**Universidade de Lisboa**

**Faculdade de Farmácia**


# Genomic analysis of a household tuberculosis transmission cluster over a ten-year period

**Mariana Sousa Dias**

**Mestrado Integrado em Ciências Farmacêuticas**

**2020**

**Universidade de Lisboa**
**Faculdade de Farmácia**

# Genomic analysis of a household tuberculosis transmission cluster over a ten-year period

**Mariana Sousa Dias**

**Monografia de Mestrado Integrado em Ciências Farmacêuticas apresentada à Universidade de Lisboa através da Faculdade de Farmácia**

**Orientador: Professora Associada, Doutora Renāte Ranka**

**Co-Orientador: Professora Auxiliar, Doutora Maria Isabel Nobre Franco de Portugal Dias Jordão**

**2020**

# Abstract

Tuberculosis remains a serious public health problem even though it is a preventable and treatable disease. World Health Organization estimates that 1.5 million people die from tuberculosis year after year and roughly one quarter of the world's population is believed to be infected.

Latvia is one of Europe's high-priority countries for tuberculosis control and has one of the highest rates of multi-drug resistant tuberculosis in the world, despite having a well-established control programme. Prevention, early detection and quick and effective response to outbreaks are essential elements to control the spread of tuberculosis.

Over a period of ten years, seven samples were collected from a family of five people from Latvia. We performed genomic analysis of the seven isolates in order to unravel the chain of transmission, investigate the origin of two recurrent cases and reveal the possible existence of drug resistance.

We prepared genomic libraries and we sequenced the isolates using the Ion Proton platform. To analyze the genomic sequences, we caried out bioinformatic analysis using a pipeline for genome-wide variant detection, that included alignment of the reads against the reference H37Rv genome, local indel realignment, variant calling and structural variant detection.

Overall, 6 structural variants were found, and we detected 1029 high-quality SNPs, from which 9 were phylogenetically informative and 17 differentiated the isolates.

Based on *in silico* Spoligotyping the isolates belonged to the T1 sub-family and when using phylogenetic specific SNPs, the studied strains were determined to be part of the Haarlem sub-lineage. No robust polymorphisms in genes associated with drug resistance were found, therefore the isolates were classified susceptible to all anti-tuberculosis drugs. Two patients had recurrent cases that we defined as re-infections. We generated hypotheses in order to establish the routes of transmission, supported by the defined cut-offs in the number of SNPs and the data from the maximum likelihood phylogenetic trees.

Although we used a high-resolution method, the WGS data was not enough to determine the direction of transmission within the cluster unambiguously. The molecular epidemiology data needed to be combined with classical epidemiology and clinical information to effectively investigate this household transmission cluster.

**Keywords**: Tuberculosis; Latvia; *Mycobacterium tuberculosis;* Whole-genome sequencing

# Resumo

A tuberculose continua a ser um grave problema de saúde pública, apesar de ser uma doença evitável e tratável. A Organização Mundial de Saúde estima que 1,5 milhões de pessoas morrem de tuberculose ano após ano e acredita-se que cerca de um quarto da população mundial esteja infectada.

A Letónia é um dos países europeus de alta prioridade no controlo da tuberculose e tem uma das taxas mais altas de tuberculose multirresistente do mundo, apesar de ter um programa de controlo bem estabelecido. Prevenção, detecção precoce e resposta rápida e eficaz aos surtos são elementos essenciais para controlar a propagação da tuberculose.

Durante um período de dez anos, foram colhidas sete amostras de uma família de cinco pessoas da Letónia. Realizámos análises genómicas dos sete isolados com o intuito de desvendar a cadeia de transmissão, investigar a origem de dois casos recorrentes e revelar a possível existência de resistência aos medicamentos.

Preparámos bibliotecas genómicas e sequenciámos os isolados com o Ion Proton. Para analisar as sequências genómicas, efectuámos uma análise bioinformática para a detecção de variantes em todo o genoma, que incluiu o alinhamento das reads contra o genoma de referência H37Rv, o local indel realignment, variant calling e a detecção de variantes estruturais.

No total, foram encontrados 6 variantes estruturais, e detectámos 1029 SNPs de alta qualidade, dos quais 9 eram filogeneticamente informativos e 17 diferenciavam os isolados. Com base Spoligotyping *in silico*, os isolados pertenciam à sub-família T1. Ao comparar os nossos dados com os da lista de SNPs filogeneticamente específicos, as estirpes estudadas faziam parte da sub-linhagem Haarlem. Não foram encontrados polimorfismos robustos nos genes associados à resistência aos medicamentos, pelo que os isolados foram classificados como susceptíveis a todas os medicamentos anti-tuberculose. Dois doentes tiveram casos recorrentes que definimos como reinfecções. Gerámos hipóteses para estabelecer a cadeia de transmissão, apoiadas pelos limites definidos no número de SNPs e pelos dados das árvores filogenéticas de máxima verossimilhança.

Embora tenhamos utilizado um método de alta resolução, os dados do WGS não foram suficientes para determinar sem ambiguidade a direcção da transmissão do surto. Os dados da epidemiologia molecular precisavam da epidemiologia clássica e da informação clínica para investigar eficazmente este surto.

**Palavras-chave**: Tuberculose; Letónia; *Mycobacterium tuberculosis;* Whole-genome sequencing

# Abbreviations

Bp – Base pair

DNA - Deoxyribonucleic acid

GC – Guanine-cytosine

MDR-TB – Multidrug-resistant tuberculosis

MTBC - Mycobacterium tuberculosis complex

NGS – Next-generation sequencing

nsSNP - Non-synonymous single nucleotide polymorphism

PCR - Polymerase Chain Reaction

SIT - Spoligo International Type

SNP – Single nucleotide polymorphism

sSNP - Synonymous single nucleotide polymorphism

SV – Structural variant

TB – Tuberculosis

VAF – Variant Allele Frequency

WGS - Whole-genome sequencing

WHO – World Health Organization

XDR-TB - Extensively drug-resistant tuberculosis

**Table of contents**

1 Introduction ..... 8
1.1 Epidemiology and pathogenesis of Tuberculosis ..... 8
1.2 Tuberculosis in Latvia ..... 9
1.3 Population Structure and Genetic Diversity of the MTBC ..... 10
1.4 Molecular genotyping ..... 10
1.4.1 Spoligotyping ..... 11
1.4.2 Next-generation sequencing (NGS) ..... 11
1.4.2.1 Whole genome sequencing (WGS) ..... 12
1.4.2.2 Single nucleotide polymorphisms (SNPs) ..... 12
2 Material and Methods ..... 14
3 Results and discussion ..... 18
4 Conclusion ..... 38
Bibliography ..... 39

**List of figures**

Figura 1 - Sequence quality histogram. Displays the mean quality value across each base position in the read – Raw data (Sourse: FastQC) ..... 19
Figure 2 - Per sequence quality scores. Shows the number of reads with average quality scores and reveals if a subset of reads has poor quality. – Raw data (*Sourse*: FastQC) ..... 19
Figure 3 – The average GC content – Raw data (*Sourse*: FastQC) ..... 20
Figura 4 - Sequence length distribution – Raw Data (Sourse: FastQC) ..... 20
Figura 5 - Sequence Quality Histogram. Displays the mean quality value across each base position in the read – Data after Trimommatic (Sourse: FastQC) ..... 21
Figura 6 - Per sequence quality scores – Data after Trimommatic (Sourse: FastQC) ..... 22
Figure 7 - Sequence lenght distribution – Data after Trimommatic (*Sourse*: FastQC) ..... 22
Figure 8 – Philogenetic tree of the Household Cluster ..... 35
Figure 9 - Maximum likelihood phylogenetic inference tree (*Sourse*: iTOL) ..... 37

**List of tables**

Table 1 - Patients and clinical isolates ..... 14
Table 2 - General statistics – Raw data (*Sourse*: FastQC) ..... 18
Table 3 - General statistics of the sequences after Trimommatic (*Sourse*: FastQC) ..... 21
Table 4 - Statistical data of the alignment ..... 23
Table 5 - Structural variants ..... 24
Table 6 - SVs detected in each sample ..... 25
Table 7 – Number and type of variants identified ..... 27
Table 8 - SNPs diferentiating each isolate ..... 28
Table 9 - *In silico* spoligotyping ..... 31
Table 10 – Phylogenetically informative SNPs ..... 33
Table 11 - Recurent cases ..... 36

# 1 Introduction

## 1.1 Epidemiology and pathogenesis of Tuberculosis

Tuberculosis (TB) is a chronic and progressive disease deemed one of the world's most serious public health problems, which since 1993 has been declared an international emergency by the World Health Organization (WHO) (1).

TB is categorized one of the top 10 causes of death and topped HIV/AIDS as the main cause from a unique infectious agent. WHO's alarming estimations of 2019 showed that worldwide 1.4 million people died from TB, and 10 million became ill, of which 1.2 million were children (2).

TB is present in all countries and age groups and it is strongly influenced by emerging and low-income economies.

TB is an airborne disease, that is, when people with active TB cough, sneeze or spit, they release droplets with the bacteria into the atmosphere, possibly spreading the disease to whomever is near them (3).

The transmission is promoted by highly dense populations and crowded indoor settings that maximize the aerosol transmission of the pathogen. Consequently, TB is portrayed as a major threat to public health, since a single person with active TB is estimated to infect between 5 and 15 other people within a year. (1). The most potentially affected individuals are those who spend more time, day in and day out, with the TB patient, addressed as close contacts, these are the family members, friends, coworkers, or schoolmates (4).

In 1882, Dr. Robert Koch discovered that TB was caused by a bacterium and named it tuberculosis bacillus or Koch's bacillus. The causative agent of TB survives within the macrophages of the host and develops more easily in the lung tissue causing a disease called Pulmonary Tuberculosis (PTB) (5), nevertheless the bacteria is also able to move through the blood and infect any part of the body inducing Extra-pulmonary Tuberculosis or disseminated disease (6).

Not every person infected with the bacteria develops the disease. Several people develop the illness rapidly, over a period of four to twelve weeks, because their immune system is not able to combat the pathogen. On the contrary, others will just develop it years later, when, for some reason, they become immunocompromised (7). As a result, there are two TB clinical outcomes: the latent TB infection (LTBI), particularly asymptomatic and non-transmissible and, in opposition, the highly transmissible progressive active TB disease (8). 90 to 95% of the recently infected develop LTBI and can eventually later develop the active disease, whereas the remaining 5 to 10% immediately develop the active TB disease (9).

Host factors, such as immune suppression, smoking, poor nutrition, diabetes, and respiratory comorbidities play an important role by increasing the risk of transitioning from latent to active TB (10).

A person with LTBI exhibits no symptoms, is not infectious and cannot transmit the TB bacteria to others. The approaches to discover if a person is indeed infected with TB are a positive result in the TB skin test reaction or with the Interferon-Gamma Release Assays. Without treatment, individuals with latent TB have a 5 to 15% risk of developing active TB disease then it can eventually be fatal (2).

When developing active PTB, patients experience symptoms in the form of unpleasantly bad cough that lasts 3 weeks or longer, with sputum and blood, chest pains, weakness, weight loss, high fever, and night sweats. The symptoms can also be mild for many months which can lead to patients dragging out the time for soliciting health care and resulting in higher transmission events. TB may also be exhibited with extrapulmonary manifestations including lymphadenitis, kidney, bone or joint involvement, meningitis or disseminated disease (11).

## 1.2 Tuberculosis in Latvia

The present study aims to unravel the transmission network of a household TB cluster from Latvia.

Latvia is a country situated in the Baltic region of Northern Europe and has an estimated population of 1.92 million people (12).

TB is one of the main infectious diseases in Latvia. The country is labeled one of the 18 high-priority countries for TB control in the WHO European Region and has consistently been ranked among the countries with the highest rates of multidrug-resistant TB (MDR-TB) in the world (13).

In 1991, the collapse of the former Soviet Union transformed the county of Latvia in a newly independent nation. Unfavorably, that drove to the disintegration of the centralized public-health system, meaning that the territory had no appropriate access to health care (14). During that period, Latvia experienced socioeconomic changes, an economic crisis, increasing poverty, unemployment, homelessness, rising of substance abuse and alcoholism, all of which led to a major negative effect on the TB national control. Taking all this into account, the tragic outcome was a rapid increase in TB morbidity and mortality (15), concomitantly with the rise of drug resistant and MDR-TB strains (16).

TB is a condition strongly impacted by the social stratum, showing a direct relationship with poverty. (17) In Latvia, the amount of people belonging to high-risk social groups is expanding and the accessibility to TB medical treatment and assistance in remote areas is insufficient. The high-risk group for TB consists of people living with HIV, prisoners, homeless, alcohol abusers, drug abusers, close contacts of TB patients, unemployed individuals, and smokers. Another major concerning group for TB is the children. The rate of childhood TB is relatively high in Latvia and incidence rates indicate continuous transmission (18). Poor socio-economic conditions like those mentioned above may predispose to TB transmission (19).

The detection and prevention of the TB spread, especially MDR-TB, is extremely important. The success for TB control is based on rapid and accurate diagnostics, effective

therapy and detection of recent transmission chains and outbreaks. Precise knowledge of the epidemiological situation in Latvia is crucial for the optimization of the local TB control. Nowadays, Latvia has a well-established TB and MDR-TB control programme and there has been substantially reduction in notification case rates since 2001 (20).

## 1.3 Population Structure and Genetic Diversity of the MTBC

TB is caused by a group of mycobacteria known as the *Mycobacterium tuberculosis* complex (MTBC). MTBC include *M. tuberculosis, Mycobacterium africanum, Mycobacterium canetti, Mycobacterium orygis, Mycobacterium bovis and the Bacillus Calmette–Guérin strain, Mycobacterium microti*, *Mycobacterium caprae*, *Mycobacterium pinnipedii*, *Mycobacterium suricattae* and *Mycobacterium mungi,* they are assumed to have all derived from the same common ancestor and are closely related at the DNA level (21). Surprisingly, this genetically monomorphic group, shows variations in virulence and immunogenicity, has adapted to different host species, and exhibits distinct phenotypes, ultimately causing variable outcomes of TB. These differences may be due to bacterial factors and strain-specific variations, in addition to host or environmental conditions (22) (23).

*M. tuberculosis* and *M. africanum* are considered the predominant etiological agents of human TB and are the most widespread in the world, thereby they have been subdivided into lineages and families, based on various strain genotyping techniques (24). Recent inferences, revealed that the human-adapted lineages of MTBC comprehend eight major phylogenetic lineages namely Indo-Oceanic (Lineage 1), East-Asian (Lineage 2), East Africa-Indian (Lineage 3), Euro-American (Lineage 4), *M. africanum* West Africa 1 (Lineage 5), *M. africanum* West Africa 2 (Lineage 6), Ethiopia (Lineage 7) (25) and the more recently discovered Lineage 8 (26).

The lineage is known to influence the *M. tuberculosis* strain's pathogenesis (27), their ability to transmit, to rapidly progress to active disease and to be associated with major outbreaks (28).

## 1.4 Molecular genotyping

The best-characterized strain of *M. tuberculosis*, H37Rv, has approximately 4.5 million bp, is estimated to have roughly 4000 potential genes, has a guanine/cytosine average content of 65% and is rich in repetitive DNA (29). H37Rv is used as a reference strain in TB phylogeny, molecular epidemiology, and drug-resistance mutations researches.

Understanding the predominant MTBC lineages, strain diversity and clustering rates is necessary to monitor the spread of TB in a given geographic region.

In molecular epidemiological studies, genotyping is used along with classical epidemiological methods, such as contact tracing investigations, to guide clinical studies, outbreak investigations, define transmission dynamics, to differentiate between relapse and re-infection cases, to support global epidemiological surveillances, for phylogenetic and evolutionary studies and for strain classification (30).

Genotyping methods are based on the direct analysis of the polymorphisms in the *M. tuberculosis* genome, these are polymorphisms in repetitive sequences, single nucleotide polymorphisms and long sequence polymorphisms.

Traditional typing techniques include IS*6110*-RFLP (31), spoligotyping (32) and mycobacterial interspersed repetitive units-variable number tandem repeats (MIRU-VNTR) (33). These methods have advantages and disadvantages and they differ in sensitivity, resolution, discriminatory power and reproducibility.

### 1.4.1 Spoligotyping

Spoligotyping or spacer oligonucleotide typing is a PCR-based genotyping approach that relies on the amplification of the Direct Repeat (DR) region of the MTBC. The DR loci consist of 36 bp copies separated by spacers, which are 34 to 41 different bp non-repeating sequences (34) The DR and the spacers together are called direct variable repeats (DVRs). The presence and order of the hybridization patterns, is translatable into a 15-digit numerical code, the so-called spoligotype (35). The resulting spoligotype pattern is specific of a given evolutionary lineage of strains and can be used for epidemiological tracking because strains belonging to the same outbreak have identical hybridization patterns (36).

Spoligotyping is used for epidemiological purposes and to classify strains into known families, but it is limited in phylogenetics because this marker changes rapidly with a tendency to converge, that is, equal or similar patterns can coincidentally emerge in strains that are not related phylogenetically (37).

### 1.4.2 Next-generation sequencing (NGS)

With the introduction of Next-generation sequencing (NGS), unprecedented and expeditious ways for genome-wide characterization have emerged. NGS, also called high-throughput sequencing, generates thousands of millions of reads of varying length, which can then be reassembled into longer or whole genome sequences using bioinformatics (38). NGS methods include single molecule-real time (SMRT) (Pacific Biosciences, US), nanopore sequencing (Oxford Nanopore Technologies, UK), sequencing by synthesis (Illumina Inc., US) and Ion Torrent semi-conductor sequencing (ThermoFisher Scientific, US) (39).

The Ion Torrent system (Thermo Fisher Scientific, US) is a second-generation sequencing platform that sequences by synthesis. Sequence templates are generated on a bead or sphere via emulsion PCR resulting in oil-water emulsion spheres  Each sphere contains one library molecule and the final result is the amplification of individual fragments to millions of identical copies (40). Basically, the DNA to be sequenced is apprehended in a microwell, and, one by one, nucleotides are released into the well. As the polymerase incorporates the correct nucleotide into the increscent complementary strand, occurs the release of a proton, consequently changing the pH of the solution. An ion sensor at the base of the well detects the

pH change and converts the chemical signal into a digital one allowing the identification of the base at that position. This allows sequencing in real time (41).

#### 1.4.2.1 Whole genome sequencing (WGS)

Whole genome sequencing (WGS) is a NGS application that determines the entire DNA sequence all at once. WGS rely on a combination of genomic DNA extraction, library preparation, template preparation and automated sequencing to determine the sequential arrangement of a DNA sequence (42). Essentially, the genome is cut into smaller fragments and sequenced in parallel to produce a large number of overlapping sequences, the reads. The information obtained is used to identify, compare, and classify the pathogenic organism, study genomes and the encoded proteins and identify changes in genes. As WGS unravels all the variation present in the strain, providing a high-resolution, accurate, and reproducible pathway to analyze evolutionary relations and mechanisms, transmission events, associate mutations to phenotypes, and determine antimicrobial resistance (9) (43) .

WGS provides higher resolution when compared to MIRU-VNTR, *IS6110*-RFLP, and Spoligotyping (44) (45) (46). WGS enables a more precise delineation of the genetic differences between strains by examining more than 90% of the genome, compared to less than 1% with traditional genotyping. (47).

#### 1.4.2.2 Single nucleotide polymorphisms (SNPs)

Single nucleotide polymorphisms (SNPs) are the most abundant form of genetic variations. SNPs can be classified as synonymous or non-synonymous, when occurring in gene coding regions, or intergenic.

Non-synonymous SNPs (nsSNPs) occur on the first or second codon positions, ultimately changing the encoded amino acid, so the resultant protein is shortened and may function improperly or not at all. nsSNPs are the principal contributors for functional mutations and are likely to have a strong effect either beneficial or deleterious (48) nsSNPs constitute approximately two thirds of the genetic variability observed in *M. tuberculosis* and there is evidence that they influence important characteristics such as transmissibility, virulence, drug resistance and immune response (49).

Synonymous SNPs (sSNPs) are variants that usually arise on the third position of the codon, and due to the redundancy of the genetic code, any change in the third position of the codon, will result in the same amino acid being incorporated in the protein sequence at that position. sSNPs provide the basis to study genetic drift and evolutionary relationships among mycobacterial strains (50).

Intergenic and sSNPs were often referred to as evolutionary neutral because their supposed effect on phenotype was considered relatively low, however they can affect regulatory regions, as some sSNPs do have indeed important phenotypic repercussions and intergenic SNPs can change the level of gene expression (51).

Despite having low discriminatory power, which limits their use, SNPs are considered the most trustworthy markers for lineage classification thanks to their low levels of homoplasy (52).

WGS is an efficient way to produce data for SNP discovery. SNP-typing informs about phylogenetic grouping, drug resistance and virulence of a strain (53).

# 2 Material and Methods

During a ten-year period (2006-2016), 7 *M. tuberculosis* clinical isolates were collected from a family of 5 people living in Latvia. The patients were all male and lived in the same house. ValP was the father of VilP, MP, OP and KP (Table 1)

**Table 1 - Patients and clinical isolates**

| Isolate ID | Patient ID | Sex | Country | Date of sample collection |
|---|---|---|---|---|
| 6 | VilP | Male | Latvia | 09/04/2010 |
| 7 | MP | Male | Latvia | 09/04/2010 |
| 8 | OP | Male | Latvia | 09/10/2012 |
| 9 | ValP | Male | Latvia | 2006 |
| 10 | VilP | Male | Latvia | 20/01/2016 |
| 11 | KP | Male | Latvia | 10/08/2009 |
| 12 | KP | Male | Latvia | 23/11/2011 |

Throughout the all procedure, the *M. tuberculosis* genomic DNA samples were always treated as potentially pathogenic, as it was strongly recommended to use suitable aseptic techniques and wear proper personal protective equipment to maintain admissible work and safety standards and limit exposure to the bacterial agent.

The samples were cultured using Lowenstein-Jensen media for DNA isolation. The genomic DNA was extracted using the CTAB method (54).

Qubit dsDNA BR Assay Kit was used to assess the initial concentration of gDNA and afterwards the samples were incubated with RNAse A, to remove RNA contaminants. Genomic DNA was physically fragmented with the Covaris S220 Sonicator and to confirm the size distribution of DNA fragments after shearing, the Qubit dsDNA BR Assay Kit was put to use, prepare as the manufacture recommended (55).

The quantification of the genomic DNA was done with Qubit dsDNA BR Assay Kit and Qubit dsDNA HS Assay Kit because it was crucial for the genomic library preparation step to know the precise concentration of the genomic DNA in each sample, as it needed to meet a minimal concentration requirement as defined by the sequence platform. The purer the DNA of the samples was, better would be the quality of the sequence data obtained. Furthermore, the selection of the library preparation method was highly important due to the *M. tuberculosis* nature and affected directly the subsequent procedures in the pipeline (42).

To ensure that the fragmentated gDNA was within a specific size range, the samples went through the DNA double size selection protocol of NucleoMag® NGS Clean-up and Size select. DNA clean-up and single size selection protocol of NucleoMag® NGS Clean-up and Size select (56) was useful to guarantee that contaminants and shorter DNA fragments were removed. Ultimately, the quantity input DNA of each sample was established using Qubit dsDNA HS Assay Kit (57).

Single-end fragment libraries were prepared using the Ion Xpress™ Plus gDNA Fragment Library Preparation (Thermo Fisher Scientific, US) (58). Firstly, the steps of End Repair and purification were done in order to blunt uneven ends of the genomic DNA. Subsequently, Adapters ligation (Ion Xpress™ Barcode Adapters 1-96 Kit), nick-repair and purification of the ligated DNA was executed to distinguish each sample, as a unique adaptor was assigned for each specimen, as well as to facilitate the attachment of the fragments to the solid surface of the sequencer. Lastly, Library PCR Amplification was carried out to supply sufficient amount of copies of each template to enable the sequencer to detect them.

After library preparation the samples underwent quality control procedures. This step consisted in measuring the final genomic DNA concentrations with Qubit dsDNA HS Assay Kit and afterwards examining the quantity and fragment size with Agilent High Sensitivity DNA Kit Guide (59).

We sequenced the 7 isolates using the Ion Proton platform. (Thermo Fisher Scientific, US). The libraries were sequenced from one end only, creating single-end reads, of a max length of 400 bp. The raw sequence reads were generated in fastq format, a text-based file of every nucleotide base sequenced with a corresponding base quality score.

Library preparation and sequencing reactions were performed according to the instructions provided by the Ion Torrent system (Thermo Fisher Scientific, US) manufacturers. The equipment and reagents employed were platform specific.

Bioinformatic analysis of raw sequence reads was carried out initially using an in-house pipeline for genome-wide variant calling.

*FastQC* (60) was applied directly on the raw sequence data, to check the run metrics summary from the sequencer. *Trimmomatic* (61) was used to remove low quality (below quality 3) bases from the start and the end of the reads, drop reads below the 36 bases long and to scan the reads with a 4-base wide sliding window, cutting when the average quality per base dropped below 20. *FastQC* was run once more to compare the quality control stats before and after Trimmomatic.

Subsequently, the sequencing data was aligned against the reference *M. tuberculosis* H37Rv complete genome (GenBank Accession NC_000962.3) using *BWA-MEM*. The output files were in Sam format. Afterwards, the Sam files were converted into the Bam binary version using *SamFormatConverter*. Depth of coverage was assessed using *Qualimap* (62) and genome coverage was known using The R Project for Statistical Computing manually to evaluate the number of positions of the reference genome which had coverage greater than zero.

The BAM files needed to be necessarily sorted because the alignment files produced were in arbitrary order comparatively to their position in the reference genome and they are

required to occur in positionally based order of their alignment coordinates on each chromosome. The *Picardtools Sort* bam coordinate program was employed to do such task. The duplicates were marked as true with *MarkDuplicates*. Read groups were added in order to associate each sequence data file to its respective sample. *Picardtools* Sort bam coordinate was employed another time to guarantee the proper order of the sequences in every file. *BuildBamIndex* was performed to index the bam files.

With the intent to enhance preciseness of the following processing steps, *Local Indel Realigner* was executed to local realign around the indels. This was a two-stage methodology, were primarily the *GATK Realigner Target Creator* detected which regions in particular needed to be realigned and then generated the Target Intervals files. Afterwards, the literal realignment was accomplished by *GATK Indel Realigner*, that employed the Target Intervals and the Bam files as inputs. The conceived Bam file was indexed with *BuildBamIndex*. Variants were called using *GATK UnifiedGenotyper*, with ploidy 1, originating outputs in Variant Call Format (VCF) format. We made use of *Samtools flagstat* for the purpose of perceiving the statistical data of the samples, with the intention of certify once more the quality level. Further, we resorted to *Samtools rmdp* to remove duplicates. *Samtools index* was utilized to index the Bam files.

Variant functional annotation was performed with *SnpEff* (63). By means of bcftools view the VCF files were filtered to achieve quality score equal to 20 Phred and DP (high-quality read depth) more or over 10, in other words, intending for 10 reads to validate each variant. Afterwards we used *Delly* to call the structural variants. Variant allele frequency was calculated as RV/(RR+RV) for precise variants and only SVs longer than 50bp and with VAF near 50% or higher were kept.

Multiple comparisons between the SNPs from the different isolates were performed using an in-house script written in R. The VCF files were analyzed and all positions were check in order to distinguish which genomic positions showed variations. SNP sites having an excess of 10% missing calls were removed from the analysis. Positions corresponding to PE/PPE genes and with Kmer lower than 49 were excluded (64). All the alleles of variant positions of the dataset were concatenated into a pseudo-DNA molecule in fasta format.

*SpoTyping* (35) was used to predict from the sequence reads the octal code of the spoligotype of the isolates. The SIT numbers and genetic families were then assigned using the SITVIT2 database (65).

Using an in-house script written in The R Project for Statistical Computing, the definition of *M. tuberculosis* lineages based in SNP-typing method was performed according to the 62 SNPs barcode proposed by Coll et al. 2014 (25).

The maximum-likelihood phylogenies were produced with the alignment of the concatenated SNPs as input. With *PHYLOViZ Online* we built a maximum-likelihood phylogenetic tree of the entire dataset using the pairwise comparison approach. Then, with the aid of jModelTest we gained knowledge about the sample's Nucleotide Substitution Model. Using the obtained information, further maximum-likelihood phylogenetic trees were constructed using Seaview (66), implementing PhyML, the Generalised Time Reversible

(GTR) model and using the Bootstrap with 1000 replicates as a branch support metric. Tree visualization and annotation was performed using the Interactive Tree of Life tool (67). We used a 5 and 12 SNPs cut-off to delineate the genomic cluster.

# 3 Results and discussion

Over a period of 10 years, 7 samples were collected from a family of 5 people who lived in Latvia. We carried out WGS of the 7 isolates and implemented bioinformatic analysis to examine and identify variants, SNPs and indels.

Considering that NGS is an elaborated method of producing large amounts of data, quality control techniques are exceedingly more intricate than customary laboratory procedures. These procedures are necessary to generate accurate variant calls from sequencing data.

So initially, the potential contamination and the sequencing quality were examined. We analyzed and compared the quality scores and length distribution of the raw data and the output files after using *Trimmomatic*. The general statistics of the raw sequence reads (Table 2) showed an average sequence length of 200 bp per isolate and nearly 5.5 million sequences per sample. Approximately 42.4% of the reads were duplicate reads

As shown in Figure 1, the Phred quality scores ranged from 27 to 5 across all positions, because as would be expected, the average quality scores decreased practically at end of the run progress due to intrinsic characteristics of the sequencing platform (68). Figure 2 illustrates that the mean quality score of each read varied between 15 and 30.

Nevertheless, the greater amount of reads had quality scores higher than 20, which is verry good considering that a Phred score of 20 corresponds to only a 1% error rate in base calling (69)

**Table 2 - General statistics – Raw data (*Sourse*: FastQC)**

| Isolate ID | Duplicate Reads (%) | GC (%) | Average Sequence Length (bp) | Total Seqs |
|---|---|---|---|---|
| **6** | 41.5 | 63 | 199 | 5189222 |
| **7** | 39.8 | 63 | 200 | 3574671 |
| **8** | 33.2 | 63 | 200 | 1801290 |
| **9** | 36.3 | 63 | 201 | 2146841 |
| **10** | 44.0 | 63 | 198 | 6393380 |
| **11** | 44.7 | 63 | 203 | 7297929 |
| **12** | 57.2 | 62 | 199 | 12511066 |

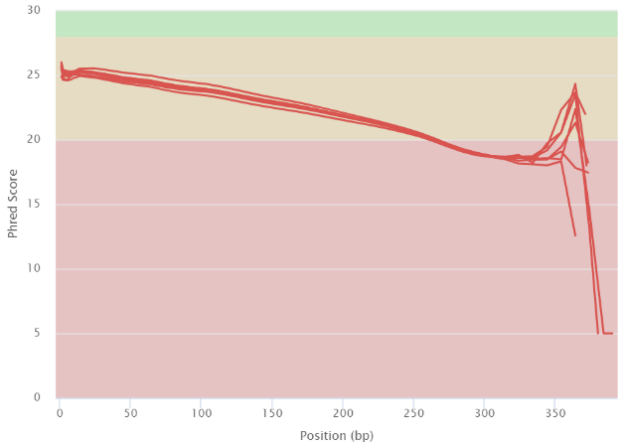


**Figura 1 - Sequence quality histogram. Displays the mean quality value across each base position in the read – Raw data (Sourse: FastQC)**

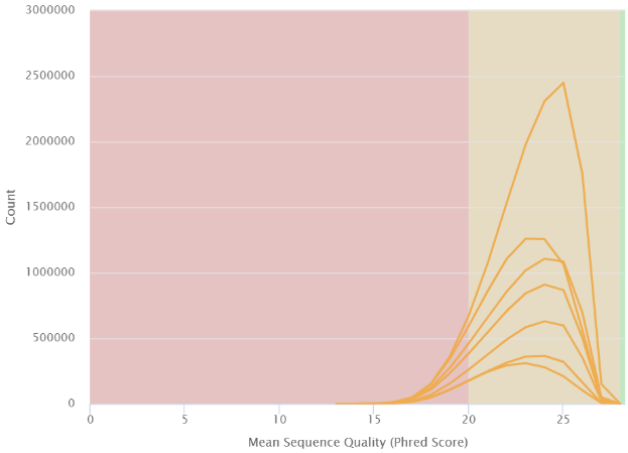


**Figure 2 - Per sequence quality scores. Shows the number of reads with average quality scores and reveals if a subset of reads has poor quality. – Raw data (*Sourse*: FastQC)**

The reads had an average of 63% of GC content (Figure 3), that is accepted as a normal percentage respecting that *M. tuberculosis* is a high GC-content organism. The sequence length distribution ranged between 25 and 376 bp (Figure 4).

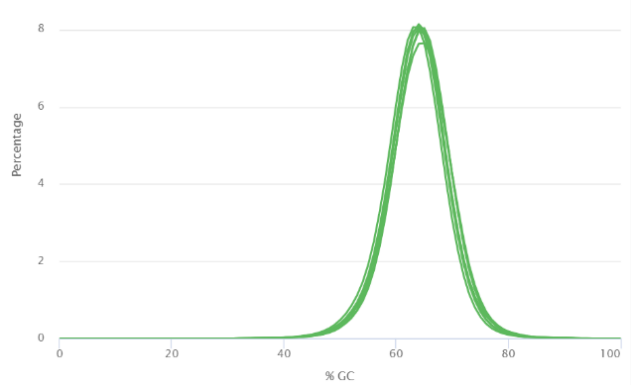


**Figure 3 – The average GC content – Raw data (*Sourse*: FastQC)**

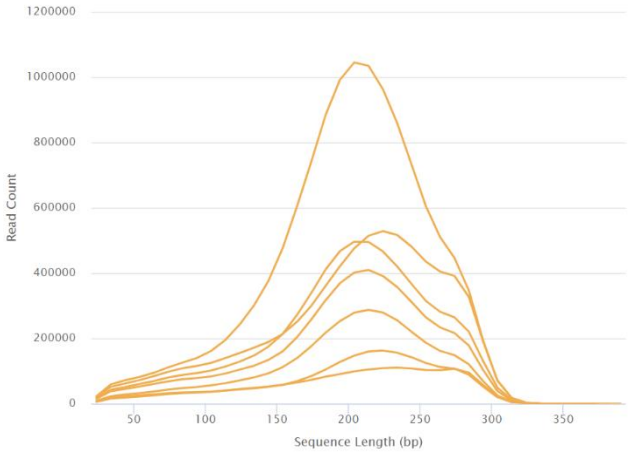


**Figura 4 - Sequence length distribution – Raw Data (Sourse: FastQC)**

NGS, allows the rapid generation of a large volume of DNA fragments, nevertheless not all reads produced are of high quality, and even reads whose average qualities are high may have bases with low quality. Therefore, *Trimmomatic* was used to filter the low-quality reads and to trim poor-quality bases from the dataset to reduce the false positive rate due to sequencing error. The general statistics output showed an average sequence length of 104 bp and approximately 3.4 million sequences per isolate (Table 3). As we intended, the mean quality score was 25 (Figures 5 and 6) and the sequence length distribution ranged from 36 to 316 bp (Figure 7).

**Table 3 - General statistics of the sequences after Trimommatic (*Sourse*: FastQC)**

| Isolate ID | Duplicate Reads (%) | GC (%) | Average Sequence Length (bp) | Total Sequences |
|---|---|---|---|---|
| **6** | 25.5 | 63 | 105 | 3188931 |
| **7** | 24.1 | 63 | 106 | 2180267 |
| **8** | 19.2 | 63 | 99 | 999147 |
| **9** | 20.9 | 63 | 102 | 1249731 |
| **10** | 27.4 | 63 | 106 | 4035772 |
| **11** | 26.5 | 62 | 103 | 4398157 |
| **12** | 36.7 | 62 | 109 | 8193071 |

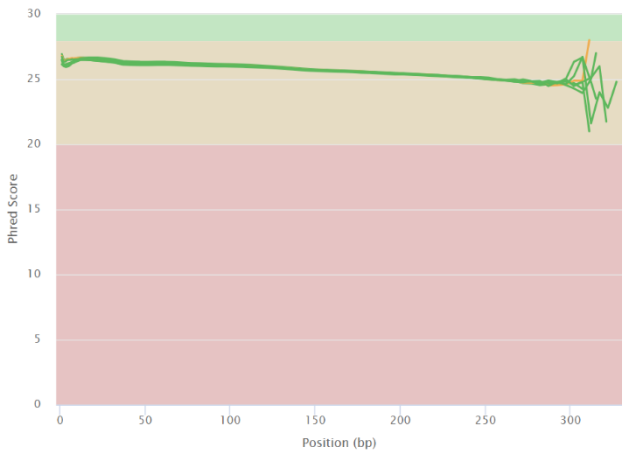


**Figura 5 - Sequence Quality Histogram. Displays the mean quality value across each base position in the read – Data after Trimommatic (Sourse: FastQC)**

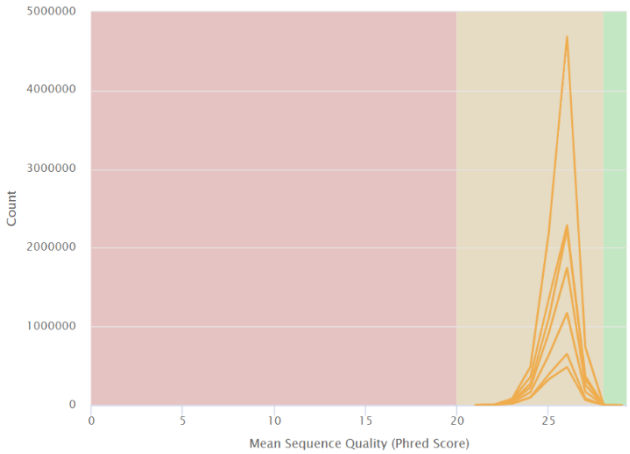


**Figura 6 - Per sequence quality scores – Data after Trimommatic (Sourse: FastQC)**

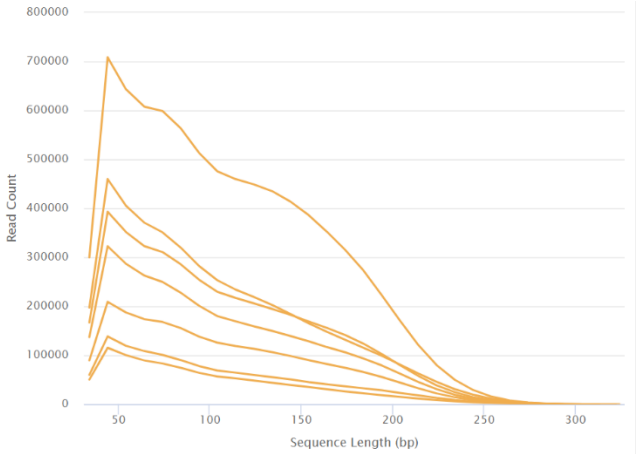


**Figure 7 - Sequence lenght distribution – Data after Trimommatic (*Sourse*: FastQC)**

Mapping our sequence data against the *M. tuberculosis* H37Rv genome (GenBank Accession NC_000962.3) was done with the intent of finding the optimum alignment position for each read and to quickly assemble and annotate the sequences. Fundamentally, mapping isolates were retained if quality degrees were achieved, and alternatively reads were excluded if they mapped to multiple locus in the genome.

The overall alignment rate was greater than 98%, we achieved extensive genome coverage and the depth of coverage varied between isolates (Table 4).

**Table 4 - Statistical data of the alignment**

| Isolate ID | Total reads | Mapped reads | Overall alignment rate (%) | Genome coverage (%) | Average depth of coverage (x) |
|---|---|---|---|---|---|
| 6 | 1798786 | 1779812 | 98.95 | 98.90 | 47.08 |
| 7 | 1297658 | 1284027 | 98.95 | 98.73 | 34.01 |
| 8 | 654181 | 646927 | 98.89 | 98.21 | 15.61 |
| 9 | 781347 | 770656 | 98.63 | 98.77 | 19.39 |
| 10 | 2174481 | 2144724 | 98.63 | 98.88 | 57.99 |
| 11 | 2340248 | 2304221 | 98.46 | 99.09 | 60.99 |
| 12 | 3316799 | 3245901 | 97.86 | 99.44 | 93.95 |

Genome coverage is the percentage of the reference genome that had at least one read mapped to it and is a representation of sequencing quality for each nucleotide sequenced, because if hundreds of fragments give the same signal for a specific position there is high confidence in that specific nucleotide call. Depth of coverage corresponds to the average number of reads mapped to each locus. The higher depth of coverage, the greater the confidence in the final results, as deep coverage helps differentiating sequencing errors from SNPs. Depth of coverage varies across genomes and is negatively affected by regions of high GC content and repetitive elements (70).

The rigor of the alignment has a critical influence on the detection of variants, since SNP and genotype calling depend on the accuracy of the align reads and quality values (71).

Structural variations are large structural alterations in the genomic DNA that are inherited and polymorphic. SVs are extremely diverse in type and size, are at least 50 bp long and are categorized as deletions, insertions, inversions, mobile-element transpositions, translocations, tandem repeats, and copy number variants. SVs can have a marked phenotypic effect, disrupting gene function and regulation or modifying gene dosage (72). SVs are studied in order to know the characterization of the copy, content and structure of genomic variants and to investigate their phenotypic impact (73).

We identified a total of 6 SVs (Table 5). Categorized as deletions these variants ranged from 78 to 439 bp.

**Table 5 - Structural variants**

| SV ID | Position[1] | Size (bp) | SVTYPE | SR | RV | RR | VAF |
|---|---|---|---|---|---|---|---|
| **SV001** | 580615 | 180 | DEL | 7 | 7 | 11 | 0.61 |
| **SV002** | 704269 | 78 | DEL | 10 | 0 | 10 | 1 |
| **SV003** | 2133182 | 287 | DEL | 8 | 35 | 54 | 0.6 |
| **SV004** | 3119994 | 439 | DEL | 9 | 7 | 6 | 0.46 |
| **SV005** | 3194705 | 87 | DEL | 10 | 0 | 55 | 1 |
| **SV006** | 4375625 | 84 | DEL | 10 | 0 | 12 | 1 |

[1]Relative to the genome position of *M. tuberculosis* H37Rv (GenBank Accession NC_000962.3)

SR - Split-read support, SVTYPE – Structural Variant type, VAF - Variant allele frequency, RV – Reads supporting the variant, RR – Reads supporting the reference

The allele frequency represents the incidence of a genetic variant in a population. Heterozygous loci are positions with evidence for more than one allele and are considered highly polymorphic positions (74) A position is called heterozygous when the less frequent allele is supported by at least 5 reads.

Variant Allele Frequency (VAF) is the percentage of sequence reads that correspond to a specific variant divided by the overall coverage at a given locus. VAF is a measure of genetic diversity, wherein heterozygous loci are near 50%, homozygous loci correspond to 100%, and 0% represents the reference loci. (75). On the initial determination we found SVs with VAF scores with discrepancies from these three expected values, they were discarded as they are regarded as dubious because they are considered to be potential errors due to incorrect base calls or alignment.

Based on the VAF scores, SV03, SV05 and SV06 were in homozygous positions and SV01, SV03, SV04 were found in heterozygous positions.

The genomic locations of the SVs were searched in the MycoBrowser database (76), we seeked to know the function of the genes in question and we looked for studies with mutants in these regions.

All SVs were found in non-essential genes according to DeJesus et al. (77) (Table 6**)**.

**Table 6 - SVs detected in each sample**

| SV ID | Isolate ID | | | | | | | Locus[1] | Gene Essentiality[2] | Product |
|---|---|---|---|---|---|---|---|---|---|---|
| | 6 | 7 | 8 | 9 | 10 | 11 | 12 | | | |
| **SV001** | x | | | x | | | | Intergenic | - | - |
| **SV002** | x | x | x | x | x | x | x | Rv0064 | Non-essential[3] | Probable conserved transmembrane protein |
| **SV003** | | | | | | | x | Rv1882c and Rv1883c | Non-essential | Rv1882c: Probable short-chain type dehydrogenase/reductase Rv1883c: Conserved hypothetical protein |
| **SV004** | | | | | | | x | Intergenic | - | - |
| **SV005** | x | x | x | x | x | x | x | Rv2885c | Non-essential | Probable transposase |
| **SV006** | x | x | x | x | x | x | x | Rv3892c (PPE69) | Non-essential | PPE family protein PPE69 |

x- SV detected in the sample

[1]Relative to the genome position of *M. tuberculosis* H37Rv (GenBank Accession NC_000962.3)

[2]Gene Essentiality according to DeJesus et al. (2017) using saturated Himar1 transposon libraries (77)

[3]Non-essential gene by DeJesus *et al*. (77) and essential gene by Lamichhane *et al*. (78)

SV01 and SV04 were variants located in intergenic regions. SV01 was amongst senX3 and regX3 genes and SV04 was located in the middle of Rv2813 and Rv2814c loci.

SV02 was a 78 bp deletion located on the Rv0064 gene, common to all isolates. Rv0064 encodes for a probable conserved transmembrane protein and was considered to be positioned in a hot spot region of the *M. tuberculosis* genome, which is considered a region with high frequency of variants (79).

SV03, a 287 bp deletion, was found on the Rv1882c and Rv1883c loci and was only present in isolate 12. Rv1882c codes for a probable short-chain type dehydrogenase/reductase involved in the intermediary metabolism and respiration, and Rv1883c's product is a conserved hypothetical protein without known function (76).

Rv1883c was found deleted in isoniazid-resistant *M. tuberculosis* mutants derived from four strains and the deletion was said to have a negative impact in the intracellular growth and showed varying sensitivities concerning oxidative stress (80).

The disruption of the Rv1882c and Rv1883c genes confer advantages for *in vitro* growth of H37Rv (77).

SV05 was an 87 bp deletion identified in all the isolates and located in the Rv2885c locus. Rv2885c codes for a transposase in the IS*1539* insertion sequence that is essential for *in vivo* survival (81)

The 84 bp deletion SV06 located in the Rv3892c locus, was discovered in all the isolates. This locus codes for the PPE69 protein, member of the *M. tuberculosis* PPE family of conserved proteins (76). PPE69 protein was associated with the cell wall and cell processes functional category (82). PPE69 was detected in or on the mycobacterial membrane and/or cell wall suggesting that its specific function may be dependent on cell wall or extracellular location (83) In a study to analyze the variation dynamics of the *M. tuberculosis* genome, PPE69 was designated to be a gene involved in strain-specific deletions in a Haarlem strain. (84) PPE69 was associated with ESX-2 type VII secretion system, which may not be functional in *M. tuberculosis* (85).

Although we used DELLY, the SVs detection had limitations due to the fact that our dataset was constituted of single-end DNA reads, therefore there was no paired-end support corroborating our results. Alone, split-read analysis has a weakness for the detection of small SVs and SVs in unique genomic regions. In addition, there is a lack of sensitivity to determine certain classes of SVs, such as tandem duplications, inversions, translocations (86).

It must also be noted that is much more difficult to detect SVs than SNPs or small indels. Sequencing and mapping errors can confuse the patterns created by the SVs because the patterns produced by the SVs can be very similar, and multiple SVs can overlap, creating complex mapping patterns (72).

Another limitation of the SV detection is the fact that is complicated to identify SVs accurately from short DNA sequence reads from next-generation DNA sequencing as each read may not comprise an entire SV, seeing that it can extend a large part of a read or can be larger than the read itself.

These limitations could be overcome using *de novo* assembly methods or third-generation sequencing (TGS) technology because TGS produce single-end reads with read length up to thousands bp (87).

Variant calling was performed to obtain information on the positions of the genetic variants and their corresponding genotypes.

We obtained a total of 1029 high quality SNPs, from which we identified 262 synonymous (sSNPs), 482 non-synonymous (nsSNPs) and 104 intergenic SNPs (**Table 7**).

**Table 7 – Number and type of variants identified**

| Type of variant | No. |
|---|---|
| Synonymous variants (sSNPs) | 262 |
| Missense variant | 412 |
| Intergenic variant | 104 |
| Frameshift variant | 65 |
| Disruptive in-frame insertion | 2 |
| Disruptive in-frame deletion | 3 |
| Splice region variant & stop retained variant | 1 |
| Stop gained | 5 |
| Stop lost & splice region variant | 1 |

Missense, frameshift, disruptive in-frame deletions and insertions, and stop gained, are all nsSNPs.

The analysis identified more nsSNPs than sSNPs. nsSNPs account for more than two thirds of SNPs in MTBC, which goes against normal sSNPs predominance, like in other organisms. nsSNPs should be naturally removed from the population by purifying selection because they are on average slightly deleterious, but surprisingly that does not occur on the MTBC.

Fleischmann et al. sequenced the complete genome of the *M. tuberculosis* clinical strain CDC1551 and performed a whole-genome comparison with H37Rv strain in order to identify polymorphic sequences. The author concluded that the substitution ratio observed in *M. tuberculosis* indicated that extra selective pressure is present on synonymous substitutions or there is reduced selective pressure against nonsynonymous mutations. That could be caused by codon bias that aims to maintain a high G/C content and thus limiting the number of synonymous substitutions, or as a result of *M.* tuberculosis' low recombination frequency (88)

Other explanation for this unusual finding was the effect of the MTBC's short evolutionary age and very rapid expansion in population size, so that there has not passed sufficient evolutionary time to permit the loss of deleterious mutations by purifying selection (89).

Later in 2008, Hershberg et al. evaluated the synonymous and nonsynonymous MTBC differences in 89 genes and did genome-wide pairwise comparison between *M. tuberculosis* H37Rv and *M. tuberculosis* H37Ra, *M. tuberculosis* F11, *M. tuberculosis* CDC1551, *M. bovis*

and *M. bovis* BCG Pasteur 1173P2 to obtain the number of synonymous and nonsynonymous differences for all orthologs protein pairs across the genome. Results showed more nonsynonymous than synonymous differences. The author concluded that the reduced natural selection resulted from high clonality, that is, the absence of horizontal gene exchange, and due to the serial transmission bottlenecks, which occurs because only a few individual pathogens are transmitted from one infected host to another in the initiation of a new infection, that is characteristic of the MTBC. Also mentioned that these factors tend to reduce an organism's effective population size. (90)

From the list of SNPs, we excluded the variants in common with the reference and analyzed 17 SNPs that differentiated the isolates (Table 8)

The genomic positions of the SNPs were searched using the MycoBrowser database (76), we sought to know the function of the genes in question and we looked for studies on mutants in those loci.

**Table 8 - SNPs diferentiating each isolate**

| SNP ID | Position[1] | Locus (gene) | AA exchange | SNP type | Gene Essentiality[2] | Product | Function | Isolate ID | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | 9 | 11 | 6 | 8 | 12 | 7 | 10 |
| **SNP001** | 725107 | Rv0631c (recC) | - | Frameshift variant | Non-essential | Probable exonuclease V (gamma chain) RecC (exodeoxyribonuclease V gamma chain)(exodeoxyribonuclease V polypeptide) | Involved in homologous recombination | x | | | | | | |
| **SNP002** | 819128 | Rv0726c | Trp171STOP | Stop gained | Non-essential | Possible S-adenosylmethionine-dependent methyltransferase | Exhibits S-adenosyl-L-methionine-dependent methyltransferase activity | | | | x | x | x | x |
| **SNP003** | 1146655 | Rv1025 | Pro32Leu | Missense variant | Essential | Conserved protein | Unknown | x | x | x | x | | | x |
| **SNP004** | 1510652 | Rv1345 (mbtM) | Gly458Arg | Missense variant | Non-essential | Probable fatty acyl-AMP ligase MbtM | Thought to be involved in the biogenesis of the hydroxyphenyloxazoline-containing siderophore mycobactins | | | x | | | | |
| **SNP005** | 1707310 | Rv1515c | Leu73 | Synonymous variant | Non-essential | Conserved hypothetical protein | Unknown | | | | | | | x |
| **SNP006** | 1894300 | Rv1668c | Pro348fs | Frameshift variant | Non-essential | Probable first part of macrolide-transport ATP-binding protein ABC transporter | Responsible for energy coupling to the transport system of macrolide antibiotics resistance by an export mechanism. | x | | | | | | |
| **SNP007** | 2266553 | Rv2020c | silent (Ser56) | Synonymous variant | Non-essential[3] | Conserved hypothetical protein | Unknown | x | x | | x | | | |

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **SNP008** | 2266583 | Rv2020c | Glu46Asp | Missense variant | Non-essential[3] | Conserved hypothetical protein | Unknown | x | x | | x | | | |
| **SNP009** | 2266604 | Rv2020c | silent (Ser39) | Synonymous variant | Non-essential[3] | Conserved hypothetical protein | Unknown | x | x | x | x | x | | |
| **SNP010** | 2266624 | Rv2020c | Leu33Ile | Missense variant | Non-essential[3] | Conserved hypothetical protein | Unknown | x | x | x | x | x | | |
| **SNP011** | 2930919 | Rv2603c | Gly214 | Synonymous variant | Non-essential | Highly conserved protein | Unknown | | | | x | | | |
| **SNP012** | 2948752 | Rv2622 | Gly39 | Synonymous variant | Non-essential | Possible methyltransferase (methylase) | Intermediary in metabolism and respiration | | x | x | x | x | x | x |
| **SNP013** | 3194705 | Rv2885c | - | Frameshift variant | Non-essential | Probable transposase | Required for the transposition of the insertion element IS1539 | x | x | x | | | | |
| **SNP014** | 3311572 | Rv2958c | Gly143 | Synonymous variant | Non-essential[3] | Possible glycosyl transferase | Probably involved in cellular metabolism and in resistance to killing by human macrophages | | x | x | x | x | x | x |
| **SNP015** | 3415180 | - | - | Intergenic | - | - | - | | x | | | x | x | x |
| **SNP016** | 3564391 | Rv3195 | Leu10Met | Missense variant | Non-essential | Conserved hypothetical protein | Unknown | | | | | x | | x |
| **SNP017** | 3964383 | - | - | Intergenic | - | - | - | | | | x | | x | x |

[1]Relative to the genome position of *M. tuberculosis* H37Rv (GenBank Accession NC_000962.3)

[2]Gene essentiality- according to DeJesus et al. (2017) using saturated Himar1 transposon libraries (77)

[3]Non-essential gene by DeJesus et al. (77) and essential gene by Sassetti et al. (81)

We detected a frameshift SNP, SNP001, in recC (Rv0631c). recC codes for a probable exonuclease V and is a member of the "information pathways" group, which are genes involved in replication, repair and transcription of nucleic acid. During survival in the human macrophages, *M. tuberculosis* actively increases the transcription of several genes, and Rv0631c (recC) was found upregulated. Rv0631c (recC) is known to be involved in homologous recombination and was related to the path of SOS response mechanism of drug resistance (91)

SNP002 resulted in a premature stop codon, leading to a shortened transcript. This variant was found in the Rv0726c gene which is associated with *M. tuberculosis* growth and with the biosynthesis of cell wall lipids (92)

SNP003 caused a Pro32Leu change and was present in isolates 6, 8, 9, 10 and 11. This variant was located in Rv1025. Several studies revealed that Rv1025 is an essential gene for *in vitro* growth (77) (93) (94) whose product is a conserved protein.

The missense SNP004 was located in the mbtM (Rv1345) gene. mbtM (Rv1345) was assumed to be a slow growth mutant by Himar1-based transposon mutagenesis in H37Rv strain (94). mbtM (Rv1345) encodes a probable fatty acyl-AMP ligase MbtM thought to be involved in the biogenesis of the hydroxyphenyloxazoline-containing siderophore mycobactins. Mycobacterial siderophores are critical components for bacterial virulence in the host and this system is believed to be utilized for iron acquisition in iron-limiting environments (95). mbtM (Rv1345) codes for a protein associated to immunological or pathogenetic features of *M. tuberculosis* infection (96)

SNP005 was a synonymous variant in the Rv1515c gene. Rv1515c encodes for a conserved hypothetical protein with unknown function. With the intention to show that the combination of cysteine with isoniazid prevented the formation of drug-tolerant and drug-resistant cells in *M. tuberculosis* cultures, the expression of several genes was studied. Among others, Rv1515c's expression was repressed upon addition of cysteine to INH-treated *M. tuberculosis* cultures (97) Rv1515c gene was acquired via horizontal gene transfer associated with the emergence of pathogenesis in *M. tuberculosis* (98)

Arising as a frameshift variant, SNP006, was situated in Rv1668c gene. Rv1668c is thought to be responsible for energy coupling the transport system involved in active transport of macrolide across the membrane, conferring macrolide antibiotics resistance by an export mechanism (99)

In the Rv2020c gene we spotted SNP007, SNP008, SNP09 and SNP010. SNP007 and SNP09 were synonymous variants, and SNP008 and SNP010 were non-synonymous. Rv2020c encodes for a conserved hypothetical protein with no known function and was found in the gene cluster of Zone-4 (Cell Wall, Cell Processes, and metabolism) on the gene expression map. (100)

SNP011 lied on gene Rv2603c. Rv2603c is codes for a highly conserved protein without known function. It is considered an essential gene for in vitro growth of H37Rv, by Himar1 transposon mutagenesis Griffin et al. (93). The *crtP* locus, from Rv2606c *to* Rv2603c, supposedly has a role in mechanisms of protection from the stresses of the intraphagosomal environment in *M. tuberculosis* (101)

Common to all isolate, except for isolate 9, SNP0012 was a synonymous variant found in Rv2622. Rv2622 encodes for a possible methyltransferase that was identified in the cell membrane fraction of *M. tuberculosis* H37Rv (102). Rv2622 was considered to be a TB reactivation-associated antigen which induced differentiate and high IFN-γ production response in infected individuals in a long-term stimulation (103).

We located the SNP013, which was a frameshift variant, in the Rv2885c locus. Rv2885c codes for a transposase in the IS*1539* insertion sequence that is essential for *in vivo* survival (81)

SNP014 was present in isolates 6, 7, 8, 10, 11 and 12 and was in Rv2958c. Studies showed that Rv2958c exhibited immunogenicity, was related to virulence (104) and demonstrated essentiality for growth, metabolism and cell wall synthesis. Rv2958c recombinant protein shown to exhibit high specificity and sensitivity in detecting immunoglobulin G antibody, notably was able to recognize TB-positive sera and was believed to be a valuable potential diagnostic antigen for TB. (105)

*In vitro* growth of H37Rv is actually enhanced when the Rv2020c and Rv2958c genes are disrupted by transposon insertion (77)

SNP016 missense variant was in Rv3195. Rv3195 was associated with resistance to mefloquine when overexpressed (106)

Understanding the types of strains and lineages of *M. tuberculosis* circulating in a country is very important for the TB control.

To portray the population structure of the 7 *M. tuberculosis* isolates, we performed *in silico* Spoligotyping (Table 9). Both the spoligotypes and lineages were inferred using SITVIT2 database (65)

**Table 9 - *In silico* spoligotyping**

| Isolate ID | Spoligotype – Octal code | Spoligo Internacional Type (SIT) | Subfamily/Clade |
|---|---|---|---|
| 6 | 777777777760771 | 53 | T1 |
| 7 | 775777777760771 | 966 | T1 |
| 8 | 777777777660771 | 167 | T1 |
| 9 | 777777775760771 | 122 | T1 |
| 10 | 777777777760771 | 53 | T1 |
| 11 | 777777777760771 | 53 | T1 |
| 12 | 777777777760771 | 53 | T1 |

Our results showed a moderate level of diversity of genotypes, taking into account the small size of the cluster, with 4 SITs discovered. All the different spoligotype patterns pointed out corresponded to a pre-existing type in the SITVIT2 database.

The predominant spoligotype pattern designated was SIT53, followed by SIT966, SIT167 and SIT122. The isolates were all appointed to be part of the T1 family. The T1 family belongs to the ill-defined T sub-lineage constituent of the Euro-American lineage.

Despite the fact that no specific spoligotypes were found with a robust local phylogeographical specificity, it is know that SIT53, which is the phylogenetic prototype of the T family, is broadly disseminated around the world (107). T is the most prevalent spoligotype family in Northern, Southern and Western Europe (108)

The Euro-American lineage is the most benign in terms of frequency and severity of TB (109). In Latvia the majority of *M. tuberculosis* isolates belonged to the Euro-American lineage (110) and T is one of the main spoligotype families in Riga (111)

Comas et al. 2009 revealed that Spoligotyping was not a phylogenetically robust marker because it showed high rates of convergent evolution and could not define phylogenetic relationships between strain groupings. This is explained by the fact that the spacers used in spoligotyping exhibit homoplasy, that is, independent mutational events that result in the loss of the same spacer. Which ultimately results in unrelated strains that belong to different evolutionary lineages to have identical patterns spoligotypes. Although Comas et al. 2009 did not considered spoligotyping to be a reliable tool for formal phylogenetic analyses, the author mentioned that specific patterns could be informative for population genetic analyses, for example, for the Beijing lineage that has a characteristic loss of 34 spacers. (37).

Kato-Maeda et al. studied the accuracy of Spoligotyping in strain classification using lineage specific-LSP/SNP. Results showed that spoligotyping was informative, classified the *M. tuberculosis* isolates into the main strain lineages in 95% of cases and there was no evidence of convergent spoligotypes (when isolates with the same spoligotype belong to different lineages). Very importantly, it is elucidated that spoligotype families are sub-lineages within the main lineages defined by SNPs and that is not possible to compare the results from both genetic markers because the spoligotype families and the corresponding SNP-based lineages are not ‘phylogenetically equivalent’, as they do not define the same branches of the *M. tuberculosis* phylogeny. Spoligotyping provided additional resolution within the lineages, nevertheless this study was not representative of the global diversity of *M. tuberculosis* as it did not include all the lineages of *M. tuberculosis*, only 3 lineages were studied. (112)

From the list of high-quality SNPs, eleven of them, 3 synonymous and 8 non-synonymous, were phylogenetically informative markers (Table 10). The SNPs were equally present in all the isolates.

**Table 10 – Phylogenetically informative SNPs**

| Position[1] | Gene (Locus) | AA Exchange | Type of SNP | Sub-lineage |
|---|---|---|---|---|
| 7585 | gyrA (Rv0006) | Ser95Thr | Missense | H37Rv ATCC |
| 62657 | dnaB (Rv0058) | silent (Pro754) | Synonymous | 4.1 (X-type) |
| 107794 | fcoT (Rv0098) | silent (Ala65) | Synonymous | 4.1.2.1 (Haarlem) |
| 491591 | fgd1 (Rv0407) | Lys270Met(s) | Missense | Haarlem |
| 575679 | mshA (Rv0486) | Asn111Ser | Missense | Haarlem |
| 648990 | mgtA (Rv0557) | Arg152Pro | Missense | Haarlem |
| 765150 | rpoC (Rv0668) | Gly594Glu | Missense | 4.1 (X-type) |
| 891756 | cfp29 (Rv0798c) | Leu(s)172Leu | Synonymous | 4.1.2 |
| 2053987 | mgtC (Rv1811) | Arg182His | Missense | Haarlem |
| 4239298 | aftA (Rv3792) | Ala456Val | Missense | 4.1.2.1 (Haarlem) |
| 4242803 | embC (Rv3793) | Val(s)981Leu | Missense | Haarlem |

[1]Relative to the genome position of *M. tuberculosis* H37Rv (GenBank Accession NC_000962.3)

The non-synonymous Ser95Thr change in *gyrA* (*Rv0006*) was assumed to be a phylogenetically informative mutation for H37Rv ATCC and the missense mutation Asn111Ser in mshA (Rv0486) was mention as being specific for the Haarlem sub-lineage (113)

As determined by Coll et al., Leu(s)172Leu in cfp29 (Rv0798c) is phylogenetically informative for lineage 4.1.2., both the silent mutation Pro754 in dnaB (Rv0058) and the non-synonymous Gly594Glu variant in rpoC (Rv0668) are specific for Lineage 4.1 (X-type). The synonymous SNP Ala65 in fcoT (Rv0098) and the Ala456Val change in aftA (Rv3792) are considered phylogenetically informative SNPs for the 4.1.2.1 (Haarlem) lineage. (25)

In a population based study, with MTBC strains from Sierra Leone, with the purpose to determine the genetic basis of first line drug resistance, the sequence analyses of *emb*C revealed a mutation in embC Val(s)981Leu which was described to being specific for the Haarlem genotype (114)

In a study about the performance of sequence based analysis for discriminatory phylogenetic classification of clinical MTBC isolates using *de novo* sequencing, it was demonstrated that the occurrence of the non-synonymous variants Arg182His in mgtC

(Rv1811), the Arg152Pro in mgtA (Rv0557) and the Lys270Met(s) mutation in fgd1 (Rv0407) were correlated with the phylogenetic strain classification of Haarlem sub-lineage. (52)

WGS enabled us to do a more robust genomic characterization of the studied isolates. Coll et al. 2014 investigated the use of SNPs as robust markers for the genetic variation of *M. tuberculosis*, and like never done before, the list of 62 SNPs to distinguish strains of *M. tuberculosis* covered all main lineages and classified a greater number of sub-lineages. (25)

We searched for lineage-specific SNPs and we classified the 7 isolates within the Euro-American lineage and in sub-lineage 4.1.2.1 (Haarlem).

With the intention of predicting drug resistance of the isolates solely from the WGS genotypic data, we relied on the accurate characterization of all sequence variants and we searched for variants in genes known to be related to phenotypic antibiotic resistance.

We found the Asp461Asn mutation in gyrB (Rv0005) in all the isolates. This SNP was listed as a high confidence SNP associated with fluoroquinolones resistance (115) The substitutions Ser95Thr, Glu21Gln and Gly668Asp in gyrA were common to all isolates. These substitutions are neutral polymorphisms not associated with FQ resistance (116) (117).

We considered the isolates to be susceptible to all the anti-TB drugs since this analysis was performed *in silico* and we lacked Drug Susceptibility Test results.

We defined our cases as being part of a household transmission cluster. A household is defined as a group of people, often a family, that live together within one residence and a household outbreak is established when a group of two or more cases occur exclusively in one house during the study period.

A prolonged stay in the same house and the duration of living in the community were associated with TB transmission. It was shown that there is an increased change of acquiring TB when there is a high effective contact rate due to prolonged and persistent exposure to *M. tuberculosis* (118)

Transmission events are primarily assessed through identification of clusters, which consist of shared *M. tuberculosis* genotypes among the TB patients. Clusters are often used as a representative for recent transmission (119) (120)

Fundamentally, over time and in a constant pace a strain will acquire new SNPs and will retain existing ones, and throughout the chain of transmission the existing SNPs will be passed on and new ones will eventually arise, accumulating in the direction of transmission (121). However, the molecular evolution of *M. tuberculosis* was characterized by periods of relative genomic stability followed by bursts of mutation (122)

Several authors have studied the inclusion thresholds for epidemiological linkage. With the intention of revealing the strain microevolution of isolates of the Harlingen outbreak, Schurch et al. 2010 found differences of 0–3 SNPs between isolates. (123). Bryant and Schürch made an attempt to estimate a molecular clock for *M. tuberculosis* by sequencing 199 isolates from epidemiologically linked TB cases, collected in the Netherlands over a period of almost 16 years. Results shown an average mutation rate of 0.3 SNPs per genome per year (124). From a retrospective study of TB outbreaks, Walker et al. 2013 derived that a cut-off fewer than six

SNPs is key to discriminate between TB patients who have suspected transmission and more than 12 SNPs to consider the cases unrelated (45)

Although we did not have contact tracing information or any relevant epidemiologic data to draw strong links, we were able to make some assumptions about the chain of transmission by analyzing both phylogenetic trees based on the SNP data of the isolates and the dates of samples collection.

Adopting the criteria defined by Walker et al. 2013, we concluded that the isolates were all clustered using the cut-off of 12 SNPs (Figure 8) and, employing the 5 SNPs threshold we divided the cases into two sub-clusters. One sub-cluster was considered to be formed by isolates 9, 6 and 11, and on the other sub-cluster we included isolates 12, 7 and 10. Isolate 8 was not included in neither of the subclusters because it had 6 SNPs difference between each sub-cluster.

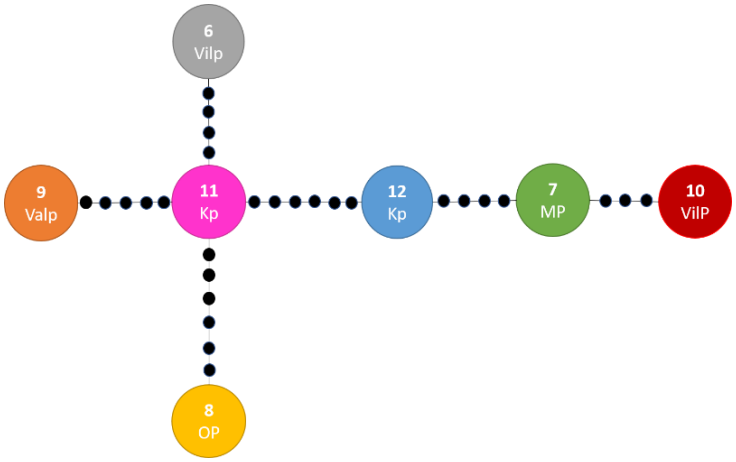


**Figure 8 – Philogenetic tree of the Household Cluster**

Each isolate from the study was represented by a node and the lengths of the branches indicate the SNP distance between isolates. The internal nodes in the tree represent the divergence of the isolates. The bootstrap values (%) were assigned to each internal node in the original tree and represents the number of times that the branch pattern of that node was reproduced in the replicate trees.

The sub-clusters defined could be also correlated in terms of temporal patterns, as samples from cases 9, 11 and 6 were collected in 2006, 2009 and 2010, respectively, and isolates 12, 7 and 10 corresponded to 2011, 2010 and 2016.

To assign the index case of the cluster, we included the dates of sample collection. The index case was assumed to be patient ValP (isolate 9) because was the one that had the earliest sample collection, in 2006. Contradictorily, the earliest identified case is not necessarily the

first source of transmission. Delays in the propitious diagnosis of TB may occur due to differences in health care-seeking behavior, difficulties in accessing health services and lack of symptoms (125). However, when analyzing the phylogenetic tree, isolate 9 was in the most basal position and near the common ancestor. All subsequently identified cases were defined as secondary cases.

Two patients, KP and VilP, had two samples collected at different times.

Recurrent episodes of TB are classified either as relapses or re-infections. A re-infection implies ongoing transmission and lack of immunity to the fight the newly infecting strain or high intensity of exposure. A relapse is suggestive of inadequate or insufficient treatment. (126)

So that it was possible to make a distinction between relapse or re-infection it was necessary to quantify the genomic differences between isolates from the first and the recurrent episodes and take under consideration the time passed between cases.

Isolates 11 and 12 were from patient KP, had 6 SNPs difference and were collected with a time gap of 2 years. 6 and 10 belonged to VilP, showed a genetic difference of 9 SNPs and were collected almost 6 years apart (Table 11).

There is no standard genomic distance to differentiate between relapses and re-infections.

**Table 11 - Recurent cases**

| Patient | Isolate ID | Time between dates of sample collection (years) | No. SNPs difference |
|---|---|---|---|
| **KP** | 11 | 2 | 6 |
| | 12 | | |
| **VilP** | 6 | 5 | 9 |
| | 10 | | |

Number of SNPs distance between isolates and time between date of collection in the patients with more than one specimen from the same episode of disease or from a relapse.

Bryant et al. considered a case as relapse if the recurrent isolates had a maximum genomic distance of 0–6 SNPs and mentions that cases were re-infections when they had more than 1306 SNPs separating each case. (127)

Guerra-Assunção et al. 2015 (128) tried to classified the recurrent cases of TB in a Malawian cohort, defining them as relapses if they differed by less than 10 SNPs from the initial

strain, and in another study (129) classified recurrent cases as re-infections with a cut-off of more than 100 SNPs.

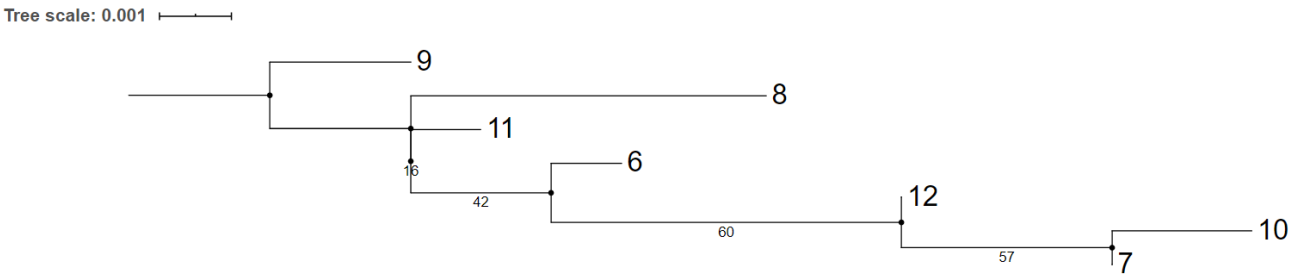


**Figure 9 - Maximum likelihood phylogenetic inference tree (*Sourse*: iTOL)**

When looking at the maximum likelihood phylogenic tree (Figure 11), the groups of isolates, 6 and 10, and 11 and 12, were too separated from each other and did not form a monophyletic group.

We assumed that these cases were re-infections caused by the same strain that infected the rest of the patients in the household because the genetic difference and the temporal interval between infections was too substantial to a possible endogenous reactivation of the disease. The patients lived in the same house which is consistent with a high intensity of exposure to TB.

This method is not that straightforward and has several limitations, and the foremost obstacle being the fact that it is possible to be re-infected with a genetically identical strain (121).

Genome sequencing allowed us to determine SNPs, distinguish between isolates and identify a transmission cluster. Even so, the SNPs were not specific as each SNP was present in any of the isolates (Table 8) so we could not infer the chain of transmission with certainty. Yet, assuming that ValP (isolate 9) was the index case, we could state that he transmitted TB to KP (isolate 11), then to VilP (isolate 6) and years later to OP (isolate 8). Patient KP (isolate 12) was re-infected probably in the course of VilP's infection. Later on, MP (isolate 7) was infected and VilP (isolate 10) got re-infected.

We lacked epidemiologic data to infer conclusions with certainty. Medical history, demographic data, socio-economic status, occupation, history of TB treatment, laboratory testing, treatment regimens and treatment outcomes, drug susceptibility testing results, size and composition of the household, contact tracing and investigation would have key information to unravel this chain of transmission.

# 4 Conclusion

Genome sequencing and subsequent analysis made it possible to determine SNPs and SVs, distinguish between isolates and identify a transmission cluster.

Six SVs were found but several limitations arose, such as not having paired-end support because the data was composed by single-end reads, the complex patterns originated by the SVs could have generated sequencing and mapping errors, and the short DNA reads may have not comprised an entire SV. The method used was considered error-prone and had low sensitivity.

We detected 1029 high-quality SNPs, from which 17 differentiated the studied isolates. Even so, the SNPs were not specific as each SNP was present in any of the isolates, not allowing us to infer unambiguously a chain of transmission.

In relation to strain classification, based on *in silico* Spoligotyping the isolates belonged to the T1 sub-family. The method used is subject to error because the spoligotype patterns were predicted from the sequencing reads instead of the standard PCR-based approach.

When employing the list of phylogenetic specific SNPs, the studied strain was determined to be part of the Haarlem sub-lineage.

We searched for polymorphisms in gene associated with drug resistance, but the SNPs present in the isolates were not robust, therefore the isolates were considered to be susceptible to all anti-tuberculosis drugs.

The two recurrent cases, from patients KP and VilP, were defined as re-infections due to the large temporal gap and the number of SNPs differentiating the first and second isolates from each patient. This method is not that straightforward, has several limitations, and the foremost obstacle being the fact that it is possible to be re-infected with a genetically identical strain. The existence of re-infection cases leads to believe that the patients were highly exposed, there was continuous transmission and they could have had low immunity to fight the bacteria.

We generated hypotheses in order to establish the routes of transmission. Phylogenetic analysis was carried out to follow the evolution of the strain, as it made it possible to compare the final genotypic data and to examine the genetic changes between the isolates. We used a 12 SNP cut-off we delineated the household transmission cluster and with a threshold of 5 SNP we separated the isolates in two subclusters. Nevertheless, even with WGS' high-resolution, the route of transmission within this household was unclear.

The present work was the perfect example to show that the molecular epidemiology data needs to be combined with classical epidemiology methods and clinical information to effectively investigate TB outbreaks.

# Bibliography


1. World Health Organization. Global tuberculosis report 2018. World Health Organization. http://www.who.int/iris/handle/10665/274453. 2018. 265 p.
2. World Health Organization. Global tuberculosis report 2020. Geneva: World Health Organization; 2020. Licence: CC BY-NC-SA 3.0 IGO. 2020. 232 p.
3. Russell DG. MYCOBACTERIUM TUBERCULOSIS: HERE TODAY, AND HERE TOMORROW. 2001;2(August):1–9.
4. Centre For Disease Control And Prevention. TB Elimination Tuberculosis: General Information. Basic TB facts. 2011;(Ic):1–2.
5. Cardona PJ. Pathogenesis of tuberculosis and other mycobacteriosis. Enferm Infecc Microbiol Clin [Internet]. 2018;36(1):38–46. Available from: http://dx.doi.org/10.1016/j.eimc.2017.10.015
6. Fry DE. Extra-Pulmonary Tuberculosis and Its Surgical Treatment. Surg Infect (Larchmt). 2016;17(4):394–401.
7. Ahmad S. Pathogenesis, immunology, and diagnosis of latent mycobacterium tuberculosis infection. Clin Dev Immunol. 2011;2011.
8. Barry III CE, Boshoff H, Dartois V, Dick T, Ehrt S, Flynn J, et al. The spectrum of latent tuberculosis: rethinking the goals of prophylaxis. Nat Rev Microbiol [Internet]. 2009;7(12):845–55. Available from: http://www.stoptb.org/globalplan/
9. Wlodarska M, Johnston JC, Gardy JL, Tang P. A microbiological revolution meets an ancient disease: Improving the management of tuberculosis with genomics. Clin Microbiol Rev. 2015;28(2):523–39.
10. Silva DR, Muñoz-Torrico M, Duarte R, Galvão T, Bonini EH, Arbex FF, et al. Risk factors for tuberculosis: Diabetes, smoking, alcohol use, and the use of other drugs. J Bras Pneumol. 2018;44(2):145–52.
11. Golden MP, Vikram HR. Extrapulmonary tuberculosis: An overview. Am Fam Physician. 2005;72(9):1761–8.
12. Eurydice. Demographic situation: Latvia. 2019;3–7. Available from: https://eacea.ec.europa.eu/national-policies/eurydice

13. World Health Organization. Division of Communicable Diseases, Surveillance. Anti-tuberculosis drug resistance in the world / the WHO/IUATLD Global Project on Anti-Tuberculosis Drug Resistance Surveillance. Report 2, Prevalence and trends. Geneva [Internet]. 2000;(document no. WHO/TB/97.229 WHO/CDS/TB/2000.278.). Available from: http://www.who.int/iris/handle/10665/66493

14. Zipperer M. Tackling tuberculosis in Latvia. PLoS Med. 2005;2(5):0380–2.

15. Leimane V LJ. Tuberculosis control in Latvia: integrated DOTS and DOTS-plus programmes. Euro Surveill. 2006;11(3):29-33. 2006; Available from: https://www.eurosurveillance.org/content/10.2807/esm.11.03.00610-en

16. Blöndal K. Barriers to reaching the targets for tuberculosis control: Multidrug-resistant tuberculosis. Bull World Health Organ. 2007;85(5):387–90.

17. Hargreaves JR, Boccia D, Evans CA, Adato M, Petticrew M, Porter JDH. The social determinants of tuberculosis: from evidence to action. Am J Public Health. 2011;101(4):654–62.

18. ECDC. European Centre for Disease Prevention and Control. Tuberculosis in Latvia. Stockholm: ECDC; 2013. 2012.

19. Lönnroth K, Jaramillo E, Williams BG, Dye C, Raviglione M. Drivers of tuberculosis epidemics: The role of risk factors and social determinants. Soc Sci Med. 2009;68(12):2240–6.

20. ECDC. European Centre for Disease Prevention and Control/WHO Regional Office for Europe. Tuberculosis surveillance and monitoring in Europe 2018 – 2016 data. 2018. 206 p.

21. Brosch R, Gordon S V., Marmiesse M, Brodin P, Buchrieser C, Eiglmeier K, et al. A new evolutionary scenario for the Mycobacterium tuberculosis complex. Proc Natl Acad Sci U S A. 2002;99(6):3684–9.

22. Gagneux S, Small PM. Global phylogeography of Mycobacterium tuberculosis and implications for tuberculosis product development. Lancet Infect Dis [Internet]. 2007 May;7(5):328–37. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1473309907701081

23. Coscolla M, Gagneux S. Does M. tuberculosis genomic diversity explain disease diversity? Drug Discov Today Dis Mech [Internet]. 2010 Mar;7(1):e43–59. Available

from: https://linkinghub.elsevier.com/retrieve/pii/S1740676510000258

24. Gagneux S. Genetic Diversity in Mycobacterium tuberculosis. In 2013. p. 1–25. Available from: http://link.springer.com/10.1007/82_2013_329

25. Coll F, McNerney R, Guerra-Assunção JA, Glynn JR, Perdigão J, Viveiros M, et al. A robust SNP barcode for typing Mycobacterium tuberculosis complex strains. Nat Commun [Internet]. 2014 Dec 1;5(1):4812. Available from: http://www.nature.com/articles/ncomms5812

26. Ngabonziza JCS, Loiseau C, Marceau M, Jouet A, Menardo F, Tzfadia O, et al. A sister lineage of the Mycobacterium tuberculosis complex discovered in the African Great Lakes region. Nat Commun [Internet]. 2020 Dec 9;11(1):2917. Available from: http://www.nature.com/articles/s41467-020-16626-6

27. Warner DF, Koch A, Mizrahi V. Diversity and disease pathogenesis in Mycobacterium tuberculosis. Trends Microbiol [Internet]. 2015 Jan;23(1):14–21. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0966842X14002157

28. Valway SE, Sanchez MPC, Shinnick TF, Orme I, Agerton T, Hoy D, et al. An Outbreak Involving Extensive Transmission of a Virulent Strain of Mycobacterium tuberculosis. N Engl J Med [Internet]. 1998 Mar 5;338(10):633–9. Available from: http://www.nejm.org/doi/abs/10.1056/NEJM199803053381001

29. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, et al. Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. Nature [Internet]. 1998 Jun;393(6685):537–44. Available from: http://www.nature.com/articles/31159

30. Gardy JL, Johnston JC, Sui SJH, Cook VJ, Shah L, Brodkin E, et al. Whole-Genome Sequencing and Social-Network Analysis of a Tuberculosis Outbreak. N Engl J Med [Internet]. 2011 Feb 24;364(8):730–9. Available from: http://www.nejm.org/doi/abs/10.1056/NEJMoa1003176

31. van Embden JD, Cave MD, Crawford JT, Dale JW, Eisenach KD, Gicquel B, et al. Strain identification of Mycobacterium tuberculosis by DNA fingerprinting: recommendations for a standardized methodology. J Clin Microbiol [Internet]. 1993;31(2):406–9. Available from: https://jcm.asm.org/content/31/2/406

32. Driscoll JR. Spoligotyping for Molecular Epidemiology of the Mycobacterium

tuberculosis Complex. In 2009. p. 117–28. Available from: http://link.springer.com/10.1007/978-1-60327-999-4_10

33. Supply P, Allix C, Lesjean S, Cardoso-Oelemann M, Rusch-Gerdes S, Willery E, et al. Proposal for Standardization of Optimized Mycobacterial Interspersed Repetitive Unit-Variable-Number Tandem Repeat Typing of Mycobacterium tuberculosis. J Clin Microbiol [Internet]. 2006 Dec 1;44(12):4498–510. Available from: https://jcm.asm.org/content/44/12/4498

34. Yang ZH, Ijaz K, Bates JH, Eisenach KD, Cave MD. Spoligotyping and polymorphic GC-rich repetitive sequence fingerprinting of mycobacterium tuberculosis strains having few copies of IS6110. J Clin Microbiol [Internet]. 2000 Oct;38(10):3572–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11015365

35. Xia E, Teo Y-Y, Ong RT-H. SpoTyping: fast and accurate in silico Mycobacterium spoligotyping from sequence reads. Genome Med [Internet]. 2016 Dec 17;8(1):19. Available from: http://genomemedicine.com/content/8/1/19

36. Ruettger A, Nieter J, Skrypnyk A, Engelmann I, Ziegler A, Moser I, et al. Rapid Spoligotyping of Mycobacterium tuberculosis Complex Bacteria by Use of a Microarray System with Automatic Data Processing and Assignment. J Clin Microbiol [Internet]. 2012 Jul 1;50(7):2492–5. Available from: https://jcm.asm.org/content/50/7/2492

37. Comas I, Homolka S, Niemann S, Gagneux S. Genotyping of Genetically Monomorphic Bacteria: DNA Sequencing in Mycobacterium tuberculosis Highlights the Limitations of Current Methodologies. Litvintseva AP, editor. PLoS One [Internet]. 2009 Nov 12;4(11):e7815. Available from: https://dx.plos.org/10.1371/journal.pone.0007815

38. Ansorge WJ. Next-generation DNA sequencing techniques. N Biotechnol [Internet]. 2009 Apr;25(4):195–203. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1871678409000089

39. Metzker ML. Sequencing technologies — the next generation. Nat Rev Genet [Internet]. 2010 Jan 8;11(1):31–46. Available from: http://www.nature.com/articles/nrg2626

40. Buermans HPJ, den Dunnen JT. Next generation sequencing technology: Advances and applications. Biochim Biophys Acta - Mol Basis Dis [Internet]. 2014 Oct;1842(10):1932–41. Available from: https://linkinghub.elsevier.com/retrieve/pii/S092544391400180X

41. Rusk N. Torrents of sequence. Nat Methods [Internet]. 2011 Jan 20;8(1):44–44. Available from: http://www.nature.com/articles/nmeth.f.330

42. Tyler AD, Christianson S, Knox NC, Mabon P, Wolfe J, Van Domselaar G, et al. Comparison of Sample Preparation Methods Used for the Next-Generation Sequencing of Mycobacterium tuberculosis. Supply P, editor. PLoS One [Internet]. 2016 Feb 5;11(2):e0148676. Available from: https://dx.plos.org/10.1371/journal.pone.0148676

43. Gilchrist CA, Turner SD, Riley MF, Petri WA, Hewlett EL. Whole-Genome Sequencing in Outbreak Analysis. Clin Microbiol Rev [Internet]. 2015 Jul 15;28(3):541–63. Available from: https://cmr.asm.org/content/28/3/541

44. Niemann S, Köser CU, Gagneux S, Plinke C, Homolka S, Bignell H, et al. Genomic Diversity among Drug Sensitive and Multidrug Resistant Isolates of Mycobacterium tuberculosis with Identical DNA Fingerprints. Ahmed N, editor. PLoS One [Internet]. 2009 Oct 12;4(10):e7407. Available from: https://dx.plos.org/10.1371/journal.pone.0007407

45. Walker TM, Ip CL, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, et al. Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. Lancet Infect Dis [Internet]. 2013 Feb;13(2):137–46. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1473309912702773

46. Roetzer A, Diel R, Kohl TA, Rückert C, Nübel U, Blom J, et al. Whole Genome Sequencing versus Traditional Genotyping for Investigation of a Mycobacterium tuberculosis Outbreak: A Longitudinal Molecular Epidemiological Study. Neyrolles O, editor. PLoS Med [Internet]. 2013 Feb 12;10(2):e1001387. Available from: https://dx.plos.org/10.1371/journal.pmed.1001387

47. Satta G, Lipman M, Smith GP, Arnold C, Kon OM, McHugh TD. Mycobacterium tuberculosis and whole-genome sequencing: how close are we to unleashing its full potential? Clin Microbiol Infect [Internet]. 2018 Jun;24(6):604–9. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1198743X17306237

48. Coscolla M, Gagneux S. Consequences of genomic diversity in Mycobacterium tuberculosis. Semin Immunol [Internet]. 2014 Dec;26(6):431–44. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1044532314000967

49. Niemann S, Supply P. Diversity and Evolution of Mycobacterium tuberculosis: Moving

to Whole-Genome-Based Approaches. Cold Spring Harb Perspect Med [Internet]. 2014 Dec 1;4(12):a021188–a021188. Available from: http://perspectivesinmedicine.cshlp.org/lookup/doi/10.1101/cshperspect.a021188

50. Kontsevaya IS, Nikolayevsky V V., Balabanova YM. Molecular epidemiology of tuberculosis: Objectives, methods, and prospects. Mol Genet Microbiol Virol [Internet]. 2011 Mar 7;26(1):1–9. Available from: http://link.springer.com/10.3103/S0891416811010034

51. Shastry BS. SNPs: Impact on Gene Function and Phenotype. In 2009. p. 3–22. Available from: http://link.springer.com/10.1007/978-1-60327-411-1_1

52. Homolka S, Projahn M, Feuerriegel S, Ubben T, Diel R, Nübel U, et al. High Resolution Discrimination of Clinical Mycobacterium tuberculosis Complex Strains Based on Single Nucleotide Polymorphisms. Manganelli R, editor. PLoS One [Internet]. 2012 Jul 2;7(7):e39855. Available from: https://dx.plos.org/10.1371/journal.pone.0039855

53. Narayanan S, Desikan S. Genetic markers, genotyping methods & next generation sequencing in Mycobacterium tuberculosis. Indian J Med Res [Internet]. 2015;141(6):761. Available from: http://www.ijmr.org.in/text.asp?2015/141/6/761/160695

54. van Soolingen D, Hermans PW, de Haas PE, Soll DR, van Embden JD. Occurrence and stability of insertion sequences in Mycobacterium tuberculosis complex strains: evaluation of an insertion sequence-dependent DNA polymorphism as a tool in the epidemiology of tuberculosis. J Clin Microbiol [Internet]. 1991;29(11):2578–86. Available from: https://jcm.asm.org/content/29/11/2578

55. Thermo Fisher Scientific. LifeTech: Qubit dsDNA BR Assay Kit. Manual. 2015;1–10.

56. Clean-up NNGS. NGS clean-up and size selection User manual. 2014;(May).

57. QtfXna L. LifeTech: Qubit ™ dsDNA HS Assay Kits. Manual [Internet]. 2015;1–9. Available from: www.lifetechnologies.com/support

58. Fisher Scientific T. Ion Xpress™ Plus gDNA Fragment Library Preparation USER GUIDE. Fish Sci Thermo. 2016;(4471989):1–102.

59. Technologies A. Agilent High Sensitivity DNA Kit Guide Agilent High Sensitivity DNA Agilent High Sensitivity DNA Notices Manual Part Number Resarch Use Only Not for use in Diagnostic Procedures. Technology Licenses Restricted Rights Legend. 2013;

Available from: http://www.agilent.com/cs/library/usermanuals/Public/G2938-90321_SensitivityDNA_KG_EN.pdf

60. Andrews S. FastQC: a quality control tool for high throughput sequence data. 2010; Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

61. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–20.

62. García-Alcalde F, Okonechnikov K, Carbonell J, Cruz LM, Götz S, Tarazona S, et al. Qualimap: evaluating next-generation sequencing alignment data. Bioinformatics [Internet]. 2012 Oct 15;28(20):2678–9. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bts503

63. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly (Austin) [Internet]. 2012 Apr 27;6(2):80–92. Available from: http://www.tandfonline.com/doi/abs/10.4161/fly.19695

64. Perdigão J, Gomes P, Miranda A, Maltez F, Machado D, Silva C, et al. Using genomics to understand the origin and dispersion of multidrug and extensively drug resistant tuberculosis in Portugal. Sci Rep [Internet]. 2020 Dec 13;10(1):2600. Available from: http://www.nature.com/articles/s41598-020-59558-3

65. Institut Pasteur de la Guadeloupe. SITVIT2. 2018; Available from: http://www.pasteur-guadeloupe.fr:8081/SITVIT2/description.jsp)

66. Gouy M, Guindon S, Gascuel O. SeaView Version 4: A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building. Mol Biol Evol [Internet]. 2010 Feb 1;27(2):221–4. Available from: https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msp259

67. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res [Internet]. 2019 Jul 2;47(W1):W256–9. Available from: https://academic.oup.com/nar/article/47/W1/W256/5424068

68. Kircher M, Kelso J. High-throughput DNA sequencing - concepts and limitations. BioEssays [Internet]. 2010 May 18;32(6):524–36. Available from: http://doi.wiley.com/10.1002/bies.200900181

69. Liao P, Satten GA, Hu Y-J. PhredEM: a phred-score-informed genotype-calling approach for next-generation sequencing studies. Genet Epidemiol [Internet]. 2017 Jul;41(5):375–87. Available from: http://doi.wiley.com/10.1002/gepi.22048

70. McNerney R, Clark TG, Campino S, Rodrigues C, Dolinger D, Smith L, et al. Removing the bottleneck in whole genome sequencing of Mycobacterium tuberculosis for rapid drug resistance analysis: a call to action. Int J Infect Dis [Internet]. 2017 Mar;56:130–5. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1201971216316423

71. Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet [Internet]. 2011 Jun 18;12(6):443–51. Available from: http://www.nature.com/articles/nrg2986

72. Mahmoud M, Gobet N, Cruz-Dávalos DI, Mounier N, Dessimoz C, Sedlazeck FJ. Structural variant calling: the long and the short of it. Genome Biol [Internet]. 2019 Dec 20;20(1):246. Available from: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1828-7

73. Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. Nat Rev Genet [Internet]. 2011 May 1;12(5):363–76. Available from: http://www.nature.com/articles/nrg2958

74. Anyansi C, Keo A, Walker BJ, Straub TJ, Manson AL, Earl AM, et al. QuantTB – a method to classify mixed Mycobacterium tuberculosis infections within whole genome sequencing data. BMC Genomics [Internet]. 2020 Dec 28;21(1):80. Available from: https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-020-6486-3

75. Strom SP. Current practices and guidelines for clinical next-generation sequencing oncology testing. Cancer Biol Med [Internet]. 2016 Mar;13(1):3–11. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27144058

76. Kapopoulou A, Lew JM, Cole ST. The MycoBrowser portal: A comprehensive and manually annotated resource for mycobacterial genomes. Tuberculosis [Internet]. 2011 Jan;91(1):8–13. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1472979210001095

77. DeJesus MA, Gerrick ER, Xu W, Park SW, Long JE, Boutte CC, et al. Comprehensive Essentiality Analysis of the Mycobacterium tuberculosis Genome via Saturating Transposon Mutagenesis. Stallings CL, editor. MBio [Internet]. 2017 Mar 8;8(1).

Available from: https://mbio.asm.org/content/8/1/e02133-16

78. Lamichhane G, Zignol M, Blades NJ, Geiman DE, Dougherty A, Grosset J, et al. A postgenomic method for predicting essential genes at subsaturation levels of mutagenesis: Application to Mycobacterium tuberculosis. Proc Natl Acad Sci [Internet]. 2003 Jun 10;100(12):7213–8. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.1231432100

79. Das S, Duggal P, Roy R, Myneedu VP, Behera D, Prasad HK, et al. Identification of Hot and Cold spots in genome of Mycobacterium tuberculosis using Shewhart Control Charts. Sci Rep [Internet]. 2012 Dec 2;2(1):297. Available from: http://www.nature.com/articles/srep00297

80. Vilchèze C, Saranathan R, Weinrick B, Jacobs WR. Characterization of Large Deletion Mutants of Mycobacterium tuberculosis Selected for Isoniazid Resistance. Antimicrob Agents Chemother [Internet]. 2020 Jul 6;64(9). Available from: https://aac.asm.org/content/64/9/e00792-20

81. Sassetti CM, Rubin EJ. Genetic requirements for mycobacterial survival during infection. Proc Natl Acad Sci [Internet]. 2003 Oct 28;100(22):12989–94. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.2134250100

82. Mazandu GK, Mulder NJ. Function Prediction and Analysis of Mycobacterium tuberculosis Hypothetical Proteins. Int J Mol Sci [Internet]. 2012 Jun 13;13(6):7283–302. Available from: http://www.mdpi.com/1422-0067/13/6/7283

83. Fishbein S, van Wyk N, Warren RM, Sampson SL. Phylogeny to function: PE/PPE protein evolution and impact on M ycobacterium tuberculosis pathogenicity. Mol Microbiol [Internet]. 2015 Jun;96(5):901–16. Available from: http://doi.wiley.com/10.1111/mmi.12981

84. Cubillos-Ruiz A, Morales J, Zambrano M. Analysis of the genetic variation in Mycobacterium tuberculosis strains by multiple genome alignments. BMC Res Notes [Internet]. 2008;1(1):110. Available from: http://bmcresnotes.biomedcentral.com/articles/10.1186/1756-0500-1-110

85. Ates LS. New insights into the mycobacterial PE and PPE proteins provide a framework for future research. Mol Microbiol [Internet]. 2020 Jan 24;113(1):4–21. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/mmi.14409

86. Rausch T, Zichner T, Schlattl A, Stutz AM, Benes V, Korbel JO. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics [Internet]. 2012 Sep 15;28(18):i333–9. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bts378

87. Tattini L, D'Aurizio R, Magi A. Detection of Genomic Structural Variants from Next-Generation Sequencing Data. Front Bioeng Biotechnol [Internet]. 2015 Jun 25;3. Available from: http://journal.frontiersin.org/Article/10.3389/fbioe.2015.00092/abstract

88. Fleischmann RD, Alland D, Eisen JA, Carpenter L, White O, Peterson J, et al. Whole-Genome Comparison of Mycobacterium tuberculosis Clinical and Laboratory Strains. J Bacteriol [Internet]. 2002 Oct 1;184(19):5479–90. Available from: https://jb.asm.org/content/184/19/5479

89. Rocha EPC, Smith JM, Hurst LD, Holden MTG, Cooper JE, Smith NH, et al. Comparisons of dN/dS are time dependent for closely related bacterial genomes. J Theor Biol [Internet]. 2006 Mar;239(2):226–35. Available from: https://linkinghub.elsevier.com/retrieve/pii/S002251930500384X

90. Hershberg R, Lipatov M, Small PM, Sheffer H, Niemann S, Homolka S, et al. High Functional Diversity in Mycobacterium tuberculosis Driven by Genetic Drift and Human Demography. Blaser MJ, editor. PLoS Biol [Internet]. 2008 Dec 16;6(12):e311. Available from: https://dx.plos.org/10.1371/journal.pbio.0060311

91. Raman K, Chandra N. Mycobacterium tuberculosis interactome analysis unravels potential pathways to drug resistance. BMC Microbiol [Internet]. 2008;8(1):234. Available from: http://bmcmicrobiol.biomedcentral.com/articles/10.1186/1471-2180-8-234

92. Ley SD, de Vos M, Van Rie A, Warren RM. Deciphering Within-Host Microevolution of Mycobacterium tuberculosis through Whole-Genome Sequencing: the Phenotypic Impact and Way Forward. Microbiol Mol Biol Rev [Internet]. 2019 Mar 27;83(2). Available from: https://mmbr.asm.org/content/83/2/e00062-18

93. Griffin JE, Gawronski JD, DeJesus MA, Ioerger TR, Akerley BJ, Sassetti CM. High-Resolution Phenotypic Profiling Defines Genes Essential for Mycobacterial Growth and Cholesterol Catabolism. Ramakrishnan L, editor. PLoS Pathog [Internet]. 2011 Sep 29;7(9):e1002251. Available from: https://dx.plos.org/10.1371/journal.ppat.1002251

94. Sassetti CM, Boyd DH, Rubin EJ. Genes required for mycobacterial growand th defined by high density mutagenesis. Mol Microbiol [Internet]. 2003 Mar 25;48(1):77–84. Available from: http://doi.wiley.com/10.1046/j.1365-2958.2003.03425.x

95. Chavadi SS, Stirrett KL, Edupuganti UR, Vergnolle O, Sadhanandan G, Marchiano E, et al. Mutational and Phylogenetic Analyses of the Mycobacterial mbt Gene Cluster. J Bacteriol [Internet]. 2011 Nov 1;193(21):5905–13. Available from: https://jb.asm.org/content/193/21/5905

96. Rindi L, Lari N, Garzelli C. Genes of Mycobacterium tuberculosis H37Rv downregulated in the attenuated strain H37Ra are restricted to M. tuberculosis complex species. New Microbiol [Internet]. 2001 Jul;24(3):289–94. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11497087

97. Vilchèze C, Hartman T, Weinrick B, Jain P, Weisbrod TR, Leung LW, et al. Enhanced respiration prevents drug tolerance and drug resistance in Mycobacterium tuberculosis. Proc Natl Acad Sci [Internet]. 2017 Apr 25;114(17):4495–500. Available from: http://www.pnas.org/lookup/doi/10.1073/pnas.1704376114

98. Veyrier F, Pletzer D, Turenne C, Behr MA. Phylogenetic detection of horizontal gene transfer during the step-wise genesis of Mycobacterium tuberculosis. BMC Evol Biol [Internet]. 2009;9(1):196. Available from: http://bmcevolbiol.biomedcentral.com/articles/10.1186/1471-2148-9-196

99. Braibant M, Gilot P, Content J. The ATP binding cassette (ABC) transport systems of Mycobacterium tuberculosis. FEMS Microbiol Rev [Internet]. 2000 Oct;24(4):449–67. Available from: https://academic.oup.com/femsre/article-lookup/doi/10.1111/j.1574-6976.2000.tb00550.x

100. Fu LM, Fu-Liu CS. The gene expression data of Mycobacterium tuberculosis based on Affymetrix gene chips provide insight into regulatory and hypothetical genes. BMC Microbiol [Internet]. 2007;7(1):37. Available from: http://bmcmicrobiol.biomedcentral.com/articles/10.1186/1471-2180-7-37

101. Gao L-Y, Groger R, Cox JS, Beverley SM, Lawson EH, Brown EJ. Transposon Mutagenesis of Mycobacterium marinum Identifies a Locus Linking Pigmentation and Intracellular Survival. Infect Immun [Internet]. 2003 Feb;71(2):922–9. Available from: https://iai.asm.org/content/71/2/922

102. Mawuenyega KG, Forst C V., Dobos KM, Belisle JT, Chen J, Bradbury EM, et al. Mycobacterium tuberculosis Functional Network Analysis by Global Subcellular Protein Profiling. Mol Biol Cell [Internet]. 2005 Jan;16(1):396–404. Available from: https://www.molbiolcell.org/doi/10.1091/mbc.e04-04-0329

103. Serra-Vidal MM, Latorre I, Franken KLCM, DÃ-az J, de Souza-GalvÃ£o ML, Casas I, et al. Immunogenicity of 60 novel latency-related antigens of Mycobacterium tuberculosis. Front Microbiol [Internet]. 2014 Oct 8;5. Available from: http://journal.frontiersin.org/article/10.3389/fmicb.2014.00517/abstract

104. Miller BH, Shinnick TM. Evaluation of Mycobacterium tuberculosisGenes Involved in Resistance to Killing by Human Macrophages. Kaufmann SHE, editor. Infect Immun [Internet]. 2000 Jan 1;68(1):387–90. Available from: https://iai.asm.org/content/68/1/387

105. You X, Li R, Wan K, Liu L, Xie X, Zhao L, et al. Evaluation of Rv0220, Rv2958c, Rv2994 and Rv3347c of Mycobacterium tuberculosis for serodiagnosis of tuberculosis. Microb Biotechnol [Internet]. 2017 May;10(3):604–11. Available from: http://doi.wiley.com/10.1111/1751-7915.12697

106. Danelishvili L, Wu M, Young LS, Bermudez LE. Genomic Approach to Identifying the Putative Target of and Mechanisms of Resistance to Mefloquine in Mycobacteria. Antimicrob Agents Chemother [Internet]. 2005 Sep;49(9):3707–14. Available from: https://aac.asm.org/content/49/9/3707

107. Groenheit R, Ghebremichael S, Pennhag A, Jonsson J, Hoffner S, Couvin D, et al. Mycobacterium tuberculosis Strains Potentially Involved in the TB Epidemic in Sweden a Century Ago. Supply P, editor. PLoS One [Internet]. 2012 Oct 8;7(10):e46848. Available from: https://dx.plos.org/10.1371/journal.pone.0046848

108. Demay C, Liens B, Burguière T, Hill V, Couvin D, Millet J, et al. SITVITWEB – A publicly available international multimarker database for studying Mycobacterium tuberculosis genetic diversity and molecular epidemiology. Infect Genet Evol [Internet]. 2012 Jun;12(4):755–66. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1567134812000317

109. Caws M, Thwaites G, Dunstan S, Hawn TR, Thi Ngoc Lan N, Thuong NTT, et al. The Influence of Host and Bacterial Genotype on the Development of Disseminated Disease with Mycobacterium tuberculosis. Bishai W, editor. PLoS Pathog [Internet]. 2008 Mar

28;4(3):e1000034. Available from: https://dx.plos.org/10.1371/journal.ppat.1000034

110. Ranka R, Pole I, Markovska S, Ozere I, Riekstina V, Norvaisa I. MOLECULAR EPIDEMIOLOGY OF TUBERCULOSIS IN LATVIA. Russ J Infect Immun [Internet]. 2019 Jan 16;8(4):578. Available from: https://www.iimmun.ru/iimm/article/view/1036

111. Pole I, Trofimova J, Norvaisa I, Supply P, Skenders G, Nodieva A, et al. Analysis of Mycobacterium tuberculosis genetic lineages circulating in Riga and Riga region, Latvia, isolated between 2008 and 2012. Infect Genet Evol [Internet]. 2020 Mar;78:104126. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1567134819303521

112. Kato-Maeda M, Gagneux S, Flores LL, Kim EY, Small PM, Desmond EP, et al. Strain classification of Mycobacterium tuberculosis: congruence between large sequence polymorphisms and spoligotypes. Int J Tuberc Lung Dis [Internet]. 2011 Jan;15(1):131–3. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21276309

113. Feuerriegel S, Koser CU, Niemann S. Phylogenetic polymorphisms in antibiotic resistance genes of the Mycobacterium tuberculosis complex. J Antimicrob Chemother [Internet]. 2014 May 1;69(5):1205–10. Available from: https://academic.oup.com/jac/article-lookup/doi/10.1093/jac/dkt535

114. Feuerriegel S, Oberhauser B, George A, Dafae F, Richter E, Rüsch-Gerdes S, et al. Sequence analysis for detection of first-line drug resistance in Mycobacterium tuberculosis strains from a high-incidence setting. BMC Microbiol [Internet]. 2012;12(1):90. Available from: http://bmcmicrobiol.biomedcentral.com/articles/10.1186/1471-2180-12-90

115. Cancino-Muñoz I, Moreno-Molina M, Furió V, Goig GA, Torres-Puente M, Chiner-Oms Á, et al. Cryptic Resistance Mutations Associated With Misdiagnoses of Multidrug-Resistant Tuberculosis. J Infect Dis [Internet]. 2019 Jun 19;220(2):316–20. Available from: https://academic.oup.com/jid/article/220/2/316/5381710

116. Lau RWT, Ho P-L, Kao RYT, Yew W-W, Lau TCK, Cheng VCC, et al. Molecular Characterization of Fluoroquinolone Resistance in Mycobacterium tuberculosis : Functional Analysis of gyrA Mutation at Position 74. Antimicrob Agents Chemother [Internet]. 2011 Feb;55(2):608–14. Available from: https://aac.asm.org/content/55/2/608

117. Sreevatsan S, Pan X, Stockbauer KE, Connell ND, Kreiswirth BN, Whittam TS, et al.

Restricted structural gene polymorphism in the Mycobacterium tuberculosis complex indicates evolutionarily recent global dissemination. Proc Natl Acad Sci [Internet]. 1997 Sep 2;94(18):9869–74. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.94.18.9869

118. Tadokera R, Bekker L-G, Kreiswirth BN, Mathema B, Middelkoop K. TB transmission is associated with prolonged stay in a low socio-economic, high burdened TB and HIV community in Cape Town, South Africa. BMC Infect Dis [Internet]. 2020 Dec 10;20(1):120. Available from: https://bmcinfectdis.biomedcentral.com/articles/10.1186/s12879-020-4828-z

119. Xu G, Mao X, Wang J, Pan H. Clustering and recent transmission of Mycobacterium tuberculosis in a Chinese population. Infect Drug Resist [Internet]. 2018 Mar;Volume 11:323–30. Available from: https://www.dovepress.com/clustering-and-recent-transmission-of-mycobacterium-tuberculosis-in-a--peer-reviewed-article-IDR

120. Teeter LD, Ha NP, Ma X, Wenger J, Cronin WA, Musser JM, et al. Evaluation of large genotypic Mycobacterium tuberculosis clusters: contributions from remote and recent transmission. Tuberculosis [Internet]. 2013 Dec;93:S38–46. Available from: https://linkinghub.elsevier.com/retrieve/pii/S147297921370009X

121. Hatherell H-A, Colijn C, Stagg HR, Jackson C, Winter JR, Abubakar I. Interpreting whole genome sequencing for investigating tuberculosis transmission: a systematic review. BMC Med [Internet]. 2016 Dec 23;14(1):21. Available from: http://www.biomedcentral.com/1741-7015/14/21

122. Schürch AC, Kremer K, Kiers A, Daviena O, Boeree MJ, Siezen RJ, et al. The tempo and mode of molecular evolution of Mycobacterium tuberculosis at patient-to-patient scale. Infect Genet Evol [Internet]. 2010 Jan;10(1):108–14. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1567134809002172

123. Schurch AC, Kremer K, Daviena O, Kiers A, Boeree MJ, Siezen RJ, et al. High-Resolution Typing by Integration of Genome Sequencing Data in a Large Tuberculosis Cluster. J Clin Microbiol [Internet]. 2010 Sep 1;48(9):3403–6. Available from: https://jcm.asm.org/content/48/9/3403

124. Bryant JM, Schürch AC, van Deutekom H, Harris SR, de Beer JL, de Jager V, et al. Inferring patient to patient transmission of Mycobacterium tuberculosisfrom whole genome sequencing data. BMC Infect Dis [Internet]. 2013 Dec 27;13(1):110. Available

from: http://bmcinfectdis.biomedcentral.com/articles/10.1186/1471-2334-13-110

125. Mbuthia GW, Olungah CO, Ondicho TG. Health-seeking pathway and factors leading to delays in tuberculosis diagnosis in West Pokot County, Kenya: A grounded theory study. Fair E, editor. PLoS One [Internet]. 2018 Nov 28;13(11):e0207995. Available from: https://dx.plos.org/10.1371/journal.pone.0207995

126. Lambert M-L, Hasker E, Deun A Van, Roberfroid D, Boelaert M, Van der Stuyft P. Recurrence in tuberculosis: relapse or reinfection? Lancet Infect Dis [Internet]. 2003 May;3(5):282–7. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1473309903006078

127. Bryant JM, Harris SR, Parkhill J, Dawson R, Diacon AH, van Helden P, et al. Whole-genome sequencing to establish relapse or re-infection with Mycobacterium tuberculosis: a retrospective observational study. Lancet Respir Med [Internet]. 2013 Dec;1(10):786–92. Available from: https://linkinghub.elsevier.com/retrieve/pii/S2213260013702315

128. Guerra-Assunção J, Crampin A, Houben R, Mzembe T, Mallard K, Coll F, et al. Large-scale whole genome sequencing of M. tuberculosis provides insights into transmission in a high prevalence area. Elife [Internet]. 2015 Mar 3;4. Available from: https://elifesciences.org/articles/05166

129. Guerra-Assunção JA, Houben RMGJ, Crampin AC, Mzembe T, Mallard K, Coll F, et al. Recurrence due to Relapse or Reinfection With Mycobacterium tuberculosis : A Whole-Genome Sequencing Approach in a Large, Population-Based Cohort With a High HIV Infection Prevalence and Active Follow-up. J Infect Dis [Internet]. 2015 Apr 1;211(7):1154–63. Available from: https://academic.oup.com/jid/article-lookup/doi/10.1093/infdis/jiu574