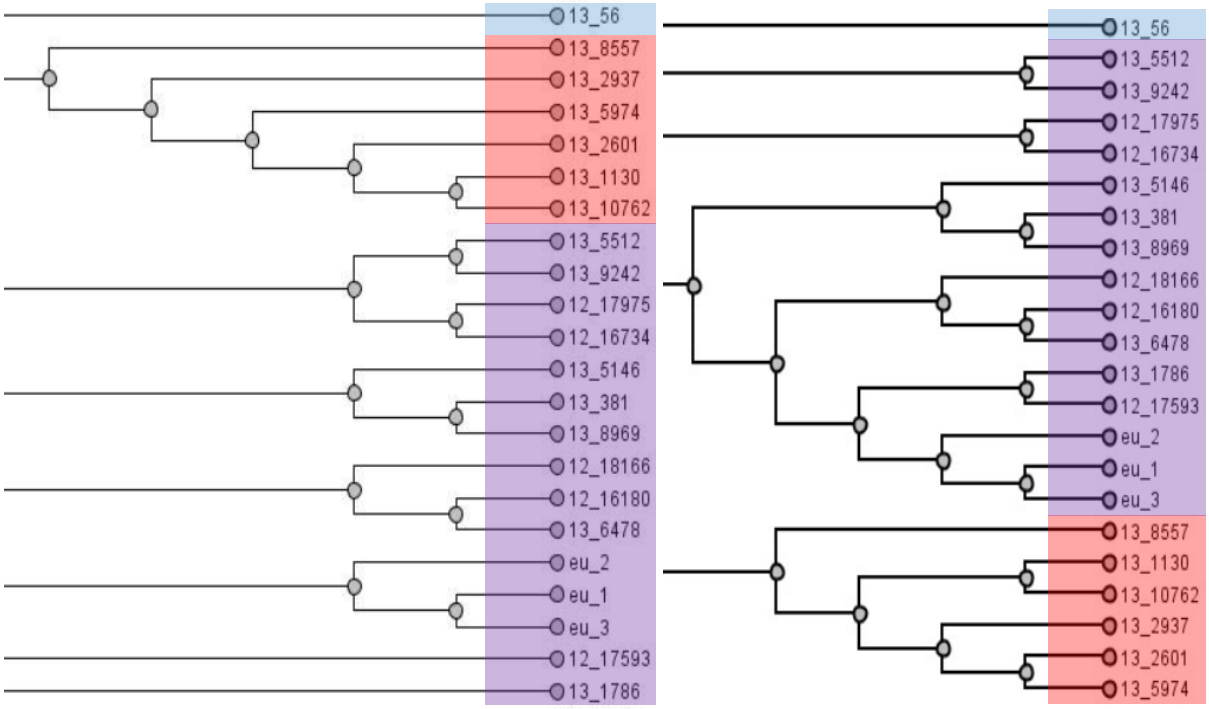

Figure 10 - Subset IV: *23 samples. Three groups of samples were considered. The samples inside each group are the same in both cases. Inside each group there are some minor differences between the orders of samples. Both subsets start with the blue group but the order of the purple and the red group is the opposite. Despite this variances we assume that no major divergences should be considered since the relationship between samples is the same.*

Subset V: 15 samples. Three groups were considered. Both ML and BI start with the red group but end with opposite groups, blue and orange, respectively. Some minor changes between samples inside groups can be observed but again we assume no major differences between them since the relationships between samples it's intact.

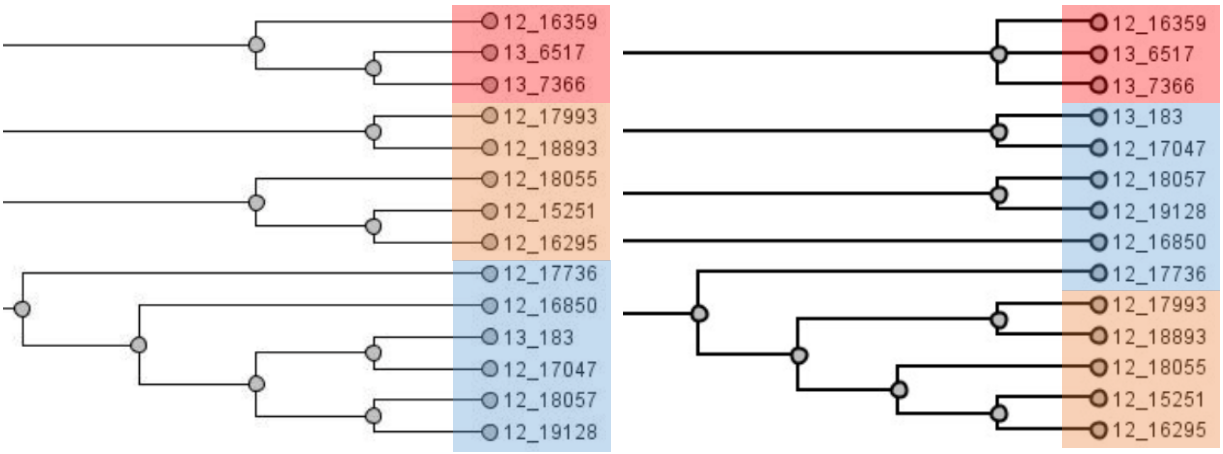

Figure 11 - Subset V: *15 samples. Three groups were considered. Both ML and BI start with the red group but end with opposite groups, blue and orange, respectively. Some minor changes between samples inside groups can be observed but again we assume no major differences between them since the relationships between samples it's intact.*

Based on this “raw” comparisons we can assume that ML and BI produce identical approximations of molecular phylogenies with some mild differences.

One of the most important features of the Bayesian likelihood methods is how simple the calculated posterior probabilities can be inferred. Making the assumption that every step of the algorithm was optimal, the value of the probability obtained can be considered as the correct one for the data analyzed (Huelsenbeck et al. 2002).

One of the most appealing aspects of Bayesian phylogenetic inference is its presentation and comparison of multiple optimal hypotheses, that later can be converged in order to have an optimal hypothesis, called the posterior output (Ronquist et al. 2012).While maximum likelihood usually converges on a single hypothesis, BI produces a range of solutions, each with a corresponding overall posterior probability as well as comparable node support values for alternative topologies within each tree hypothesis (Mau et al. 1999).

Another dissimilarity between both methods is the Bayesian method also using a prior calculated probability density distribution of the summation of all the possible combinations of the model parameters and branch lengths, making the parameters adjustable to the MCMC sampling. Consequently, although the calculations are made before the model, the values can change in order to obtain better results (Huelsenbeck et al. 2002).

Taking this into account we decided to use the BI tree for further analysis.

4.3.2 Lineages and Sub-Lineages

A crucial factor in the pathogenesis of MTB that might influence the transmissibility, host response, virulence and consequently the current emergence of drug resistance strains it is the genomic diversity specific for each strain in the MTB complex.

MTB is classified into seven phylogenetic lineages each of which can be divided into sublineages. Sublineages of the same lineage have phenotypic differences, including their pathogenicity (Anderson et al. 2013; Blouin et al. 2012).

- Lineage 1 - Indo-Oceanic;
- Lineage 2 - East Asian;
- Lineage 3 - East African-Indian;
- Lineage 4 - Euro-American;
- Lineage 5 - West African 1;
- Lineage 6 - West African 2;

- Lineage 7 - Recently discovered in north-western Ethiopia and among Ethiopian immigrants in Djibouti (Yimer et al. 2015).

For a long time, multiple methods have been proposed to categorize MTB strains into different lineages, i.e., spoligotyping and polyphasic genotyping (Gori et al. 2005; Weniger et al. 2010). In 2014, Coll et al. proposed a new system to categorize all the lineages and families, which are currently described in the literature, using SNPs as stable markers of genetic variation for phylogenetic analysis (Coll et al. 2014).

Since our dataset was a group of 5249 SNP per sample, we used their method to provide us insight into the lineages of our samples (Benavente et al. 2015).

According to the results, each sample contains several SNPs matching different lineages and sub-lineages of the MTB complex, which leads to the conclusion that our population was exposed to more than one strain. In order to consider a unique lineage per sample, the one who had the most lineage-specific SNPs was the one selected (Table 9).

A table with all the specific lineage SNP matches is provided in the Supplementary Appendix.

Considering the results, 83,53% of our samples belong to the *Beijing* lineage, described for the first time in 1995(van Soolingen et al. 1995; Kremer et al. 2004). Together with the Haarlem lineage, known as “modern” lineages, both are associated with the massive spread of multidrug-resistant strains especially in Eurasia (Merker et al. 2013; Mokrousov 2013).

All the samples belonging to the Beijing lineage exhibit SNPs matching the Euro-American family, however the opposite is not observed. None of the samples belonging to the Euro-American family (*Harleem, Ural, LAM, H37Rv-Like, mainly T*) exhibit SNPs belonging to another family.

Table 9 – Summary of number of samples per Lineage.

Lineage	Number of samples
East-Asian (Beijing)	71
Euro-American (Haarlem)	5
Euro-American (Ural)	3
Euro-American (LAM)	3
Euro-American (H37Rv-like)	1
East African-Indian	1
Euro-American (mainly T)	1

As said above, the BI tree was used for the following analysis. In Figure 12 it is possible to see that samples belonging to the same lineage are neighboring, as expected (Lagos et al. 2016). The green sample belongs to the Euro-American T lineage; pink samples belong to the Euro-American LAM lineage; blue samples belong to the Euro-American Ural lineage; grey sample belongs to Euro-American H37Rv-like lineage followed by the purple samples belonging to the Euro-American Haarlem lineage. It's clear that the Euro-American family is neighboring and clearly separated from the East-Asian family by the yellow sample which belongs to East-African-Indian lineage. All the samples in red belong to the East Asian family, specifically the Beijing lineage.

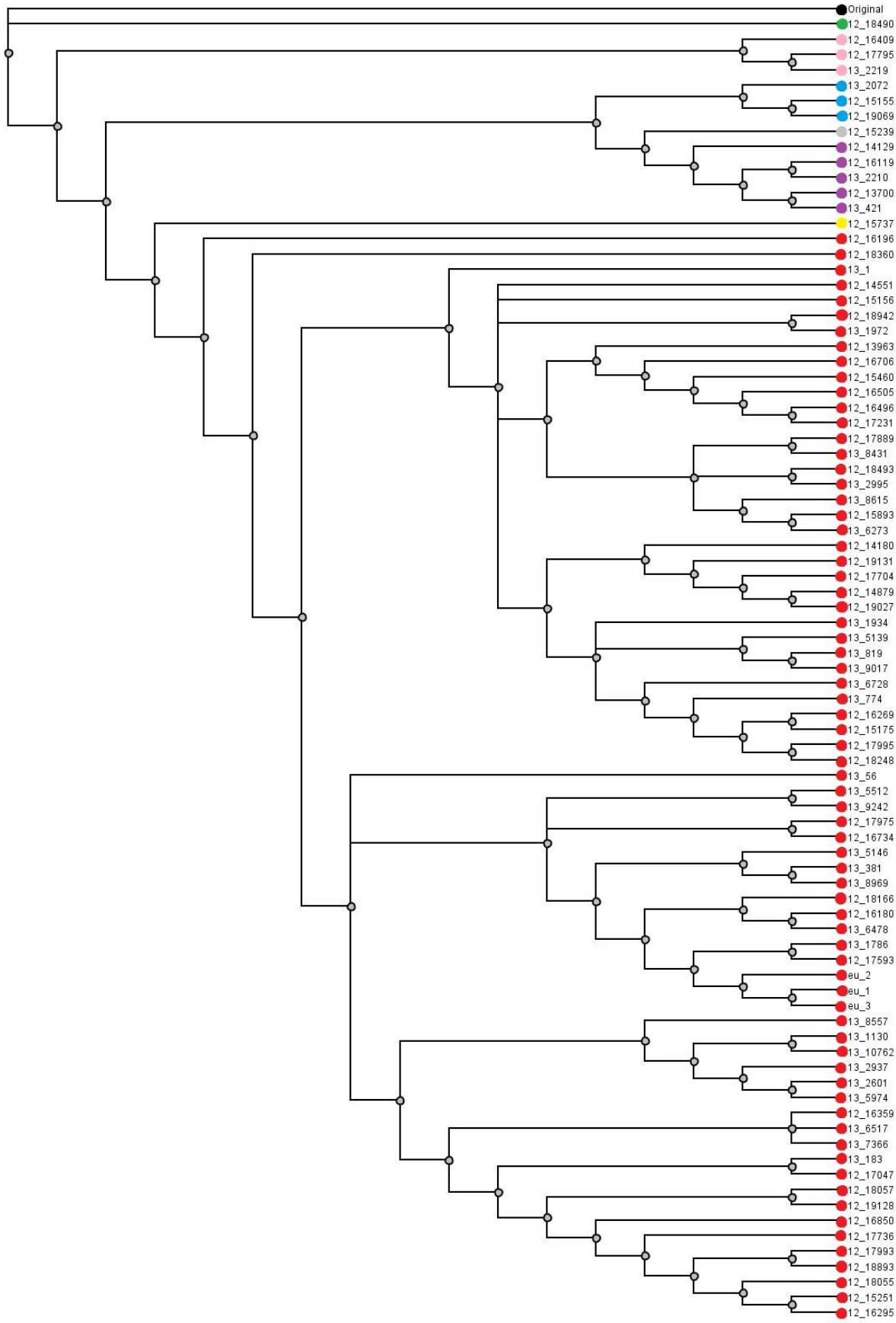

Figure 12 – BI tree with lineages distribution. *Green: Euro-American T lineage; pink: Euro-American LAM lineage; blue: Euro-American Ural lineage; grey: Euro-American H37Rv-like lineage; purple: Euro-American Haarlem lineage; yellow: East-African-Indian lineage; red: East Asian family, specifically the Beijing lineage.*

4.4 Population Structure

The population structure of isolates based on SNP information separated samples into previously described lineages (Figure 12). With three exceptions, little variation was seen in the follow up samples from the same patient over time, fluctuating between 4-14 SNPs.

Regarding the patient samples, in order to infer events of transmission, reinfection or evolution we did another tree using only the follow ups samples (Figure 13).

Looking at the beginning of the tree, it starts with a sample from Patient 11, baseline sample and then it separates into three different clades. The Clade 1 is composed by two samples from Patient 1. Looking closely at Patient 1 it is clearly a case of evolution, with two samples clustering with a difference of 6 SNPs, over 3 months. Both samples belong to Euro-American (Haarlem) lineage and they present the same rpoB and rpoC polymorphisms. Also, they both have resistance SNP matches for Ryfamicin, Streptomycin and Ethambutol.

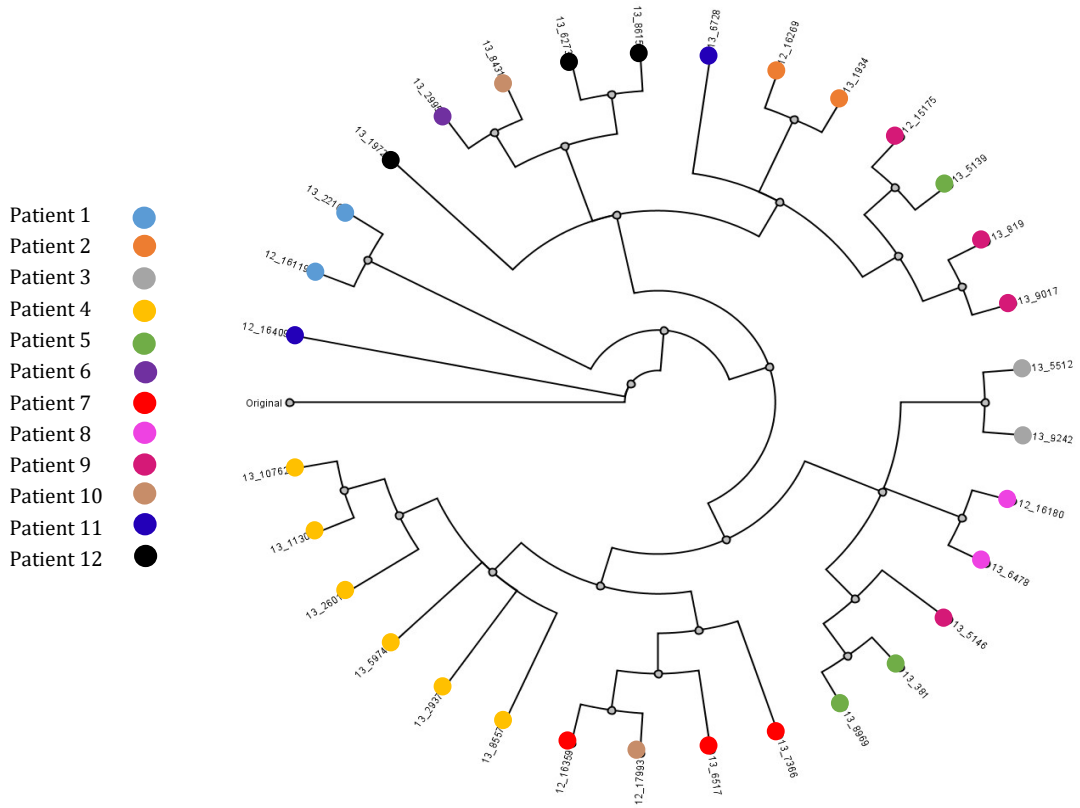

Figure 13. Circular tree with all the follow up samples from our 12 patients. With seven exceptions, follow ups from the same patient are close within the tree. We assume that these seven exceptions might be cases of reinfection while the other being cases of evolution.

Clade 2 starts with a sample from Patient 12, baseline sample and then the other two samples from Patient 12 appear clustering with Patient 6 and a sample from Patient 10 (6th month follow up). From this we assume it is a case of reinfection since the baseline sample from Patient 10 belong to a different and distant Clade 3, being the difference between the

samples on the order of >1000 SNPs. The next cluster inside Clade 2 is composed by 4 different Patients, 2, 5, 9 and 11. It all starts with Patient 11 (5^{th} month) and then Patient 2 with both samples clustering. Patient 2 is another case of evolution, with two samples clustering with a difference of 12 SNPs, over 3 months. Both samples belong to Beijing lineage and they present the same rpoB and rpoC polymorphisms. Also, they both have resistance SNP matches for Fluoroquinolones, Ryfamicin, Isoniazid, Pyrazinamide and Ethambutol. The last cluster is composed by Patient 5 (6^{th} month) and Patient 9 (baseline, 3^{rd}, 9^{th} month). We consider Patient 5 as a case of reinfection since the previous samples from it belong to a different and distant Clade 3 being the difference between the samples on the order of >200 SNPs. Patient 9 is also missing a sample from the 6^{th} month which is clustering with the remaining samples from Patient 5 in Clade 3.

Clade 3 is composed by Patient 3, 4, 5, 7, 8 and 10. Inside this clade there are two clearly separated clusters. The first one starts with Patient 3. Patient 3 is another case of evolution, with two samples clustering with a difference of 10 SNPs, over 3 months. Both samples belong to Beijing lineage and they present the same rpoB and rpoC polymorphisms. Also, they both have resistance SNP matches for Ryfamicin. Streptomycin, Isoniazid and Kanamycin. After that we Patient 8. Patient 8 is another case of evolution, with two samples clustering with a difference of 13 SNPs, over 3 months. Both samples belong to Beijing lineage and they present the same rpoB and rpoC polymorphisms. Also, they both have resistance SNP matches for Ryfamicin. Streptomycin, Isoniazid, Kanamycin and Fluoroquinolones. Then the case already discussed above with Patient 5 and 9.

The other cluster inside Clade 3, starts with Patient 7 and 10 clustering. The Patient 10 (baseline) seems to be primarily very close to Patient 7, with a difference between 25-18 SNPs. Patient 10 seems to be a case of reinfection since the 6^{th} month is clustering in Clade 2 with Patient 6 and 12.

Lastly, we have Patient 4. Patient 4 is another case of evolution, with six samples clustering with a difference of 15-1 SNPs, over 7 months. All samples belong to Beijing lineage and they present the same rpoB polymorphism, with no rpoC. Also, they all have resistance SNP matches for Ryfamicin. Streptomycin and Pyrazinamide.

Moreover in the samples collected from the same patient, it was observed that the number of SNPs tended to increase over time.

After focusing only on Patient, we decided to focus again in the whole population. Eleven clusters (53 samples) with variation less than 30 SNPs were examined to ascertain the relatedness of the strains; we examined the genes associated with drug resistance. Results are summarized in Table 10 and show that the majority of the clusters have mutations in genes associated with drug resistance. However there are two exceptions, Cluster C, F, H, I and K. Cluster C has an isolate that doesn't present Fluoroquinolones resistance like the other two. Cluster F has one isolate with Rifampicin resistance. Cluster H presents one isolate with Pyrazinamide resistance. Cluster I has 3 isolates that differ from others in the Fluoroquinolones and Streptomycin resistance. Cluster K characterized by resistance to Streptomycin has one isolate which besides that also presents Rifampicin, Streptomycin and Fluoroquinolones resistance.

Table 10 – Distribution of mutations in genes associated with drug resistance within each cluster. *In the drug column R: rifampicin, S: streptomycin, E: ethambutol, I: isoniazid, K: kanamycin, F: fluoroquinolones, P: pyrazinamide.*

Drugs	Cluster	Sample ID	Drugs	Cluster	Sample ID
R, S, E	A	12_16119	S, P, K	F	12_18893
R, S, E		13_2210	S, P, K		12_16295
R, S, I, K	B	13_5512	S, P, K		12_18055
R, S, I, K		13_9242	R, S		12_17593
R, S, E, K, F	C	12_16180	R, S, K, F	G	12_15893
R, S, E, K		12_18166	R, S, K, F		13_6273
R, S, E, K, F		13_6478	R, S, K, F		13_8615
R, S	D	13_381	R, S, P, K,E	H	12_17889
R, S		13_5146	R, S, K, E		13_8431
R, S		13_8969	R, E, I, F, P	I	12_16269
R, S, P	E	13_1130	R, E, I, P		12_17995
R, S, P		13_10762	R, S, E, I, P		12_18248
R, S, P		13_8557	R, E, I, P		12_15175
R, S, P		13_2601	R, E, I, P		13_819
R, S, P		13_2937	R, E, I, P		13_5139
R, S, P		13_5974	R, E, I, P		13_9017
S, P, K	F	12_16359	R, E, I, F, P		13_1934
S, P, K		13_6517	R, E, I, P		13_774

S, P, K		13_7366	R, E, I, P		13_6728
S, P, K		12_15251	R, I		12_19131
S, P, K		12_18057	R, I		12_14879
S, P, K		12_17736	R, I	J	12_19027
S, P, K		12_19128	R, I		12_17704
S, P, K		13_183	S		12_16505
S, P, K		12_16850	S	K	12_16496
S, P, K		12_17047	R, S, E, F		12_17231
S, P, K		12_17993			

4.5 Compensatory mutations *versus* drug resistance

Of the 85 isolates, 75 had matches for SNPs predictive of drug resistance and only 6 were resistant to one single drug (Figure 13).

Currently more than twenty drugs have been developed for the treatment of TB. The drugs are usually used in different combinations taking into account the circumstances of the patient, i.e., for "new" patients there is a specific group of drugs to be used and for patients with drug resistant TB another set of drugs is used.

The five "first line" TB drugs are generally the ones with the greatest activity against TB bacteria and are mostly used for someone with active TB disease who has not had TB drug treatment before. These are Isoniazid, Rifampicin, Pyrazinamide, Ethambutol and Streptomycin. The resistance to this drugs is commonly associated with treatment failure and poor clinical response to therapy (Hershfield 1999; Cox et al. 2007).

We decided to analyze the Rifampicin and Pyrazinamide resistance since most of our samples had SNPs matches for resistance to these drugs.

Resistance to rifampicin is predominantly acquired by mutations in *rpoB* (critical for cell viability), coding for the β-subunit of RNA polymerase. In an early stage, when rifampicin is administrated, most of the mutations on this gene provide selective advantage while having a negative impact on the bacteria's fitness. However the fitness may be restored with subsequent adaptations after the acquisition of primary resistance (Bergval et al. 2007).

By comparing mutations in *rpoB* in all the clusters with the results from Table 10, we found on the one hand the ones that didn't present any Rifampicin resistance are the ones with

only this specific A1075A (GCT→GCC) polymorphism; on the other hand the ones with S450L (TCG→TTG) polymorphism always presented Rifampicin resistance. Besides that it is also possible to have isolates in the same cluster with different *rpoB* polymorphisms, for example in cluster C, F, H and J, suggesting independent emergence of MDR within these clusters.

Pyrazinamide is an important sterilizing drug that shortens TB therapy. However, the mechanism of action of pyrazinamide is poorly understood because of its unusual properties (Zhang 2003). Besides that mutations on the *pncA* gene are associated with this drug (Zhang 1996).

The *pncA* polymorphisms were also analyzed. Despite the fact that all the clusters had mutations on the pncA gene, only four clusters E, F, H and I presented polymorphisms predictive of resistance to Pyrazinamide, T135P (ACC→CCC), C14R (TGC→CGC), Q141P (CAG→CCG), I6L (ATC→CTC) and H71R (CAT→CGT). Besides that it is also possible to have isolates in the same cluster with different *pncA* polymorphisms, for example cluster C, F and H suggesting independent emergence of MDR within these clusters.

A summary table with all the specific polymorphisms is provided in the supplementary appendix.

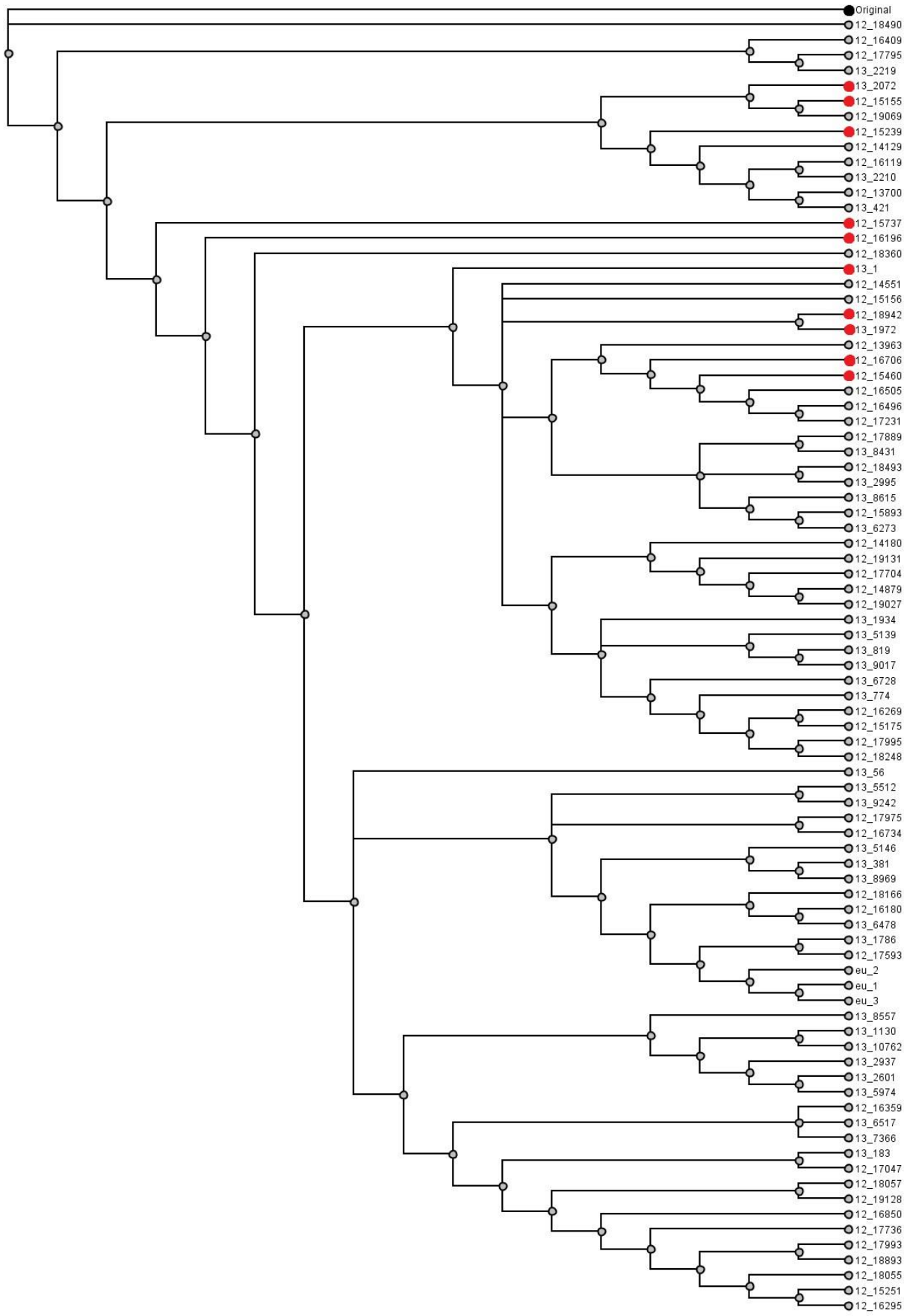

Figure 14 – Distribution of samples with and without SNPs matching for drug resistance. *In red are the samples without SNPs matching for drug resistance; in grey samples with SNPs matching for drug resistance.*

Little is known about the epidemiological relevance of compensatory evolution in MDR-TB. In order to achieve similar fitness as the wild-type strain, compensatory mutations ease the fitness cost associated with drug resistance mutations, e.g., *rpoC* mutation. Some *rpoC* mutations have a compensatory effect regarding the fitness cost associated with mutations in rifampicin-resistant bacteria carrying mutations in *rpoB*, which even in the absence of antibiotic pressure increase the fitness cost (De Vos et al. 2013).

With this in mind we decided to assess the distribution of *rpoB* and *rpoC* in all the isolates using the BI tree as a form of representation and we can highlight these six features:

1. 6 samples without any mutation in *rpoB* and 2 of them with *rpoC* mutation, leading to assume that it is not necessary to have the *rpoB* mutation in order to have *rpoC* mutation;
2. 32 samples without any mutation in *rpoC*;
3. 16 unique polymorphisms; 6 *rpoB* and 10 *rpoC;*
4. "Neighboring" isolates with different polymorphisms;
5. Only 51 samples with actual resistance to rifampicin despite the fact that 79 samples presented mutation in *rpoB*;
6. We assume that the last 19 isolates in the tree don't have mutations in *rpoC* because the *rpoB* mutation is not conferring resistance to rifampicin.

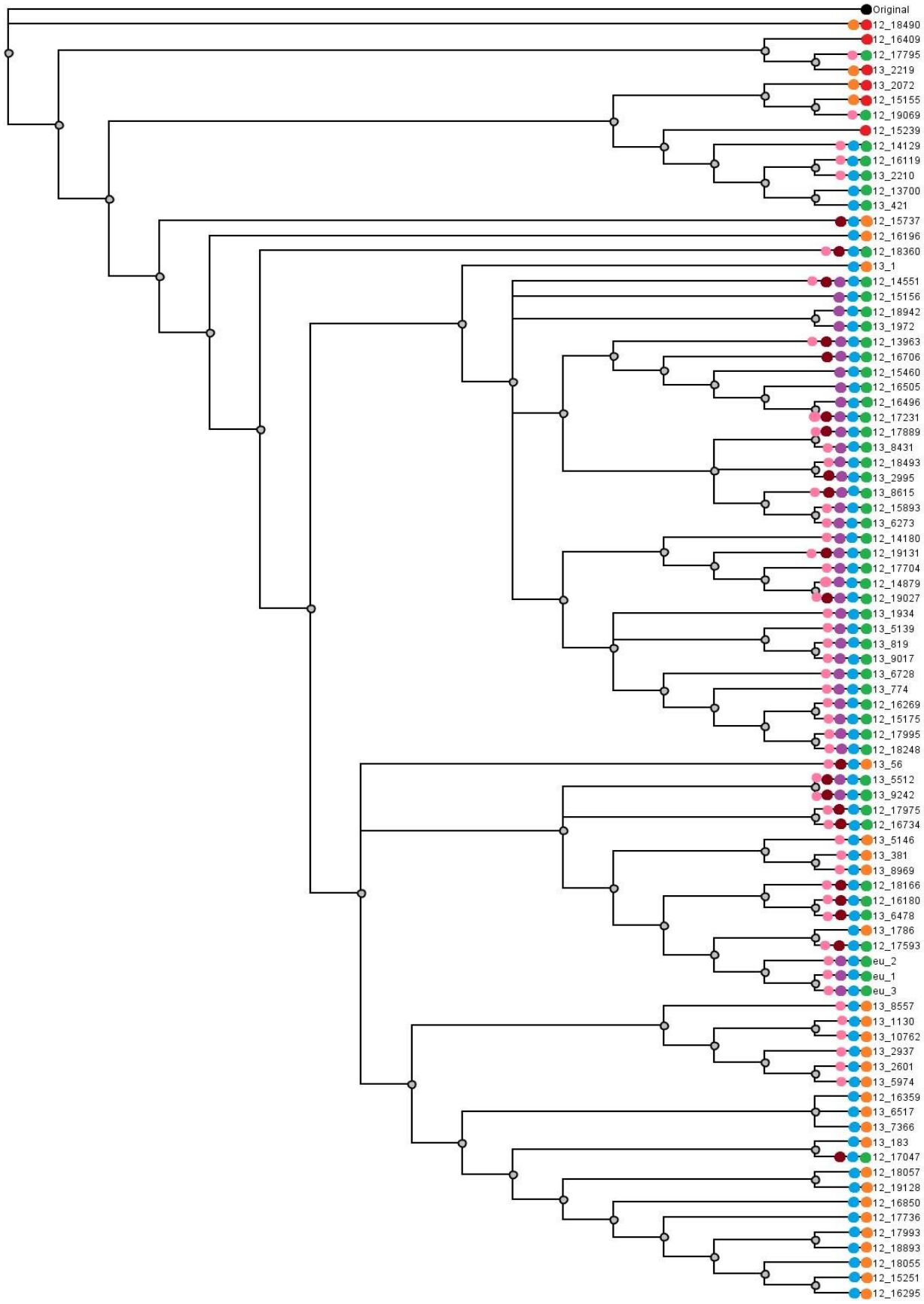

Figure 15 – BI tree with the distribution of rpoB and rpoC mutations. *Red: no rpoB mutation; Orange: no rpoC mutation; Green: with rpoB & rpoC mutation; Blue: same rpoB mutation as nearest neighbors; Purple: same rpoC mutation as nearest neighbors; Deep Red: single event; Pink: rifampicin resistance.*

5. CONCLUSIONS AND FUTURE PERSPECTIVES

In this study we used WGS in order to understand the genotypic and epidemiological factors that might influence the spread and fitness of this MTB by analyzing deep –sequencing data of 85 isolates from Central Asia.

We found that the amount of variation accumulated within a patient can be as high as that observed between patients along, what we assume to be, a chain of transmission. Intrapatient diversity was found in all of the follow up patients. The analysis of the mechanisms responsible for this microevolution, i.e., the genetic variability of MTB in a short period of time, of a parental strain into clonal variants is a relevant issue that needs to be addressed.

Regarding the eleven epidemic clusters, ten belonged to the Beijing lineage and one to the Harleem lineage, both associated with the massive spread of MDR strains. Within clusters independent emergence of MDR was observed.

Relationship between mutations on *rpoB* and *rpoC* were associated with drug resistance to rifampicin and compensatory evolution, thereby contributing to the spread of drug resistance. Not all mutations in *rpoB* confer resistance to rifampicin, we were able to find that the S450L polymorphism confers but A1075A polymorphism does not. Moreover, our data confirms the convergent evolution of specific compensatory *rpoC* mutations, indicating its positive selection.

Mutations on *pncA* demonstrated to be related with drug resistance to pyrazinamide but there was not enough evidence to relate it with a mechanism of compensatory evolution. We consider that this issue should be addressed in the future.

We believe that our study adds new data to the understandings of the variability among MTB strains in an intra and interpatient microevolution scenario. Moreover we suggest that alternative mechanisms of fitness compensation might exist.

Last of all we strongly believe that the improved understanding of key success factors in growth and transmission of MTB will provide additional or improved therapeutic approaches.

BIBLIOGRAPHY

Akaki, T., Tomioka, H. & Shimizu, T., 2000. Comparative roles of free fatty acids with reactive nitrogen intermediates and reactive oxygen intermediates in expression of the anti-microbial activity of macrophages against Mycobacterium tuberculosis. *Clin exp Immunol*, 121, pp.302–310.

Anderson, J. et al., 2013. Sublineages of lineage 4 (Euro-American) Mycobacterium tuberculosis differ in genotypic clustering SUMMARY. , 17(February), pp.885–891.

Anisimova, M. & Gascuel, O., 2006. Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Systematic biology*, 55(4), pp.539–552.

Barrick, J.E. et al., 2014. Identifying structural variation in haploid microbial genomes from short-read resequencing data using breseq. , pp.1–17.

Benavente, E.D. et al., 2015. PhyTB : Phylogenetic tree visualisation and sample positioning for M . tuberculosis. , pp.0–4.

Bergval, I.L. et al., 2007. Specific mutations in the Mycobacterium tuberculosis rpoB gene are associated with increased dnaE2 expression. *FEMS Microbiology Letters*, 275(2), pp.338–343.

Black, P.A. et al., 2015. Whole genome sequencing reveals genomic heterogeneity and antibiotic purification in Mycobacterium tuberculosis isolates. *BMC genomics*, 16(1), p.857. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4619333&tool=pmcentrez&rendertype=abstract.

Blouin, Y. et al., 2012. Significance of the Identification in the Horn of Africa of an Exceptionally Deep Branching Mycobacterium tuberculosis Clade. , 7(12).

Boshoff, H., Lun, D., 2011. Systems biology approaches to understanding mycobacterial survival mechanisms. *Drug Discovery Today*, 7(1), pp.1–13.

Bozzano, F., Marras, F. & De Maria, A., 2014. Immunology of tuberculosis. *Mediterranean journal of hematology and infectious diseases*, 6(1), p.e2014027.

Brooks, D.R. et al., 2007. Quantitative Phylogenetic Analysis in the 21 st Century. *Revista Mexicana de Biodiversidad*, 78, pp.225–252.

Carmona, J. et al., 2013. Mycobacterium tuberculosis Strains Are Differentially Recognized by TLRs with an Impact on the Immune Response. , 8(6), pp.1–10.

Cole, S.T. et al., 1998. Deciphering the biology of Mycobacterium tuberculosis from the

complete genome sequence. *Nature*, 396(September), pp.537–544.

Coll, F. et al., 2014. PolyTB: A genomic variation map for Mycobacterium tuberculosis. *Tuberculosis*, 94(3), pp.346–354. Available at: http://dx.doi.org/10.1016/j.tube.2014.02.005.

Comas, I. et al., 2011. Whole-genome sequencing of rifampicin-resistant M. tuberculosis strains identifies compensatory mutations in RNA polymerase. *Nature Genetics*, 44(1), pp.106–110. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3246538/%5Cnhttp://www.ncbi.nlm.nih.gov/pmc/articles/PMC3246538/pdf/nihms339089.pdf.

Cooper, A.M., 2009. Cell mediated immune responses in Tuberculosis Andrea. *Annu Rev Immunol.*, 27, pp.393–422.

da Costa, A.C. et al., 2014. Recombinant BCG: Innovations on an old vaccine. Scope of BCG strains and strategies to improve long-lasting memory. *Frontiers in Immunology*, 5(APR), pp.1–9.

Cox, H.S. et al., 2007. Risk of acquired drug resistance during short-course directly observed treatment of tuberculosis in an area with high levels of drug resistance. *Clinical Infectious Diseases*, 44(11), pp.1421–1427.

Denamur, E. & Matic, I., 2006. MicroReview Evolution of mutation rates in bacteria. *Molecular Microbiology*, 60(March), pp.820–827.

Didelot, X. et al., 2016. Within-host evolution of bacterial pathogens. *Nature Publishing Group*, 14(3), pp.150–162. Available at: http://dx.doi.org/10.1038/nrmicro.2015.13.

Edgar, R.C., Drive, R.M. & Valley, M., 2004. MUSCLE : multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids research*, 32(5), pp.1792–1797.

Elena, S.F. & Lenski, R.E., 2003. EVOLUTION EXPERIMENTS WITH MICROORGANISMS : THE DYNAMICS AND GENETIC BASES OF ADAPTATION. , 4(June).

Farhat, M.R. et al., 2013. Genomic analysis identifies targets of convergent positive selection in drug-resistant Mycobacterium tuberculosis. *Nature genetics*, 45(10), pp.1183–9. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3887553&tool=pmcentrez&rendertype=abstract.

Flynn, JoAnne L., Chan, J., 2001. Immunology of Tuberculosis. *Annual Review of Immunology*, 19, pp.93–129.

Flynn, J.L. & Chan, J., 2001. Mmunology of. , pp.93–129.

Ford, C.B. et al., 2014. Mycobacterium tuberculosis mutation rate estimates from different lineages predict substantial differences in the emergence of drug resistant tuberculosis. *Nature genetics*, 45(7), pp.784–790.

Gardy, J.L. et al., 2011. Whole-Genome Sequencing and Social-Network Analysis of a Tuberculosis Outbreak. *The New England Journal of Medicine*, 364(8), pp.730–739.

Garra, A.O. et al., 2013. *The Immune Response in Tuberculosis*,

Gori, A. et al., 2005. Spoligotyping and Mycobacterium tuberculosis. *Emerging Infectious Diseases*, 11(8), pp.1242–1248. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3320497/.

Guindon, S., Dufayard, J. & Lefort, V., 2010. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies : Assessing the Performance of PhyML 3 . 0. *Syst. Biol.*, 59(3), pp.307–321.

Hasegawa, M., Kishino, H. & Yano, T., 1985. Dating of the Human-Ape Splitting by a Molecular Clock of Mitochondrial DNA. *Journal of Molecular Evolution*, 22, pp.160–174.

Hastings, W.K., 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1), pp.97–109.

Hershfield, E., 1999. Tuberculosis 9. Treatment. *Canadian Medical Association Journal*, 161(4), pp.405–411.

Huelsenbeck, J.P. et al., 2002. Potential applications and pitfalls of Bayesian inference of phylogeny. *Systematic biology*, 51(5), pp.673–688.

Jarher, V. & Nikaido, H., 1994. Mycobacterial cell wall : Structure and role in natural resistance to antibiotics. *FEMS Microbiology Letters*, 123, pp.11–18.

Kleinnijenhuis, J. et al., 2011. Innate immune recognition of Mycobacterium tuberculosis. *Clinical & developmental immunology*, 2011, p.405310.

Kremer, K. et al., 2004. Definition of the Beijing / W Lineage of Mycobacterium tuberculosis on the Basis of Genetic Markers. , 42(9), pp.4040–4049.

Krishnan, N. et al., 2011. Mycobacterium tuberculosis Lineage Influences Innate Immune Response and Virulence and Is Associated with Distinct Cell Envelope Lipid Profiles. , 6(9).

Lagos, J. et al., 2016. Analysis of Mycobacterium tuberculosis Genotypic Lineage Distribution in Chile and Neighboring Countries R. Manganelli, ed. *PLoS ONE*, 11(8), p.e0160434.

Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4982630/.

Langmead, B. & Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. *Nature methods*, 9(4), pp.357–360.

Lee, R.S. & Behr, M.A., 2016. The implications of whole-genome sequencing in the control of tuberculosis. *Therapeutic Advances in Infectious Disease*, pp.47–62.

Lew, J.M. et al., 2011. TubercuList e 10 years after. *Tuberculosis*, 91(1), pp.1–7. Available at: http://dx.doi.org/10.1016/j.tube.2010.09.008.

Lewandowski, C.M., Co-investigator, N. & Lewandowski, C.M., 2015. WHO Global tuberculosis report 2015. *The effects of brief mindfulness intervention on acute pain experience: An examination of individual difference*, 1, pp.1689–1699.

Lin, P.L. et al., 2009. Quantitative comparison of active and latent tuberculosis in the cynomolgus macaque model. *Infection and Immunity*, 77(10), pp.4631–4642.

Luo, C. et al., 2012. Direct Comparisons of Illumina vs . Roche 454 Sequencing Technologies on the Same Microbial Community DNA Sample. , 7(2).

MacMicking, J.D. et al., 1997. Identification of nitric oxide synthase as a protective locus. *Proc. Natl. Acad. Sci*, 94(May), pp.5243–5248.

Maeda, N. et al., 2003. The cell surface receptor DC-SIGN discriminates between Mycobacterium species through selective recognition of the mannose caps on lipoarabinomannan. *Journal of Biological Chemistry*, 278(8), pp.5513–5516.

Mau, B. et al., 1999. Bayesian Phylogenetic Inference via Markov Chain Monte Carlo Methods. *Biometrics*, 6(March), pp.1–12.

Medzhitov, R., 2007. Recognition of microorganisms and activation of the immune response. , 449(October), pp.819–826.

Mena, A. et al., 2008. Genetic adaptation of Pseudomonas aeruginosa to the airways of cystic fibrosis patients is catalyzed by hypermutation. *Journal of Bacteriology*, 190(24), pp.7910–7917.

Merker, M. et al., 2013. Whole genome sequencing reveals complex evolution patterns of multidrug-resistant Mycobacterium tuberculosis Beijing strains in patients. *PLoS ONE*, 8(12), pp.1–11.

Metropolis, N. et al., 1953. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6), pp.1087–1092.

Mokrousov, I., 2013. Insights into the origin, emergence, and current spread of a successful Russian clone of Mycobacterium tuberculosis. *Clinical Microbiology Reviews*, 26(2),

pp.342–360.

North, R. & Jung, Y., 2004. Immunity to Tuberculosis. *Annual Review of Immunology*, 22, pp.599–623. Available at: http://arjournals.annualreviews.org/doi/full/10.1146/annurev.immunol.22.012703.104635%5Cnpapers2://publication/uuid/98AE2274-FFAD-4A54-B5EA-1DAC28517F5C.

Pérez-Lago, L. et al., 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), pp.98–108.

Portevin, D. & Gagneux, S., 2011. Human Macrophage Responses to Clinical Isolates from the Mycobacterium tuberculosis Complex Discriminate between Ancient and Modern Lineages. , 7(3).

Prezzemolo, T. et al., 2014. Functional Signatures of Human CD4 and CD8 T Cell Responses to Mycobacterium tuberculosis. *Frontiers in immunology*, 5(April), p.180. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4001014&tool=pmcentrez&rendertype=abstract.

R Development Core Team, 2008. R: A Language and Environment for Statistical Computing. Available at: http://www.r-project.org.

Raja, A., 2004. Immunology of tuberculosis. *Indian Journal of Medical Research*, 120(4), pp.213–232.

Reyes, F. et al., 2013. Immunogenicity and cross-reactivity against Mycobacterium tuberculosis of proteoliposomes derived from Mycobacterium bovis BCG. *BMC Immunology*, 14(Suppl 1), p.S7. Available at: http://www.biomedcentral.com/1471-2172/14/S1/S7.

Rich, E.A. et al., 1997. Mycobacterium tuberculosis (MTB). stimulated production of nitric oxide by human alveolar macrophages and relationship of nitric oxide production to growth inhibition of MTB. *Tubercle and Lung Disease*, 78, pp.247–255.

Roitt I, Brostoff J, M.D., 1998. *Immunology*,

Ronquist, F. et al., 2012. MrBayes 3 . 2 : Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. *Sys*, 61(3), pp.539–542.

Sanger, F., Nicklen, S. & Coulson, A.R., 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12), pp.5463–5467. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC431765/.

Saunders, N.J. et al., 2011. Deep resequencing of serial sputum isolates of Mycobacterium tuberculosis during therapeutic failure due to poor compliance reveals stepwise mutation of key resistance genes on an otherwise stable genetic background. *Journal of Infection*, 62(3), pp.212–217. Available at: http://dx.doi.org/10.1016/j.jinf.2011.01.003.

Schrag, J.D. et al., 1997. The open conformation of a Pseudomonas lipase. *Structure*, 5, pp.187–202.

Seo, K.W. et al., 2014. Persistently Retained Interferon-Gamma Responsiveness in Individuals with a History of Pulmonary Tuberculosis. , pp.123–128.

van Soolingen, D. et al., 1995. Predominance of a single genotype of Mycobacterium tuberculosis in countries of east Asia. *Journal of Clinical Microbiology*, 33(12), pp.3234–3238. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC228679/.

Springer, B. et al., 2004. Lack of mismatch correction facilitates genome evolution in mycobacteria. *Molecular microbiology*, 53(6), pp.1601–1609.

Sun, G. et al., 2012. Dynamic Population Changes in Mycobacterium tuberculosis During Acquisition and Fixation of Drug Resistance in Patients. *Journal of Infectious Diseases*, 206.

Torrado, E. & Cooper, A.M., 2010. IL-17 and Th17 cells in tuberculosis. *Cytokine Groeth Factor Rev.*, 21(6), pp.455–462.

Torrado, E. & Copper, A., 2013. Cytokines in balance of protection and pathology during mycobacterial infections. *Adv Exp Med Biol*, 783, pp.121–140.

Torrelles, J.B. & Schlesinger, L.S., 2011. Diversity in *M. tuberculosis* mannosylated cell wall dterminants impact adaptation to the host. *Tuberculosis (Edinburgh, Scotland)*, 90(2), pp.84–93.

Tukvadze, N., Bergval, I. & Bablishvil, N., 2016. Evaluation of SNP-based genotyping to monitor tuberculosis control in a high MDR-TB setting. *bioRxiv*, 31(0), pp.1–22.

De Vos, M. et al., 2013. Putative compensatory mutations in the rpoc gene of rifampin-resistant mycobacterium tuberculosis are associated with ongoing transmission. *Antimicrobial Agents and Chemotherapy*, 57(2), pp.827–832.

Dos Vultos, T. et al., 2009. DNA repair in Mycobacterium tuberculosis revisited. *FEMS Microbiology Reviews*, 33(3), pp.471–487.

Dos Vultos, T. et al., 2008. Evolution and diversity of clonal bacteria: The paradigm of Mycobacterium tuberculosis. *PLoS ONE*, 3(2), pp.1–10.

Warner, D.F. & Mizrahi, V., 2006. Tuberculosis Chemotherapy : the Influence of Bacillary

Stress and Damage Response Pathways on Drug Efficacy. *Clinical Microbiology Reviews*, 19(3), pp.558–570.

Weniger, T. et al., 2010. MIRU-VNTRplus: a web tool for polyphasic genotyping of Mycobacterium tuberculosis complex bacteria. *Nucleic Acids Research*, 38(Web Server issue), pp.W326–W331. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2896200/.

Whitlock, M.C., 2003. Fixation Probability and Time in Subdivided Populations. *Genetics Society of America*, 779(June), pp.767–779.

WHO, 2014. Antimicrobial Resitance - Global Report on Surveillance.

Yimer, S.A. et al., 2015. Mycobacterium tuberculosis Lineage 7 Strains Are Associated with Prolonged Patient Delay in Seeking Treatment for Pulmonary Tuberculosis in Amhara Region , Ethiopia. , 53(4), pp.1301–1309.

SUPPLEMENTARY APPENDIX

Table 11 - Summary of before and after filter of all mutations.

ID	before	after	differences	ID	before	after	differences
12_13700	4256	1059	3197	12_18893	5175	1678	3497
12_13963	5236	1677	3559	12_18942	5266	1627	3639
12_14129	4176	1031	3145	12_19027	5167	1683	3484
12_14180	5063	1611	3452	12_19069	4253	1093	3160
12_14551	5080	1598	3482	12_19128	5065	1613	3452
12_14879	5206	1687	3519	12_19131	5199	1595	3604
12_15155	4358	1064	3294	13_1	5025	1617	3408
12_15156	5215	1586	3629	13_10762	5373	1670	3703
12_15175	5132	1690	3442	13_1130	5438	1709	3729
12_15239	4076	1033	3043	13_1786	5303	1723	3580
12_15251	5262	1686	3576	13_183	5302	1602	3700
12_15460	5286	1672	3614	13_1934	5248	1600	3648
12_15737	5119	1509	3610	13_1972	5116	1634	3482
12_15893	5258	1695	3563	13_2072	4249	1116	3133
12_16119	3862	1049	2813	13_2210	4279	1056	3223
12_16180	5266	1713	3553	13_2219	3879	950	2929
12_16196	5146	1566	3580	13_2601	5109	1643	3466
12_16269	5283	1704	3579	13_2937	5221	1620	3601
12_16295	5292	1712	3580	13_2995	5178	1608	3570
12_16359	5245	1701	3544	13_381	5205	1686	3519
12_16409	4155	960	3195	13_421	4186	1060	3126
12_16496	5255	1673	3582	13_5139	5319	1607	3712
12_16505	5202	1652	3550	13_5146	5302	1626	3676
12_16706	5195	1668	3527	13_5512	5167	1647	3520
12_16734	4996	1625	3371	13_56	5090	1625	3465
12_16850	5212	1615	3597	13_5974	5053	1618	3435
12_17047	5186	1630	3556	13_6273	5205	1614	3591
12_17231	5200	1715	3485	13_6478	5331	1641	3690
12_17593	5285	1715	3570	13_6517	5237	1612	3625
12_17704	5575	1877	3698	13_6728	5285	1625	3660
12_17736	5123	1616	3507	13_7366	5262	1624	3638
12_17795	4001	941	3060	13_774	5037	1597	3440
12_17889	5278	1675	3603	13_819	5179	1609	3570
12_17975	5179	1623	3556	13_8431	5367	1627	3740
12_17993	5317	1677	3640	13_8557	5386	1639	3747
12_17995	5255	1693	3562	13_8615	5266	1622	3644
12_18055	5242	1709	3533	13_8969	5190	1624	3566
12_18057	5028	1617	3411	13_9017	5227	1627	3600

12_18166	5205	1712	3493	13_9242	6880	1641	5239
12_18248	5196	1690	3506	eu_1	6387	1657	4730
12_18360	5148	1566	3582	eu_2	7106	1539	5567
12_18490	2808	607	2201	eu_3	6337	1650	4687
12_18493	5074	1623	3451				

Table 12 - Summary of before and after filter of SNPs.

ID	before	after	differences	ID	before	after	differences
12_13700	922	766	156	12_18893	1435	1173	262
12_13963	1477	1188	289	12_18942	1414	1189	225
12_14129	911	775	136	12_19027	1455	1176	279
12_14180	1405	1193	212	12_19069	965	819	146
12_14551	1386	1178	208	12_19128	1394	1182	212
12_14879	1459	1176	283	12_19131	1397	1186	211
12_15155	927	798	129	13_1	1406	1192	214
12_15156	1372	1172	200	13_10762	1447	1191	256
12_15175	1455	776	679	13_1130	1469	1188	281
12_15239	900	775	125	13_1786	1471	1187	284
12_15251	1445	1180	265	13_183	1391	1179	212
12_15460	1434	1174	260	13_1934	1390	1187	203
12_15737	1328	1139	189	13_1972	1419	1192	227
12_15893	1453	1183	270	13_2072	981	833	148
12_16119	915	771	144	13_2210	922	775	147
12_16180	1426	1190	236	13_2219	838	717	121
12_16196	1344	1134	210	13_2601	1417	1189	228
12_16269	1465	1179	286	13_2937	1400	1186	214
12_16295	1462	1178	284	13_2995	1399	1191	208
12_16359	1458	1175	283	13_381	1459	1187	272
12_16409	853	720	133	13_421	931	782	149
12_16496	1444	1179	265	13_5139	1398	1190	208
12_16505	1427	1168	259	13_5146	1410	1186	224
12_16706	1433	1167	266	13_5512	1427	1200	227
12_16734	1410	1189	221	13_56	1412	1189	223
12_16850	1396	1180	216	13_5974	1399	1187	212
12_17047	1408	1181	227	13_6273	1399	1188	211
12_17231	1473	1181	292	13_6478	1419	1191	228
12_17593	1462	1172	290	13_6517	1393	1191	202
12_17704	1588	1183	405	13_6728	1410	1187	223
12_17736	1395	1180	215	13_7366	140244	1180	222
12_17795	836	715	121	13_774	1396	1185	211
12_17889	1441	1181	260	13_819	1394	1192	202
12_17975	1415	1191	224	13_8431	1416	1190	226
12_17993	1436	1173	263	13_8557	1414	1189	225
12_17995	1455	1183	272	13_8615	1409	1188	221
12_18055	1461	1179	282	13_8969	1406	1185	221
12_18057	1405	1185	220	13_9017	1415	1190	225
12_18166	1462	1189	273	13_9242	1428	1198	230
12_18248	1452	1181	271	eu_1	1449	1186	263
12_18360	1355	1151	204	eu_2	1357	1173	184
12_18490	533	451	82	eu_3	1433	1184	249
12_18493	1411	1195	216				

Script used for filtering the mutations of our samples.

Input: Matrix with all the positions and mutations of all the samples.

Output: Tables with only the SNPs of interest of all the samples.

```
############## FILTERS ##############


##### Converting .gd to .csv #####

setwd("C:/Users/marlichimi/Desktop/MBINF/THESIS/FILTER/GD")
library(dplyr)
library(data.table)

filenames <- list.files("C:/Users/marlichimi/Desktop/MBINF/THESIS/FILTER/GD", pattern = ".gd",full.names=FALSE)
col = 0
for (file in filenames){
  temp = max(na.omit(count.fields(file, sep = "\t")))
  if (temp>col){
    col = temp
  }
}

all_samples = data.frame(matrix(nrow = 1,ncol = 5))
colnames(all_samples) = c("id_sample","ref_genome","position","type","new_seq")
mutations = c("SNP", "SUB", "DEL", "INS", "MOB", "AMP", "CON", "INV")

for (file in filenames){

  gd = fread(file,sep = "\n",sep2 = "\t",stringsAsFactors = F)

  temp = data.frame(matrix(nrow = dim(gd)[1], ncol = col))

  for (i in 1:dim(gd)[1]){
    tstring = strsplit(as.character(gd[i,]),split = '\t')
    for (j in 1:length(tstring[[1]])){
      temp[i,j] = tstring[[1]][j]
    }
```

```
        }

        idx_to_mantain = c()

        for (mutation in mutations){
          idx = which(temp$X1==mutation)
          idx_to_mantain = append(idx_to_mantain,idx)
        }

        temp = temp[idx_to_mantain,]
        temp = temp[,c(2,4,5,1,6)]
        temp[,1] = substr(file,1,nchar(file)-3)
        colnames(temp)                                                    =
c("id_sample","ref_genome","position","type","new_seq")

        temp_name                                                         =
paste("C:/Users/marlichimi/Desktop/MBINF/THESIS/FILTER/Original_csv/",subst
r(file,1,nchar(file)-3),sep="")
        temp_name = paste(temp_name,".csv",sep="")


        write.csv(temp,temp_name,row.names = F)

        all_samples = rbind(all_samples,temp)

      }

      write.csv(all_samples[2:dim(all_samples)[1],],
"C:/Users/marlichimi/Desktop/MBINF/THESIS/FILTER/All_Samples/all_samples.cs
v",row.names = F)

      ##### Filtering By Family #####


      setwd("C:/Users/marlichimi/Desktop/MBINF/THESIS/FILTER/Original_csv")

      family                                                              =
read.csv("C:/Users/marlichimi/Desktop/MBINF/THESIS/FILTER/familia.csv",sep=
";",header = F)
      family = family[,1:2]
      interval = c()
```

```
      for(i in 1:dim(family)[1]){
        temp = c(family[i,1] : family[i,2])
        interval = append(interval, temp)
      }

      filenames <-
list.files("C:/Users/marlichimi/Desktop/MBINF/THESIS/FILTER/Original_csv",
pattern = ".csv",full.names=FALSE)
      all_samples_filtred = data.frame(matrix(nrow = 1,ncol = 5))
      colnames(all_samples_filtred) =
c("id_sample","ref_genome","position","type","new_seq")

      for (file in filenames){
        idx_to_remove = c()
        temp = read.csv(file,stringsAsFactors = F, sep = ",")
        for(i in 1:dim(temp)[1]){
          if (temp[i,'position'] %in% interval){
            idx_to_remove = append(idx_to_remove,i)
          }
        }

        temp_filtred = temp[-idx_to_remove,]

        temp_name =
paste("C:/Users/marlichimi/Desktop/MBINF/THESIS/FILTER/Filtred_By_Family/",
substr(file,1,nchar(file)-4),sep="")
        temp_name = paste(temp_name,"_filtred_by_family.csv",sep = "")

        write.csv(temp_filtred,temp_name,row.names = F)

        all_samples_filtred = rbind(all_samples_filtred,temp_filtred)

        # print(file)
      }

      write.csv(all_samples_filtred[2:dim(all_samples_filtred)[1],],"C:/Use
rs/marlichimi/Desktop/MBINF/THESIS/FILTER/All_Samples/all_samples_filtred.c
sv",row.names = F)


      ##### Filtering By Distance #####
```

```
setwd("C:/Users/marlichimi/Desktop/MBINF/THESIS/FILTER/Filtred_By_Family/")

all_samples_filtred = read.csv("C:/Users/marlichimi/Desktop/MBINF/THESIS/FILTER/All_Samples/all_samples_filtred.csv")

#transposing matrix

samples = unique(all_samples_filtred[,1])

# to create an empty data frame
## data.frame(matrix(nrow = x,ncol= y))
### x in this case is the number of all positions and y is the number of samples
trans_all_samples_filtred = data.frame(matrix(nrow = dim(all_samples_filtred)[1],ncol = length(samples)))
colnames(trans_all_samples_filtred) = samples
for (sample in samples){
  idx = which(all_samples_filtred[,1]==sample) #where row are related to the 'sample' variable
  for(i in 1:length(idx)){
    trans_all_samples_filtred[i,sample] = all_samples_filtred[idx[i],'position'] #only getting the position column
  }
}

#write a csv for each filtred sample in a new folder
for (i in 1:dim(trans_all_samples_filtred)[2]){
  name = paste("./Filtred_By_Family(Only_Position)/",colnames(trans_all_samples_filtred)[i],sep="")
  name = paste(name,".csv",sep = "")
  print(name)

write.csv(trans_all_samples_filtred[order((trans_all_samples_filtred[,i])), i], file = name,row.names = F,na = "")
}
```

```
      filenames <-
list.files("C:/Users/marlichimi/Desktop/MBINF/THESIS/FILTER/Filtred_By_Fami
ly/Filtred_By_Family(Only_Position)", pattern = ".csv",full.names=FALSE)

      values_to_remove = c()
      setwd("C:/Users/marlichimi/Desktop/MBINF/THESIS/FILTER/Filtred_By_Fam
ily/Filtred_By_Family(Only_Position)")

      for (file in filenames){ # for each sample

        temp = read.csv(file) #temporary object with the positions of the
sample
        temp[,1] = sort(temp[,1]) # sort the list to make the sum
        for(i in 1:dim(temp)[1]){
          sum_pos = 0 # variable to add if the sum of the next positions
          if(temp[i,1] %in% values_to_remove){ # if the position was already
meant to be removed
            next
          }
          if (i < dim(temp)[1]){

            k = 1 #number of row(s) that you will search

            sum_pos = sum_pos + temp[i+k,1]-temp[i,1] # temporary variable
to have the sum of the position
            print(sum_pos)

            if (sum_pos<12){
              values_to_remove = append(values_to_remove, temp[i,1]) # to add
the 1st one in which we are looking for
            }

            while (sum_pos < 12){
              if(temp[i+k,1] %in% values_to_remove){
                print("already exists")
              }
              else{
                values_to_remove = append(values_to_remove, temp[i+k,1])
              }
              sum_pos = sum_pos + temp[i+k+1,1]-temp[i+k,1]
              k = k+1
```

```
            }
          }
        }
        print(length(values_to_remove))
        print(sort(values_to_remove))
      }



      idx_to_remove = c()
      for (value in values_to_remove){
        if (value %in% all_samples_filtred[,'position']){
          idx_to_remove
=append(idx_to_remove,which(all_samples_filtred[,'position']==value))
        }
      }
      all_samples_final_filtred = all_samples_filtred[-idx_to_remove,]

      write.csv(all_samples_final_filtred,"C:/Users/marlichimi/Desktop/MBIN
F/THESIS/FILTER/All_Samples/final_all_samples_filtred.csv",row.names = F)

      setwd("C:/Users/marlichimi/Desktop/MBINF/THESIS/FILTER/Filtred_by_Fam
ily/Filtred_By_Family(Only_Position)/Filtred_By_Family_And_Distance")

      samples = unique(all_samples_final_filtred[,1]) # vector with the 81
samples
      for (sample in samples){
        idx = which(all_samples_final_filtred[,1]==sample)# vector with the
row numbers related to that sample
        idx                          =                          intersect(idx,
which(all_samples_final_filtred[,'type']=='SNP'))
        filen = paste(as.character(sample),".csv",sep="") # concatenate the
sample name with ".csv" to write the file
        write.table(all_samples_final_filtred[idx,],file = filen,row.names =
F,sep=",")
      }
```

Table 13 – Specific lineage SNP matches obtained for each sample using PhyTB.

SAMPLES	SPECIFIC LINEAGE SNP MATCHES	MAIN LINEAGE
12_13700	**##lineage4.1.2.1 = 20**	**Euro-American (Haarlem)**
	##lineage4.1 = 12	
	##lineage4.1.2 = 3	
	##lineage4.9 = 2	
12_13963	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	
12_14129	**##lineage4.1.2.1 = 20**	**Euro-American (Haarlem)**
	##lineage4.1 = 12	
	##lineage4.1.2 = 3	
	##lineage4.9 = 2	
12_14180	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	
12_14551	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	
12_14879	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	
12_15155	**##lineage4.2 = 61**	**Euro-American (Ural)**
	##lineage4.2.1 = 29	
	##lineage4.9 = 2	
12_15156	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	
12_15175	**##lineage2.2 = 47**	**Beijing**

	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	
12_15239	**##lineage4.1 = 12**	**Euro-American (H37Rv-like)**
	##lineage4.9 = 2	
	##lineage7 = 1	
12_15251	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
12_15460	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	
12_15737	**##lineage3 = 123**	**East-African-Indian**
	##lineage4 = 15	
	##lineage4.9 = 2	
12_15893	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	
12_16119	**##lineage4.1.2.1 = 20**	**Euro-American (Haarlem)**
	##lineage4.1 = 12	
	##lineage4.1.2 = 3	
	##lineage4.9 = 2	
12_16180	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	
12_16196	**##lineage2.2.2 = 70**	**Beijing**
	##lineage2.2 = 47	
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	

12_16269	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	
12_16295	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	
12_16359	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	
12_16409	**##lineage4.3.3 = 48**	**Euro-American (LAM)**
	##lineage4.3 = 13	
	##lineage4.9 = 2	
12_16496	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	
12_16505	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	
12_16706	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	
12_16734	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	

	##lineage4.1.1.2 = 1	
12_16850	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	
12_17047	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	
12_17231	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	
12_17593	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	
12_17704	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	
12_17736	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	
12_17795	**##lineage4.3.3 = 49**	**Euro-American (LAM)**
	##lineage4.3 = 13	
	##lineage4.9 = 2	
12_17889	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	

	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	
12_17975	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	
12_17993	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	
12_17995	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	
12_18055	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	
12_18057	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	
12_18166	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	
12_18248	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	

12_18360	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	
12_18490	**##lineage4.8 = 14**	**Euro-American (mainly T)**
	##lineage4.9 = 2	
12_18493	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	
12_18893	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	
12_18942	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	
12_19027	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	
12_19069	**##lineage4.2 = 61**	**Euro-American (Ural)**
	##lineage4.2.1 = 29	
	##lineage4.9 = 2	
12_19128	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	
12_19131	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	

	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	
13_1	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	
13_10762	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	
13_1130	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	
13_1786	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	
13_183	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	
13_1934	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	
13_1972	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	

	##lineage4.1.1.2 = 1	
13_2072	**##lineage4.2 = 61**	**Euro-American (Ural)**
	##lineage4.2.1 = 29	
	##lineage4.9 = 2	
13_2210	**##lineage4.1.2.1 = 20**	**Euro-American (Haarlem)**
	##lineage4.1 = 12	
	##lineage4.1.2 = 3	
	##lineage4.9 = 2	
13_2219	**##lineage4.3.3 = 49**	**Euro-American (LAM)**
	##lineage4.3 = 13	
	##lineage4.9 = 2	
13_2601	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	
13_2937	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	
13_2995	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	
13_381	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	
13_421	**##lineage4.1.2.1 = 20**	**Euro-American (Haarlem)**
	##lineage4.1 = 12	
	##lineage4.1.2 = 3	
	##lineage4.9 = 2	
13_5139	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	

	##lineage4.1.1.2 = 1	
13_5146	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	
13_5512	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	
13_56	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	
13_5974	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	
13_6273	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	
13_6478	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	
13_6517	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	
13_6728	**##lineage2.2 = 47**	**Beijing**

	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	
13_7366	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	
13_774	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	
13_819	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	
13_8431	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	
13_8557	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	
13_8615	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	
13_8969	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	

	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	
13_9017	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	
13_9242	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	
eu_1	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	
eu_2	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
eu_3	**##lineage2.2 = 47**	**Beijing**
	##lineage2 = 36	
	##lineage4 = 15	
	##lineage2.2.1 = 10	
	##lineage4.9 = 2	
	##lineage4.1.1.2 = 1	

Table 14 - Drug resistance SNP matches obtained using PhyTB.

Sample	Drug	Region	Position
12_13700	Isoniazid	*fabG1_promoter*	1673425
12_13963	Rifampicin	*rpoB*	761155
	Streptomycin	*rrs*	1473246
12_14129	Rifampicin	*rpoB*	761109
	Streptomycin	*rrs*	1473246
	Ethambutol	*embB*	4248003
	Ethambutol	*embB*	4249583
12_14180	Fluoroquinolones	*gyrA*	7570
	Rifampicin	*rpoB*	761155
	Isoniazid	*fabG1_promoter*	1673425
	Kanamycin	*eis_promoter*	2715344
12_14551	Rifampicin	*rpoB*	761155
	Streptomycin	*rpsL*	781687
	Kanamycin	*eis_promoter*	2715342
	Ethambutol	*embA_promoter*	4243217
12_14879	Rifampicin	*rpoB*	761155
	Isoniazid	*fabG1_promoter*	1673425
12_15155			
12_15156	Streptomycin	*rpsL*	781687
	Ethambutol	*embA_promoter*	4243217
12_15175	Rifampicin	*rpoB*	761155
	Isoniazid	*fabG1_promoter*	1673425
	Pyrazinamide	*pncA*	2289030
	Ethambutol	*embA_promoter*	4243221
	Ethambutol	*embB*	4247513
12_15239			
12_15251	Streptomycin	*rpsL*	781687
	Pyrazinamide	*pncA*	2288820
	Kanamycin	*eis_promoter*	2715369
12_15460			
12_15737			
12_15893	Fluoroquinolones	*gyrB*	6750
	Rifampicin	*rpoB*	761155
	Streptomycin	*rrs*	1472362
	Kanamycin	*eis_promoter*	2715346
12_16119	Rifampicin	*rpoB*	761155
	Streptomycin	*rrs*	1473246
	Ethambutol	*embB*	4248003
12_16180	Fluoroquinolones	*gyrA*	7570
	Rifampicin	*rpoB*	761155
	Streptomycin	*rpsL*	781687
	Kanamycin	*eis_promoter*	2715347
	Ethambutol	*embB*	4248003
12_16196			
12_16269	Fluoroquinolones	*gyrA*	7563

	Rifampicin	*rpoB*	761155
	Isoniazid	*fabG1_promoter*	1673425
	Pyrazinamide	*pncA*	2289030
	Ethambutol	*embA_promoter*	4243221
	Ethambutol	*embB*	4247513
12_16295	Streptomycin	*rpsL*	781687
	Pyrazinamide	*pncA*	2288820
	Kanamycin	*eis_promoter*	2715369
12_16359	Streptomycin	*rpsL*	781687
	Pyrazinamide	*pncA*	2289202
	Kanamycin	*eis_promoter*	2715369
12_16409	Isoniazid	*kasA*	2518919
	Para-Aminosalisylic-Acid	*thyA*	3073868
12_16496	Streptomycin	*rpsL*	781687
12_16505	Streptomycin	*rpsL*	781687
12_16706			
12_16734	Rifampicin	*rpoB*	761155
	Rifampicin	*rpoC*	764841
	Streptomycin	*rpsL*	781687
	Pyrazinamide	*pncA*	2289096
	Kanamycin	*eis_promoter*	2715369
12_16850	Streptomycin	*rpsL*	781687
	Pyrazinamide	*pncA*	2288820
	Kanamycin	*eis_promoter*	2715369
12_17047	Streptomycin	*rpsL*	781687
	Pyrazinamide	*pncA*	2288820
	Kanamycin	*eis_promoter*	2715369
12_17231	Fluoroquinolones	*gyrA*	7570
	Rifampicin	*rpoB*	761155
	Streptomycin	*rpsL*	781687
	Ethambutol	*embB*	4247495
12_17593	Rifampicin	*rpoB*	761155
	Streptomycin	*rpsL*	781687
12_17704	Rifampicin	*rpoB*	761155
	Isoniazid	*fabG1_promoter*	1673425
12_17736	Streptomycin	*rpsL*	781687
	Pyrazinamide	*pncA*	2288820
	Kanamycin	*eis_promoter*	2715369
12_17795	Rifampicin	*rpoB*	761155
	Isoniazid	*fabG1_promoter*	1673425
	Isoniazid	*kasA*	2518919
	Para-Aminosalisylic-Acid	*thyA*	3073868
	Ethambutol	*embB*	4247574
12_17889	Rifampicin	*rpoB*	761155
	Streptomycin	*rrs*	1472362
	Pyrazinamide	*pncA*	2289226
	Kanamycin	*eis_promoter*	2715369
	Ethambutol	*embB*	4247574

12_17975	Rifampicin	*rpoB*	761155
	Streptomycin	*rpsL*	781687
	Pyrazinamide	*pncA*	2288852
	Ethambutol	*embB*	4247399
12_17993	Streptomycin	*rpsL*	781687
	Pyrazinamide	*pncA*	2288820
	Kanamycin	*eis_promoter*	2715369
12_17995	Rifampicin	*rpoB*	761155
	Isoniazid	*fabG1_promoter*	1673425
	Pyrazinamide	*pncA*	2289030
	Ethambutol	*embA_promoter*	4243221
	Ethambutol	*embB*	4247513
12_18055	Streptomycin	*rpsL*	781687
	Pyrazinamide	*pncA*	2288820
	Kanamycin	*eis_promoter*	2715369
12_18057	Streptomycin	*rpsL*	781687
	Pyrazinamide	*pncA*	2288820
	Kanamycin	*eis_promoter*	2715369
12_18166	Rifampicin	*rpoB*	761155
	Streptomycin	*rpsL*	781687
	Kanamycin	*eis_promoter*	2715344
	Ethambutol	*embB*	4248003
12_18248	Rifampicin	*rpoB*	761155
	Streptomycin	*rrs*	1473246
	Isoniazid	*fabG1_promoter*	1673425
	Pyrazinamide	*pncA*	2289030
	Ethambutol	*embA_promoter*	4243221
	Ethambutol	*embB*	4247513
12_18360	Rifampicin	*rpoB*	761155
	Rifampicin	*rpoC*	764822
	Streptomycin	*rpsL*	781687
	Isoniazid	*katG*	2155109
	Ethambutol	*embB*	4248003
12_18490	Isoniazid	*fabG1_promoter*	1673425
12_18493	Rifampicin	*rpoB*	761155
	Streptomycin	*rrs*	1472362
12_18893	Streptomycin	*rpsL*	781687
	Pyrazinamide	*pncA*	2288820
	Kanamycin	*eis_promoter*	2715369
12_18942			
12_19027	Rifampicin	*rpoB*	761155
	Rifampicin	*rpoB*	761244
	Isoniazid	*fabG1_promoter*	1673425
12_19069	Rifampicin	*rpoB*	761155
	Rifampicin	*rpoC*	764363
12_19128	Streptomycin	*rpsL*	781687
	Pyrazinamide	*pncA*	2288820
	Kanamycin	*eis_promoter*	2715369

12_19131	Ryfamycin	*rpoB*	761155
	Ryfamycin	*rpoB*	761244
	Isoniazid	*fabG1_promoter*	1673425
13_1			
13_10762	Ryfamycin	*rpoB*	76115
	Streptomycin	*rpsL*	781687
	Pyrazinamide	*pncA*	2288839
13_1130	Ryfamycin	*rpoB*	76115
	Streptomycin	*rpsL*	781687
	Pyrazinamide	*pncA*	2288839
13_1786	Fluoroquinolones	*gyrB*	6735
	Streptomycin	*rpsL*	781687
	Streptomycin	*rrs*	1473246
	Pyrazinamide	*pncA*	2289201
	Kanamycin	*eis_promoter*	2715342
	Ethambutol	*embA_promoter*	4243222
	Ethambutol	*embB*	4247393
	Ethionamide	*ethA*	4326770
13_183	Streptomycin	*rpsL*	781687
	Pyrazinamide	*pncA*	2288820
	Kanamycin	*eis_promoter*	2715369
13_1934	Fluoroquinolones	*gyrA*	7563
	Ryfamycin	*rpoB*	761155
	Isoniazid	*fabG1_promoter*	1673425
	Pyrazinamide	*pncA*	2289030
	Ethambutol	*embA_promoter*	4243221
	Ethambutol	*embB*	4247513
13_1972			
13_2072			
13_2210	Ryfamycin	*rpoB*	761155
	Streptomycin	*rrs*	1473246
	Ethambutol	*embB*	4248003
13_2219	Isoniazid	*kasA*	2518919
	Para-Aminosalisylic-Acid	*thyA*	3073868
13_2601	Ryfamycin	*rpoB*	761155
	Streptomycin	*rpsL*	781687
	Pyrazinamide	*pncA*	2288839
13_2937	Ryfamycin	*rpoB*	761155
	Streptomycin	*rpsL*	781687
	Pyrazinamide	*pncA*	2288839
13_2995	Streptomycin	*rpsL*	781687
	Streptomycin	*rrs*	1472362
	Isoniazid	*fabG1_promoter*	1673425
	Pyrazinamide	*pncA*	2289030
	Kanamycin	*eis_promoter*	2715369
	Ethambutol	*embB*	4247553

13_381	Ryfamycin	*rpoB*	761155
	Streptomycin	*rpsL*	781687
	Streptomycin	*rrs*	1473246
13_421	Isoniazid	*fabG1_promoter*	1673425
13_5139	Ryfamycin	*rpoB*	761155
	Isoniazid	*fabG1_promoter*	1673425
	Pyrazinamide	*pncA*	2289030
	Ethambutol	*embA_promoter*	4243221
	Ethambutol	*embB*	4247513
13_5146	Ryfamycin	*rpoB*	761155
	Streptomycin	*rpsL*	781687
	Streptomycin	*rrs*	1473246
13_5512	Ryfamycin	*rpoB*	761155
	Streptomycin	*rpsL*	781687
	Isoniazid	*fabG1_promoter*	1673425
	Kanamycin	*eis_promoter*	2715344
13_56	Ryfamycin	*rpoB*	761139
	Streptomycin	*rpsL*	781687
13_5974	Ryfamycin	*rpoB*	761155
	Streptomycin	*rpsL*	781687
	Pyrazinamide	*pncA*	2288839
13_6273	Fluoroquinolones	*gyrB*	6750
	Ryfamycin	*rpoB*	761155
	Streptomycin	*rrs*	1472362
	Kanamycin	*eis_promoter*	2715346
13_6478	Fluoroquinolones	*gyrA*	7570
	Ryfamycin	*rpoB*	761155
	Streptomycin	*rpsL*	781687
	Kanamycin	*eis_promoter*	2715347
	Ethambutol	*embB*	4248003
13_6517	Streptomycin	*rpsL*	781687
	Pyrazinamide	*pncA*	2289202
	Kanamycin	*eis_promoter*	2715369
13_6728	Ryfamycin	*rpoB*	761155
	Isoniazid	*fabG1_promoter*	1673425
	Pyrazinamide	*pncA*	2289030
	Ethambutol	*embA_promoter*	4243221
	Ethambutol	*embB*	4247513
13_7366	Streptomycin	*rpsL*	781687
	Pyrazinamide	*pncA*	2289202
	Kanamycin	*eis_promoter*	2715369
13_774	Ryfamycin	*rpoB*	761155
	Isoniazid	*fabG1_promoter*	1673425
	Pyrazinamide	*pncA*	2289030
	Ethambutol	*embA_promoter*	4243221

	Ethambutol	*embB*	4247513
13_819	Ryfamycin	*rpoB*	761155
	Isoniazid	*fabG1_promoter*	1673425
	Pyrazinamide	*pncA*	2289030
	Ethambutol	*embA_promoter*	4243221
	Ethambutol	*embB*	4247513
13_8431	Ryfamycin	*rpoB*	761155
	Streptomycin	*rrs*	1472362
	Kanamycin	*eis_promoter*	2715369
	Ethambutol	*embA_promoter*	4243222
	Ethambutol	*embB*	4247574
13_8557	Ryfamycin	*rpoB*	761155
	Streptomycin	*rpsL*	781687
	Pyrazinamide	*pncA*	2288839
13_8615	Fluoroquinolones	*gyrB*	6750
	Ryfamycin	*rpoB*	761155
	Streptomycin	*rrs*	1472362
	Kanamycin	*eis_promoter*	2715346
13_8969	Ryfamycin	*rpoB*	761155
	Streptomycin	*rpsL*	781687
	Streptomycin	*rrs*	1473246
13_9017	Ryfamycin	*rpoB*	761155
	Isoniazid	*fabG1_promoter*	1673425
	Pyrazinamide	*pncA*	2289030
	Ethambutol	*embA_promoter*	4243221
	Ethambutol	*embB*	4247513
13_9242	Ryfamycin	*rpoB*	761155
	Streptomycin	*rpsL*	781687
	Isoniazid	*fabG1_promoter*	1673425
	Kanamycin	*eis_promoter*	2715344
eu_1	Ryfamycin	*rpoB*	761155
	Ryfamycin	*rpoC*	764724
	Streptomycin	*rpsL*	781687
	Pyrazinamide	*pncA*	2288850
	Kanamycin	*eis_promoter*	2715342
eu_2	Ryfamycin	*rpoB*	761155
	Ryfamycin	*rpoC*	764724
	Streptomycin	*rpsL*	781687
	Kanamycin	*eis_promoter*	2715342
eu_3	Ryfamycin	*rpoB*	761155
	Ryfamycin	*rpoC*	764724
	Streptomycin	*rpsL*	781687
	Pyrazinamide	*pncA*	2288953
	Kanamycin	*eis_promoter*	2715342

Table 15 – Specific rpoB and rpoC polymorphisms. *Single events are highlighted in yellow.*

samples	rpoB	rpoC
12_13700	D103D (GAC→GAT)	G594E (GGG→GAG)
12_13963	S450L (TCG→TTG)	V483G (GTG→GGG)
	A1075A (GCT→GCC)	E1092D (GAA→GAC)
12_14129	D103D (GAC→GAT)	G594E (GGG→GAG)
	D435Y (GAC→TAC)	
12_14180	S450L (TCG→TTG)	G332C (GGC→TGC)
	A1075A (GCT→GCC)	E1092D (GAA→GAC)
12_14551	S450L (TCG→TTG)	V483G (GTG→GGG)
	A1075A (GCT→GCC)	E1092D (GAA→GAC)
12_14879	S450L (TCG→TTG)	E1092D (GAA→GAC)
	A1075A (GCT→GCC)	
12_15155		
12_15156	A1075A (GCT→GCC)	E1092D (GAA→GAC)
12_15175	S450L (TCG→TTG)	V483G (GTG→GGG)
	A1075A (GCT→GCC)	E1092D (GAA→GAC)
12_15239		G594E (GGG→GAG)
12_15251	A1075A (GCT→GCC)	
12_15460	A1075A (GCT→GCC)	E1092D (GAA→GAC)
12_15737	G876G (GGT→GGG)	
	A1075A (GCT→GCC)	
12_15893	S450L (TCG→TTG)	V483G (GTG→GGG)
	A1075A (GCT→GCC)	E1092D (GAA→GAC)
12_16119	D103D (GAC→GAT)	G594E (GGG→GAG)
	S450L (TCG→TTG)	
12_16180	S450L (TCG→TTG)	G332S (GGC→AGC)
	A1075A (GCT→GCC)	
12_16196	A1075A (GCT→GCC)	
12_16269	S450L (TCG→TTG)	V483G (GTG→GGG)
	A1075A (GCT→GCC)	E1092D (GAA→GAC)
12_16295	A1075A (GCT→GCC)	
12_16359	A1075A (GCT→GCC)	
12_16409		A542A (GCC→GCG)
12_16496	A1075A (GCT→GCC)	E1092D (GAA→GAC)
12_16505	A1075A (GCT→GCC)	E1092D (GAA→GAC)
12_16706	A1075A (GCT→GCC)	E1092D (GAA→GAC)
12_16734	S450L (TCG→TTG)	I491T (ATC→ACC)
	A1075A (GCT→GCC)	
12_16850	A1075A (GCT→GCC)	
12_17047	A1075A (GCT→GCC)	
12_17231	S450L (TCG→TTG)	I491V (ATC→GTC)
	A1075A (GCT→GCC)	E1092D (GAA→GAC)

12_17593	S450L (TCG→TTG)	Q523K (CAG→AAG)
	A1075A (GCT→GCC)	
12_17704	S450L (TCG→TTG)	E1092D (GAA→GAC)
	A1075A (GCT→GCC)	
12_17736	A1075A (GCT→GCC)	
12_17795	S450L (TCG→TTG)	W484G (TGG→GGG)
		A542A (GCC→GCG)
12_17889	S450L (TCG→TTG)	E1092D (GAA→GAC)
	E761D (GAG→GAC)	
	A1075A (GCT→GCC)	
12_17975	S450L (TCG→TTG)	D485Y (GAT→TAT)
	A1075A (GCT→GCC)	
12_17993	A1075A (GCT→GCC)	
12_17995	S450L (TCG→TTG)	V483G (GTG→GGG)
	A1075A (GCT→GCC)	E1092D (GAA→GAC)
12_18055	A1075A (GCT→GCC)	
12_18057	A1075A (GCT→GCC)	
12_18166	S450L (TCG→TTG)	F452C (TTC→TGC)
	A1075A (GCT→GCC)	
12_18248	S450L (TCG→TTG)	V483G (GTG→GGG)
	A1075A (GCT→GCC)	E1092D (GAA→GAC)
12_18360	S450L (TCG→TTG)	D485N (GAT→AAT)
	A1075A (GCT→GCC)	
12_18490		
12_18493	S450L (TCG→TTG)	V483A (GTG→GCG)
	A1075A (GCT→GCC)	E1092D (GAA→GAC)
12_18893	A1075A (GCT→GCC)	
12_18942	A1075A (GCT→GCC)	E1092D (GAA→GAC)
12_19027	S450L (TCG→TTG)	E1092D (GAA→GAC)
	I480V (ATC→GTC)	
	A1075A (GCT→GCC)	
12_19069	S450L (TCG→TTG)	G332R (GGC→CGC)
12_19128	A1075A (GCT→GCC)	
12_19131	S450L (TCG→TTG)	E1092D (GAA→GAC)
	I480V (ATC→GTC)	
	A1075A (GCT→GCC)	
13_1	A1075A (GCT→GCC)	
13_10762	S450L (TCG→TTG)	
	L731P (CTG→CCG)	
	A1075A (GCT→GCC)	
13_1130	S450L (TCG→TTG)	
	L731P (CTG→CCG)	
	A1075A (GCT→GCC)	
13_1786	A1075A (GCT→GCC)	

13_183	A1075A (GCT→GCC)	
	S450L (TCG→TTG)	V483G (GTG→GGG)
13_1934	A1075A (GCT→GCC)	E1092D (GAA→GAC)
13_1972	I925V (ATT→GTT)	E1092D (GAA→GAC)
	A1075A (GCT→GCC)	
13_2072		
13_2210	D103D (GAC→GAT)	G594E (GGG→GAG)
	S450L (TCG→TTG)	
13_2219		
13_2601	S450L (TCG→TTG)	
	L731P (CTG→CCG)	
	A1075A (GCT→GCC)	
	S450L (TCG→TTG)	
	L731P (CTG→CCG)	
13_2937	A1075A (GCT→GCC)	
13_2995	L430P (CTG→CCG)	E1092D (GAA→GAC)
	H445N (CAC→AAC)	
	A1075A (GCT→GCC)	
13_381	S450L (TCG→TTG)	
	A1075A (GCT→GCC)	
13_421	D103D (GAC→GAT)	G594E (GGG→GAG)
13_5139	S450L (TCG→TTG)	V483G (GTG→GGG)
	A1075A (GCT→GCC)	E1092D (GAA→GAC)
13_5146	S450L (TCG→TTG)	
	A1075A (GCT→GCC)	
13_5512	S450L (TCG→TTG)	N698S (AAC→AGC)
	A1075A (GCT→GCC)	
13_56	H445Y (CAC→TAC)	
	A1075A (GCT→GCC)	
13_5974	S450L (TCG→TTG)	
	L731P (CTG→CCG)	
	A1075A (GCT→GCC)	
13_6273	S450L (TCG→TTG)	V483G (GTG→GGG)
	A1075A (GCT→GCC)	E1092D (GAA→GAC)
13_6478	S450L (TCG→TTG)	G332S (GGC→AGC)
	A1075A (GCT→GCC)	
13_6517	A1075A (GCT→GCC)	
	S450L (TCG→TTG)	V483G (GTG→GGG)
13_6728	A1075A (GCT→GCC)	E1092D (GAA→GAC)
13_7366	A1075A (GCT→GCC)	
13_774	S450L (TCG→TTG)	V483G (GTG→GGG)
	A1075A (GCT→GCC)	E1092D (GAA→GAC)
	S450L (TCG→TTG)	V483G (GTG→GGG)
13_819	A1075A (GCT→GCC)	E1092D (GAA→GAC)

13_8431	S450L (TCG→TTG)	E1092D (GAA→GAC)
	E761D (GAG→GAC)	
	A1075A (GCT→GCC)	
13_8557	S450L (TCG→TTG)	
	L731P (CTG→CCG)	
	A1075A (GCT→GCC)	
13_8615	S450L (TCG→TTG)	E1092D (GAA→GAC)
	A1075A (GCT→GCC)	V483G (GTG→GGG)
13_8969	S450L (TCG→TTG)	
	A1075A (GCT→GCC)	
13_9017	S450L (TCG→TTG)	V483G (GTG→GGG)
	A1075A (GCT→GCC)	E1092D (GAA→GAC)
13_9242	S450L (TCG→TTG)	N698S (AAC→AGC)
	A1075A (GCT→GCC)	
eu_1	S450L (TCG→TTG)	F452S (TTC→TCC)
	A1075A (GCT→GCC)	
eu_2	S450L (TCG→TTG)	F452S (TTC→TCC)
	A1075A (GCT→GCC)	
eu_3	S450L (TCG→TTG)	F452S (TTC→TCC)
	A1075A (GCT→GCC)	

Table 16 - pncA polymorphisms for each sample. *Single events are highlighted in yellow.*

***pncA* (2,288,681 → 2,289,241)**	
12_13963	**L85R (CTG→CGG)**
12_18166	2,288,794 (+C)
13_5512	2,288,852 (+T)
13_9242	2,288,852 (+T)
12_16180	C138R (TGT→CGT)
13_6478	C138R (TGT→CGT)
13_6517	C14R (TGC→CGC)
12_16359	C14R (TGC→CGC)
13_7366	C14R (TGC→CGC)
13_1786	**C14Y (TGC→TAC)**
12_14129	**D242D (GAT→GAC)**
12_16734	**D49A (GAC→GCC)**
12_14879	H71P (CAT→CCT)
12_17704	H71P (CAT→CCT)
12_19027	H71P (CAT→CCT)
12_19131	H71P (CAT→CCT)
13_819	H71R (CAT→CGT)
13_9017	H71R (CAT→CGT)
12_15175	H71R (CAT→CGT)
12_16269	H71R (CAT→CGT)
12_17995	H71R (CAT→CGT)
12_18248	H71R (CAT→CGT)
13_1934	H71R (CAT→CGT)
13_2995	**H71R (CAT→CGT)**
13_5139	H71R (CAT→CGT)
13_6728	H71R (CAT→CGT)
13_774	H71R (CAT→CGT)
12_17593	**H71Y (CAT→TAT)**
12_17889	**I6L (ATC→CTC)**
12_15251	Q141P (CAG→CCG)
12_16295	Q141P (CAG→CCG)
12_16850	Q141P (CAG→CCG)
12_17047	Q141P (CAG→CCG)
12_17736	Q141P (CAG→CCG)
12_17993	Q141P (CAG→CCG)
12_18055	Q141P (CAG→CCG)
12_18057	Q141P (CAG→CCG)
12_18893	Q141P (CAG→CCG)
12_19128	Q141P (CAG→CCG)
13_183	Q141P (CAG→CCG)
13_2937	T135P (ACC→CCC)
13_10762	T135P (ACC→CCC)

13_1130	T135P (ACC→CCC)
13_2601	T135P (ACC→CCC)
13_5974	T135P (ACC→CCC)
13_8557	T135P (ACC→CCC)
12_14180	**T142M (ACG→ATG)**
13_381	V128G (GTC→GGC)
13_5146	V128G (GTC→GGC)
13_8969	V128G (GTC→GGC)
12_17975	**V130G (GTG→GGG)**
eu_1	**V131F (GTC→TTC)**
12_15893	V7A (GTC→GCC)
13_6273	V7A (GTC→GCC)
13_8615	V7A (GTC→GCC)
12_14551	**W119L (TGG→TTG)**
12_16119	W68C (TGG→TGT)
13_2210	W68C (TGG→TGT)
12_13700	[redacted]
12_15155	
12_15156	
12_15239	
12_15460	
12_15737	
12_16196	
12_16409	
12_16496	
12_16505	
12_16706	
12_17231	
12_17795	
12_18360	
12_18490	
12_18493	
12_18942	
12_19069	
13_1	
13_1972	
13_2072	
13_2219	
13_421	
13_56	
13_8431	
eu_2	
eu_3	