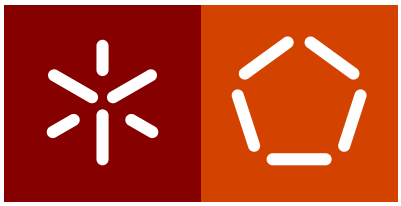
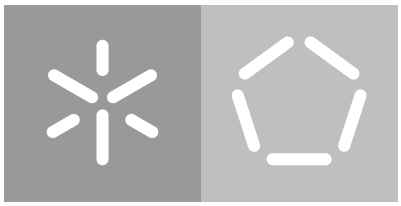


University of Minho
School of Engineering

Daniela Sofia Gaspar Pereira

A comprehensive phylogenetic analysis of *Mycobacterium tuberculosis* protein-coding genes: insights into evolution and virulence

October 2018



University of Minho
School of Engineering

Daniela Sofia Gaspar Pereira

A comprehensive phylogenetic analysis of *Mycobacterium tuberculosis* protein-coding genes: insights into evolution and virulence

Master Degree in Bioinformatics

Dissertation supervised by
Teresa Sofia Teixeira Rito, PhD
Pedro Alexandre Dias Soares, PhD

October 2018

DECLARAÇÃO

Nome: Daniela Sofia Gaspar Pereira

Endereço electrónico: danielapereira_cambra@hotmail.com

Telefone: 939423474

Cartão de Cidadão:14325707

Título da dissertação: A comprehensive phylogenetic analysis of Mycobacterium tuberculosis protein-coding genes: insights into evolution and virulence

Orientador(es):

Teresa Sofia Teixeira Rito, PhD

Pedro Alexandre Dias Soares, PhD

Ano de conclusão: 2018

Mestrado em Bioinformática

Universidade do Minho, 29/10/2018

Assinatura: Daniela Pereira

ACKNOWLEDGEMENTS

First of all, I would like to thank my supervisers, Professor Teresa Sofia Teixeira Rito and Professor Pedro Alexandre Dias Soares, for all the advice and technical guidance, highlighting also their sympathy and constant concern with the development of this work.

I can not fail to express my thanks to Daniel Almeida and Raquel Silva for the availability they have always had to help me with whatever it takes.

To my family, especially my parents, my brother and my grandparents, a huge thank you for always believing in me, what I do and all the teachings of life. I hope that this stage, which I now conclude, can somehow reciprocate and compensate for all the affection, support and dedication that, constantly, they offer me. I dedicate all this work to them.

A thank you to my boyfriend and friends, for the support in the difficult times. Although they were not directly involved, their motivation was imperative.

Finally, I thank all those who have helped me who, although not mentioned here, I express to them all my gratitude.

ABSTRACT

Understanding the evolutionary relationships between organisms is a complex issue that has gained importance not only in evolution but also in clinical and biological inferences, only possible with technological advances that require new analytical tools.

In this work, the main focus is to study the genome of *Mycobacterium tuberculosis*, the causal agent of tuberculosis, establishing a detailed evolutionary framework. The secondary objective, regardless of the focus on the evolution of Mtb, is that the pipeline created can be applied to any other organism.

This dissertation presents some basic facts about tuberculosis, an infectious bacterial disease caused by the *Mycobacterium tuberculosis* complex, its constitution, evolution, pathology, drug resistance and genetic variation. Our objectives were contextualized considering the most up-to-date tools of alignment and phylogenetics, an area in constant progress due to the growing needs of bioinformatic tools in the area of genomics and evolution.

Taxonomic and genetic data were compiled from all organisms with complete genomes in NCBI. This database was subsequently cured by eliminating redundant genomes, i.e., containing only one representative element of each species with the complete proteome. A search of each *Mycobacterium tuberculosis* protein in this local database using BLAST allowed the detection of probable homologs in a large number of taxonomically informative organisms. The search results were limited to two hundred homologues, which were aligned using MUSCLE. Phylogenetic trees, based on maximum likelihood were constructed for the approximately four thousand *Mycobacterium tuberculosis* proteins. The phylogenetic relationship and monophyly of *Mycobacterium tuberculosis* with the remaining bacteria of the same genus (*Mycobacterium*) and the same family (*Corynebacteriaceae*) were studied to understand possible processes of acquisition of genes by horizontal transference.

Finally, positive selection processes were studied by searching for excess or deficit of non-synonymous mutations in relation to the synonymous (Ka / Ks) using the CODEML software, in order to identify branches with an accelerated evolution in the establishment of the pathogenic species *Mycobacterium tuberculosis*. These genes may form the basis of the physiological and biochemical characteristics that make this bacterium pathogenic to humans.

Keywords: evolution of *Mycobacterium tuberculosis*, bioinformatics, phylogenetics, Horizontal gene transfer(HGT)

RESUMO

Compreender as relações evolutivas entre os organismos é um questão complexa que ganhou cada vez mais relevo no âmbito não apenas da evolução mas para inferências clínicas e biológicas, só possíveis com os avanços tecnológicos, que exigem novas ferramentas analíticas.

Neste trabalho, o foco principal é estudar o genoma da *Mycobacterium tuberculosis*, o agente causal da tuberculose, estabelecendo-se um quadro evolutivo detalhado. O objetivo secundário, independentemente do foco na evolução da Mtb, é que a pipeline criada possa ser aplicada a qualquer outro organismo.

Esta dissertação apresenta alguns fatos básicos sobre a tuberculose, uma doença infeciosa bacteriana causada pelo complexo *Mycobacterium tuberculosis*, sobre a sua constituição, evolução, patologia, resistência aos medicamentos e variação genética. Os nossos objetivos foram contextualizados considerando as mais atualizadas ferramentas de alinhamento e de filogenética, uma área em constante progresso devido às crescentes necessidades de ferramentas bioinformáticas na área da genómica e evolução.

Foram compilados dados taxonómicos e genéticos de todos os organismos com genomas completos no NCBI. Esta base de dados foi posteriormente curada eliminando-se genomas redundantes, ou seja, contendo apenas um elemento representativo de cada espécie com o proteoma completo. Uma busca de cada proteína da *Mycobacterium tuberculosis* nesta base de dados local utilizando o BLAST permitiu a deteção de prováveis homólogos num vasto número de organismos taxonomicamente informativos. Os resultados da busca foram limitados a duzentos homólogos, que foram alinhados recorrendo ao MUSCLE. Árvores filogenéticas, baseadas em máxima verossimilhança foram construídas para as cerca de quatro mil proteínas da *Mycobacterium tuberculosis*. A relação filogenética e monofilia da *Mycobacterium tuberculosis* com as restantes bactérias do mesmo género (*Mycobacterium*) e da mesma família (*Corynebacteriaceae*) foram estudadas para compreender possíveis processos de aquisição de genes por transferência horizontal.

Finalmente, processos de seleção positiva foram estudados através da procura de excesso ou défice de mutações não sinónimas em relação às sinónimas (Ka/Ks) usando o software CODEML, de modo a identificar ramos com uma evolução acelerada no estabelecimento da espécie patogénica *Mycobacterium tuberculosis*. Esses genes podem estar na base das características fisiológicas e bioquímicas que tornam esta bactéria patogénica para o Homem.

Palavras Chave: evolução de *Mycobacterium tuberculosis*, bioinformática, filogenética, Transferência Horizontal de Genes.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

BCG	Bacillus Calmette Guerin
BI	Bayesian Inference
CDSs	Coding sequences
EMB	Ethambutol
GC	Guanine + Cytosine
HTUs	Hypothetical Taxonomic Units
IFN	Interferon
INH	Isoniazid
Ka	Non-synonyms
Ks	Synonyms
LSPs	Large Sequence Polymorphisms
MCMC	Markov chain Monte Carlo algorithm
MDR	Multidrug-resistance

ML	Maximum Likelihood
MLST	Multilocus Sequence Typing
MP	Maximum Parsimony
Mtbc	*Mycobacterium tuberculosis* complex
Mtb	*Mycobacterium tuberculosis*
NJ	Neighbor Joining
OTUs	Operacional Taxonomic Units
PCR	Polymerase Chain Reaction
PZA	Pyrazinamide
RDs	Regions of Difference
RFLP	Restriction Fragment Length Polymorphism
RIF	Rifampin
SNPs	Single Nucleotide Polymorphisms
TB	*Tuberculosis*
TNF	Tumor necrosis factor
UPGMA	Unweighted Pair Group Method with Arithmetic Mean
WHO/IUATLD	World Health Organization/International Union Against TB and Lung Diseases
XDR	Extensively drug resistance

1

INTRODUCTION

Tuberculosis (TB) is caused by any of the organisms in the *Mycobacterium tuberculosis* complex (Mtbc), of which *M. tuberculosis* (Mtb), the causative agente of TB in humans, is the most frequent [5]. Recent studies suggest a scenario of co-evolution between man and Mtb during which mycobacteria developed strategies to overcome the immune response and persist in the human host [7].

Archaeological findings show that TB was already present in ancient times [8, 9]. It would later take epidemic proportions in Europe during the Middle Ages. Currently, TB is considered to be one of the classic infectious diseases, but recognition of the disease's clinical manifestations has evolved considerably over the last two millennia.

The term "tuberculosis" was first applied at the beginning of the 19th century, having been derived from tubers characteristic of the lesions caused by the disease. The unique biological properties of this organism allow a long latency phase between the time of infection and the development of symptoms. Latent TB can persist for decades before it causes disease, or even for the entire life of an infected person, without ever causing the disease to be clinically evident. TB occurs in a more aggravated way among disadvantaged populations, such as the malnourished, homeless, and those living in overcrowded and precarious housing. There is also an increase in the occurrence of TB in HIV-positive individuals [10].

The fight against TB, through the development of vaccines, drugs, and new methods of diagnosis, was one of the main objectives of biomedical research in the 19th and 20th centuries. The first description of Mtb bacillus as the cause of TB was by Robert Koch in 1882. He also developed microscopic and culture methods for the detection of TB bacilli, which are still widely used today [11]. Calmette and Guérin developed an effective vaccine for TB in the early 20th century. In the 1940s the first antibiotics against TB were discovered [12] and all these events, linked to changes in house and hygienic conditions of the metropolitan areas in the developed World, as well as strong governmental control programs led to the decrease of TB numbers.

In recent years, the TB situation worsened fueled by the increased number of cases due to co-infection with HIV-AIDS and the emergence of drug-resistant Mtb strains (clinical Mtb strains that are resistant to the first line antitubercular compounds). The latest estimates point to 9.6 million new cases and 1.6 million TB deaths in 2016, the majority of them occurring in problematic areas of Africa and Asia. TB is the ninth leading cause of death worldwide, and in the last five years (2012 – 2016) it has been the main cause of death of a single infectious agent, above HIV/AIDS (Figure 1) [1].

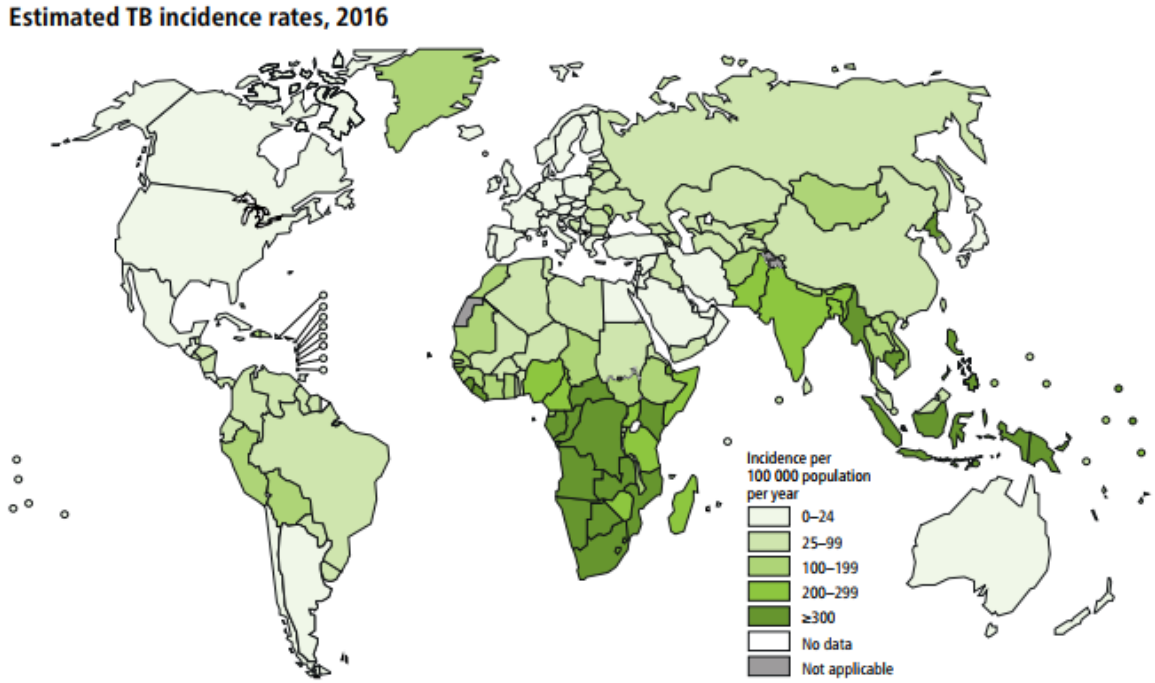


Figure 1: Annual number of tuberculosis incidents in relation to population size (the incidence rate). This varied widely among countries in 2016, from less than 10 per 100,000 inhabitants in most of the high-income countries to 150 - 300 in most of the 30 countries to 150 - 300 in most of the 30 countries with high TB burden and over 500 in some countries, including the Democratic People's Republic of Korea, Lesotho, Mozambique, the Philippines and South Africa. Retrieved directly from [1].

Resistance to anti-TB drugs is a major challenge in the overall control of TB. Mutations in the wild-type Mtb strain, causing the natural occurrence of resistance (primary resistance), have become clinically significant under selective pressure from misuse of anti-TB drugs. Subsequently, through transmission of resistant microorganisms, such mutations have merged into the TB epidemic scenario and are now being transmitted between individuals (secondary resistance). The proportions and trends of the phenomenon of resistance to anti-TB drugs have been identified and monitored through the Global Drug Resistance Survey organized by the World Health Organization / International Union Against TB

and Lung Diseases (WHO / IUATLD) [13]. Poor management of TB cases is consistently associated with drug resistance.

1.1 THE *mycobacterium tuberculosis* COMPLEX

Bacteria of the genus *Mycobacterium* belong to the family Mycobacteriaceae, family that is integrated in the order Actinomycetales, class Actinobacteria, phylum Actinobacteria, and Bacteria domain [14]. The genus *Mycobacterium* encompasses more than 170 species, some of which are grouped into complexes, such as Mtbc [15]. The Mtbc consists of several species of importance to human and animal health, namely: *M. canettii; M. africanum; M. pinnipedii; M. microti; M. caprae; M. bovis;* and *M. tuberculosis*. Mtbc members present tropism to a given host or group of hosts. For example, *M. tuberculosis, M. africanum* and *M. canettii* are associated with infection in humans; *M. bovis* in cattle, *M. caprae* in goats, *M. pinnipedii* in seals and *M. microti* in rodents [16]. Despite the high homology between their genomes, Mtbc species present phenotypic differences that are thought to result from the adaptation of the microorganisms to different hosts, evidencing differences in virulence (Table 2) [16].

Table 1: Members of the *Mycobacterium tuberculosis* complex. (Adapted from [5])

		Virulence			
Mtbc Member	**Host**	**Mouse**	**Guinea pig**	**Bunny**	**Single attribute**
M. canettii	Human	+	+	-	Oldest member phylogenetically
M. tuberculosis	Human	+	+	-	Predominant cause of human tuberculosis
M. africanum WA1	Human	+	+	-	Rarely isolated
M. africanum WA2	Human	+	+	-	Phenotypically heterogenic
M. orygis	Oryx (Bovidae)	?	?	?	Etiologic agent of tuberculosis in animals and humans in Africa and South Asia
M. mungi	Mungos mungo	?	?	?	Non-respiratory transmission
M. pinnipeii	Pinipedes	+	+	+	Closely related to M. microti
M. microti	Shrew	-	-	-	Attenuated, it has already been used as a vaccine against tuberculosis
Bacilo de Dassie	Petromus typicus	-	-	-	More attenuated than M. microti
M.caprae	Goat	+	+	+	Only described in Europe
M. bovis	Cow	+	+	+	Dynamic pathogen with wild reservoirs
M.bovis BCG	None	-	-	-	M. bovis laboratorially attenuated, used as a vaccine against tuberculosis

The evolutionary processes of members of the Mtbc complex have been characterized by successive and irreversible eliminations of genome regions (dubbed Regions of Difference

(RDs)). Mtb strains can be subdivided into ancestral and recent, based on the elimination (or presence) of a specific region designated TbD1. The successive elimination of these chromosomal regions made it possible to distinguish the ecotypes from the complex and define their phylogeny. The evolutionary lineage represented by *M. africanum, M. microti*, and *M. bovis* is characterized by the elimination of RD9 [16].

1.1.1 *Morphological and structural characteristics*

Bacteria of this genus present the morphology of rods ($0.2 - 0.6um$ wide $\times$ $1,0 - 10um$ in length) rectilinear or slightly curved, immobile, encapsulated and non-sporulated. The morphology of the colonies in solid culture medium is variable, being able to present yellow pigmentation and rough aspect (Figure 2) [14].

Figure 2: Photograph obtained by electron microscopy of Mtb, the etiological agent of TB [CDC/ Elizabeth "Libby" White (PHIL #8433)].

Mycobacteria have a thick cell wall, are rich in lipids and mycolic acids, presenting a cytoplasmic membrane to which a layer of peptidoglycan that is covalently bound to arabinogalactan overlaps. Mycolic acids are produced by all mycobacteria and give the cells the property of alcohol-acid resistance, being commonly termed alcohol-acid resistant bacilli. They also contribute to the maintenance of rigid cellular structure and low permeability, allowing intrinsic resistance to hydrophilic compounds [17]. Mycobacteria are classified as Gram-positive, although Gram staining is not very informative [14].

Slow growth rate is amongst one of the remarkable features of Mtb. TB is characterized by two distinct phases an acute phase where the bacteria are actively growing and a persistent phase where the bacteria are in a slow growing or non-growing state. As a human pathogen, Mtb has an optimum growth at 37 degrees Celsius and generation time of 24 hours [18].

1.1.2 *Genome of Mycobacterium tuberculosis complex*

The specific genomic diversity for strains in Mtbc is an important factor in pathogenesis that may affect virulence, transmissibility, host response and drug resistance [19].

Bacteria belonging to the genus Mycobacterium are classified according to their morphological characteristics, cell envelope composition and high guanine + cytosine content (GC content) (between 60% and 70%) [20].

Mycobacterium tuberculosis

Mtb is a well-studied organism, with currently 141 complete genomes available at ncbi (www.ncbi.nlm.nih.gov/genome/genomes/166).

Since its isolation in 1905, the H37Rv strain of Mtb has found extensive, worldwide application in biomedical research because it has retained full virulence in animal models of TB, unlike some clinical isolates; it is also susceptible to drugs and amenable to genetic manipulation. Mtb was first sequenced in 1998 by Cole and colleagues, its genome consists of about 4.4 million base pairs and in approximately 4,000 genes, having a very high GC content (65%) which is reflected in the tendentious amino-acid content of the proteins [21].

Mycobacterium bovis and Mycobacterium bovis BCG

M. bovis was differentiated from Mtb by Theobald Smith in 1896, for small but constant cultural differences [22].

The *M. bovis* genome sequence is 4,345,492 bp in length, arranged on a single circular chromosome with an average GC content of 65.63%. The genome contains 3,952 genes encoding proteins. Surprisingly, the genome is >99.95% identical at the nucleotide level to Mtb. Prior to the availability of the M. bovis genome sequence, the comparative genomics

of Mtbc were performed using hybridization-based methods, exploring this high degree of sequential identity [23]. This revealed 11 deletions of the *M. bovis* genome, ranging in size from 1 to 12.7 kb.

Surprisingly, the sequence contains only one *M. bovis* locus, termed TbD1, which is absent in most existing Mtb strains. Therefore, deletion has been the dominant mechanism in the formation of the M. bovis genome.

There is also a non-virulent *M. bovis* strain called Bacillus Calmette Guerin (BCG), which originates from a virulent strain of *M. bovis*, this strain is important as it was the strain used for vaccination against TB. BCG is a well-studied vaccine [24].

Mycobacterium africanum

M. africanum was first described in 1968 in a Senegalese patient [25], after which it was found almost exclusively in West Africa.

In general, the genome of *M. africanum* is not very different from other members of Mtbc with a content typical of %GC (65.6%) and genome size (4,389,314 bp) from the usual values for *M. bovis* and Mtb. This is also colinear with those of *M. bovis* and Mtb and shares the majority of coding sequences (CDSs)[26].

The *M. africanum* genome is, as expected, highly homologous to that of other members of Mtbc, but contains a single sequence, RD9, which was lost independently during the evolution of the complex [27].

Mycobacterium canettii

M. canettii, a rare variant of Mtbc was first isolated in a Somali-born patient in 1969 by Canetti [28].

Although it shares identical 16S rRNA sequences with the other members of Mtbc, *M. canettii* differs in many respects, including polymorphisms in certain maintenance genes, IS1081 copy number, colony morphology, and cell wall lipid content. *M. canettii* conserves the RD, RvD and TbD1 regions in the genome, presenting 26 unique spacer sequences in the DR region (14) that are not present in any other member of Mtbc [16].

Mycobacterium microti

M. microti is the causative agent of TB in rats, wood mice and shrews and can also cause disease in a limited number of other mammalian species. It was first described by Wells in 1946 from rats (*Microtus agrestis*) in Britain [29]. In humans, it was reported for the first time in 1998 in immunocompromised patients although the transmission of human infection by *M. microti* seems to be rare [30].

Based on biochemical properties, it is difficult to distinguish this bacterium from Mtb, *M. africanum* or *M. bovis*, but *M. microti* strains differ in their IS6110 and spoligotypes profiles, distinct from other strains of the Mtbc. In addition, their chromosomal information, such as RD1mic and MiD1, have been reported as characteristic of *M. microti* strains [16].

Mycobacterium pinnipedii

M. pinnipedii was reported for the first time in 1993, when seals captured on the Argentine coast presented isolates that had a characteristic pattern in their IS6110 [31] profiles. Transmission of *M. pinnipedii* to humans has been reported in individuals who are in close contact with marine mammals [32].

The isolates of *M. pinnipedii* present a distinct pattern of spoligotypes when compared to other members of the Mtbc [31].

Mycobacterium caprae

M. caprae was first isolated in goats in Spain [Mycobacterium tuberculosis subsp. caprae subsp. nov.: a taxonomic study of a new member of the Mycobacterium tuberculosis complex isolated from goats in Spain]. It has also been isolated in humans [33].

Based on biochemical tests, the results are similar to those of *M. bovis* and *M. bovis BCG*. By spoligotyping, *M. caprae* species form a homogeneous cluster easily recognizable by the absence of spacers 1,3-16, 30-33 and 39-43. The lack of spacers 39-43 has also been described in *M. bovis* and *M. microti* [34].

1.2 EVOLUTION OF MTBC

A growing body of evidence suggests that Mtbc members evolved from a common ancestor, diversifying through successive deletions/insertions of DNA, resulting in the speciation and differences in pathogenesis currently observed. Through the genetic analysis of the members of this complex it was possible to determine the existence of 14 RDs. The analysis of these RDs together with single nucleotide polymorphisms allowed the establishment of phylogenetic relationships among the members of this group (Figure 3) [16]. Additionally, these regions are present in the laboratory reference strain of Mtb H37Rv and are absent in the attenuated strain used as *M. bovis* BCG vaccine, highlighting the chromosomal location of genes associated with virulence.

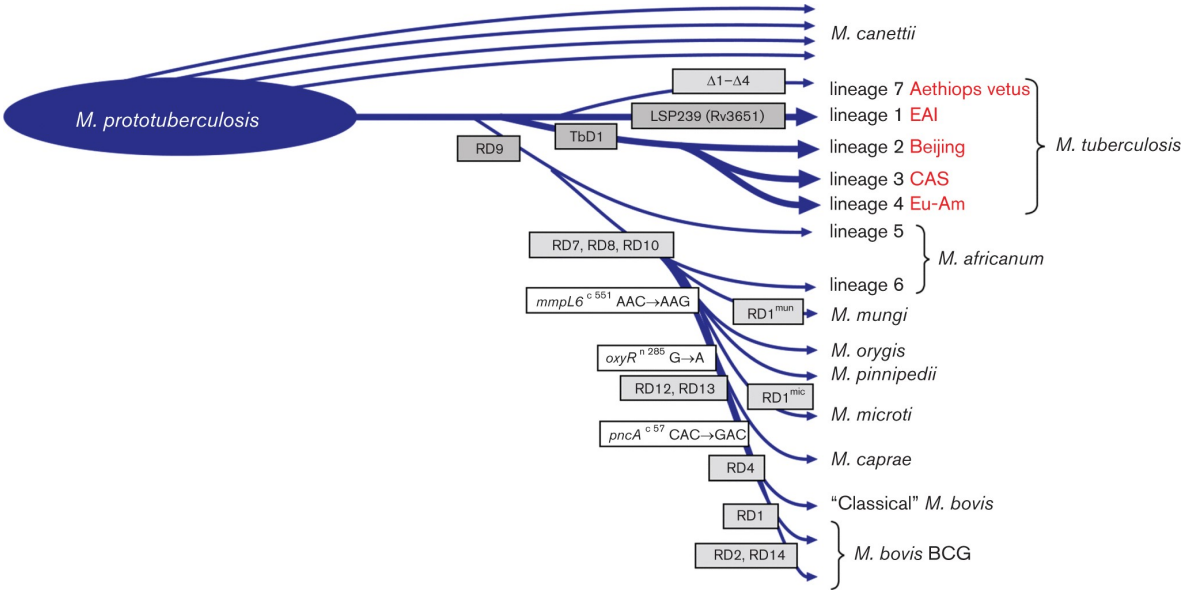


Figure 3: Scheme of the proposed evolutionary pathway of the tubercle bacilli illustrating successive loss of DNA in certain lineages (gray boxes). The scheme is based on the presence or absence of conserved deleted regions and on sequence polymorphisms in five selected genes. Note that the distances between certain branches may not correspond to actual phylogenetic differences calculated by other methods. Retrieved directly from [2].

Phylogenetic studies on Mtbc species adapted to humans (i.e. Mtb and *M. africanum*) have been carried out to establish 7 main lines (1 Indo-Oceanic, 2, East-Asian including Beijing, 3 East-African-Indian, 4 Euro- American, 5 West Africa or *M. africanum I*, 6 West Africa or *M. africanum II*, 7 Aethiops vetus). Lineages 1, 5 and 6 are considered "ancient", and 2 to 4 "modern". A novel phylogenetic lineage of Mtbc which appears to be intermediate between the ancient and modern has recently been described in Ethiopia and in the Horn of Africa [35, 2] to be associated with specific human populations (Figure 4). A growing body of evidence suggests that the variation of Mtb strains has biological significance and

that specific genotypes are associated with specific phenotypes. Although genetic diversity seems to influence the clinical presentation of TB in humans, the specific factors for this phenomenon are still unknown. One of the most studied Mtb strains due to its association with outbreaks of TB and virulence is the Beijing/W family [36].

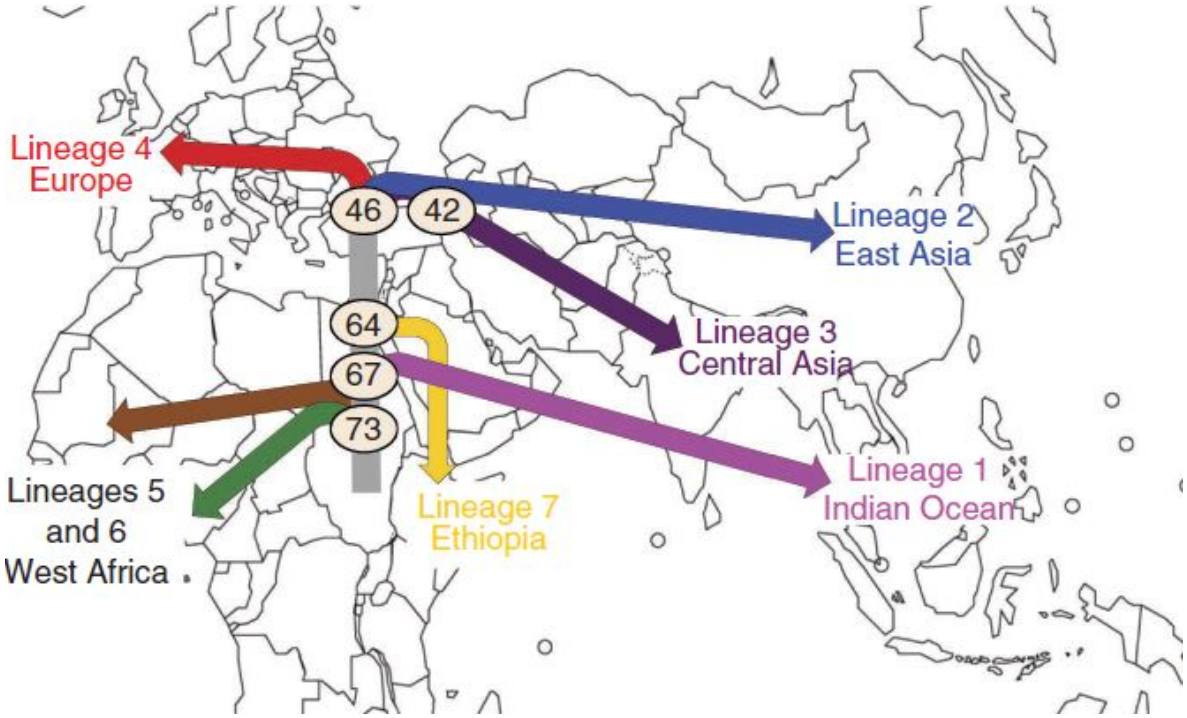


Figure 4: Map summarizing the results of the phylogeographic and dating analyses for Mtbc. Major splits are annotated with the median value (in thousands of years) of the dating of the relevant node [2].

It is now widely accepted that Mtbc strains originated from environmental mycobacteria. Phylogenetic studies indicate that Mtbc emerged about 70,000 years ago and that Mtbc species adapted to humans accompanying the migrations of modern humans out of Africa and co expanded with the increasing density of human populations during the Neolithic period originating the current strains and their phylogeographic distribution [37]. This timeline needs further revision. These ancient strains would have evolved to persist in low density populations, causing active TB followed by a long period of latent infection. The introduction of agriculture and increasing population densities in urban areas led to the selection of modern Mtb strains much more virulent and with more efficient transmissibility. Modern Mtb strains have spread throughout the globe causing TB epidemics that have plagued mankind for centuries and are responsible for most TB cases today [38].

1.3 Pathogenesis of Mtb

Infection of Mtb follows a relatively well-defined sequence of events (Figure 5).

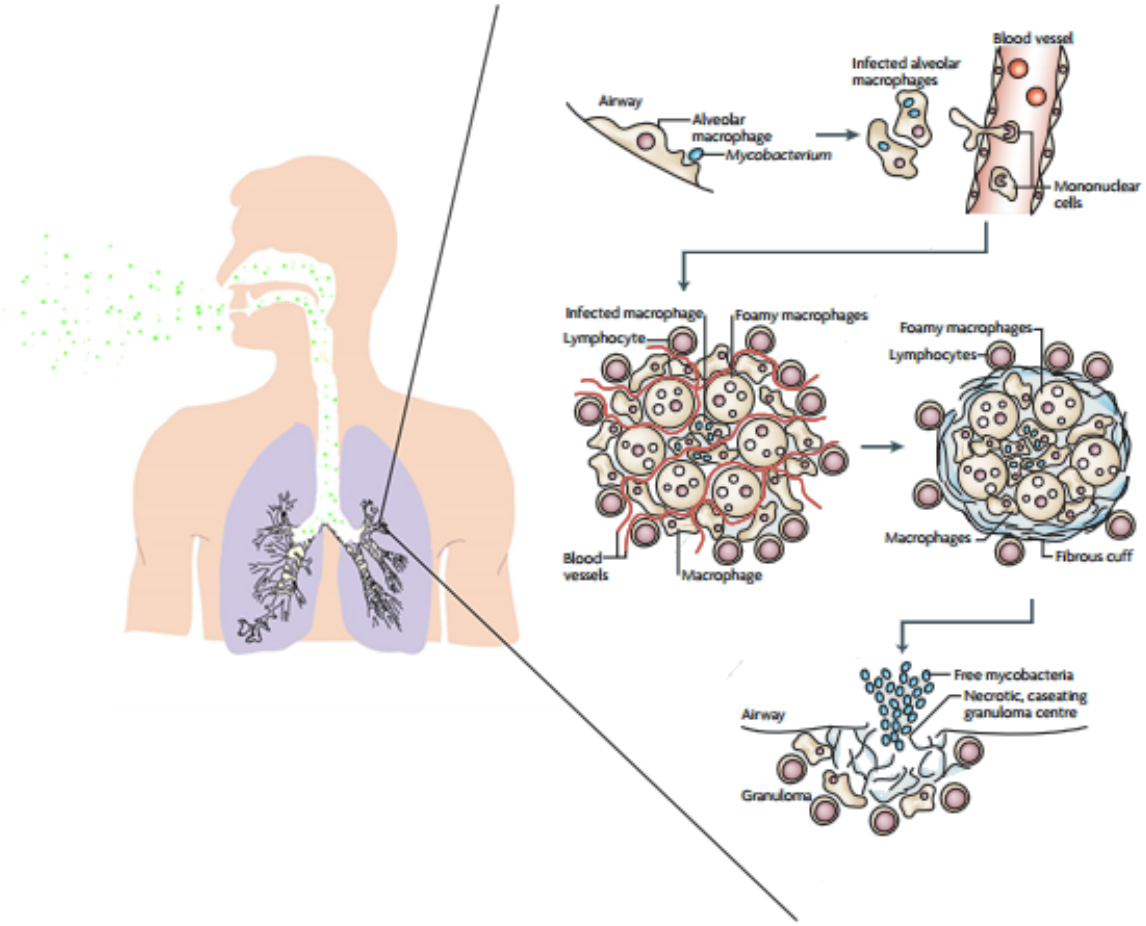


Figure 5: Mtb infection cycle. Alveolar macrophages phagocyte the inhaled bacteria. After phagocytosis cells of the innate and adaptive immune system are recruited into the infection zone giving rise to the formation of a granuloma. Bacilli can be contained within these structures for long periods of time. However, when the immune system weakens the bacilli restart replication, and a caseous necrotic granuloma is formed allowing the release of Mtb into the airways. Adapted from [3].

TB is transmitted through aerosols containing Mtb from an infected person. When inhaled, most bacilli are retained by the body's physical defense barriers but some bacteria are aspirated into the lungs. Once in the lung, bacilli are rapidly detected and phagocytosed by alveolar macrophages and dendritic cells [39]. During this process, cytokines and chemokines are secreted inducing a local inflammatory response and migration of monocytes from the bloodstream to the site of infection [3]. The presentation of Mtb antigens by dendritic cells to T lymphocytes in the lymph nodes induces the migration of the lymphocytes to the site

of infection inducing the formation of the granuloma. This granulomatous lesion or tubercle, comprises a central area of mycobacterial-infected cells surrounded by other, non-infected phagocytic cells and foamy giant cells, with lymphocytes found at the periphery. The lesion is sealed off from surrounding tissue by a fibrotic capsule [40].

In granulomas, macrophages are activated by T lymphocytes through the production of interferon (IFN) and tumor necrosis factor (TNF). These cytokines have the function of containing Mtb in the granuloma [39, 3]. This structure is characterized by low levels of oxygen, pH and nutrients, restricting the growth of the TB bacillus and establishing latency. Granulomas may persist for years and contain the pathogen efficiently while the individual remains immunocompetent. In these cases of latent TB, the control of this chronic infection results from a permanent balance between the host and mycobacteria.

In a small number of individuals (about 5%) with latent infection, changes in the immune system due to aging, malnutrition, immunosuppressive medication or HIV infection, re-activation or secondary TB occurs. The granuloma loses its structure which leads to the replication of the viable bacteria and the spread of the disease (reactivation of TB) [41, 3]. Macrophages do not survive in these lesions, being ineffective in controlling the outburst of bacteria, now also with an extracellular localization [42]. T lymphocytes are capable of lysing infected macrophages, releasing Mtb. This process can result in the release of viable bacteria into the bloodstream, becoming available to be phagocytosed and infected other macrophages, causing outbreaks of infection in various organs. Also, the bacteria can be released into the airways where they can be expelled from the lungs and infect other individuals, this way representing a high risk of contagion.

1.4 Anti-TB therapy and drug-resistance

Pathogenic species of the genus *Mycobacterium* cause a variety of infectious diseases, such as tuberculosis, leprosy and skin ulcers. The availability of whole genomic sequences has opened up the possibility of using various techniques to identify vaccines and therapeutic targets against these pathogenic agents, as well as identifying determinantsn for virulence or other functional features. A wide variety of techniques is currently available, including: Pan-genomics, that analyzes the genome of various organisms of the same species to detect an antigenic target representing the diversity of an organism; Transcriptomics, that is the study of expressed genes by an organism under specific conditions; Proteomics, that similarly to transcriptomics, directly analyzes the expression of protein pools under specific

conditions; Functional genomics, that identifies the candidate genes required for the survival of an organism and comparative genomics that is a powerful tool that can identify virulence genes present in pathogens but absent in non-pathogenic agents [43], among other features. The comparisons that can be made are endless and flexible.

Several approaches have been used to elucidate the molecular mechanisms employed by bacteria to survive antibiotic treatment.

The standart anti-TB therapy for susceptible TB with treatment regimens consist of an ”intensive” phase of 2 months, where it is administered isoniazid (INH) (discovered in 1952), rifampin (RIF) (1966), pyrazinamide (PZA) (1952) and ethambutol (EMB) (1961) followed by a 4-month ”continuation” phase with only INH and RIF [44]. However, the effectiveness of these regimens is threatened by the increasing number of Mtb strains resistant to these first-line drugs. The WHO, in 2016, estimated that 4.1% of new and 19% of previously treated TB cases had multidrug-resistance (MDR)-TB. [1] MDR-TB is caused by Mtb isolates resistant to the two most potent anti-TB drugs, RIF and INH with overall success rates of about 54% [1]. Patients with MDR-TB usually persist for 2 years or more and their treatment include the use of second-line drugs (such as fluoroquinolones and injectable aminoglycosides) that are less effective, more toxic and more costly [45]. The additional acquisition of resistance to these second-line drugs defines extensively drug resistant (XDR)-TB. The prognosis of patients infected with XDR-TB is poor [46] with virtually no successful standard treatment to date.

Re-emergence of drug resistance in an individual patient may occur as a result of poor adherence to treatment, inadequate drug regimen (e.g. wrong antibiotic choices or doses, poor drug quality), and patient-dependent pharmacodynamic and pharmacokinetic properties of drugs administered [47].

In recent years, several genetic determinants of drug resistance have been elucidated. These are associated with spontaneous mutations that interfere with binding of target drugs (for example, for RIF in the *rpo*B gene, for fluoroquinolones in the *gyr*A/B genes), activation of a prodrug of engagement (for example, for INH in the *kat*G gene, for PA-824 in the *fgd* gene), or causes overexpression of the target (for example, for INH / ethionamide in the *inh*A promoter region) [48]. However, resistance phenotypes of a significant proportion of clinical isolates of Mtb cannot be explained by these mutations alone: and it is estimated that up to 30% of INH-resistant isolates and approximately 5% of those resistant to RIF do not have mutations in the genes previously associated with drug-resistance [49].

Bacteria have great plasticity and are able to develop a series of mechanisms that facilitate rapid adaptation to changes in their environment (such as exposure to drugs) and modulate the effects of drug resistance [50, 51, 52, 53]. These observations illustrate the complexity of drug resistance in Mtb and highlight the need for a deeper exploration of the repertoire of strategies that lead to the emergence and subsequent fixation of resistant strains of Mtb.

1.5 Genetic variation in Mtbc: Single Nucleotide Polymorphisms

An important factor in pathogenesis that may affect virulence, transmissibility, host response, and drug resistance emergence is the specific genomic diversity found in Mtbc. As a way of classifying Mtbc strains, several systems have been proposed. One of the systems used is single nucleotide polymorphisms (SNPs), used as robust (stable) markers of genetic variation for phylogenetic analysis.

SNPs are the most common form of genetic variation in Mtbc. Generally, SNPs represent unique nucleotide differences between at least two DNA sequences. Recent advances in DNA sequencing have led to the discovery of thousands of SNPs in clinical isolates of Mtbc [19]. This genomic variation is used to infer both inter- and intra-lineage phylogenetic relationships at an unprecedented level of resolution, and lead to the development of a nomenclature extension for both lineages and sub-lineages.

Depending on their position in the genome, the SNPs may be encoded or not. Most SNPs in Mtbc (90-96%) are in coding regions of the genome [41]. The coding SNPs may further be divided into synonyms (Ks) and non-synonyms (Ka), depending on whether they lead to changes in the corresponding amino acid sequence. Although, on average, Ka may have a stronger effect on body fitness (beneficial or deleterious) and therefore they will be under a selective pressure stronger than Ks, they are not necessarily selectively neutral. Likewise, non-coding SNPs are often considered to be selectively "neutral", but increasingly the importance of the non-coding (i.e., untranslated) regions of the bacterial genome for gene regulation is becoming evident [42].

This genetic variation has changed our understanding of differences and phylogenetic relationships among strains.

One way to detect adaptive changes in amino acid sequences is by comparing Ka and

Ks. If the sequence of a gene was free to vary as a result of a mutation, being, as in a pseudogene, under no selective restriction, then the ratio of Ka / Ks should be one. In almost all evolutionary comparisons, however, the Ka / Ks ratios are much smaller than one, implying that the "purifying" selection is operating to preserve the amino acid sequence. This provides the basis of a test to see if the genes are evolving adaptively. If base changes that cause amino acid substitutions in a gene occur faster than the silent rate, this is certainly strong evidence for natural selection that changes the amino acid sequence [54].

1.6 PHYLOGENETIC STUDIES

Evolutionary history is the explanation for the origin of biodiversity and the best way to present it is using phylogenetic analysis. In molecular phylogenetic analysis, the sequence of a common gene or protein can be used to evaluate the evolutionary relationship of species, including the order of branching, allowing the drawing of a phylogenetic tree to which we can associate times of divergence, with an estimation of a time depth with the usage of a molecular clock [55]. The phylogenetic tree indicates how closely organisms are related to each other [4].

1.7 STRUCTURE OF THE PHYLOGENETIC TREE

Phylogenetic trees can be defined as graphs that have a hierarchical structure and are formed by links or branches that intersect and terminate at nodes. In these trees, nodes are called taxonomic units, which may represent, depending on the data analyzed, species, populations, genes or proteins. The nodes are classified in terminals (leaves), when these are in the end of the tree, or internal, when from them depart one or more descending branches. The terminal nodes represent the same samples used for inference of the tree and are therefore also called operational taxonomic units (OTUs), which in turn correspond to the basic unit (species, population, gene or protein) to be studied and compared (Figure 6) [56]. Internal nodes represent evolutionary events that depict the divergence of the taxonomic unit under analysis, such as speciation events, if the taxonomic unit is population, or gene duplication events, if the taxonomic unit is a gene or a protein. As the determination of internal nodes are products of a phylogenetic inference, they are also called hypothetical taxonomic units (HTUs). The branches are elements that connect the nodes. The smaller the distance of the branches between the samples or organisms being compared, the closer they are in

evolutionarily terms.

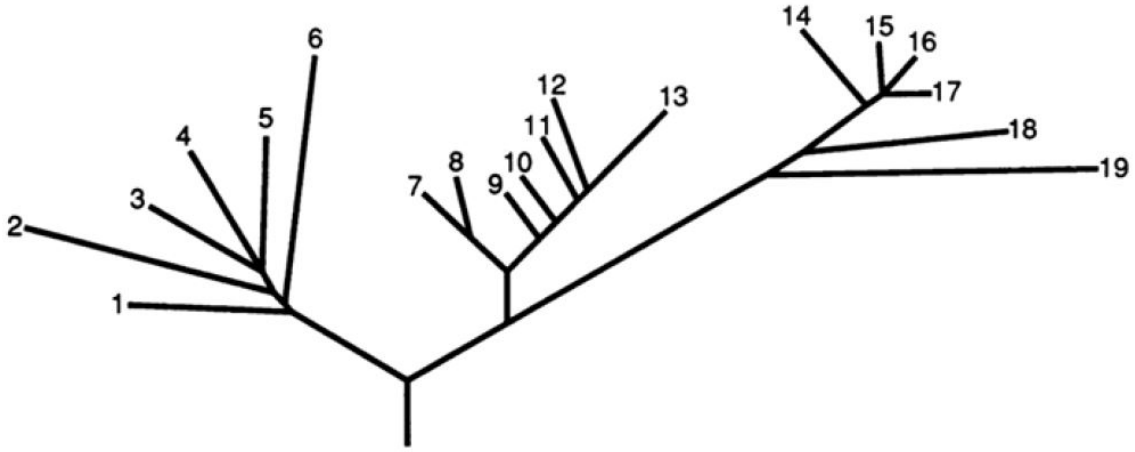


Figure 6: Schematic of a phylogenetic tree. OTUs represented by numbers, positioned at the terminal nodes. Adapted from [4].

Phylogenetic trees may assume different conformations depending on the arrangement of the branches along the tree. These different conformations are called topologies. The concept of topology is of great importance in phylogenetic studies because it represents the basis of all interpretation of evolutionary histories among the samples under analysis. Different topologies imply different evolutionary events, and it is up to phylogeny programs to determine the topology that best fits the data provided by the user.

1.8 PHYLOGENETIC TREE INFERENCE METHODS

Phylogenetic analyzes from DNA or protein sequence data start with alignment of orthologous sequences and can be divided according to the type of data used for inference. Some methods generate from the alignment data a distance matrix, which provides a measure of dissimilarity between the pairs of sequences present in the alignment. These methods based on distance matrices are characterized by being computationally fast and providing the user with a single tree (Examples include the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) [57] and the NeighborJoining method (NJ) [58]).

Other molecular phylogeny methods use algorithms based on optimization criteria. These

methods evaluate several topologies of the tree and choose the one that presents the best optimization score (i.e., the tree that best fits the alignment data). Examples of methods that use this approach are the maximum parsimony (MP) [59], the maximum likelihood (ML) [60] and Bayesian inference (BI) [61].

In the MP method, the different topologies of the trees are punctuated according to the number of characters change (substitutions) necessary for the data to fit in the topology. The algorithm will choose the topology that explains the alignment data with a lower number of substitutions.

The ML method evaluates the different tree topologies probabilistically. The probability of tree topologies is calculated based on a substitution model and the algorithm will determine which tree presents the highest value of this probability [62].

Bayesian methods start from the same mathematical structure as the ML, however, the ML maximizes the probability of observing the data given the tree, while BI infers the tree with higher probability given the data and the evolution model. Before the data analysis, the parameters are assigned to a distribution, which is then combined with the data to generate the posterior distribution that serves as the basis for inferences about the parameters that must be estimated. The main algorithm used for this type of approach is based on the Markov chain Monte Carlo algorithm (MCMC). The 'Markov chains' are a stochastic sequence (or chain) of states already reached, only of the current state. Thus, the choice of the resulting tree will depend on the smaller number of changes in the case of the maximum parsimony methods, the log-likelihood value in the maximum likelihood methods and the posterior probability in BI methods [53].

2

STATE OF THE ART

As the genetic analysis progressed, further studies were developed to determine the population structure and phylogeny of Mtb and it was demonstrated that all members of Mtbc share more than 99.95% genetic similarity. In addition, it became clear that modern TB is not caused by a single strain, but rather that the population structure of Mtb exhibits genetic diversity. This diversity may be linked to human demographics and migratory events and therefore suggests that the various Mtb strains have co-evolved with their human hosts [16, 63, 64].

Several studies have investigated this divergence based on the combined analysis of large sequence polymorphisms (LSPs), multilocus sequence typing (MLST) and SNPs, and seven major phylogenetic lines can be recognized within the Mtbc adapted to humans [65, 36]. Each of these lineages is named after the geographical region in which they were first identified and / or are most prevalent [63].

The first molecular method that was standardized was the IS6110 RFLP (restriction fragment length polymorphism) analysis by Van Embden et al. [66], which involves determining the number of copies and position of IS6110, a specific insertion sequence Mtbc (mobile genetic element) with high diversity in the copy numbers and insertion sites between the different members of Mtbc. Although the transposition of the IS6110 elements is to some extent random (at time and at the genome site), this method proved to be quite adequate for the study of transmission dynamics and has been widely applied for cluster analysis. However, for strains carrying less than six IS6110 copies, discriminatory power is low and a second method is generally required for more accurate genotyping. Consequently, the IS6110 RFLP analysis went out of fashion as new methods were developed and had the potential to replace it.

Since the first genome, several others have been sequenced [21]. The existing gaps being filled by extensive data sets generated by functional and comparative genomics, and by new disciplines employing powerful techniques covering:

Physical and integrated maps. An essential feature of mycobacterial genomics has been

the construction of integrated genomic maps. These were obtained by binding the physical map of the chromosome, obtained by pulsed gel electrophoresis, to the contiguous map, comprising sets of ordered cosmid clones (BACs), through various reference points [67, 68]. Physical maps were useful for estimating the size and structure of the genome (eg circular chromosome and absence of plasmids) and provided a framework for the establishment of integrated maps of cosmids and ordered artificial chromosome (BAC) libraries [68] which is essential for the rapid completion of genome sequences [21]. To construct BAC matrices a minimal overlapping set of BAC clones spanning the entire genome may be used or they can be used to compare individual BAC clones using libraries of different strains. This combined approach served to identify several deleted regions (RD1-10) in the genome of M. bovis BCG Pasteur compared to M. tuberculosis H37Rv, as well as loci (RvD1 and RvD2) that were deleted from the sequenced reference strain M. tuberculosis H37Rv relative to other strains of the M. tuberculosis complex [69].

Microarrays. Behr et al. [70] to better understand the differences between Mtb, *M. bovis* and the various daughter strains of BCG, studied their genomic compositions by performing comparative hybridization experiments on a microarray of DNA, being able to identify 14 regions (RD1-14) that were absent from BCG Pasteur over Mtb H37Rv, and two deletions (RD15 and 16) specific for particular subgroups of BCG, by hybridization of whole genome probes to a microarray of stained PCR (Polymerase chain reaction) products representing the majority of 4000 Mtb H37Rv ORFs (open reading frame). Because of cross-hybridization, the ability of microarrays using stained PCR products to detect deletions in multi-gene families is somewhat limited. This is particularly important in the Mtbc, since 10% of the genome contains repetitive DNA [19], including the PE and EPP gene families, which often comprise multiple tandem repeats. Reorganizations, insertions, inversions and genetic duplications are also difficult to detect using microarrays.

Subtractive Hybridization. Genomic subtraction, another powerful technique for complete genomic comparisons, has been used to compare the genomes of *M. bovis* and *M. bovis BCG* Connaught; three regions (RD1-3) were identified that were eliminated during the attenuation of *M. bovis BCG* [71]. Unlike microarrays, the strength of this technique is its ability to identify regions that are present in some members of a species but absent from the genome of the sequenced lineage [70].

Proteomics. Jungblut et al. used comparative proteome analysis to compare the protease of Mtb H37Rv with that of BCG, this technique is based on two-dimensional (2D) protein electrophoresis, mass spectrometry and database comparisons [72]. Comparative proteomics has the potential to detect very subtle differences.

Bioinformatics. Alignment of complete sequences of the in silico genome is the DNA-based comparative strategy. With the increasing number of complete genomic sequences and advances in bioinformatics, highly refined comparisons of sequence variation between

two lineages are possible using genome alignment tools [73, 74, 75, 76]. This is by far the most informative approach and its importance increases as more sequences become available.

Currently, the most widely used typing method is the typing of spacer oligonucleotides, or spoligotyping. This is a broad PCR-based method for the genotyping of Mtbc organisms. Barbara al. evaluated the performance of a microsphere-based spoligotyping assay using samples extracted from smear microscopy preparations stained with Ziehl-Neelsen and described the genetic diversity of Mycobacterium tuberculosis. Microbola-based spoligotyping is a high-throughput, easy-to-perform assay that can generate genotyping information using material obtained from smear microscopy preparations. The method provides an opportunity to obtain data from Mtb genetic epidemiology in environments with limited laboratory resources [77].

As more genetic markers become available, it becomes more difficult to compare DNA typing methods and choose the appropriate method. Typing methods should preferably be reproducible, quick, inexpensive, easy to perform and directly applicable to clinical material.

2.1 RELEVANT SOFTWARE

Due to tremendous advances that are being made in sequencing and phylogenetic analysis, and to keep up with the influx of genomic information there is the need for bioinformatics tools.

2.1.1 *Phylogenetic analysis by maximum likelihood*

Maximum Likelihood Phylogenetic Analysis (PAML) is a package of programs for phylogenetic analysis of DNA and protein sequences using ML. These programs can be used to compare and test phylogenetic trees, but their main strengths lie in the rich repertoire of implemented evolutionary models that can be used to estimate parameters in sequence evolution models and to test interesting biological hypotheses. The use of the programs includes estimation of synonymous and non-synonymous rates (Ks and Ka) between two protein-coding DNA sequences, positive Darwinian selection inference by phylogenetic comparison of protein coding genes, reconstruction of ancestral genes and proteins for studies molecular restoration of extinct life forms, combined analysis of heterogeneous multiple

gene locus datasets, and estimation of species divergence times incorporating uncertainties in fossil calibrations [62].

2.1.2 *MrBayes*

MrBayes is a free software that performs BI of phylogeny. Originally written by John P. Huelsenbeck and Frederik Ronquist in 2001[78]. As Bayesian methods increased in popularity MrBayes became one of the software of choice for many molecular phylogeneticists. The program uses the standard MCMC algorithm as well as the Metropolis coupled MCMC variant (MCMCMC). MrBayes reads aligned matrices of sequences (DNA or amino acids) in the standard NEXUS format [79].

These software uses MCMC to approximate the posterior probabilities of trees [80]. The user can change assumptions of the substitution model, priors and the details of the MCMCMC analysis. It also allows the user to remove and add taxa and characters to the analysis. MrBayes is also able to infer ancestral states accommodating uncertainty to the phylogenetic tree and model parameters. It uses the MCMCMC by default [78].

MrBayes allows the users to run multiple analyses in parallel. It also provides faster likelihood calculations and allow these calculations to be delegated to graphics processing unites (GPUs). The new version provides wider outputs options compatible with FigTree and other tree viewers [81].

2.1.3 *Randomized Axelerated Maximum Likelihood*

Randomized Axelerated Maximum Likelihood (RAxML) is a popular program for phylogenetic analyses of large datasets under maximum likelihood. Its major strength is a fast maximum likelihood tree search algorithm that returns trees with good likelihood scores. It can also be used for postanalyses of sets of phylogenetic trees, analyses of alignments and, evolutionary placement of short reads. It has originally been derived from fastDNAml which in turn was derived from Joe Felsentein's dnaml which is part of the PHYLIP package [76].

In RAxML, the topology of the tree is obtained by a step addition algorithm employing

the maximum parsimony criterion. Parameters and lengths of branches are also optimized in this initial step. Subsequently, modifications of the SPR type are applied, where a subtree is removed and then reinserted into another position of the tree. The RAxML applies such movements so that the distance between the place where the tree is removed and the place where it is inserted does not exceed a certain limit. In addition, only the branch lengths that are affected by reinsertion of the tree are optimized. The top 20 topologies resulting from the topological modifications are later optimized and the best of them is the new topology for a new iteration of the algorithm [82].

2.1.4 *Phylogenetic Analysis Using Parsimony**

Phylogenetic Analysis Using Parsimony (PAUP) is a program of computational phylogeny to infer evolutionary trees (phylogenies), written by David L. Swofford. Originally, as the name implies, PAUP implemented only the maximum parsimony method, but from version 4.0 (when the program became known as PAUP* (Phylogenetic Analysis Using Parsimony * and other methods) it started to support matrix methods distances and maximum likelihood [83].

The program features an extensive selection of analysis options and model choices, and accommodates DNA, RNA, protein and general data types. Among the many strengths of the program are the rich array of options for dealing with phylogenetic trees including importing, combining, comparing, constraining, rooting and testing hypotheses [83].

2.1.5 *Molecular evolutionary genetics analysis*

Molecular evolutionary genetics analysis (MEGA) is a bioinformatics tool used for genomic analysis of molecular sequences to measure evolutionary distance for the construction of phylogenies [74].

For purposes of comparative genome analysis, MEGA performs the following: sequence alignments; evolutionary distance measures; phylogenetic tree construction methods; phylogenetic tree evaluation; identification of genes / domains; selection confirmation; implementing sequential statistics [84].

MEGA extracts valuable information from double or multiple alignments of protein or DNA sequences, employing a series of statistical techniques to determine specific nucleotide or nucleotide physiognomy for the prediction of evolutionary relationships [74].

There are several methods of estimating trees in MEGA. Distance-based methods, such as UPGMA and NJ, begin by placing all rates on a single node, and then separate with each repetition. In this way, pairs of nodes are selected that are grouped in each iteration to reduce the total length of the branch. However, mutation rates are not constant. On the other hand, character-based methods, such as parsimony and probalistic methods, select the tree with the minimum number of changes as the preferred tree, identifying and estimating the total number of changes in each informational site for each possible tree. Specifically speaking, methods of estimating trees based on probalistic characters, such as ML and BI, find a set of trees most likely evaluating the possibility of a particular evolutionary model [84].

3

OBJECTIVES

In this project we aim to study the genome and evolution of Mtb, the causal agent of TB. For that, we performed a detailed analysis of all the gene encoding proteins in the genome of this pathogen using a pipeline developed for this purpose. Homologous proteins were searched in a well-defined database of complete genomes compiled previously, the closest homologues were extracted in a nucleotide sequence format and a phylogenetic tree were constructed (using Maximum likelihood) for each gene, where evolutionary parameters were estimated. The goal was to find genes that might have been acquired by horizontal transfer, in some cases previously pointed out as important in Mtb evolution, and genes that underwent relevant evolutionary changes in the evolution of Mtb or the Mtb complex (Mtbc), leading to the identification of underlying genes responsible for processes like virulence acquisition.

This work focused on establishing a detailed evolutionary picture of the genome of pathogenic Mtb. Independently of the focus on Mtb evolution the secondary purpose was to develop an analytic pipeline that could be applied to any other organism.

4

METHODOLOGIES

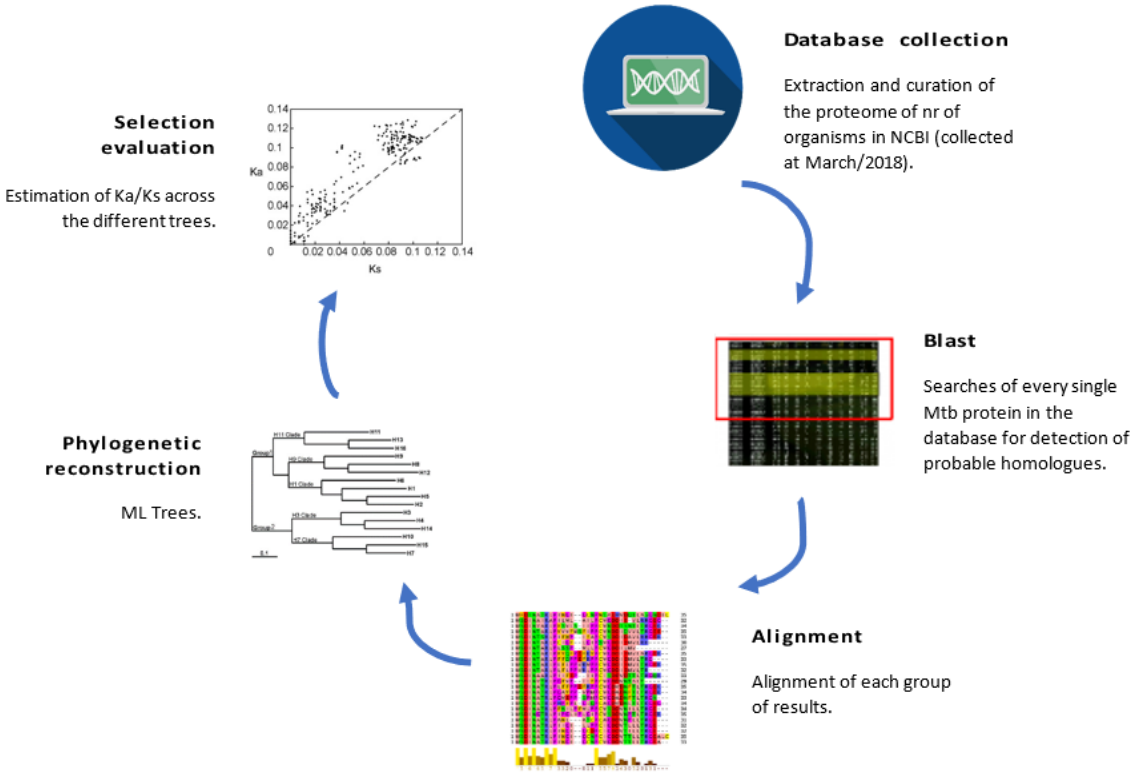


Figure 7: Schematic of the steps carried out by the pipeline.

The pipeline developed in this work (Figure 7) used Python as the main programming language. This was tested in Windows environment using the Python interpreter in version 3.5.

The following libraries in Python were used:

Biopython - manipulation of biological data;

- Bio.SeqIO - reading and manipulating files in several formats with sequences;

- Bio.Align.Applications (MuscleCommandline) - access to alignment tools;
- Bio.AlignIO - reading and manipulating files with alignments;
- Bio.Phylo - manipulate phylogenetic trees;

requests - performs HTTP requests;
os - allows the use of operating system-dependent features in a portable way;
Subprocess – allows us to generate new processes, connecting to their input / output / error channels and get the return code;
ete3 - aid in the reconstruction, manipulation, analysis and visualization of phylogenetic trees;

- ete3.Tree – stores a tree structure;
- ete3.Evoltree - adds attributes and methods to working with phylogenetic trees;

Xlswriter - allows creating xls files.

4.1 DATABASE

A local database was used, which was collected in March 2018 through the NCBI Assembly with the filter: complete genome, chromosome. In NCBI assembly a search was made for all the proteomes of organisms whose genome was complete. The local database consisted of organisms that had a reference sequence (RefSeq) or, in the case of not having a single wild-type or reference, to avoid ambiguities, only one of the strains was selected. In the latter case the criterion was to maintain the strain that was more similar to the wild-type (if this information was available) or the one with the highest number of proteins.

4.2 PROTEOME OF *mycobacterium tuberculosis*

In order to obtain the proteome, strain 6548, the strain with the largest number of proteins (4077) was chosen based on the principle that would give access to more information. The proteome includes predicted sequences using software tools and many have unknown functions.

4.3 BLASTS

A search of each Mtb protein in the database was performed to detect probable homologues, for which Blast + (Basic local alignment search tool) was used. This allows to perform alignments between DNA sequences or proteins, that are called query, against the sequences deposited in the database and these sequences are called subject.

The search carried out by Blast uses sequences of the query sequence against the sequences in the database. The query sequence is divided into small subsequences that will be used in the search. These are compared to the subject sequences until they find identical sequences. After an identical region is found there is an extension of the subsequence at both ends with other parts of the query sequence, to verify whether the similarity between the two sequences continues, producing an alignment between the two sequences.

As input, the input sequences (query and subject) are used, in the FASTA or Genbank format, they begin with the ">" sign followed by the identifier and other optional information, the next line contains the sequence. As output, we obtain files that show the hits found and a table containing data about subject id, subject title, % identity, alignment length, evalue, subject seq and subject tax ids.

Number of hits were limited to the first 200 with highest e-values. This value was established so that the computational burden for running the remaining pipeline was not too intense. Since we were using a local database containing a simple subject from each species, 200 hits were generally enough to generate a large spectrum of homologues within the tree of life, mostly bacteria.

4.4 ALIGNMENTS

The sequences used for protein and DNA alignments were obtained in two ways. The first way is through HTTP requests on NCBI servers. The second is through a database that contains this information. The first approach may be more time-consuming in the analysis, but it allows us not to have to maintain a relatively large local database. The second approach allows quick querying of most of the requested information, making HTTP requests only necessary for missing information in the local database.

To compose the pipeline, MUSCLE [73] was selected as the aligner of the sequences.

MUSCLE is a tool that performs multiple sequence alignment with good computational performance and significantly accurate biological results [73]. It implements, by default, the UPGMA method. This method is heuristic and follows an iterative procedure in an attempt to improve final alignment.

As input this tool uses a file fasta that contains for the query and each associated subject a line that starts with the ">" sign following the identifier and in the next line the protein sequence. This file only serves as input if you have 200 sequences or less, it was the limit set for this work. At the end of the alignment, a fasta file containing the aligned sequences, typically represented as lines of an array, is obtained. Gaps are inserted between the residues so that similar characters are aligned in successive columns.

In order to obtain the alignments of the DNA sequences, a script was developed that has as input the file obtained through Muscle with the sequences of the aligned proteins and another file with the DNA sequences, containing for the query and each associated subject a line that begins with the ">" sign following the identifier and, in the line, following the DNA sequence, both in fasta format. Through the aligned sequence and the DNA sequence the script makes the changes from protein to DNA considering the passage of amino acids to codons, that is, for each amino acid in the aligned sequence it reads three amino acids from the DNA sequence and if there is any gap (-) adds three gaps (—) in the DNA sequence, thus obtaining the DNA alignment in the fasta format corresponding to the alignment obtained from proteins.

4.5 MEGA

The reconstruction of the phylogenetic trees of the pipeline is performed by the MEGA-CC program, this is the command line interface to use the MEGA [84]. This software is known for being fast and for being able to handle a lot of sequences.

To run MEGA-CC,it is needed at least two input files - a file with the scanning options for MEGA and one or more data files that will be scanned. The file with the analysis options for MEGA specifies the calculation and the desired settings. It is created using MEGA-GUI (graphical interface) and has a .mao file extension. The data file contains multiple sequence alignment in the MEGA format. To convert the alignment files into MEGA files a script has been developed which searches the file for a line that begins with the ">" sign and replaces it for the "#" sign, uses the word after the signal as the sequence name and the

following lines (up to the next line beginning with ">") as the sequence data.

Using the options file, we chose to make phylogenetic inferences using the maximum likelihood method, searching for the most probable tree, assuming a probabilistic evolution model, with a cutoff of 80%.

In general, two types of output files are produced by MEGA-CC, a Newick file containing the tree (ids and distances), and a summary file with additional information (Data Type, No. of Taxa, No. of Sites, etc).

All phylogenetic methods can produce a phylogenetic tree, either a graph of a tree that shows the evolutionary relationships of its terminals, but many of them are silent on the polarity of the tree. Thus, to circumvent this problem, the midpoint root method was applied to each Newick file, through the program R (library: phytools). In this method, the longest distance between two terminals of the tree is identified and the root is placed precisely in the middle of this distance, obtaining a new Newick file with the new rooted tree.

4.6 MONOPHYLY DETECTION

The monophyly of three groups within the trees: the Mtbc, the Mycobacterium and the Corynebacteriales are studied. As input the developed script receives the file with the tree and checks the monophyly for each group, verifying if the group has all the same common ancestor, obtaining a True or False in response. With these answers an excel is created to be able to analyse each tree in more detail, to this excel are also added the number of elements of Mtbc, Mycobacterium, Corynebacteriales and the size of the tree. This information allows to identify which proteins had an expected evolution following the taxonomic grouping (basically True for the three performed monophyly tests), which could have been obtained through horizontal transfer in the Mycobacterium lineage (False for Corynebacteriales, and True for the other two) or in the Mtbc lineage (False for Corynebacteriales, False for Mycobacterium and True for Mtbc). Although this test allows to analyze without problems most of the proteins, trees where two homologues exist would need to be verified by the user.

For some of the trees, a graphical analysis was necessary for manual verification of the tree. With that objective, in each tree, through analysis of the lineage information, the identifier was changed to a name composed of the identifier, followed by cor||, if it were

Corynebacteriales, by cor||myco||, if it were Mycobacterium and cor||myco||mtc||, if it belongs to Mtbc, followed by the name of the organism (example:OMH53788||cor||myco||mtc|| Mycobacterium tuberculosis). The graphic analysis was done using the program FigTree.

Thus, with the results obtained by the monophyly, it was decided which trees would serve directly as input to Codeml (without manual verification), which were trees where the monophyly of Mtbc is True and trees that were not only composed by Mtbc or Mycobacterium.

4.7 PAML

In addition to the Newick file containing the tree, Codeml needs another input to run, that is the file with DNA alignment. During the process, several files containing the analysis data are generated. The objective of the analysis is to determine which branches of the Mtb complex are under faster evolution. This is performed by comparing values of the non-synonymous to synonymous ration throughout the tree against the specific Mtbc branch.

As output a file with the results is obtained, with variations depending on the branch model used. The branching models allow the ratio of ω (ω = Ka / Ks) to vary between branches in the phylogeny. They are specified using the model variable, model = 0 means an index of ω for all branches; 1 means a ratio for each branch and 2 means an arbitrary number of proportions. Only the model = 0 (M0) and 2 (M2) were tested. When model = 2 is used, we must group the tree branches belonging to Mtbc into a group of branches using the #1 tag.

With the results obtained by Codeml we created a table in the excel presenting the values of lnL (logarithm of the likelihood of the analysis), a measure of certainty of the data, and of ω for M0 (one ω) and M2(two ω). The ω ratio is a measure of natural selection that acts on the protein, values of $\omega < 1$, $= 1$ and > 1 mean negative selection, neutral evolution and positive selection, however those values are only useful on a theoretical framework. The objective is to identify specific branches where evolution is faster than the general tree.

In order to identify an adaptive evolution, the values of ω obtained in M2 are compare. Only the proteins where "ω of Mtbc" > "ω of general tree" are of interest as they represent the faster evolution in Mtb. The values of lnL and ω serve to make the likelihood ratio test (LRT), a statistical test used to compare the adjustment adequacy of two statistical models [85], through the function:

$$D = 2 * (-log_{10}(-L1) - (-log_{10}(-L2)))$$

A chi-square test is used to evaluate the value obtained. This function in excel is composed by three parameters, the value in which the distribution is to be evaluated, the number of degrees of freedom (calculated through the table, #p of M2 minus #p of M0 (4 - 1 = 3)) and a value logical that determines the form of the function, in this case we will use False for retrieving a probability distribution value. A *p-value* below 0.05 is the LRT was considered as a signal of statistically differential evolution in the Mtb clade.

Table 2: Parameters in the site models [6].

Model	NSsites	#p	Parameters
M0 (one ratio)	0	1	ω
M1a (neutral)	1	2	p_0 ($p_1 = 1 - p_0$), $\omega_0 < 1$, $\omega_1 = 1$
M2a (selection)	2	4	p_0, p_1 ($p_2 = 1 - p_0 - p_1$), $\omega_0 < 1$, $\omega_1 = 1$, $\omega_2 > 1$
M3 (discrete)	3	5	p_0, p_1 ($p_2 = 1 - p_0 - p_1$) ω_0, ω_1, ω_2
M7 (beta)	7	2	p, q
M8 (beta&ω)	8	4	p_0 ($p_1 = 1 - p_0$), p, q, $\omega_s > 1$

NOTE - #p is the number of free parameters in the ω distribution.

5

RESULTS

A pipeline was developed with the objective of studying the genome and evolution of Mtb. For this, it was necessary to follow the steps described in the methodology and obtain intermediate results that will also be described in this chapter along with the overall results.

5.1 DATABASE COLLECTION

A local database was used; the data was collected in March of 2018 through the NCBI Assembly with the filter: complete genome, chromosome. A total of 14899 organisms were obtained. In order to avoid ambiguities, only one strain per organism (preferably wild type) was selected. Only complete genomes fully annotated were used (from RefSeq). The final dataset was composed by 2863 organisms (30-animal, 193-archaea, 2489-bacteria, 82-fungi, 34-plants, 45-protista) and 11969222 sequences of proteins.

5.2 BLAST

We performed BLAST searches for 4077 queries, corresponding to the proteome of Mtb. 4077 files were retrieved each with a set of sequences that capture the close homologues from the taxonomy-related groups (Figure 8). Files with 1 hit up to 3625 hits were obtained. We limited the number of analyzed sequences to the first 200 hits. This will allow the overall pipeline to run for the various following steps with relative good speed.

x0000_OMH53778.out

```
# BLASTP 2.2.29+
# Query: OMH53778.1 putative PPE family protein PPE29 [Mycobacterium tuberculosis]
# Database: all
# Fields: subject id, subject title, % identity, alignment length, evalue, subject seq, subject tax ids, subject com nam
# 27 hits found
gb|OMH53778.1|  putative PPE family protein PPE29 [Mycobacterium tuberculosis]  100.00  131 3e-88   MTINNQFDDADTHGATSDFV
gb|AMC61213.1|  hypothetical protein RN08_3786 [Mycobacterium microti]  100.00  131 3e-88   MTINNQFDDADTHGATSDFWCDAEWAGI
gb|APU27268.1|  hypothetical protein BBG46_17845 [Mycobacterium caprae] 100.00  131 3e-88   MTINNQFDDADTHGATSDFWCDAEWAGI
gb|AMQ40082.1|  hypothetical protein AZ248_18215 [Mycobacterium africanum]  100.00  131 3e-88   MTINNQFDDADTHGATSDFWCDAE
emb|CCC45779.1| putative uncharacterized protein [Mycobacterium canettii CIPT 140010059]    89.31   131 3e-78   MTISNQFI
gb|AMC58766.1|  PPE family protein PPE19 [Mycobacterium microti]    45.00   100 2e-12   VAANLGRAASVGSLSVPQAWAAANQAVTPAAE
gb|APU25479.1|  hypothetical protein BBG46_07300 [Mycobacterium caprae] 42.00   100 1e-11   VAANLGRAASVGSLSVPQAWAAANQAVT
emb|CCC43711.1| PPE family protein [Mycobacterium canettii CIPT 140010059]  42.00   100 1e-11   VAANLGRAASVGSLSVPQAWAAAN
gb|AMC58953.1|  PPE family protein PPE19 [Mycobacterium microti]    42.00   100 1e-11   VAANLGRAASVGSLSVPQAWAAANQAVTPAAE
gb|AMQ38279.1|  hypothetical protein AZ248_07585 [Mycobacterium africanum]  42.00   100 1e-11   VAANLGRAASVGSLSVPQAWAAAN
gb|OMH59280.1|  putative PPE family protein PPE29 [Mycobacterium tuberculosis]  42.00   100 2e-11   VAANLGRAASVGSLSVPQAV
gb|OMH59112.1|  putative PPE family protein PPE29 [Mycobacterium tuberculosis]  44.00   100 3e-11   VAANLGRAASVGSLSVPQAV
emb|CCC43543.1| PPE family protein [Mycobacterium canettii CIPT 140010059]  44.00   100 3e-11   VAANLGRAASVGSLSVPQAWAAAN
gb|ANG87101.1|  hypothetical protein SZ58_001175 [Mycobacterium bovis]  44.00   100 4e-11   VAANLGRAASVGSLSVPQAWAAANQAVT
gb|AMQ38134.1|  hypothetical protein AZ248_06705 [Mycobacterium africanum]  44.00   100 4e-11   VAANLGRAASVGSLSVPQAWAAAN
gb|APU25333.1|  hypothetical protein BBG46_06430 [Mycobacterium caprae] 44.00   100 4e-11   VAANLGRAASVGSLSVPQAWAAANQAVT
gb|ANG87252.1|  hypothetical protein SZ58_002060 [Mycobacterium bovis]  41.00   100 1e-10   VAANLGRAASVGSLSVPPAWAAANQAVT
gb|ARG91815.1|  PPE family protein [Mycobacterium kansasii] 41.00   100 2e-09   VAAGLGRALSIGSLSVPHGWAAANQTVVPAARA-------
gb|AKN18512.2|  hypothetical protein B586_06060 [Mycobacterium haemophilum DSM 44634]   42.17   83  8e-09   ADFWDTTDWTGI
gb|AKN16456.1|  hypothetical protein B586_07625 [Mycobacterium haemophilum DSM 44634]   41.51   106 1e-07   GVGGPELSAGLG
gb|ASW89487.1|  hypothetical protein CKJ54_05975 [Mycobacterium marseillense]   42.73   110 3e-07   EIAAIRGPVLGGLGRATLIG
emb|CCC45823.1| PE family protein [Mycobacterium canettii CIPT 140010059]   39.00   100 3e-06   VAANLGRAASVGSLSVPPAWAAAN
gb|OMH61450.1|  putative PPE family protein PPE29 [Mycobacterium tuberculosis]  39.00   100 3e-06   VAANLGRAASVGSLSVPPAV
gb|AMC61263.1|  PE family protein PPE60 [Mycobacterium microti] 39.00   100 3e-06   VAANLGRAASVGSLSVPPAWAAANQAVTPAARA---
gb|AMQ40120.1|  hypothetical protein AZ248_18445 [Mycobacterium africanum]  39.00   100 3e-06   VAANLGRAASVGSLSVPPAWAAAN
gb|APU27311.1|  hypothetical protein BBG46_18105 [Mycobacterium caprae] 39.00   100 3e-06   VAANLGRAASVGSLSVPPAWAAANQAVT
gb|ANG89097.1|  hypothetical protein SZ58_012850 [Mycobacterium bovis]  39.00   100 3e-06   VAANLGRAASVGSLSVPPAWAAANQAVT
# BLAST processed 1 queries
```

Figure 8: Example of a file developed by Blast, showing the hits found and a table containing data about the subject.

5.3 Alignment and phylogenetic reconstruction

Initially, protein alignments were made using Muscle, and a total of 4042 alignment files were obtained (Figure 9). No more alignments were obtained because the rest of the files only had 1 hit. The number of aligned sequences in Muscle was limited to a maximum of 200 as stated above.

aln_prot_x0000.out.aln

```
MUSCLE (3.8) multiple sequence alignment


AKN18512_2      MTTNYEFGDLDP----------LGDLDGYEAAAQAASLESEYPISADDMRADF-------
OMH53778_1      ----------------------------MTINNQFDD--ADTHGATSDF-----------
AMC61213_1      ----------------------------MTINNQFDD--ADTHGATSDF-----------
APU27268_1      ----------------------------MTINNQFDD--ADTHGATSDF-----------
AMQ40082_1      ----------------------------MTINNQFDD--ADTHGATSDF-----------
CCC45779_1      ----------------------------MTISNQFDD--ADTHGATDDF-----------
ASW89487_1      ---MLDFGALPPEINSARMYAGPGSAPMSAAASAWDALAAQLESHAAGYSATLSELRGRA
AKN16456_1      ----MDFGALPPEVNSAKMYSGPGAGPMLAAAAAWDSLAADLYSVASSIQSIISELTLGL
ARG91815_1      ----MDFGALPPEINSARMYTGPGSGSLLAAAQVWDGVAIDLYNAASAVQSVIWGLLVGP
APU25479_1      ---MVDFGALPPEINSARMYAGPGSASLVAAAQMWDSVASDLFSAASAFQSVVWGLTVGS
APU27311_1      ---MVDFGALPPEINSARMYAGPGSASLVAAAKMWDSVASDLFSAASAFQSVVWGLTVGS
ANG89097_1      ---MVDFGALPPEINSARMYAGPGSASLVAAAKMWDSVASDLFSAASAFQSVVWGLTVGS
CCC45823_1      ---MVDFGALPPEINSARMYAGPGSASLVAAAKMWDSVASDLFSAASAFQSVVWGLTTGS
OMH61450_1      ---MVDFGALPPEINSARMYAGPGSASLVAAAKMWDSVASDLFSAASAFQSVVWGLTVGS
AMC61263_1      ---MVDFGALPPEINSARMYAGPGSASLVAAAKMWDSVASDLFSAASAFQSVVWGLTVGS
AMQ40120_1      ---MVDFGALPPEINSARMYAGPGSASLVAAAKMWDSVASDLFSAASAFQSVVWGLTVGS
APU25333_1      ---MVDFGALPPEINSARMYAGPGSASLVAAAQMWDSVASDLFSAASAFQSVVWGLTVGS
ANG87101_1      ---MVDFGALPPEINSARMYAGPGSASLVAAAQMWDSVASDLFSAASAFQSVVWGLTVGS
CCC43543_1      ---MVDFGALPPEINSARMYAGPGSASLVAAAQMWDSVASDLFSAASAFQSVVWGLTTGS
OMH59112_1      ---MVDFGALPPEINSARMYAGPGSASLVAAAQMWDSVASDLFSAASAFQSVVWGLTVGS
AMQ38134_1      ---MVDFGALPPEINSARMYAGPGSASLVAAAQMWDSVASDLFSAASAFQSVVWGLTVGS
AMC58766_1      ---MVDFGALPPEINSARMYAGPGSASLVAAAQMWDSVASDLFSAASAFQSVVWGLTVGS
ANG87252_1      ---MVDFGALPPEINSARMYAGPGSASLVAAAQMWDSVASDLFSAASAFQSVVWGLTVGS
CCC43711_1      ---MVDFGALPPEINSARMYAGPGSASLVAAAQMWDSVASDLFSAASAFQSVVWGLTTGS
```

Figure 9: Example of a file obtained by Muscle.

These alignments after being converted to the MEGA format were used to obtain the phylogenetic trees for the various proteins. With the help of MEGA, 3920 trees were obtained. The remaining alignments were not converted to trees because they only had two sequences, and any tree can only be built with at least three sequences. With the midpoint rooting method, 3893 new trees were obtained. The method of midpoint rooting was not applied to trees that were composed only of Mtbc genes (Figure 10). These genes corresponded to widely repeated sequences across the Mtb genome (PE and PPE families) that are generally excluded from any Mtb genomic analysis [86] and again in this case the results would not be analyzable in an intelligent fashion using this pipeline.

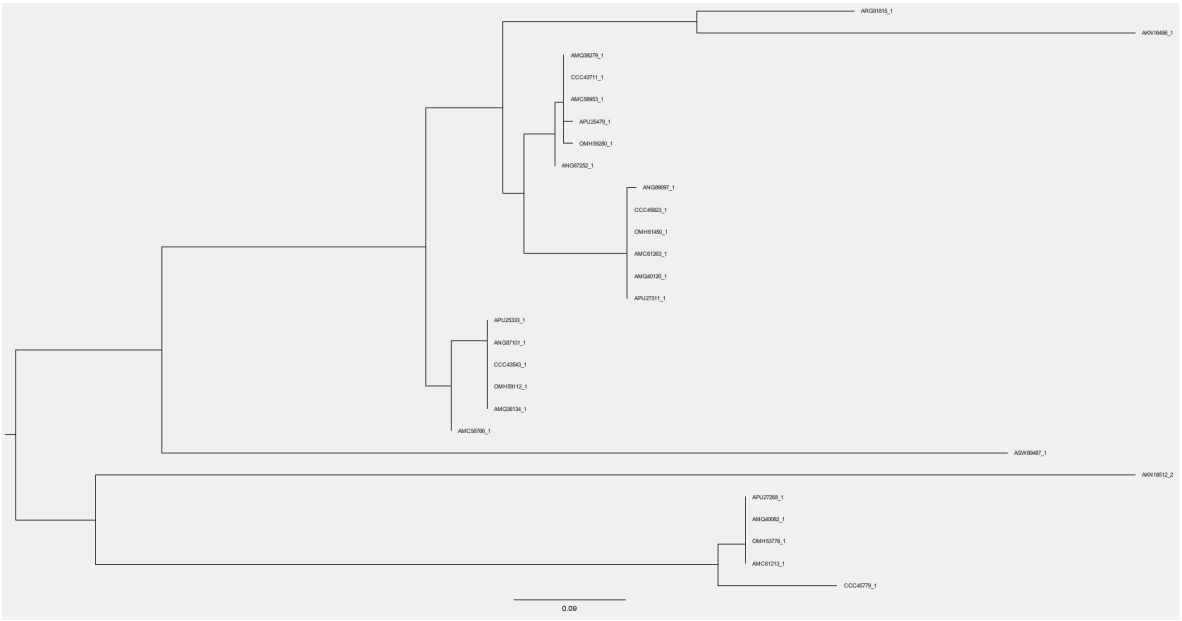


Figure 10: Example of a file obtained by Mega where the midpoint root method was applied. Image obtained through the FigTree program.

As mentioned in the methodology, it became evident that the detection of monophyly would only be reasonable if only a single homologue of the analyzed species would be present. It was necessary to analyze the trees manually. In order to make that task easier a script was done so that Mtbc, Mycobacterium (non-Mtbc) and Corynebacteriales (non-Mycobacterium) would be indicated in the tree. The script that changes the names of organisms using the information on their taxonomy makes it easier to graphically analyse the trees of most interest as displayed in the example (Figure 11). The tree displays a case where Mtbc was monophyletic however none of the other groups were. Homologues are non-existent in other Mycobacteria, suggesting a case of horizontal transfer.

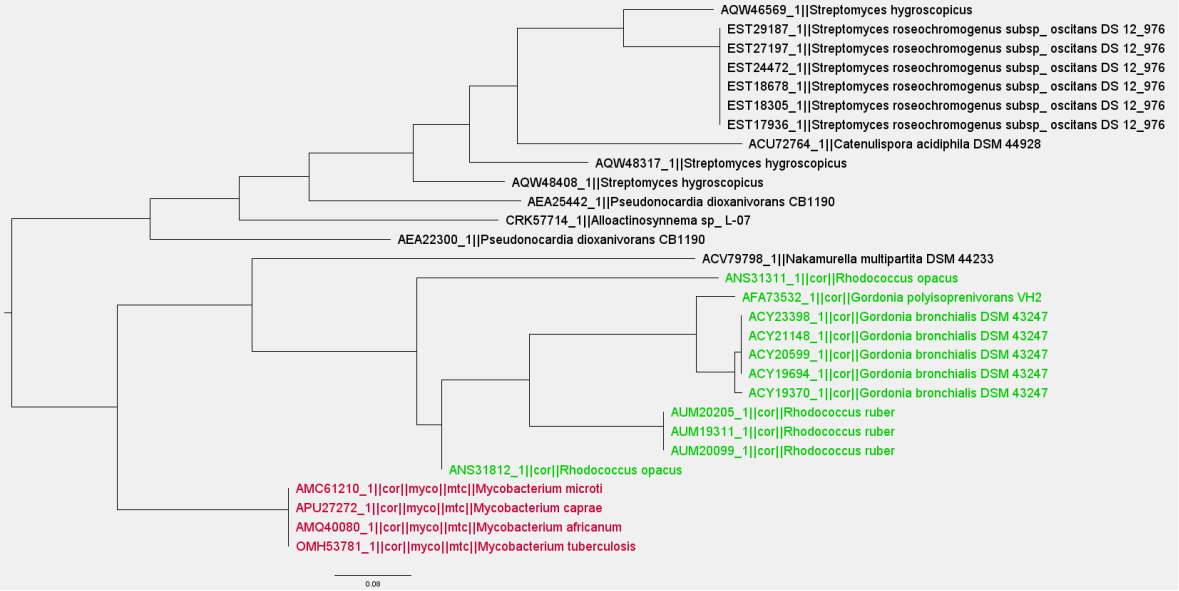


Figure 11: Example of a tree with the name of the organisms altered through its lineage. In green can be seen the organisms belonging to the Corynebacteriaceae and in red the organisms belonging to the Mtbc, being no organisms belonging to the Mycobacterium but that are not inserted in Mtbc.

With the trees obtained, monophyly was studied for three groups: Mtbc, Mycobacterium and Corynebacteriales (Figure 12).

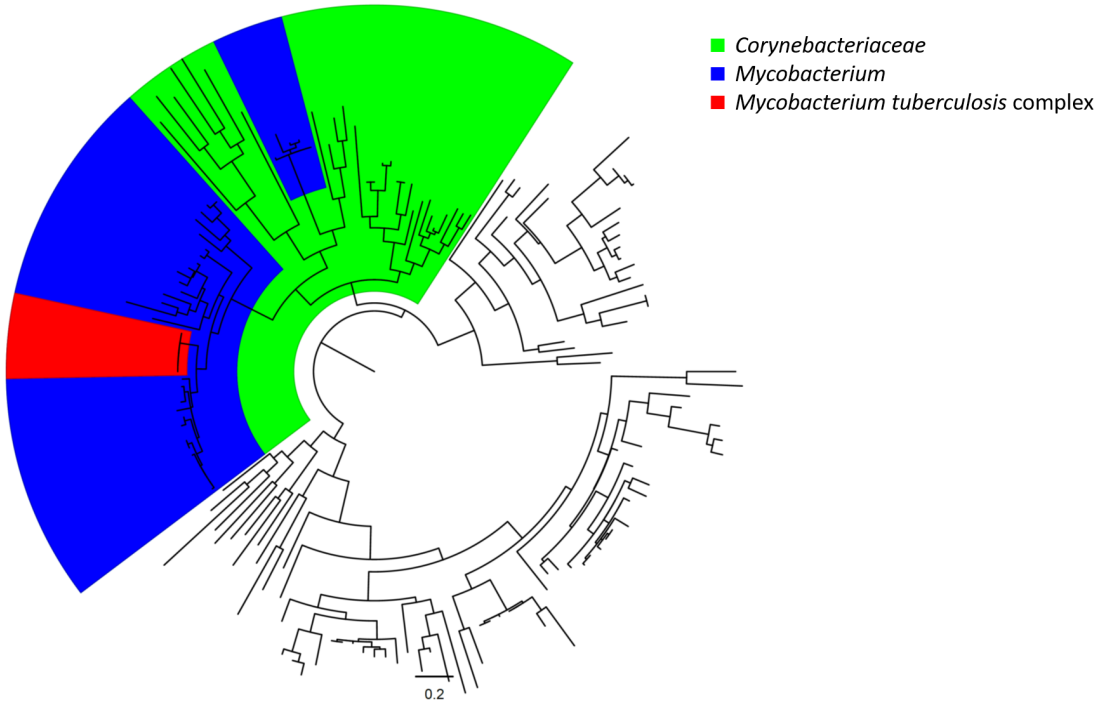


Figure 12: Example tree where the monophyly can be observed graphically for the three groups under study.

The example in figure 12 displays a case that was inspected manually. The model tree is a good example as it shows a monophyletic branch for Corynebacteriaceae and Mtbc. Mycobacterium are not monophyletic but are clearly grouped into a main branch with only a few off and Mtbc is clearly included in this larger branch of Mycobacterium, so although Mycobacterium is not monophyletic, Mtbc clearly evolved from the general Mycobacterium branch so for the purpose of this work monophyly could be considered as evidence of inexistence of horizontal gene transfer.

With all the results obtained on the monophyly for the three taxonomic groups and for all the organisms a graph was created make it easier to visualize the overall results (Figure 13).

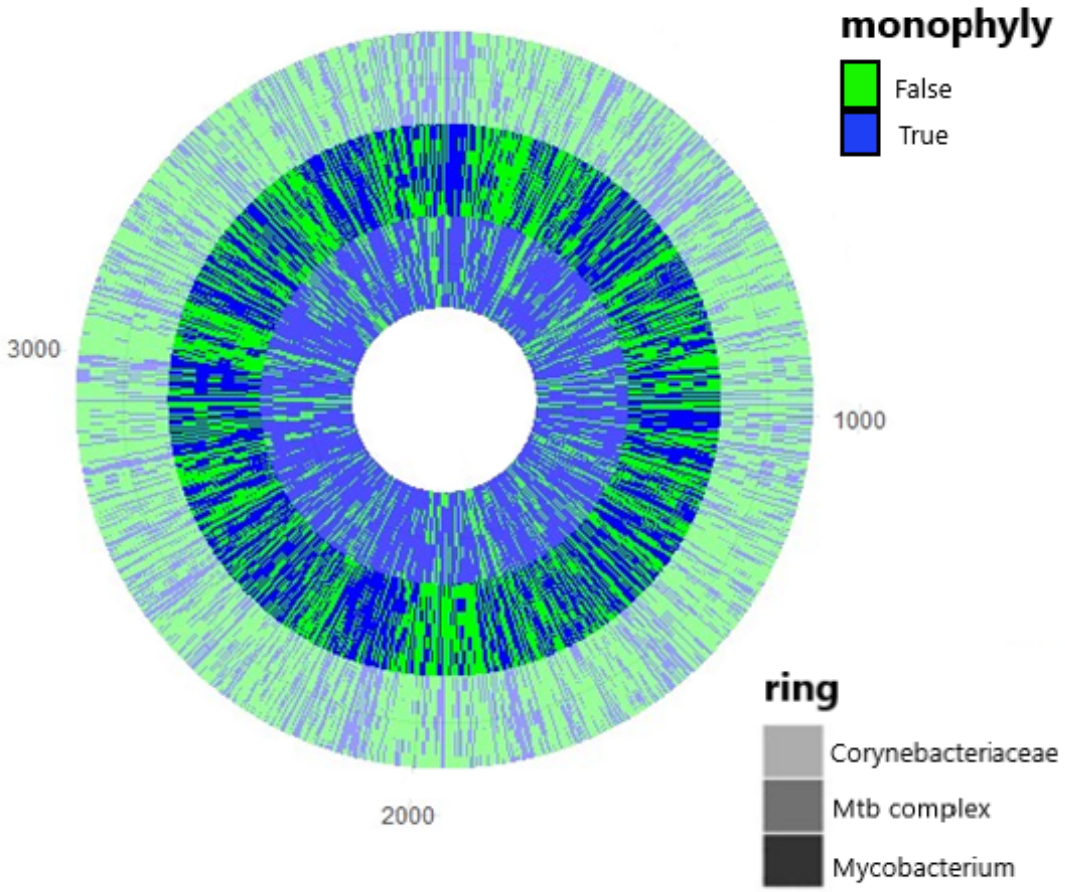


Figure 13: Monophyly of Mtb divided into three groups Mtbc, Mycobacterium and Corynebacteriaceae. The result for monophyly is obtained in true and false.

It is possible to see some trends in the figure. Most of the Mtc displat monophyletic branches. This basically reflects the fact that the Mtbc is composed of specimens whose genomes are very similar between them. Many of these proteins do not display monophyly for the Mycobacterium group suggesting it could be the result of horizontal transfer into the Mtbc lineage. The Corynebacteriaceae group is the one that displays lower level of monophyly. Such result is expected since we are looking into deeper evolutionary time scales and it is more probable that other issues occur, namely more uncertain phylogenetic reconstruction, existence of more than one homologue, horizontal transfer both for the outside to the Corynebacteriaceae and the other way around. Nevertheless this offers a first good picture of the overall trends and a first draft that will need to be checked in many proteins in the near future.

4042 files were obtained through the script created to align the alignment of the protein sequences with the DNA sequences (Figure 14).

aln_dna_x0000.out.aln.phylip

```
 27 1281 I
ANG87101_1   ---------A TGGTGGATTT CGGGGCGTTA CCACCGGAGA TCAACTCCGC
AMQ40082_1   ---------- ---------- ---------- ---------- ----------
OMH59280_1   ---------G TGGTGGACTT CGGGGCGTTA CCACCGGAGA TCAACTCCGC
ARG91815_1   ---------- --ATGGATTT CGGAGCTTTG CCGCCGGAGA TCAACTCCGC
AMQ38279_1   ---------G TGGTGGACTT CGGGGCGTTA CCACCGGAGA TCAACTCCGC
OMH53778_1   ---------- ---------- ---------- ---------- ----------
APU27268_1   ---------- ---------- ---------- ---------- ----------
AMQ40120_1   ---------G TGGTGGATTT CGGGGCGTTA CCACCGGAGA TCAACTCCGC
CCC43711_1   ---------G TGGTGGACTT CGGGGCGTTA CCACCGGAGA TCAACTCCGC
CCC43543_1   ---------A TGGTGGATTT CGGGGCGTTA CCACCGGAGA TCAACTCCGC
AMQ38134_1   ---------A TGGTGGATTT CGGGGCGTTA CCACCGGAGA TCAACTCCGC
CCC45823_1   ---------G TGGTGGATTT CGGGGCGTTA CCACCGGAGA TCAACTCCGC
AMC61213_1   ---------- ---------- ---------- ---------- ----------
ANG87252_1   ---------G TGGTGGATTT CGGGGCGTTA CCACCGGAGA TCAACTCCGC
OMH59112_1   ---------A TGGTGGATTT CGGGGCGTTA CCACCGGAGA TCAACTCCGC
ANG89097_1   ---------G TGGTGGATTT CGGGGCGTTA CCACCGGAGA TCAACTCCGC
AMC58953_1   ---------G TGGTGGACTT CGGGGCGTTA CCACCGGAGA TCAACTCCGC
AMC61263_1   ---------G TGGTGGATTT CGGGGCGTTA CCACCGGAGA TCAACTCCGC
OMH61450_1   ---------G TGGTGGATTT CGGGGCGTTA CCACCGGAGA TCAACTCCGC
APU25479_1   ---------G TGGTGGACTT CGGGGCGTTA CCACCGGAGA TCAACTCCGC
AKN16456_1   ---------- --ATGGATTT CGGGGCGTTA CCACCAGAGG TCAATTCGGC
AKN18512_2   ATGACGACCA ACTACGAGTT CGGGGACCTG GACCCG---- ----------
```

Figure 14: Example of a file with DNA alignment obtained through the script.

As mentioned in the methodology this script for each amino acid of the protein sequence advances three nucleotides (corresponding to a codon) in the DNA sequence and when it finds a gap (-) it adds three gaps (—) (Figure 15). This way the DNA alignment correspond exactly to the protein alignment that nevertheless is more reliable to perform on the deeper evolutionary levels.

```
AKN18512_2  M   T   T   N   Y   E   F   G   D   L   D   P   -   -        ==> protein
            |   |   |   |   |   |   |   |   |   |   |   |   |   |
AKN18512_2  ATG ACG ACC AAC TAC GAG TTC GGG GAC CTG GAC CCG --- ---       ==> DNA


-   -   -   -   -   -   -   -   L   G   D   L   D   G   Y   E   A   A     ==> protein
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
--- --- --- --- --- --- --- --- CTT GGT GAC TTG GAC GGT TAC GAG GCC GCC    ==>DNA
```

Figure 15: Schematic showing how the script works to align to DNA.

5.4 SELECTION EVALUATION

Later, having the phylogenetic trees and the alignments of DNA, Codeml was used to estimate the values of lnL and Ka / Ks for each tree, a ratio also known as omega.

This analysis was also performed in two ways. One was calculating an overall omega for the tree and the second was through the calculation of two omega values, one specifically for the Mtbc branch and a second for the remaining tree. A statistical comparison (LRT) between both analysis allows to indicate if the Mtbc branch has a differential omega. This is a well-defined way to perform a statistical evaluation on selection for the around 4000 protein available. Nevertheless this is a much more time-consuming step and by the time we reached it, it was impossible to perform the full analysis. For that reason, a few trees were picked as example.

We created a table (Table 3) where the results can be seen and analysed.

Table 3: Table presenting the values of lnL (logarithm of the likelihood of the analysis) and of ω for Mo (one ω) and M2 (two ω).

	Model M0:		Model M2:	general	mtbc		
gene	log likelihood	omega	log likelihood	omega	#1	D	chi square distribution
OMH53781	-14322,9079	0,1126	-14321,8779	0,1117	0,3936	2,060044	0,204416262
OMH57858	-113111,66	0,0675	-113108,512	0,0671	0,11	6,295858	0,042984248
OMH57895	-20189,9102	0,0899	-20188,6577	0,0909	0,0227	2,50495	0,180454042
OMH57903	-6308,04562	0,1153	-6307,30817	0,1133	0,1936	1,474904	0,231750462
OMH57909	-47031,9737	0,144	-47027,9024	0,1464	0,0672	8,14271	0,019414559
OMH57914	-16399,1149	0,064	-16365,5431	0,0633	0,0242	67,14362	8,59727E-15
OMH57916	-13634,9199	0,0486	-13626,7473	0,0490	0,0160	16,345244	0,000455283
OMH57927	-1019,69692	0,1128	-1019,45033	0,0977	0,2742	0,493176	0,218937031
OMH58349	-1537,73547	0,3082	-1537,55878	0,3900	0,0001	0,35338	0,1987447
OMH58355	-26917,9534	0,0748	-26912,324	0,0740	0,2451	11,25883	0,004806543
OMH58356	-30211,7363	0,1412	-30211,1366	0,1420	0,0902	1,19948	0,239851699
OMH58358	-37331,1946	0,1113	-37330,8615	0,1109	0,1308	0,66624	0,233374414
OMH58360	-58306,5648	0,1219	-58304,8829	0,1224	0,0639	3,363798	0,136108599
OMH58372	-1539,49765	0,0956	-1539,48012	0,0946	0,1249	0,03506	0,073401146
OMH58373	-26386,7907	0,1111	-26386,6451	0,1105	0,1289	0,291182	0,186107108
OMH58377	-66888,9876	0,2401	-66887,1261	0,2418	0,1496	3,72315	0,119644732
Virulence							
OMH58354	-37883,7832	0,1008	-37881,9328	0,1020	0,0510	3,700672	0,1206312
OMH58651	-67258,9022	0,0456	-67258,6352	0,0457	0,0351	0,534	0,223215356
OMH59737	-73114,4468	0,1011	-73112,4322	0,1015	0,053	4,029172	0,106805689
OMH59837	-180531,145	0,0866	-180524,632	0,0869	0,0447	13,025748	0,002137008
OMH60377	-37484,9038	0,0501	-37479,8985	0,0505	0,0153	10,010436	0,008460538
OMH61789	-35394,1788	0,2057	-35394,1782	0,2057	0,2008	0,001076	0,013079241

As it was described in the methodology only the values in which ω of Mtbc $> \omega$ general were of primary interest, since we are interested in proteins were the evolution was accelerated in the Mtbc branch. Proteins with significantly *p-values* but with lower values of ω in Mtbc are not so direct to interpret but they might grant a future evaluation. The chi-square test was used to evaluate the obtained value, obtaining in this case five significant values (red and green) (Table 3).

However only a case is presented (green) where the omega has a significantly higher value in the mtc than in the general one. This protein (OMH58355) is of function unknown although it is clearly a well-defined protein present across various groups of organisms are very conserved in Actionbacteria in general.

We also tested specifically six proteins previously associated with virulence in Mtbc, however none of them showed significant accelerated evolution (positive selection). In opposite, two of them showed lower omega than the general tree.

6

DISCUSSION AND FUTURE WORK

The objective of this work was to study the evolution of Mtb, the causal agent of TB, through a detailed analysis of all protein-coding genes in the genome of this pathogen using a pipeline developed for this purpose. This pipeline was successfully developed and implemented (appended with the code), including all the steps involved: from the blast searches performed for each protein of the proteome on a local database, to the evolutionary analysis of selection effect obtained by PAML. This pipeline has been developed so that it can be applied to any other organism (bacteria, eukaryotes, etc.) and not exclusively to Mtb, that just functions as a model/example organism here.

With technological developments, the volume of genomic data is extremely large nowadays and it is continuously growing. Thus, the local database used in this work proved to be of great importance and a crucial point in the analysis because, unlike the general large online databases (ncbi, Uniprot, ENA), which are redundant and only allow us to access what is relatively close to the organism we want to study, this database is well defined, promoting support to the storage of data and efficient access to them, presenting only one strain of each organism, this way avoiding redundancies. This database is representative of the tree of life, being composed of several organisms: animals, archaea, bacteria, fungi, plants and protista. In the developed pipeline, the search for the 200 top hits in BLAST would allow a wide variety of organisms to be selected, while in an online database the first 200 results would result, in this case, probably in 200 Mtb genes from other strains. Even curated databases and ones containing only a single representative genome (RefSeq) displays an uncertain definition of representative genome in some organisms corresponding to the species level and to strains in other organisms. Another main advantage of the local database is that, since only one organism of each species is included, in the case of two homologues (or more) of the same species within a tree, these cases can be immediately interpreted as representing duplicated homologues within the same genome. An additional feature is that since we know exactly which species are included, we can assess the prevalence of a protein or family of proteins within a given taxonomic group. The pipeline is used here to investigate genetic patterns of the different protein-coding genes but it could also be a powerful tool for genomic annotation.

In addition to the main objectives fulfilled, this pipeline presents another smaller advantage over other methods for an overall analysis in different cases and organisms. This relates with the passage of amino acids to codons. While deeper evolutionary relations are easier to be analysed through the multiple alignment of proteins instead of DNA (due to the degeneracy of the genetic code), most of the tools use the direct correspondence or the passage of amino acids to codons using a table of the genetic code under study and generating a sequence of DNA representing the most likely coding sequence. This becomes problematic when different organisms with different genetic codes are displayed in the same alignment (bacteria, nuclear eukaryotic DNA, mitochondrial DNA) where a single genetic is not adequate. In here, the amino acid sequence and the codon sequence of the gene are used directly and following the correspondence between both, each amino acid is replaced by three nucleotides. This way the DNA alignment correspond exactly to the protein alignment that nevertheless is more reliable to perform on the deeper evolutionary levels and it works whatever the multiple genetic codes that are involved in each alignment.

The final objective of the pipeline in this work consisted of estimating the evolutionary parameters for each gene in order to find genes that could have been acquired by horizontal transfer or genes that underwent relevant evolutionary changes in the evolution of Mtb or Mtbc.

Mycobacterium contains some of the deadliest pathogens known to humans so far. Becq et al. proposed that extensive genome reduction followed by many episodes of horizontal gene transfer contributed effectively to the evolution of pathogenic strains of *Mycobacterium* [87]. Horizontal gene transfer (HGT) can be detected by methods based on unrelated genomic signatures or atypical sequence composition. However, such methods have been shown to be ineffective in detecting relatively old HGTs with very high false positive and false negative detection rates and are considered to be poor indicators [88]. Phylogenetic methods, on the other hand, detect genes that were supposedly transferred considering their similarity with unrelated taxa. The main advantage of phylogenetic methods is that the point of acquisition of the exogenous gene can be traced from the topology of the tree. Panda et al. considered proteins representative of all the fully sequenced available fully sequenced *Mycobacterium* genomes and identified putative HGTs through their phylogenetic trees [89]. In this case, maximum likelihood phylogenetic trees were constructed and phylogenetic signals indicative of likely horizontal gene transfer were manually searched in tree topology [Panda2018]. While the approach is very similar to the one employed here, ours has the mentioned advantage that can be applied directly to any proteome, while theirs relates with a specific analysis for Mtb. Our results already point out many genes across the Mtb

genome that could have been acquired by HGT however some manual checking points are still required.

Another important point in our pipeline is that we also look, beyond tree topology, genes that might be under positive selection in the Mtb lineage. The test of the likelihood ratio between the two analyses obtained by Codeml allows to indicate if the branch Mtbc has a differential omega, either significantly higher or lower than the general tree of its homologues. This is a well-defined way to carry out a statistical evaluation in the selection of the approximately 4000 proteins available. However, this is a much more time-consuming step, and full analysis has been impossible in the timeframe of the Master. For this reason, only some trees were chosen as examples, hopefully an adequate illustration of the full potential of the work developed.

Since we were interested in protein-coding genes and most genes that could show signals of adaptation generating the biological features of Mtb, the main evolutionary interest was on accelerated (positive selection) in the Mtbc branch. In the examples performed only one protein-coding gene (OMH58355) displays this pattern, corresponding to a protein of unknown function. Nevertheless, it is clearly a well-defined protein present in several groups of organisms and actually very conserved in Actinobacteria in general. Proteins with significant p-values, but with smaller values for Mtbc omega are not as straightforward to interpret but may allow future evaluation. In an evolutionary perspective they could be interpreted as genes where the evolutionary constraints were stricter than in the remaining tree (meaning, more conserved). The chi-square test was used to evaluate the value obtained, obtaining in this case five significant values.

In order to have more accurate data, four proteins previously associated to virulence in Mtbc were specifically studied, but none of them presented a significant accelerated evolution (positive selection). On the contrary, one of them presented omega inferior to the general tree. The results obtained suggest that these genes were not under evolutionary positive selection in the Mtbc lineage.

As previously stated, it was not possible to perform the complete analysis given the time constraints of a Master degree. For future work, it is necessary to finish the evolutionary analysis of non-synonymous and synonymous ratios following a careful checking of the tree topology for various of the genes. Another format of the analysis that would be interesting would be not to limit the CODEML analysis to the Mtbc branch but extend it to the analysis of the ancestral *Mycobacterium* branch as pathogenicity and other important

features of the Biology of Mtb may have arisen earlier in evolution and are shared across the full genus *Mycobacterium*.

BIBLIOGRAPHY

[1] W. H. Organization *et al.*, "Global tuberculosis report 2017," in *Global tuberculosis report 2017*, 2017.

[2] H. Nebenzahl-Guimaraes, S. A. Yimer, D. van Soolingen, R. Brosch, C. Holm-Hansen, and J. de Beer, "Genomic characterization of mycobacterium tuberculosis lineage 7 and a proposed name: 'aethiops vetus'," *Microbial Genomics*, vol. 2, no. 6, jun 2016.

[3] D. G. Russell, "Who puts the tubercle in tuberculosis?" *Nature Reviews Microbiology*, vol. 5, no. 1, p. 39, 2007.

[4] D. A. Morrison, "Phylogenetic analysis of pathogens," in *Genetics and Evolution of Infectious Disease*. Elsevier, 2011, pp. 203–231.

[5] M. A. Forrellad, L. I. Klepp, A. Gioffre, J. Sabio y Garcia, H. R. Morbidoni, M. d. l. P. Santangelo, A. A. Cataldi, and F. Bigi, "Virulence factors of the mycobacterium tuberculosis complex," *Virulence*, vol. 4, no. 1, pp. 3–66, 2013.

[6] *PAML 4: a program package for phylogenetic analysis by maximum likelihood*. Molecular Biology and Evolution, 2007, vol. 24.

[7] D. Brites and S. Gagneux, "Co-evolution of mycobacterium tuberculosis and homo sapiens," *Immunological reviews*, vol. 264, no. 1, pp. 6–24, 2015.

[8] T. M. Daniel, "The history of tuberculosis," *Respiratory Medicine*, vol. 100, no. 11, pp. 1862–1870, nov 2006.

[9] R. Müller, C. A. Roberts, and T. A. Brown, "Complications in the study of ancient tuberculosis: Presence of environmental bacteria in human archaeological remains," *Journal of Archaeological Science*, vol. 68, pp. 5–11, 2016.

[10] D. B. Tierney, M. F. Franke, M. C. Becerra, F. A. A. Viru, C. A. Bonilla, E. Sanchez, D. Guerra, M. Munoz, K. Llaro, E. Palacios *et al.*, "Time to culture conversion and regimen composition in multidrug-resistant tuberculosis treatment," *PloS one*, vol. 9, no. 9, p. e108035, 2014.

[11] E. Cambau and M. Drancourt, "Steps towards the discovery of mycobacterium tuberculosis by robert koch, 1882," *Clinical Microbiology and Infection*, vol. 20, no. 3, pp. 196–201, 2014.

[12] A. Sakula, "Robert koch: centenary of the discovery of the tubercle bacillus, 1882." *Thorax*, vol. 37, no. 4, pp. 246–251, 1982.

[13] J. A. Caminero, A. Matteelli, and R. Loddenkemper, "Tuberculosis: are we making it incurable?" 2013.

[14] B. Saviola and W. Bishai, "The genus mycobacterium–medical," in *The Prokaryotes*. Springer, 2006, pp. 919–933.

[15] J. P. EUZéBY, "List of bacterial names with standing in nomenclature: a folder available on the internet," *International Journal of Systematic and Evolutionary Microbiology*, vol. 47, no. 2, pp. 590–592, 1997.

[16] R. Brosch, S. V. Gordon, M. Marmiesse, P. Brodin, C. Buchrieser, K. Eiglmeier, T. Garnier, C. Gutierrez, G. Hewinson, K. Kremer *et al.*, "A new evolutionary scenario for the mycobacterium tuberculosis complex," *Proceedings of the national academy of Sciences*, vol. 99, no. 6, pp. 3684–3689, 2002.

[17] J. Liu, C. E. Barry, G. S. Besra, and H. Nikaido, "Mycolic acid structure determines the fluidity of the mycobacterial cell wall," *Journal of Biological Chemistry*, vol. 271, no. 47, pp. 29 545–29 551, 1996.

[18] A. J. Caulfield and N. L. Wengenack, "Diagnosis of active tuberculosis disease: From microscopy to molecular techniques," *Journal of Clinical Tuberculosis and Other Mycobacterial Diseases*, vol. 4, pp. 33–43, aug 2016.

[19] F. Coll, R. McNerney, J. A. Guerra-Assunção, J. R. G. J. Perdigao, M. Viveiros, I. Portugal, A. Pain, N. Martin, and T. G. Clark, "A robust snp barcode for typing mycobacterium tuberculosis complex strains," *Nature Communications*, vol. 5, p. 4812, Sep 2014. [Online]. Available: https://www.nature.com/articles/ncomms5812#supplementary-information

[20] I. K. Neonakis, Z. Gitti, E. Petinaki, S. Maraki, and D. A. Spandidos, "Evaluation of the genotype mtbc assay for differentiating 120 clinical mycobacterium tuberculosis complex isolates," *European Journal of Clinical Microbiology & Infectious Diseases*, vol. 26, no. 2, pp. 151–152, Feb 2007. [Online]. Available: https://doi.org/10.1007/s10096-007-0255-y

[21] S. Cole, R. Brosch, J. Parkhill, T. Garnier, C. Churcher, D. Harris, S. Gordon, K. Eiglmeier, S. Gas, C. Barry Iii *et al.*, "Deciphering the biology of mycobacterium tuberculosis from the complete genome sequence," *Nature*, vol. 393, no. 6685, p. 537, 1998.

[22] J. M. Grange and C. H. Collins, "Tuberculosis and the cow," *Journal of the Royal Society of Health*, vol. 117, no. 2, pp. 119–122, 1997. [Online]. Available: https://doi.org/10.1177/146642409711700210

[23] G. B. Migliori, G. De Iaco, G. Besozzi, R. Centis, and D. M. Cirillo, "First tuberculosis cases in italy resistant to all tested drugs," *Weekly releases (1997–2007)*, vol. 12, no. 20, 2007. [Online]. Available: https://www.eurosurveillance.org/content/10.2807/esw.12.20.03194-en

[24] T. F. Brewer, "Preventing tuberculosis with bacillus calmette-guérin vaccine: A meta-analysis of the literature," *Clinical Infectious Diseases*, vol. 31, no. 3, pp. S64–S67, 2000. [Online]. Available: http://dx.doi.org/10.1086/314072

[25] M. Castets, H. Boisvert, F. Grumbach, M. Brunel, and N. Rist, "Tuberculosis bacilli of the african type: preliminary note," *Revue de tuberculose et de pneumologie*, vol. 32, no. 2, p. 179—184, March 1968. [Online]. Available: http://europepmc.org/abstract/MED/4985104

[26] S. D. Bentley, I. Comas, J. M. Bryant, D. Walker, N. H. Smith, S. R. Harris, S. Thurston, S. Gagneux, J. Wood, M. Antonio, M. A. Quail, F. Gehre, R. A. Adegbola, J. Parkhill, and B. C. de Jong, "The genome of mycobacterium africanum west african 2 reveals a lineage-specific locus and genome erosion common to the m. tuberculosis complex," *PLOS Neglected Tropical Diseases*, vol. 6, no. 2, pp. 1–11, 02 2012. [Online]. Available: https://doi.org/10.1371/journal.pntd.0001552

[27] S. Mostowy, A. Onipede, S. Gagneux, S. Niemann, K. Kremer, E. P. Desmond, M. Kato-Maeda, and M. Behr, "Genomic analysis distinguishes mycobacterium africanum," *Journal of Clinical Microbiology*, vol. 42, no. 8, pp. 3594–3599, 2004. [Online]. Available: https://jcm.asm.org/content/42/8/3594

[28] D. Van Soolingen, T. Hoogenboezem, P. E. W. De Haas, P. W. M. Hermans, M. A. Koedam, K. S. Teppema, P. J. Brennan, G. S. Besra, F. Portaels, J. Top, L. M. Schouls, and J. D. A. Van Embden, "A novel pathogenic taxon of the mycobacterium tuberculosis complex, canetti: Characterization of an exceptional isolate from africa," *International Journal of Systematic and Evolutionary Microbiology*, vol. 47, no. 4, pp. 1236–1245, 1997. [Online]. Available: http://ijs.microbiologyresearch.org/content/journal/ijsem/10.1099/00207713-47-4-1236

[29] C. C. Frota, D. M. Hunt, R. S. Buxton, L. Rickman, J. Hinds, K. Kremer, D. van Soolingen, and M. J. Colston, "Genome structure in the vole bacillus, mycobacterium microti, a member of the mycobacterium tuberculosis complex with a low virulence for humans," *Microbiology*, vol. 150, no. 5, pp. 1519–1527, 2004. [Online]. Available: http://mic.microbiologyresearch.org/content/journal/micro/10.1099/mic.0.26660-0

[30] D. van Soolingen, A. G. M. van der Zanden, P. E. W. de Haas, G. T. Noordhoek, A. Kiers, N. A. Foudraine, F. Portaels, A. H. J. Kolk, K. Kremer, and J. D. A. van

Embden, "Diagnosis of mycobacterium microtiinfections among humans by using novel genetic markers," *Journal of Clinical Microbiology*, vol. 36, no. 7, pp. 1840–1845, 1998. [Online]. Available: https://jcm.asm.org/content/36/7/1840

[31] D. Cousins, S. Williams, R. Reuter, D. Forshaw, B. Chadwick, D. Coughran, P. Collins, and N. Gales, "Tuberculosis in wild seals and characterisation of the seal bacillus," *Australian Veterinary Journal*, vol. 70, no. 3, pp. 92–97, 1993.

[32] P. J. Thompson, D. V. Cousins, B. L. Gow, D. M. Collins, B. H. Williamson, and H. T. Dagnia, "Seals, seal trainers, and mycobacterial infection," *American Journal of Respiratory and Critical Care Medicine*, vol. 147, no. 1, pp. 164–167, 1993.

[33] T. Kubica, S. Rüsch-Gerdes, and S. Niemann, "Mycobacterium bovis subsp. caprae caused one-third of human m. bovis-associated tuberculosis cases reported in germany between 1999 and 2001," *Journal of Clinical Microbiology*, vol. 41, no. 7, pp. 3070–3077, 2003. [Online]. Available: https://jcm.asm.org/content/41/7/3070

[34] A. [U+263A]Aranaz, D. Cousins, A. Mateos, and L. Domínguez, "Elevation of mycobacterium tuberculosis subsp. caprae aranaz et al. 1999 to species rank as mycobacterium caprae comb. nov., sp. nov." *International Journal of Systematic and Evolutionary Microbiology*, vol. 53, no. 6, pp. 1785–1789, 2003. [Online]. Available: http://ijs.microbiologyresearch.org/content/journal/ijsem/10.1099/ijs.0.02532-0

[35] R. Firdessa, S. Berg, E. Hailu, E. Schelling, B. Gumi, G. Erenso, E. Gadisa, T. Kiros, M. Habtamu, J. Hussein *et al.*, "Mycobacterial lineages causing pulmonary and extrapulmonary tuberculosis, ethiopia," *Emerging infectious diseases*, vol. 19, no. 3, p. 460, 2013.

[36] N. E. Mikheecheva, M. V. Zaychikova, A. V. Melerzanov, and V. N. Danilenko, "A nonsynonymous snp catalog of mycobacterium tuberculosis virulence genes and its use for detecting new potentially virulent sublineages," *Genome biology and evolution*, vol. 9, no. 4, pp. 887–899, 2017.

[37] I. Comas, M. Coscolla, T. Luo, S. Borrell, K. E. Holt, M. Kato-Maeda, J. Parkhill, B. Malla, S. Berg, G. Thwaites *et al.*, "Out-of-africa migration and neolithic coexpansion of mycobacterium tuberculosis with modern humans," *Nature genetics*, vol. 45, no. 10, p. 1176, 2013.

[38] G. Delogu, M. Sali, and G. Fadda, "The biology of mycobacterium tuberculosis infection," *Mediterranean journal of hematology and infectious diseases*, vol. 5, no. 1, 2013.

[39] S. H. Kaufmann and A. J. McMichael, "Annulling a dangerous liaison: vaccination strategies against aids and tuberculosis," *Nature medicine*, vol. 11, no. 4s, p. S33, 2005.

[40] G. R. Stewart, B. D. Robertson, and D. B. Young, "Tuberculosis: a problem with persistence," *Nature Reviews Microbiology*, vol. 1, no. 2, pp. 97–105, nov 2003.

[41] S. Verver, R. M. Warren, N. Beyers, M. Richardson, G. D. van der Spuy, M. W. Borgdorff, D. A. Enarson, M. A. Behr, and P. D. van Helden, "Rate of reinfection tuberculosis after successful treatment is higher than rate of new tuberculosis," *American journal of respiratory and critical care medicine*, vol. 171, no. 12, pp. 1430–1435, 2005.

[42] M. Silva Miranda, A. Breiman, S. Allain, F. Deknuydt, and F. Altare, "The tuberculous granuloma: an unsuccessful host defence mechanism providing a safety shelter for the bacteria?" *Clinical and Developmental Immunology*, vol. 2012, 2012.

[43] D. A. Rasko, M. J. Rosovitz, G. S. A. Myers, E. F. Mongodin, W. F. Fricke, P. Gajer, J. Crabtree, M. Sebaihia, N. R. Thomson, R. Chaudhuri, I. R. Henderson, V. Sperandio, and J. Ravel, "The pangenome structure of escherichia coli: Comparative genomic analysis of e. coli commensal and pathogenic isolates," *Journal of Bacteriology*, vol. 190, no. 20, pp. 6881–6893, aug 2008.

[44] M. A. Espinal, S. J. Kim, P. G. Suarez, K. M. Kam, A. G. Khomenko, G. B. Migliori, J. Baéz, A. Kochi, C. Dye, and M. C. Raviglione, "Standard short-course chemotherapy for drug-resistant tuberculosis," *JAMA*, vol. 283, no. 19, p. 2537, may 2000.

[45] N. R. Gandhi, P. Nunn, K. Dheda, H. S. Schaaf, M. Zignol, D. van Soolingen, P. Jensen, and J. Bayona, "Multidrug-resistant and extensively drug-resistant tuberculosis: a threat to global control of tuberculosis," *The Lancet*, vol. 375, no. 9728, pp. 1830–1843, may 2010.

[46] N. A. Martinson and R. E. Chaisson, "Survival in XDR TB: Shifting the curve and shifting the paradigm," *JAIDS Journal of Acquired Immune Deficiency Syndromes*, vol. 57, no. 2, pp. 89–91, jun 2011.

[47] J. G. Pasipanodya and T. Gumbo, "A new evolutionary and pharmacokinetic–pharmacodynamic scenario for rapid emergence of resistance to single and multiple anti-tuberculosis drugs," *Current Opinion in Pharmacology*, vol. 11, no. 5, pp. 457–463, oct 2011.

[48] P. E. A. D. Silva and J. C. Palomino, "Molecular basis and mechanisms of drug resistance in mycobacterium tuberculosis: classical and new drugs," *Journal of Antimicrobial Chemotherapy*, vol. 66, no. 7, pp. 1417–1430, may 2011.

[49] G. E. Louw, R. M. Warren, N. C. G. van Pittius, C. R. E. McEvoy, P. D. V. Helden, and T. C. Victor, "A balancing act: Efflux/influx in mycobacterial drug resistance," *Antimicrobial Agents and Chemotherapy*, vol. 53, no. 8, pp. 3181–3189, may 2009.

[50] R. C. MacLean, A. R. Hall, G. G. Perron, and A. Buckling, "The population genetics of antibiotic resistance: integrating molecular mechanisms and treatment contexts," *Nature Reviews Genetics*, vol. 11, no. 6, pp. 405–414, jun 2010.

[51] S. Gagneux, "The competitive cost of antibiotic resistance in mycobacterium tuberculosis," *Science*, vol. 312, no. 5782, pp. 1944–1946, jun 2006.

[52] S. Borrell and S. Gagneux, "Strain diversity, epistasis and the evolution of drug resistance in mycobacterium tuberculosis," *Clinical Microbiology and Infection*, vol. 17, no. 6, pp. 815–820, jun 2011.

[53] M. A. Behr, "Evolution of mycobacterium tuberculosis," in *The New Paradigm of Immunity to Tuberculosis*. Springer, 2013, pp. 81–91.

[54] J. Brookfield, "Evolution: What determines the rate of sequence evolution?" *Current Biology*, vol. 10, no. 11, pp. R410 – R411, 2000. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0960982200005066

[55] S. Choudhuri, "Phylogenetic analysis," in *Bioinformatics for Beginners*. Elsevier, 2014, pp. 209–218.

[56] R. R. Sokal, "Numerical taxonomy," *Scientific American*, vol. 215, no. 6, pp. 106–117, 1966. [Online]. Available: http://www.jstor.org/stable/24931358

[57] R. R. Sokal and F. J. Rohlf, "The comparison of dendrograms by objective methods," *Taxon*, vol. 11, no. 2, p. 33, feb 1962.

[58] M. N. N Saitou, "The neighbor-joining method: a new method for reconstructing phylogenetic trees." *Molecular Biology and Evolution*, jul 1987.

[59] J. S. Farris, "Methods for computing wagner trees," *Systematic Biology*, vol. 19, no. 1, pp. 83–92, mar 1970.

[60] J. Aldrich, "R.a. fisher and the making of maximum likelihood 1912-1922," *Statistical Science*, vol. 12, no. 3, pp. 162–176, sep 1997.

[61] E. T. Jaynes, "Bayesian methods: General background," 1986.

[62] Z. Yang, "PAML 4: Phylogenetic analysis by maximum likelihood," *Molecular Biology and Evolution*, vol. 24, no. 8, pp. 1586–1591, apr 2007.

[63] R. Hershberg, M. Lipatov, P. M. Small, H. Sheffer, S. Niemann, S. Homolka, J. C. Roach, K. Kremer, D. A. Petrov, M. W. Feldman, and S. Gagneux, "High functional diversity in mycobacterium tuberculosis driven by genetic drift and human demography," *PLOS Biology*, vol. 6, no. 12, pp. 1–14, 12 2008. [Online]. Available: https://doi.org/10.1371/journal.pbio.0060311

[64] T. Wirth, F. Hildebrand, C. Allix-Béguec, F. Wölbeling, T. Kubica, K. Kremer, D. van Soolingen, S. Rüsch-Gerdes, C. Locht, S. Brisse, A. Meyer, P. Supply, and S. Niemann, "Origin, spread and demography of the mycobacterium tuberculosis complex," *PLOS Pathogens*, vol. 4, no. 9, pp. 1–10, 09 2008. [Online]. Available: https://doi.org/10.1371/journal.ppat.1000160

[65] I. Filliol, A. S. Motiwala, M. Cavatore, W. Qi, M. H. Hazbón, M. Bobadilla del Valle, J. Fyfe, L. García-García, N. Rastogi, C. Sola, T. Zozio, M. I. Guerrero, C. I. León, J. Crabtree, S. Angiuoli, K. D. Eisenach, R. Durmaz, M. L. Joloba, A. Rendón, J. Sifuentes-Osornio, A. Ponce de León, M. D. Cave, R. Fleischmann, T. S. Whittam, and D. Alland, "Global phylogeny of mycobacterium tuberculosis based on single nucleotide polymorphism (snp) analysis: Insights into tuberculosis evolution, phylogenetic accuracy of other dna fingerprinting systems, and recommendations for a minimal standard snp set," *Journal of Bacteriology*, vol. 188, no. 2, pp. 759–772, 2006. [Online]. Available: https://jb.asm.org/content/188/2/759

[66] J. van Embden, M. Cave, J. Crawford, J. Dale, K. Eisenach, B. Gicquel, P. Hermans, C. Martin, R. McAdam, and T. Shinnick, "Strain identification of mycobacterium tuberculosis by dna fingerprinting: recommendations for a standardized methodology," *Journal of clinical microbiology*, vol. 31, no. 2, p. 406—409, February 1993. [Online]. Available: http://europepmc.org/articles/PMC262774

[67] S. Singh, H. Salamon, C. J. Lahti, M. Farid-Moyer, and P. M. Small, "Use of pulsed-field gel electrophoresis for molecular epidemiologic and population genetic studies of mycobacterium tuberculosis." *Journal of clinical microbiology*, vol. 37 6, pp. 1927–31, 1999.

[68] S. T. Cole, "Learning from the genome sequence of mycobacterium tuberculosis h37rv," *FEBS Letters*, vol. 452, no. 1, pp. 7–10, 1999. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0014579399005360

[69] S. Gordon, R. Brosch, A. Billault, T. Garnier, K. Eiglmeier, and S. Cole, "Identification of variable regions in the genomes of tubercle bacilli using bacterial artificial chromosome arrays," *Molecular microbiology*, vol. 32, pp. 643–55, 1999.

[70] M. A. Behr, M. A. Wilson, W. P. Gill, H. Salamon, G. K. Schoolnik, S. Rane, and P. M. Small, "Comparative genomics of bcg vaccines by whole-genome dna microarray," *Science*, vol. 284, no. 5419, pp. 1520–1523, 1999. [Online]. Available: http://science.sciencemag.org/content/284/5419/1520

[71] C. Huang, R.-Q. Li, and W.-J. Zhang, "Screening of mycobacterium tuberculosis distinctive genes by suppression subtractive hybridization technique," vol. 29, pp. 507–512, 01 2009.

[72] P. R. Jungblut, U. E. Schaible, H.-J. Mollenkopf, U. Zimny-Arndt, B. Raupach, J. Mattow, P. Halada, S. Lamer, K. Hagens, and S. H. E. Kaufmann, "Comparative proteome analysis of mycobacterium tuberculosis and mycobacterium bovis bcg strains: towards functional genomics of microbial pathogens," *Molecular Microbiology*, vol. 33, no. 6, pp. 1103–1117. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1365-2958.1999.01549.x

[73] R. C Edgar, "Muscle: Multiple sequence alignment with high accuracy and high throughput," vol. 32, pp. 1792–7, 02 2004.

[74] S. Kumar, M. Nei, J. Dudley, and K. Tamura, "Mega: A biologist-centric software for evolutionary analysis of dna and protein sequences," *Briefings in Bioinformatics*, vol. 9, no. 4, pp. 299–306, 2008.

[75] S. Capella-Gutiérrez, J. M. Silla-Martínez, and T. Gabaldón, "trimal: a tool for automated alignment trimming in large-scale phylogenetic analyses," *Bioinformatics*, vol. 25, no. 15, pp. 1972–1973, 2009. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btp348

[76] A. Stamatakis, "Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies," *Bioinformatics*, vol. 30, no. 9, pp. 1312–1313, 2014. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btu033

[77] B. Molina-Moya, M. Agonafir, S. R. Blanco, R. Dacombe, M. K. Gomgnimbou, L. Spinasse, M. Gomes-Fernandes, D. G. Datiko, T. Edwards, L. E. Cuevas, J. Dominguez, and C. Sola, "Microbead-based spoligotyping of mycobacterium tuberculosis from ziehl-neelsen-stained microscopy preparations in ethiopia," in *Scientific Reports*, 2018.

[78] J. P. Huelsenbeck and F. Ronquist, "MRBAYES: Bayesian inference of phylogenetic trees," *Bioinformatics*, vol. 17, no. 8, pp. 754–755, aug 2001.

[79] D. R. Maddison, D. L. Swofford, and W. P. Maddison, "Nexus: An extensible file format for systematic information," *Systematic Biology*, vol. 46, no. 4, pp. 590–621, dec 1997.

[80] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of state calculations by fast computing machines," *The Journal of Chemical Physics*, vol. 21, no. 6, pp. 1087–1092, jun 1953.

[81] F. Ronquist, M. Teslenko, P. van der Mark, D. L. Ayres, A. Darling, S. Höhna, B. Larget, L. Liu, M. A. Suchard, and J. P. Huelsenbeck, "MrBayes 3.2: Efficient bayesian phylogenetic inference and model choice across a large model space," *Systematic Biology*, vol. 61, no. 3, pp. 539–542, feb 2012.

[82] A. Stamatakis, "Raxml-vi-hpc: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models," *Bioinformatics*, vol. 22, no. 21, pp. 2688–2690, 2006. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btl446

[83] D. Swofford, "Paup*. phylogenetic analysis using parsimony (*and other methods)," *Nature Biotechnology*, vol. 18, pp. 233–234, 2003.

[84] N. T. Khan, "Mega - core of phylogenetic analysis in molecular evolutionary genetics," *Journal of Phylogenetics Evolutionary Biology*, vol. 5, no. 2, pp. 1–4, 2017. [Online]. Available: https://www.omicsonline.org/open-access/mega--core-of-phylogenetic-analysis-in-molecular-evolutionary-genetics-2329-9002-1000183.php?aid=93268

[85] J. Felsenstein, "Evolutionary trees from dna sequences: A maximum likelihood approach," *Journal of Molecular Evolution*, vol. 17, no. 6, pp. 368–376, Nov 1981. [Online]. Available: https://doi.org/10.1007/BF01734359

[86] J. Phelan, F. Coll, I. Bergval, R. Anthony, R. Warren, S. Sampson, N. Gey van Pittius, J. R Glynn, A. Crampin, A. Alves, T. Barbosa, S. Campino, K. Dheda, L. Grandjean, R. Hasan, Z. Hasan, A. Miranda, D. Moore, S. Panaiotov, and T. Clark, "Recombination in pe/ppe genes contributes to genetic variation in mycobacterium tuberculosis lineages," *BMC genomics*, vol. 17, p. 151, 02 2016.

[87] J. Becq, M. C. Gutierrez, V. Rosas-Magallanes, J. Rauzier, B. Gicquel, O. Neyrolles, and P. Deschavanne, "Contribution of horizontally acquired genomic islands to the evolution of the tubercle bacilli," *Molecular biology and evolution*, vol. 24, no. 8, pp. 1861–1871, 2007.

[88] E. V. Koonin, K. S. Makarova, and L. Aravind, "Horizontal gene transfer in prokaryotes: Quantification and classification," *Annual Review of Microbiology*, vol. 55, no. 1, pp. 709–742, 2001. [Online]. Available: https://doi.org/10.1146/annurev.micro.55.1.709

[89] A. Panda, M. Drancourt, T. Tuller, and P. Pontarotti, "Genome-wide analysis of horizontally acquired genes in the genus mycobacterium," *Scientific Reports*, vol. 8, no. 1, pp. 2045–2322, 2018. [Online]. Available: https://doi.org/10.1038/s41598-018-33261-w

A

PYTHON CODE

The methods created throughout this work for the pipeline are listed below. The inputs and outputs of the most relevant methods were described throughout the methodology.

```
#libraries
import os
import requests
import subprocess
import re
from Bio.Align.Applications import MuscleCommandline
from Bio import AlignIO
from Bio import SeqIO
from ete3 import NCBITaxa
from ete3 import Tree
from ete3 import EvolTree
ncbi = NCBITaxa()
import xlsxwriter as xls


#################     Prot and DNA
##open ids
#creates a file with the ids obtained from blast
#filename - file obtained from blast with hits found
def open_id_blast(filename, sep = "\t"):
    file = open(filename)
    lines = file.readline()
    ids = []
    while lines[0] == "#":
        lines = file.readline()
    outFile = open('ids_' + filename, 'w')
    while lines[0] != "#":
        lines = lines.split(sep)
        id_s = lines[0]            #gb|OMH53778.1|
        #idd = id_s[-11:-3]        #OMH53778
        lines = file.readline()
        if id_s not in ids:
            print(id_s)
            ids.append(id_s)
            outFile.write(id_s + '\n')
    return(ids)
```

```
#get the seqs dna from NCBI
#fasta - file with the ids of which the sequences are wanted
#file - only serves to name the output
def get_dna(fasta, file):
    with open('dna_ruim_' + file, 'w') as out_file:
        id_list = open (fasta)
        id_list = [i for i in id_list if i != '']  # lista de IDs
        for uni_id in id_list:
            data = requests.get("https://www.ncbi.nlm.nih.gov/sviewer/viewer.cgi?
                                tool=portal&save=file&log$=seqview&db=
                                protein&report=fasta_cds_na&id="
                                + uni_id + "&extrafeat=984&conwithfeat=on")
            x = data.text
            uni_id = uni_id.replace('.', '_')
            out_file.write("#" + uni_id + x)

#only >ids and seqs
#ruim_dna - file obtained from ncbi with the sequences
#file - only serves to name the output
def dna_fasta(ruim_dna, file):
    with open(ruim_dna,"r") as input:
        with open('dna_' + file,"w") as output:
            for line in input:
                if line[0]!=">" or "#" in line:
                    line=line.replace("#",">")
                    output.write(line)
```

```
#get the seqs prot from NCBI
#fasta - file with the ids of which the sequences are wanted;
#file - only serves to name the output;
def get_prot(fasta, file):
    with open('prot_ruim_' + file, 'w') as out_file:
        id_list = open (fasta)
        id_list = [i for i in id_list if i != '']  # lista de IDs
        for uni_id in id_list:
            data = requests.get("https://www.ncbi.nlm.nih.gov/sviewer/viewer.cgi?
                                tool=portal&save=file&log$=seqview&db=protein&report=
                                fasta&id=" + uni_id + "&extrafeat=984&conwithfeat=on")
            x = data.text
            out_file.write(x)

#only >ids and seqs
#ruim_prot - file obtained from ncbi with the sequences;
#file - only serves to name the output;
def prot_fasta(ruim_prot, file):
    seq = SeqIO.parse(open(ruim_prot), 'fasta')
    seqs = []
    with open('prot_' + file[0:5] +'.out',"w") as output:
        for fasta in seq:
            name, sequence = fasta.id, fasta.seq.tostring()
            if name not in seqs:
                seqs.append(name)
                if name in seqs:
                    name = name.replace('.','_')
                    output.write('>' + name +'\n' + sequence + '\n')
```

```
####alinhamentos
#alignes protein sequences through Muscle (.aln);
#fastaa - file fasta that contains for the query and each associated subject a line that
#starts with the ">" sign following the identifier and in the next line the protein sequence;
#file - only serves to name the output;
def muscle(fastaa, file):
    muscle_exe = r"C:\Program Files (x86)\Muscle\muscle3.8.31_i86win32.exe"
    cmdline = MuscleCommandline(muscle_exe, input=fastaa, out='aln_'+fastaa+
                                '.aln' , clw=True)
    cmdline()

##aln_prot --> phylip
#converts .aln format files to .phylip;
#aln - file obtained through Muscle with the sequences of the aligned proteins in .aln format;
def alignIO_prot_phy(aln):
        align = AlignIO.read(aln, "clustal")
        alignn = AlignIO.convert(aln, "clustal", aln+".phylip", "phylip")

##aln_prot to aln_dna
#through the aligned sequence and the DNA sequence, makes the changes from protein to DNA
#considering the passage of amino acids to codons;
#aln - file obtained through Muscle with the sequences of the aligned proteins in .aln format;
#dna - file with the DNA sequences, containing for the query and each associated subject a line
#that begins with the ">" sign following the identifier and, in the line, following the
#DNA sequence;
def aln_to_dna(aln, dna, sep = '\n'):
    file1 = SeqIO.parse(aln, "phylip")
    file2 = SeqIO.parse(dna, "fasta")
    with open('aln_' + dna + '.aln', 'w') as out_file:
        seqs1 = [(seq1.id, str(seq1.seq)) for seq1 in file1]  # protein
        seqs2 = [(seq2.id[-11:-1], str(seq2.seq)) for seq2 in file2]  # DNA
        seqs1.sort()
        seqs2.sort()
        out_dic = dict([(i[0],'') for i in seqs1])
        for i in range(len(seqs1)):
            seq1 = seqs1[i][1]
            seq2 = seqs2[i][1]
            k = 0
            for j in range(0, len(seq1)):
                if seq1[j] == '-':
                    out_dic[seqs1[i][0]] += '---'
                else:
                    out_dic[seqs1[i][0]] += seq2[k:k + 3]
                    k += 3
        for key,value in out_dic.items():
            key = key.replace('.','_')
            out_file.write('>' + key + '\n' + value + '\n')
```

```
#################     MEGA
##aln_prot -- > fasta
#converts .aln format files to .fasta;
#aln - file obtained through Muscle with the sequences of the aligned proteins in .aln format;
def alignIO_fas(aln):
    align = AlignIO.read(aln, "clustal")
    alignn = AlignIO.convert(aln, "clustal", 'fas_'+aln, "fasta")

#fasta --> File MEGA
#converts .fasta format files to .meg;
#aln - file obtained through Muscle with the sequences of the aligned proteins in .fasta format;
def fasta_mega(aln):
    f = open(aln, 'r')
    lines = f.readlines()
    lines.insert(0, '#Mega' + '\n')
    lines.insert(1, '!Title ' +aln+ ';'+ '\n')
    lines.insert(2, ' '+ '\n')
    f.close()
    f = open(aln, 'w')
    f.writelines(lines)
    f.close()
    with open(aln,"r") as input:
        with open(aln + '.meg',"w") as output:
            for line in input:
                line=line.replace(">","#")
                output.write(line)

##Mega --> tree
#reconstruction of the phylogenetic trees;
#file - file contains multiple sequence alignment in the MEGA format;
def mega_tree(file):
    file = file[12:18]
    cmd = "megacc","-a","infer_ML_amino_acid.mao","-d",file,"-o","tree_"+file+'.nwk'
    mega = subprocess.call(cmd, shell=True)
    return mega
```

```
#################     Midpoint Root
#midpoint root method was applied to each Newick file;
#out - Newick file with the tree;
def midpoint_root(out):
    R_exe= "C:\\Program Files\\R\\R-3.5.1\\bin\\Rscript"
    R_file = "C:\\Users\\Daniela Pereira\\Desktop\\tese\\mid.R"
    arg1 = "tree_"+ out[0:5] + ".nwk"
    arg2 = "mid_tree_"+out[0:5]+".nwk"
    output = subprocess.Popen([R_exe, R_file, arg1, arg2],
                              stdout=subprocess.PIPE).communicate()[0]
    output

#-------mid.R----
#library(phytools)
#setwd("D:/tese/new")

#args <- commandArgs(trailingOnly = TRUE)

#arg1 = args[1]
#arg2 = args[2]

#tree = read.newick(args[1])

#mpt1 = midpoint.root(tree)

#write.tree(mpt1,file = args[2])
#-----------------


#################     Lineage
##open tax_ids
#creates a file with the tax ids obtained from blast;
#filename - file obtained from blast with hits found;
def open_tax_id(filename):
    file = open(filename, mode = 'r')
    lines = file.readline()
    ids = {}
    outFile = open('tax_id_' + filename + '.fasta', 'w')
    while lines[0] == "#":
        lines = file.readline()
    while lines[0] != "#":
        lines = lines.split(sep = "\t")
        id_s = lines[0][-11:-1]
        id_s = id_s.replace('.', '_')
        tax_id = lines[6]
        ids[id_s] = tax_id
        lines = file.readline()
    return(ids)
```

```
#get the lineage
#algo_dic - Dictionary -> id: tax_id;
#file - only serves to name the output
def lineage(algo_dic, file):
    dic = algo_dic
    coryne = []
    myco = []
    mycoComp = []
    euka = []
    other = []
    outFile = open('lineage_' + file[0:5], 'w')
    for i in dic.values():
        lineage = ncbi.get_lineage(i)
        names = ncbi.get_taxid_translator(lineage)
        if 85007 in names.keys():  #'Corynebacteriales'
            coryne.append(i)
        if 1763 in names.keys():   #'Mycobacterium'
            myco.append(i)
        if 77643 in names.keys():  #'Mycobacterium tuberculosis complex'
            mycoComp.append(i)
        if 'Eukaryotic' in names.values():
            euka.append(i)
        else:
            other.append(i)
    for key,value in dic.items():
        taxid2name = ncbi.get_taxid_translator([value])
        for k in taxid2name.values():
            if value in coryne:
                if value in myco:
                    if value in mycoComp:
                        outFile.write(key +'='+ key +'||cor||myco||mtc||'+ k +'\n')
                    else:
                        outFile.write(key + '=' + key + '||cor||myco||' + k + '\n')
                else:
                    outFile.write(key + '=' + key + '||cor||' + k + '\n')
            else:
                outFile.write(key + '=' + key + '||' + k + '\n')
        for j in taxid2name.values():
            if value in euka:
                eukaFile = open('euka_' + file, 'w')
                eukaFile.write(key + '=' + key + '||euka||' + j + '\n')
            else:
                break
```

```
#change id's names on the tree
#tree - Newick file with the tree;
#lineage - file with the lineage information;
#file - only serves to name the output
def ids2names(tree, lineage, file):
    d = {}
    oh = open('names_' + file[0:5] + '.tree', 'w')
    fh = open(lineage)
    for line in fh:
        ids = line.split('=')[0]
        names = line.split('=')[1]
        names = names.replace('.','_')
        names = names.replace('coelicolor A3(2)', 'coelicolor A3')
        d[ids] = names
    fh.close()
    fh = open(tree)
    tree = fh.read()
    for key,value in d.items():
            tree = tree.replace(key, value)
            print(value)
    oh.write(tree)
    oh.close()
```

```
#################     Monophyly
#get monophyly
#tree - Newick file with the tree;
#algo_dic - Dictionary -> id: tax_id;
#file - only serves to name the output;
def get_monophyly(tree, algo_dic, file):
    t = Tree(tree)
    dic = algo_dic
    coryne = []
    myco = []
    mycoComp = []
    tuberculosis = []
    outFile = open('mono_' + file[0:5], 'w')
    for key,value in dic.items():
        lineage = ncbi.get_lineage(value)
        names = ncbi.get_taxid_translator(lineage)
        if 85007 in names.keys():  #'Corynebacteriales'
            coryne.append(key)
        if 1763 in names.keys():   #'Mycobacterium'
            myco.append(key)
        if 77643 in names.keys():  #'Mycobacterium tuberculosis complex'
            mycoComp.append(key)
        if 1773 in names.keys():   #'Mycobacterium tuberculosis'
            tuberculosis.append(key)
    with open(file) as f:
        line = f.readlines()[5]
        line = line.split(sep = "\t")
        idd = line[0][-11:-1]
    mono_mycoComp = t.check_monophyly(values=mycoComp, target_attr="name")
    if len(myco)==len(mycoComp):
        mono_myco = 'False'
    else:
        mono_myco = t.check_monophyly(values=myco, target_attr="name")
    if len(coryne)==len(myco):
        mono_cor= 'False'
    else:
        mono_cor = t.check_monophyly(values=coryne, target_attr="name")
    #outfile
    outFile.write(idd+'\t'+str(mono_mycoComp[0])+'\t'+str(mono_myco[0])+'\t'+
                  str(mono_cor[0])+'\t'+str(tuberculosis)+'\t'+str(mycoComp)+
                  '\t'+str(myco)+'\t'+str(coryne)+'\t'+str(len(t)))
```

```
#create a table in excel with the results of the monophyly
def creatTable(folder):
    outs = []
    outs += [each for each in os.listdir(folder) if each.endswith('.out')]

    workbook = xls.Workbook('monophyly.xlsx')
    worksheet = workbook.add_worksheet()
    worksheet.write('A1', 'out')
    worksheet.write('B1', 'gene')
    worksheet.write('C1', 'Mtc')
    worksheet.write('D1', 'Mycobacterium')
    worksheet.write('E1', 'Corynebacteriales')
    worksheet.write('F1', 'Have Mtb')
    worksheet.write('G1', 'Only cor?')
    worksheet.write('H1', 'Only myco?')
    worksheet.write('I1', 'Only Mtbc?')
    worksheet.write('J1', 'tree size')
    bold=workbook.add_format({'bold':True})
    row=1

    for out in outs:
        for line in open('mono_'+out[0:5],"r"):  #file with the result of the monophyly
            line = line.split(sep = "\t")
            gene = line[0]
            mono_mycoComp = line[1]
            mono_myco = line[2]
            mono_cor = line[3]
            tuber = line[4]
            myco = line[6]
            myco = myco.split(',')
            mycoComp = line[5]
            mycoComp = mycoComp.split(',')
            cor = line[7]
            cor = cor.split(',')
            tree = line[8]
            worksheet.write(row,0, out[0:-4])
            worksheet.write(row,1, gene)              #gene
            worksheet.write(row,2, mono_mycoComp[0]) #result of monophyly for Mtbc
            worksheet.write(row,3, mono_myco[0])     #result of monophyly for Mycobacterium
            worksheet.write(row,4, mono_cor[0])      #result of monophyly for Corynebacteriales
            worksheet.write(row,5, tuber)            #elements of Tuberculosis
            worksheet.write(row,7, len(myco))        #number of elements of Mycobacterium
            worksheet.write(row,8, len(mycoComp))    #number of elements of Mtbc
            worksheet.write(row,6, len(cor))         #number of elements of Corynebacteriales
            worksheet.write(row,9, tree)             #size of the tree
            row += 1
    workbook.close()
```

```
#################     Paml
#objective of the analysis is to determine which branches of the Mtb complex are under
#faster evolution

#Model 0
#means an index of omega for all branches
#monophyly - file with the result of the monophyly;
#tree - Newick file containing the tree;
#align - file with DNA alignment;
#workdire - workbook directory;
def codeml_M0(monophyly, tree, align, workdire):
    outFile = open('Evolutionary_Model_M0_' + tree[9:14], 'w')
    tree  =  EvolTree(tree)
    align=tree.link_to_alignment(align)
    tree.workdir = workdire
    mono = open(monophyly, mode = 'r')
    mono = mono.readline()
    mono = mono.split(sep = "\t")
    myco = mono[6]
    myco = myco.split(sep = ",")

    if mono[1] == 'True':
        if len(myco) != len(tree):
            tree.run_model ('M0')
            evol = tree.get_evol_model('M0')
            outFile.write(str(evol))
        else:
            print('only mycobacterium')
    else:
        print('monophyly of Mtbc = False')
```

```
#Model 2
#means an arbitrary number of proportions
#monophyly - file with the result of the monophyly;
#tree - Newick file containing the tree;
#align - file with DNA alignment;
#workdire - workbook directory;
#file - only serves to name the output;
def codeml_M2(monophyly, tree, align, workdire, file):
    outE = open('Evolutionary_Model_b_free_' + tree[9:14], 'w')
    tree  =  EvolTree(tree)
    outFile = open('tree_internal_' + file[0:5]+ '.nwk', 'w')
    align=tree.link_to_alignment(align)
    tree.workdir = workdire
    mono = open(monophyly, mode = 'r')
    mono = mono.readline()
    mono = mono.split(sep = "\t")
    tuber = mono[4]
    tuber = tuber.split(sep = ",")

    if mono[1] == 'True' and len(tuber) == 1:
        # mark a group of branches
        tree.mark_tree ([edge], marks=['#1'])
        outFile.write(tree.write())
        tree.run_model('b_free')
        evol = tree.get_evol_model('b_free')
        outE.write(str(evol))
    else:
        print('Has several tuberculosis')
```

```
#################     Get All
def get_all(folder):
    outs = []
    outs += [each for each in os.listdir(folder) if each.endswith('.out')]

    for out in outs:
        print(out)
        #prot and dna
        ids =open_id_blast(out)
        prot_ruim = get_prot('ids_'+out, out)
        dna_ruim = get_dna('ids_'+out, out)
        prot = prot_fasta(out+'.prot.fasta', out)
        dna = dna_fasta('dna_ruim_'+out, out)

        #alignments
        muscle_aln = muscle('prot_'+out+'.out', out)
        prot_phylip = alignIO_prot_phy('aln_prot_'+out+'.out.aln')
        aln_dna = aln_to_dna('aln_prot_'+out+'.aln.phylip','dna_'+out)

        #MEGA
        fasta = alignIO_fas('aln_prot_' + out + '.out.aln')
        meg = fasta_mega('fas_aln_prot_'+ out + '.out.aln')
        tree = mega_tree('fas_aln_prot_'+ out + '.out.aln.meg')

        #Midpoint Root
        mid_root = midpoint_root(out)

        #Lineage
        tax_id = open_tax_id(out)
        lin = lineage(tax_id, out)
        names = ids2names('mid_tree_' + out+'.nwk','lineage_' + out, out)

        #Monophyly
        get_monophyly('mid_tree_' + out +'.nwk' , tax_id, out)

        #Paml
        M0 = codeml_M0('mono_'+out, 'mid_tree_'+out+'.nwk', 'aln_dna_'+out+'.aln', r"D:/teste/teste")
        M2 = codeml_M2('mono_'+out, 'mid_tree_'+out+'.nwk', 'aln_dna_'+out+'.aln', r"D:/teste/teste", out)




if __name__ == '__main__':
    get_aln_all('out')
    creatTable('out')
```