Evolutionary systems biology of the *Mycobacterium tuberculosis* complex

Author: Álvaro Chiner Oms

SUPERVISORS
Iñaki Comas Espadas
Fernando González Candelas

PhD thesis
Doctoral Programme in Biomedicine and Biotechnology

Valencia, 2019

D. Iñaki Comas Espadas, Científico Titular del Instituto de Biomedicina de València (IBV-CSIC).

D. Fernando González Candelas, Catedrático de Genética de la Universitat de València.

CERTIFICAN: Que D. Álvaro Chiner Oms ha realizado bajo su dirección y supervisión el trabajo que lleva por título "Evolutionary systems biology of the *Mycobacterium tuberculosis* complex", para optar al Grado de doctor en Biomedicina y Biotecnología por la Universitat de València

Y para que conste, en el cumplimiento de la legislación presente, firman el presente certificado en València, a 20 de Mayo de 2019.

Dr. Iñaki Comas Espadas Director	Dr. Fernando González Candelas Director	Álvaro Chiner Oms Doctorando

Este es el resultado final de un camino de 4 años de estudio, trabajo y vivencias. No me hubiera sido posible recorrerlo (ni siquiera empezarlo) a mi solo. Asi pues, no puedo sino agradecer a aquellos que me han acompañado. Una lista detallada de agradecimientos necesitaría de varias páginas, y correría el riesgo de dejarme a alguien fuera. Directores, amigos, familiares, compañeros... solo puedo deciros GRACIAS, esto no hubiera sido posible sin vosotros.

Sin embargo, tampoco sería justo no reconocer a aquellos que han jugado un papel fundamental en el desarrollo de esta tesis.

Fernando, que fué el primero que me mostró que Informática y Biología eran dos términos que podían ir de la mano. Desde primero de carrera me empujó a ampliar mi formación fuera de las aulas de la universidad. Iñaki, que más que un jefe ha sido un maestro y un modelo a seguir. Ha sido mi guia desde el inicio del proyecto de tesis, y principal responsable de que este documento vea hoy la luz. Ninguno de los dos ha tenido nunca la puerta cerrada para mí. Me han dado la oportunidad de hacer cursos, asistir a congresos, dar docencia y establecer colaboraciones. No he podido tener mejores mentores durante estos años. Teresa, que no hizo mas que darme facilidades durante mi periodo en Londres y enseñarme los secretos del RNA-seq. Gracias a ella tuve una estancia más fructífera y cómoda de lo que inicialmente podía imaginarme.

A todos los miembros de TGU (los 'iñakis') y EPIMOL (los 'fernandos'). Los congresos, *meetings* y el día a día a vuestro lado también han formado parte de un camino que, gustosamente, estaría dispuesto a recorrer de nuevo.

A Rosa, mi otra mitad, depositaria de mis miedos, incertidumbres y alegrias durante todo este tiempo.

A mis padres, que desde siempre me han apoyado a todos los niveles.

A Dídac, que me alegra la vida.

Contents

General introduction

1.1 Tuberculosis

Tuberculosis (TB) is a highly contagious, airborne-transmitted disease that mainly affects the respiratory tract, although it can affect other body parts. According to the World Health Organization (WHO), TB is the leading cause of death by an infectious agent, ranking above AIDS/HIV. In 2017, the WHO estimated that 1.6 million people died due to the disease, and that 10 million were infected. The impact of the disease is geographically heterogeneous, with 8 countries (India, China, Indonesia, the Philippines, Pakistan, Nigeria, Bangladesh and South Africa) accounting for one third of the new annual cases [1] (Figure 1.1).

In developed countries, the incidence of the disease decreased substantially since the second half of the XXth century. As a consequence, the budget devoted to TB research diminished notably during this period. However, during the 80's and 90's, the incidence of TB in developed countries rose again, following the global expansion of AIDS. This sudden emergence of the disease was not predicted by public health offices. Thus, billions of dollars from emergency funds were used to control TB outbreaks particularly in large urban settings [2]. In addition, the low investment in basic research and development of specific treatments meant that global TB control laid on diagnostic tests, vaccines and treatments developed more than 50 years ago. The WHO stated

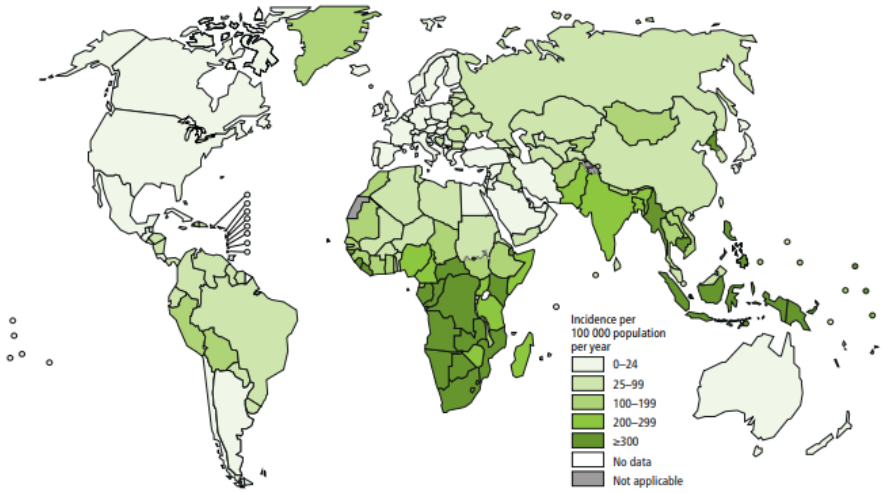


Figure 1.1: Global estimated incidence of TB in 2017. Source: WHO [1]

that the majority of TB fatalities could be prevented with an early diagnostic and an appropriate treatment. However, decades-old diagnostics and treatments are still used in many countries in which TB is one of the most important health issues [1].

Classically, the disease was clinically classified in two main forms: latent TB, which is asymptomatic and non-transmissible, and active TB, which is transmissible and can be found in two main presentations: pulmonary and extrapulmonary [3, 4]. However, recent studies suggest that classifying the clinical forms of the disease as binary (either latent or active TB) is an oversimplification of the real clinical scenarios [3, 5]. In reality, there is a whole spectrum of infection outcomes, with active TB showing different combinations of symptoms ranging from mild to severe in distinct patients. In addition, patients without TB symptoms could represent cases of latent TB or subclinical TB. Extrapulmonary tuberculosis accounted for 14% of the incident cases notified in 2017 [1] and its incidence is even higher among HIV-positive patients. In extrapulmonary tuberculosis, the disease disseminates in the patient's body, affecting many organs. In pulmonary tuberculosis, which is the most common clinical form of the disease, the disease affects the lungs. The

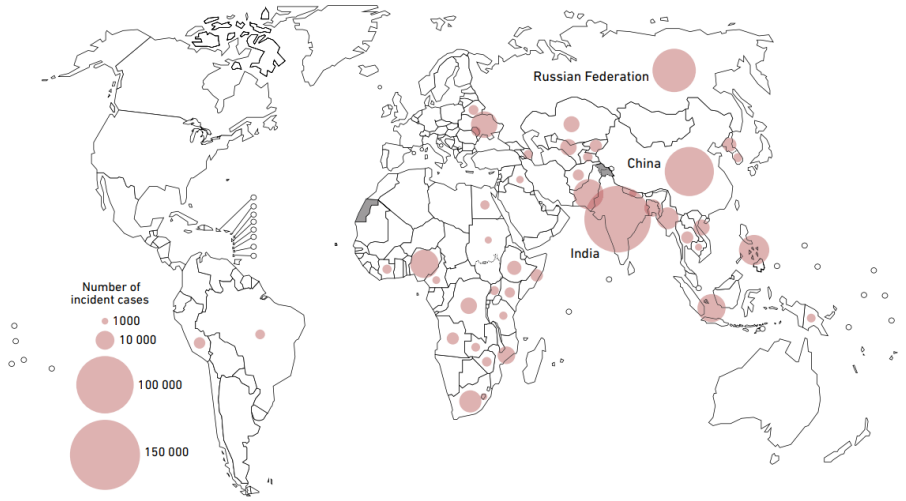


Figure 1.2: Estimated incidence of MDR and rifampicin resistance cases in 2017, for countries with at least 1,000 incident cases. Source:WHO [1]

symptoms include coughing with sputum and blood, chest pain, fever, night sweats and weight loss [6, 1].

The standard treatment for patients with active TB consists of a minimum of 6 months of therapy with a combination of four antibiotic drugs (rifampicin, isoniazid, pyrazinamide and ethambutol) [6]. Although this treatment has been used for decades, there is a current increase in TB-cases that do not respond to it. Cases not responding to rifampicin and isoniazid, the two most powerful anti-TB drugs, are classified as multidrug-resistant tuberculosis (MDR-TB) (Figure 1.2). In these cases, treatment relies on the use of second-line drugs (more expensive, toxic and less effective) and could last up to 2 years. More extreme are the extensively drug-resistance cases (XDR-TB) which, in addition to isoniazid and rifampicin, are resistant to several second-line anti-TB drugs. In those cases, treatment is personalized depending on the clinical history of the patient. Options in these cases include mixtures of more than 4 drugs and sometimes are limited to experimental therapies [7].

Latent TB is typically diagnosed by means of immunodiagnostic tests, that

report if the patient has been in contact with the infectious agent. These tests are applied to people identified as being at risk to develop active disease (previous contact with active TB patients, immunocompromised status,...) [8]. Further assays, such as smear microscopy from sputums (mainly in high burden countries) and chest X-rays, are performed if these tests are positive to confirm active TB cases [9]. At the end, the gold standard to confirm active TB is a positive culture of the bacilli isolated from the patient, typically from sputum [10]. These detection methodologies face several problems. First, immunodiagnostic tests could be positive in response to past exposures to the pathological agent, thus not reflecting recent contact. Second, sputum-based tests are not recommended in at least one third of the global cases because of several factors (extra-pulmonary infections, inability to obtain sputum from children, low bacillar concentrations,...) [9]. Third, the sputum culture could take around three weeks to show positive results. In addition, in 2017 44% of the global pulmonary infections were not bacteriologically confirmed by current methods [1]. So, the global diagnostic capacity is suboptimal. Currently, efforts are invested to solve the low sensitivity of classical methods [11], the long time needed to confirm a positive TB case [12] and the technical requirements that are not available in many of the high burden regions [13].

In 2014, the WHO started a strategy aimed at ending the TB world epidemic by 2035. Specifically, the strategy aspire to achieve a 95% reduction of mortality and 90% reduction of incidence by 2035 in comparison with 2015 [14]. This strategy is based on three main pillars: improving the protocols of prevention and patient attention, reinforcing the political actions and supporting systems, and intensifying research and innovation.

Despite the WHO goal, the global TB incidence is declining at a rate of only 2% per year. To meet the ENDTB objectives of WHO by 2025, we need to increase this rate to 4%-5% by 2020 and by 10% by 2020 to 2025 [15]. New diagnostic methods, strategies to interrupt transmission and more effective treatments are needed to fulfil this objective [16]. And all these improvements have a strong need of basic research.

1.2 The *Mycobacterium tuberculosis* complex

The main causative agent of TB in humans is a slow-growing mycobacteria called *Mycobacterium tuberculosis* [17]. In Africa, TB is also caused by a closely related bacterial lineage classically called *Mycobacterium africanum* [18]. In addition, several mammalian species are prone to be infected with mycobacteria that cause animal tuberculosis disease [19]. These mycobacteria are named in reference to their preferred host such as *M. bovis* (cows), *M. caprae* (goats), *M. microti* (rodents), *M. pinnipedii* (seals), *M. origys* (oryxes), *M. suricattae* (meerkats) and *M. mungi* (mongooses). All these mycobacteria that cause tuberculosis in animals together with *M. africanum* and *M. tuberculosis* form a monophyletic group called *Mycobacterium tuberculosis* complex (MTBC) [20].

The *Mycobacterium* genus comprises a relatively large number of species ($\sim$190) [21]. Most of them are free living organisms found in a wide range of environments, in contrast with *M. tuberculosis*, which is an obligate pathogen that is not able to survive outside the host environment [22]. Despite their saprophytic lifestyle, an important number of these mycobacterial species are capable of causing disease in humans (i.e. *M. kansasii*, *M. avium*, and *M. marinum*, among others). The increasing public availability of mycobacterial genomes has allowed us to draw a precise map of the phylogenetic relationships of this group, avoiding the limitations of classifications based on 16S rRNA reconstructions and phenotypes [21]. Average Nucleotide Identities (ANI) analyses have shown that some members previously considered as independent taxa could represent, in reality, different variants of the same species [23](understanding species as groups of individuals with genomic ANI values of 95% or higher [24]). On the light of the new genomic information available, Gupta *et al.*, [25] have proposed recently a new classification of the *Mycobacterium* genus by introducing a division of the group in 5 different genera.

Phylogenomic analysis places the MTBC near other slow-growers that are

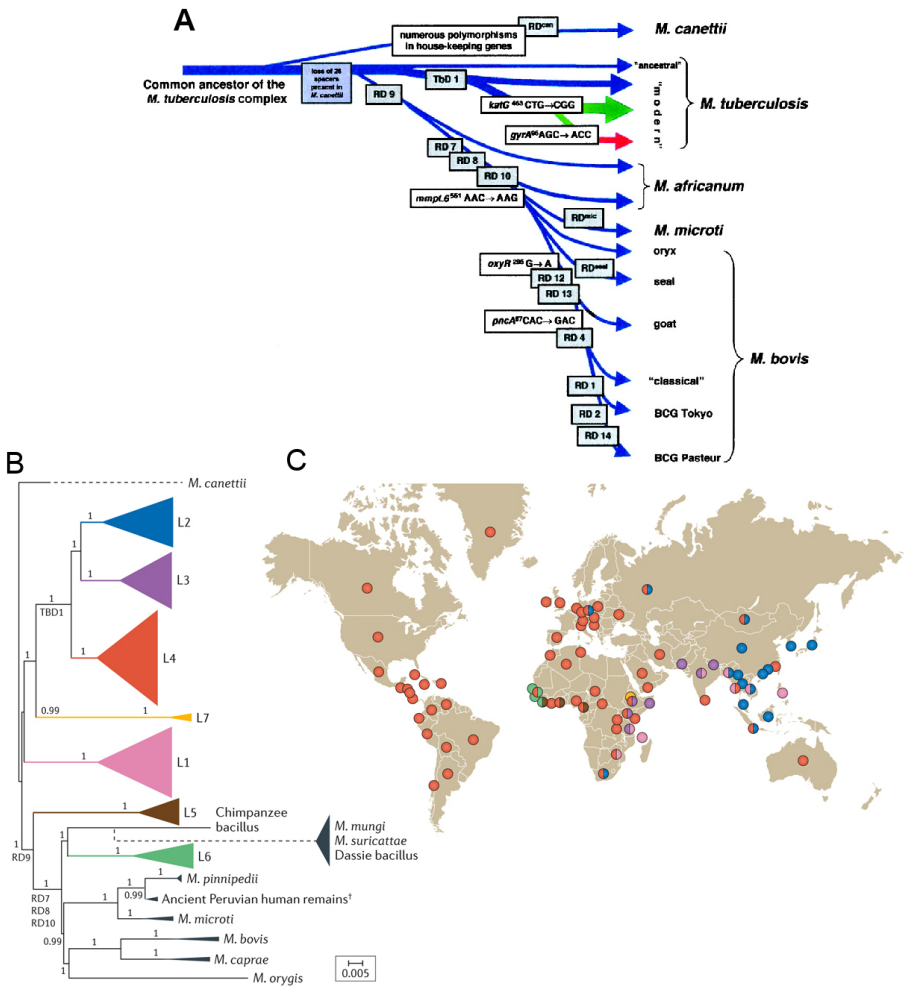


Figure 1.3: The *Mycobacterium tuberculosis* complex phylogeny. A) Proposed evolutionary pathway derived from the evaluation of the Regions of Difference between distinct MTBC members by Brosch *et al.* B) The MTBC phylogeny derived using complete genome sequences. C) Global distribution of the human-affecting clades of the MTBC. Source: Brosch *et al.*[26] 2002 and Gagneux 2018 [20]

associated with human diseases such as *M. leprae* and *M. ulcerans*. The closest known phylogenetic species to MTBC is *Mycobacterium canettii*. Despite its historical classification as a different species, the ANI between the

M. canettii group and *M. tuberculosis* is about 98% (range 97.71%-99.30%) ([27], our own data). Thus, the current consensus is to include *M. canettii* as part of the MTBC. However, in contrast with other members of the MTBC, *M. canettii* has a low epidemiological incidence, with less than 100 cases reported since its first isolation in 1969 [27] and it seems to have an environmental lifestyle [28]. In consequence, in the present thesis we will exclude *M. canettii* when we refer to the MTBC throughout the text. The MTBC (excluding *M. canettii*) comprises a group of bacteria with genome sequences having an ANI greater than 99% and sharing a single common ancestor. So, again, although the different members of the MTBC have distinct taxonomic names, all of them belong to the same genomic species [29]. Nevertheless, to facilitate readability, in the present thesis we will refer to the different strains of the MTBC by their classical names (i.e., *M. tuberculosis, M. africanum, M. bovis*, etc.)

The MTBC has a clonal population structure, consisting of seven human-adapted bacterial lineages (L1-7) and several animal-adapted strains (Figure 1.3). Focusing in the human-adapted strains, L1,L2,L3,L4 and L7 belong to *M. tuberculosis sensu stricto* whereas L5 and L6 belong to *M. africanum* [20]. The genetic diversity found among these lineages results from large genetic deletions and point mutations [30]. Classically, some of these large deleted regions (known as RD from Regions of Difference) have been used for strain typing and lineage identification as they are accurate phylogenetic markers with virtually absence of homoplasy [31, 26] (Figure 1.3A). The rationale behind this is that because of the MTBC clonal structure, some parts of the genome involved in past deletion events were never recovered by the descendants of the ancestral strain. The loss of these genomic regions had important implications in terms of bacterial pathogenicity [26]. In fact, the history of the BCG vaccine is a clear example of such a process [32, 33] where cumulative deleted regions have led to different immunogenic potential. Nowadays, the use of complete genome sequences has allowed us to drawn a more precise picture of the phylogenetic relationships of the MTBC (Figure 1.3B).

The different lineages are heterogeneously distributed around the globe

(Figure 1.3C). L1 is mainly found is southeast Africa, southeast Asia and India; L2 is widely distributed in east Asia; L3 is present in east Africa and India; L4 is the most widely distributed, affecting the whole Euro-American territory, with spots in Africa and Asia; L5 and L6 are restricted to specific regions of west Africa and L7 is only present in Ethiopia [30, 34, 35]. Regarding the animal-adapted clade, our knowledge is much more limited. The phylogenetic analyses shown that the animal-adapted strains and L6 share a common ancestor [26, 19]. However, little is known about its ecology and global distribution. In addition, the number of complete genomes available for these MTBC members is lower than those of the human-adapted strains. For these reasons, in the present thesis we have focused our analyses in the human-adapted strains of the MTBC.

Population genetics studies have shown that there has been parallel evolution between humans and the MTBC. The bacteria radiated in the distinct lineages and occupied different regions following human population changes and migrations, from ancient ages until present [36, 37]. Although the seven main lineages are deeply rooted in the phylogeny, L2, L3 and L4 are also known as 'modern lineages' because they diverged recently in comparison to the others, whereas L1, L5, L6 and L7 are called 'ancient lineages' [26]. It is also known that the maximum genetic distance between strains of different lineages is around 2,500 Single Nucleotide Polymorphisms (SNPs)(after elimination of hypervariable regions which are not usually considered in whole genome comparisons) [38, 39].

It has been hypothesized that modern lineages have evolved in scenarios with a high density of hosts. In these scenarios, the variants having lower latency time could be selected naturally. In contrast, ancient lineages appeared in low density populations and, as a result, they may have been selected for larger latency periods [40]. In 2011, Portevin *et al.* [41] selected 26 strains representing the global MTBC diversity and infected human macrophages and dendritic cells from different donors to evaluate the innate immune response to the different bacteria. They found that strains from modern lineages induced a

lower innate inflammatory response than ancient lineages, and these differences appeared even when infecting different donors. Another work, from de Jong *et al.* [42], was based on the monitorization of a cohort of patients and their relatives in The Gambia for two years. The authors found that patients infected with strains belonging to modern lineages had 3 times more chances to develop active TB than those infected with *M. africanum*. These studies, and some others [30, 38, 43, 44], show that the MTBC genetic diversity has epidemiological implications and genetic differences among lineages lead to differences in the immune response and disease progression in the host.

Besides the differences in virulence and latency time shown above, the MTBC genetic diversity seems to have implications in disease transmission too. For example, animal-adapted strains have undergone an evolutionary pathway different to the human-adapted strains [26]. The genetic differences between both type of strains, are enough to limit the transmission capacity of animal-adapted strains among humans [45].

As stated above, past studies have shown that there was likely co-evolution between the different MTBC lineages and human populations. A study by Gagneux *et al.* with epidemiological data of the city of San Francisco [30] showed that the MTBC lineages have host preferences related to their phylogeographical origin. For example, L1 is mainly found in southeast Asia, southeast Africa and India, whereas L2 is especially abundant in east Asia. Gagneux *et al.* showed that, in a mixed population with individuals from different geographical origins, L1 strains transmitted significantly better among individuals from southeast Asia (mainly The Philippines and Vietnam) whereas L2 strains transmitted better among individuals from east Asia (China). This fact was further confirmed by Reed *et al.*, [46], in a study following a similar approach in the city of Montreal. Although social factors could be responsible of such a pattern, this trend was not observed in HIV patients [47]. In consequence, the interaction of the host-immune system and the pathogen seems to be the cause of this preference for specific hosts. Moreover, a recent study by Stucki *et al.* [34] tried to identify the factors that lead to this

host-pathogen association by focusing on the differential distribution and adaptation to local human populations of L4 strains. L4 is the most widely distributed lineage and comprises 10 well-defined sublineages. Among these lineages, there is a subset that are globally distributed and showed a high transmission rate among different host populations. These lineages were called 'generalists' due to their cosmopolitan distribution. On the other hand, the so called 'specialists', are a set of L4 clades that are geographically restricted and show a high association with local human populations. These phenotypic differences seem to depend on genetic variants found in some specific antigens recognized by the host immune-system. Due to these variants, the 'generalists' clades are able to respond to a broader range of human leukocyte antigens, thus being capable of interacting with a larger diversity of hosts.

So, the distinct MTBC members show differences in their pathogenic characteristics linked to their genetic background. However, identifying specific genetic variants is much more challenging. As a result a global solution to the TB problem will probably need to incorporate the bacilli genetic diversity as a variable.

1.3 *Mycobacterium tuberculosis*, a professional pathogen

The MTBC shares several microbiological characteristics with other members of the *Mycobacterium* genus: (i) it possesses mycolic acids in its cell wall, a type of fatty acids that contain from 60 to 90 carbon atoms, (ii) it is acid fast [22], (iii) not only the MTBC, but also all the *Mycobacterium* species have a unique cell wall that confers them a high robustness to several stress conditions. In general, they are able to tolerate acidic environments, desiccation, heat, host-immune mechanisms and have a basal resistance to antibiotics [22, 48, 49]. These singular cell wall is also responsible for many virulence characteristics [50].

The infectious process starts when the bacterium gets into the host through

the respiratory tract. Once in the lung, it is phagocytosed by macrophages, triggering a slow inflammatory response, and forming a structure known as granuloma in distal sites of the lungs [51]. The bacteria can be dormant and survive inside the granuloma during months, years or even decades in an asymptomatic disease state called latency [52]. The transition from latency to an active disease state depends on biological features of the bacteria [53, 42], the host [54, 55], environmental factors [56, 57, 58] and the interactions among all of them [59].

At a certain moment, the bacterium starts to replicate inside the granuloma. For reasons not completely known, the bacterial load becomes so high that the granuloma cannot contain the infection, and the bacilli are released [6]. At this stage, the bacterial infection can provoke tissue necrosis and lung cavitation [60]. Due to these lung lesions, the host starts to cough, spreading the bacteria and potentially infecting new individuals [9]. In some cases, the bacteria can enter the bloodstream and disseminate through the body, causing extrapulmonary tuberculosis.

Early experiments with animal models suggested that the minimum infectious unit for MTBC is one single cell [61, 62]. This means that one single individual that accesses the respiratory track of the host is capable of starting an infectious process which could ultimately lead to TB disease. Although the bacterium is initially phagocytosed by the macrophages, in many cases it is able to obstruct the macrophages defenses not only to survive but also to replicate. First, the bacterium is able to resist the acidic conditions of the phagosome [48]. Second, the pathogen secretes a set of virulence factors that interfere with the phagosome maturation process (Figure 1.4) [63]. Third, the bacillus is able to inhibit the JNK, p38 and NF-κB pathways which are important mediators of anti-TB immunity [63].

In these conditions, the pathogen can either stop its proliferation (and become dormant) or start to replicate inside the macrophage. This ability to hijack the host-immune system to serve the bacterial purposes is one of the reasons why *M. tuberculosis* is considered the world's most successful human

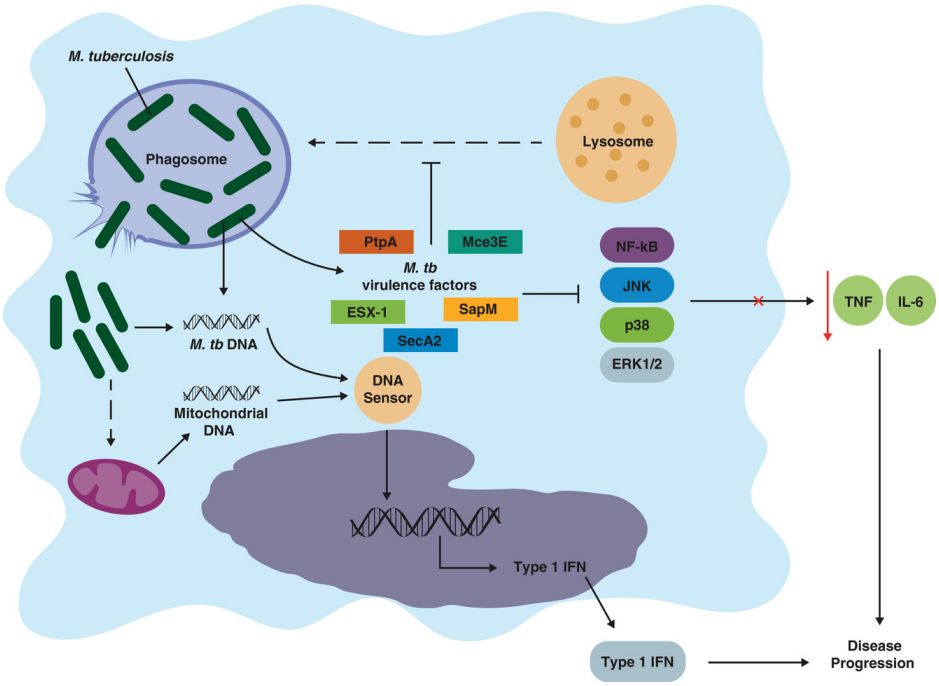


Figure 1.4: Schematic view of the macrophage infection by the pathogen. Several bacterial factors have the ability to alter the phagosome maturation as well as the inflammatory response. Source: Stutz *et al.* 2017 [63]

pathogen.

In summary, TB disease and infection are multifactorial processes. We are far from knowing all the details of these processes, as they involve different biological characteristics of the host, the pathogen and the environment. From our point of view, a crucial step to fight against TB is to fully understand the biological characteristics of M. tuberculosis, at all its different levels. So, in the present thesis, we have tried to decipher genomic determinants of the bacteria that are relevant for its capacity as a professional pathogen.

1.4 Genomic features of the MTBC

No plasmids have been reported in the MTBC so all the genetic information of the bacteria is contained in a single, circular chromosome. Its genome has a high GC content (65.6%) and a length of 4,411,532 bp [64]. The dN/dS value is a measure of the functional conservation of a gene. It represents the ratio of nonsynonymous to synonymous substitutions weighted by the number of nonsynonymous and synonymous sites in the genome. When a gene is under the action of the adaptive or diversifying selection, we expect that nonsynonymous mutations accumulate at a higher rate than synonymous ones (dN/dS $>$ 1). In the MTBC genome we found the opposite situation, with mean dN/dS around 0.5 for essential genes and around 0.66 for non-essential genes [65]. Despite the low genetic diversity, the accumulation of nonsynonymous mutations is higher than in other organisms, likely reflecting strong genetic drift and recent emergence of the clade and leading to functional diversity [40]. As stated above, the maximum genetic distance between any strain of the MTBC is around 2,500 variants (excluding repetitive regions) [38, 39]. Apart from single variants, the RD cause that different strains vary slightly in its gene content.

Some of the genomic characteristics of the MTBC have been used for phylogenetic typing. The RDs commented above, for example, are useful to determine the lineage or the major clade of an MTBC strain. But in order get more detail into the strain identification we need more information. For example, the MTBC has a region of Clustered Regulatory Short Palindromic Repeats (CRISPR) or also known as spoligotypes. In this region, there are a series of direct repeats alternated with short unique regions called spacers. More concretely, there has been reported 43 unique spacers between these repetitive regions. As not all of these spacers are present in all the MTBC strains, their pattern of presence/absence have been used for identifying the phylogenetic group of the strain (Figure 1.5) [31]. Regarding epidemiology, there are some genomic regions called Mycobacterial Interspersed Repetitive

Units (MIRU) that are enriched for Variable Number Tandem Repeats (VNTR). The number of VNTR inside each MIRU is variable across strains, so they have been used classically to identify closely related strains. For years, this MIRU-VNTR approach was the gold standard for molecular epidemiological studies.

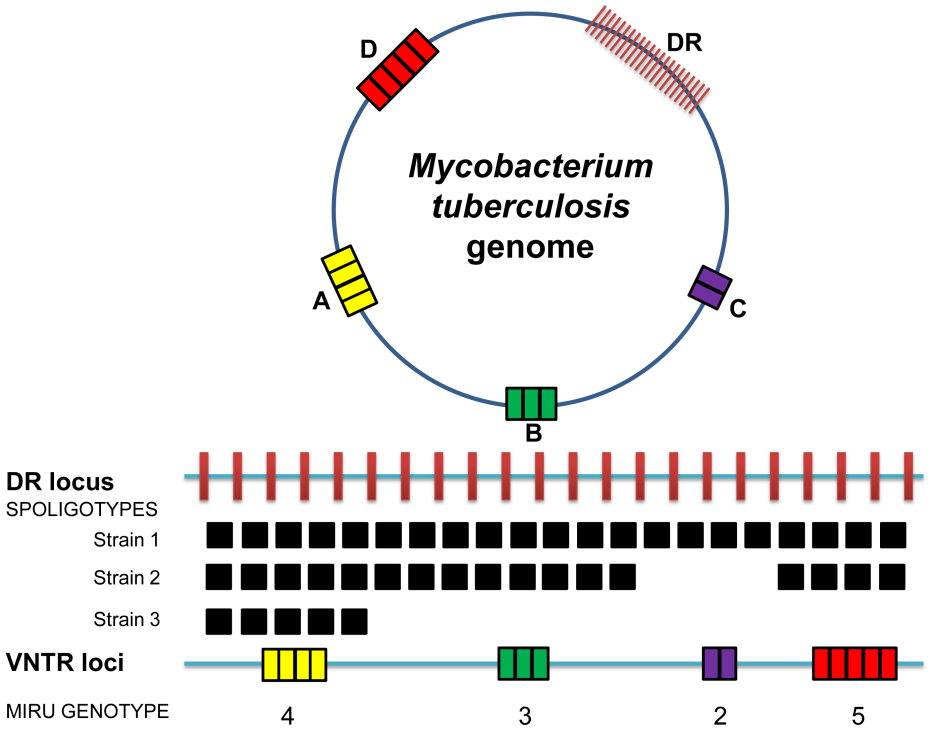


Figure 1.5: Schematic view of the spoligotype and MIRU-VNTR loci in the MTBC. The spoligotype and MIRU-VNTR patterns vary broadly across strains. Source: Comas *et al.* 2009 [31]

Despite their utility in achieving a fast result and their simplicity, these techniques have a low sensitivity and specificity, discouraging their use for population-based studies or in complex epidemiological situations [66, 67]. However, with the advent of whole-genome sequencing (WGS) techniques, new prospects for TB research appeared [68]. WGS provides researchers with

information at the genomic level, allowing to derive reliable epidemiological and evolutionary inferences at a population scale. WGS data is much more precise to detect recent transmission [69] and allows researchers to obtain important results in many other areas, such as MTBC evolution, physiology or resistance mechanism among others[70].

1.5 The impact of Computational Biology on TB research

The popularization of WGS drove the creation of public databases to store and share this type of data, the two main examples are the Sequence Read Archive (SRA) or the European Nucleotide Archive (ENA). Since its establishment in 2007, the number of MTBC WGS datasets in these databases has grown steadily (Figure 1.6). This has a main advantage: the large amount of information generated by WGS in each study can be reused. Researchers can now join genomic data from different sources to create datasets containing thousand of samples. This type of information has been used to perform population-based analysis, as for example, to identify genetic variants linked to drug resistance phenotypes [71]. Despite its indubitable usefulness, the amount of information generated by WGS overcame at the first moment the capacity of analysis with the classical methodologies.

Bioinformatics and computational biology initially emerged as an assistance for studies which deal with genetic data or in punctual statistical analyses. However, these large-scale data analyses have become so essential that bioinformaticians are current keystones for biomedical research [72]. So that, not only specialized scientific profiles but also specific equipment is needed for analyzing this type of data. As most WGS analyses exceed the processing capacity of personal computers, high-throughput computing equipment has been installed in most institutes and laboratories. In parallel to these changes in laboratory structures, facilities and staff, a new research field appeared merging biology and computer science called systems biology.

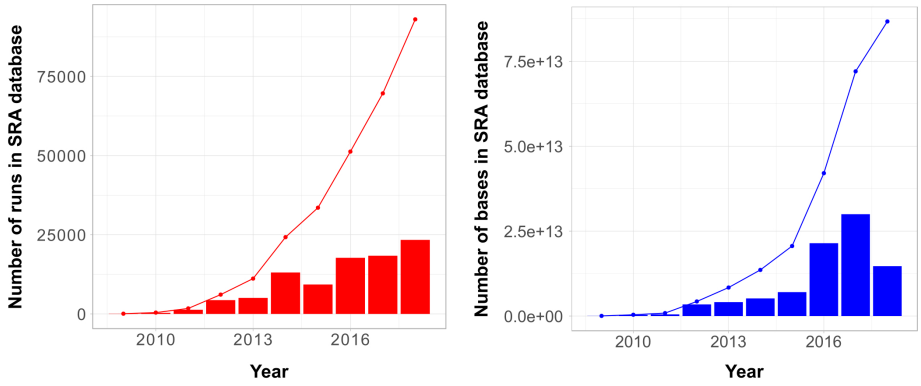


Figure 1.6: Amount of WGS data in the SRA database derived from *M. tuberculosis* samples by year (from january 2009 to october 2018). The number of deposited data increases year by year. Bars represent the data deposited each year, while the scatter plot represents the cumulative amount of data in the database.

Systems biology joins mathematicians, biologists, computer scientists and physicians to derive *in-silico* models to better understand the structure, functionality and evolution of complex biological systems. This research field has been demonstrated to be extremely useful in biomedical research, as it has the power of making reliable predictions on disease progression and outcome, reducing the amount of animal experimentation or patient samples [73, 74, 75]. In the field of TB research, systems biology approaches have produced encouraging results. For example, Dutta *et al.* [76] constructed computational algorithms that were capable of identifying genes involved in persistent infections in mice. These algorithms used as an input phenotypes from previous studies on persistence infection in mice, transcriptomics data and functional interaction networks. As a result, the models returned a list of genes potentially involved in persistence. These genes were tested *in vivo*, confirming their direct link with persistence in many of them. Other research by Pienaar *et al.* [77] derived computational models to simulate the dynamics of TB infection inside the granuloma when applying two different antitubercular drugs. They made important findings related to antibiotic penetration, concentration and efficacy inside the granuloma as well as bacterial population dynamics. In a

similar way, Lalande *et al.* [78] have tested the pharmacokinetics and pharmacodynamics of TB drugs (isoniazid in this case) during the first days of administration by using computational models. A different approach was used by Peterson *et al.*[79] to detect resistance mechanisms for a new anti-tubercular drug (bedaquiline). In this case, the authors created a regulatory network model and studied the effect on bedaquiline resistance when disrupting the network. Their model predictions were further assessed *in-vitro*. In another study, Sambarey *et al.* [80] applied a network analysis to identify biomarkers of TB infection in blood. They analyzed whole-blood transcriptomic data from healthy and TB infected patients. While comparing the different host responses and applying a systems biology approach they were able to identify 10 genes that act as biomarkers of TB infection and differentiate between latent and active TB infection cases. As a last example, Farrell *et al.*, [81] used a combination of different computational algorithms with an *M. bovis* proteome dataset to predict epitopes recognized by the host immune system (cattle in this case). The importance of this study lies in that epitope identification is key to develop new vaccines and diagnosis tools. The peptides identified as potential epitopes by the computational approaches were experimentally tested to assess their capacity of inducing an immune response, using a random selection of peptides as controls. Approximately 24% of the epitopes selected induced interferon-γ secretion by T-cells from infected cattle, proving the validity of the approach.

These new methodologies and techniques are becoming relevant tools in TB research as they allow processing huge quantities of data in a fast, reproducible and reliable way. Therefore, bioinformatics and systems biology are at the core of many analyses the present thesis.

1.6 Motivation

Despite its low genetic diversity, the MTBC is not genetically homogenous. Analyses of the diversification of *M. tuberculosis* from its common ancestor

have been carried out in the past [26, 36, 82]. However, dozens to hundreds of genomes were used in these studies, in contrast to the thousands of genomes currently available in public databases. With this in mind, we aim to analyze this huge amount of new information to: (i) obtain a more detailed view of the evolutionary processes that have led to the present MTBC population structure, and (ii) detect the genetic mechanisms and genes involved in the emergence of the MTBC as an obligate pathogen globally distributed (Chapter 4).

In addition, we have stated previously that the genetic diversity of MTBC, although modest in comparison with other pathogenic bacteria, may have implications in the disease outcome. Most of the research currently carried out on *M. tuberculosis* does not take into account the pathogen's genetic diversity, as they rely in clinical reference strains. Thus, it is of special interest to evaluate whether the conclusions derived from cutting-edge research not taking into account this diversity can be generalized to the whole MTBC or, on the contrary, are biased (Chapter 5).

Finally, the forces that drive the evolution of the MTBC have created a range of different phenotypic traits. The different evolutionary mechanisms that acted along the MTBC history to model this current phenotypic diversity are not fully catalogued. So, we intend to characterize in detail the transcriptomic and methylation diversity of the MTBC as phenotypes associated to the underlying genetic diversity of the pathogen (Chapter 6).

Part of the introduction has been published as a Review Article:

Chiner-Oms Á., Comas I. Large genomics datasets shed light on the evolution of the *Mycobacterium tuberculosis* complex. **Infection, Genetics and Evolution**. In press. DOI:10.1016/j.meegid.2019.02.028

Objectives

The present thesis aims to study the evolution and biological characteristics of the tuberculosis pathogen by using bioinformatics and cutting-edge systems biology techniques. We are going to use WGS coming from different sources as the main source of data. Specifically, we will address the following objectives:

- To study the different evolutionary processes that guide the evolution of the MTBC from an environmental reservoir to its current ecological niche as an obligate pathogen (Chapter 4).
- To depict the main genetic changes involved in the MTBC adaptation to specific mammalian hosts. (Chapter 4).
- To evaluate the capacity of current *M. tuberculosis* predictive models, based on the H37Rv strain, to make reliable predictions about other members of the MTBC (Chapter 5).
- To study the impact of the MTBC genetic diversity on biological networks such as the regulatory network and the protein-protein interaction network (Chapter 5).
- To study the transcriptomic signatures of the different MTBC members and the main evolutionary processes that lead to clade-specific regulatory patterns (Chapter 6).

General materials and methods

3.1 Computers and High Performance Computing servers

All the work presented in this thesis has been developed using bioinformatic techniques and methods. Thus, the main instruments used along this 4-year period were computing equipments. Daily analyses were run in a personal computer with the following characteristics:

- Ubuntu 14.04 (2014 to 2017) and Ubuntu 18.04 (2018 to 2019) as the basic operative system.
- CPU Intel i7 7700 (8 cores).
- 16 GB DDR4 RAM memory.
- 1 TB SATA 3 HDD.

Some of the analyses that required a higher computational capacity were performed in several High Performance Computing servers, located in FISABIO-CSISP and the IBV-CSIC.

- Cuda (FISABIO-CSISP).
  - CentOS 6.7 operative system, 64 bits.

  - 2 x Intel Xeon Family 6 Model 44 3.4GHz (24 cores).
  - 64 GB DDR4 RAM memory.
  - 44 TB SATA 6 HDD.
  - 3 x NVIDIA Quadro 6000.
- Calcul2 (FISABIO-CSISP).
  - CentOS 5.11 operative system, 64 bits.
  - 4 x Intel Xeon E7450 2.4GHz (24 cores).
  - 64 GB DDR4 RAM memory.
  - 40 TB SATA 6 HDD.
- Yersin (IBV-CSIC).
  - CentOS 7.3.1611 operative system, 64 bits.
  - 2 x Intel Xeon E5-2620v3 2.4GHz (24 cores).
  - 128 GB DDR4 RAM memory.
  - 48 TB SATA 6 HDD.
- Koch (IBV-CSIC).
  - CentOS 7.4.1708 operative system, 64 bits
  - 2 x Intel Xeon E5-2620v3 2.4GHz (24 cores).
  - 256 GB DDR4 RAM memory.
  - 48 TB SATA 6 HDD.

3.2 Analysis pipeline

An important part of the studies tackled in this thesis is the analysis of MTBC genomic data. Raw data was either downloaded from public databases or supplied directly from the original sources by the owners. In some chapters,

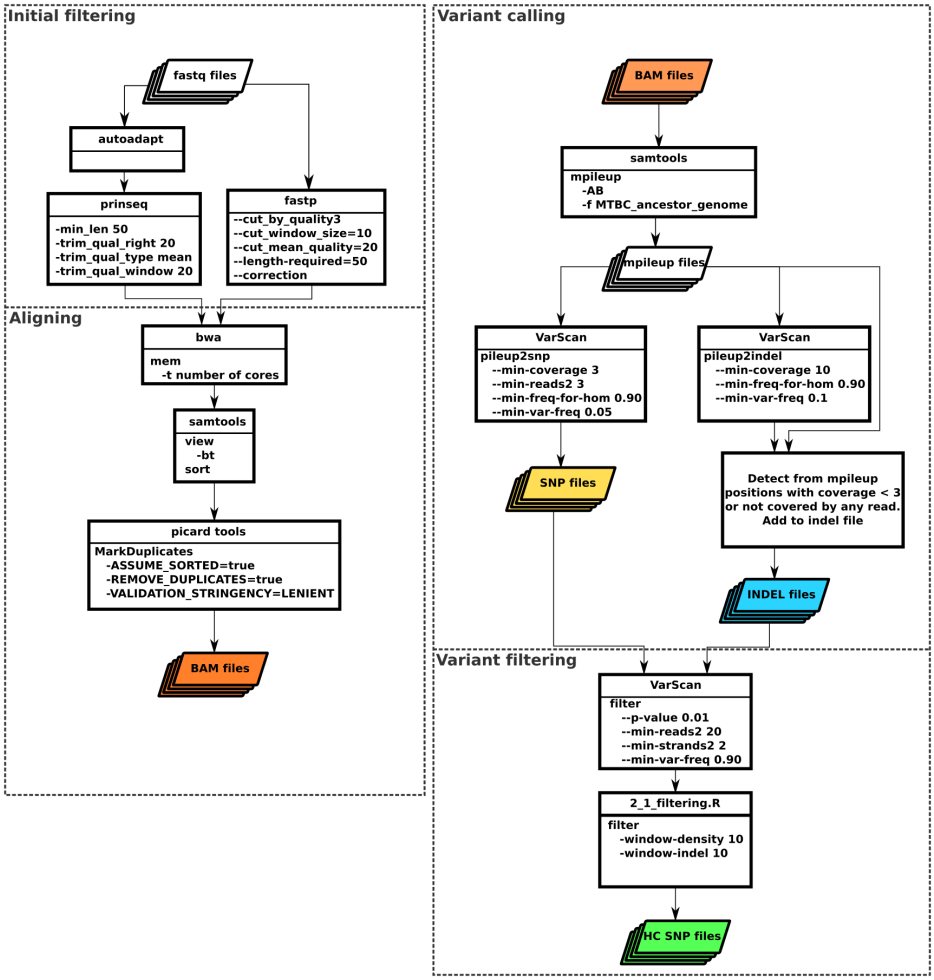
more than 7,000 genomic samples were analyzed. A custom analysis pipeline based on a previous one [36], was constructed step-by-step to automate the extraction of genomic information from the data samples. A general overview of the pipeline can be found in Figure 3.1. This pipeline was specifically set up to manage short-read sequencing data. Almost all the genomic data analyzed in the present thesis have been generated with different Illumina platforms, except part of the data analyzed in Chapter 6 (see this chapter for specific details on the methodology used). The pipeline performs the following detailed steps:

Initial filtering

Initially, FASTQ files were filtered and trimmed using autodapt [83] to remove adapters in case they were present. After that, prinseq [84] was used to trim poor quality bases. We required a minimum read length of 50 bp, and right-end bases were trimmed if their mean quality was lower than 20 in a window of 20 bp. In the last part of the thesis, the pipeline was updated and the initial filtering was performed with the fastp program [85]. Fastp was written in C code in contrast to prinseq, which was written in Perl. Hence, fastp is much faster than prinseq. Moreover, fastp includes an automatic adapter detection function so it covers the joint functionality of autoadapt+prinseq. We run fastp requiring a minimum length of 50 bp, scanning the 3' end of the reads using a window size of 10 bp and cutting the bases that had a mean quality under 20. In addition, fastp corrects bases in overlapped pair-end reads that mismatch.

Aligning

Once the FASTQ files were trimmed, we used the BWA-mem [86] algorithm to align the reads to a reference genome. The reference genome we used is the MTBC most likely ancestral genome [87], derived by maximum parsimony and likelihood methods. This ancestor is H37Rv-like in terms of genome structural variants, but H37Rv alleles were replaced by those present in the inferred common ancestor of all MTBC lineages. Samtools [88] was used to create


Initial filtering
fastq files
autoadapt
prinseq
-min_len 50
-trim_qual_right 20
-trim_qual_type mean
-trim_qual_window 20
fastp
--cut_by_quality3
--cut_window_size=10
--cut_mean_quality=20
--length-required=50
--correction
Aligning
bwa
mem
-t number of cores
samtools
view
-bt
sort
picard tools
MarkDuplicates
-ASSUME_SORTED=true
-REMOVE_DUPLICATES=true
-VALIDATION_STRINGENCY=LENIENT
BAM files
Variant calling
BAM files
samtools
mpileup
-AB
-f MTBC_ancestor_genome
mpileup files
VarScan
pileup2snp
--min-coverage 3
--min-reads2 3
--min-freq-for-hom 0.90
--min-var-freq 0.05
VarScan
pileup2indel
--min-coverage 10
--min-freq-for-hom 0.90
--min-var-freq 0.1
SNP files
Detect from mpileup positions with coverage < 3 or not covered by any read. Add to indel file
INDEL files
Variant filtering
VarScan
filter
--p-value 0.01
--min-reads2 20
--min-strands2 2
--min-var-freq 0.90
2_1_filtering.R
filter
-window-density 10
-window-indel 10
HC SNP files

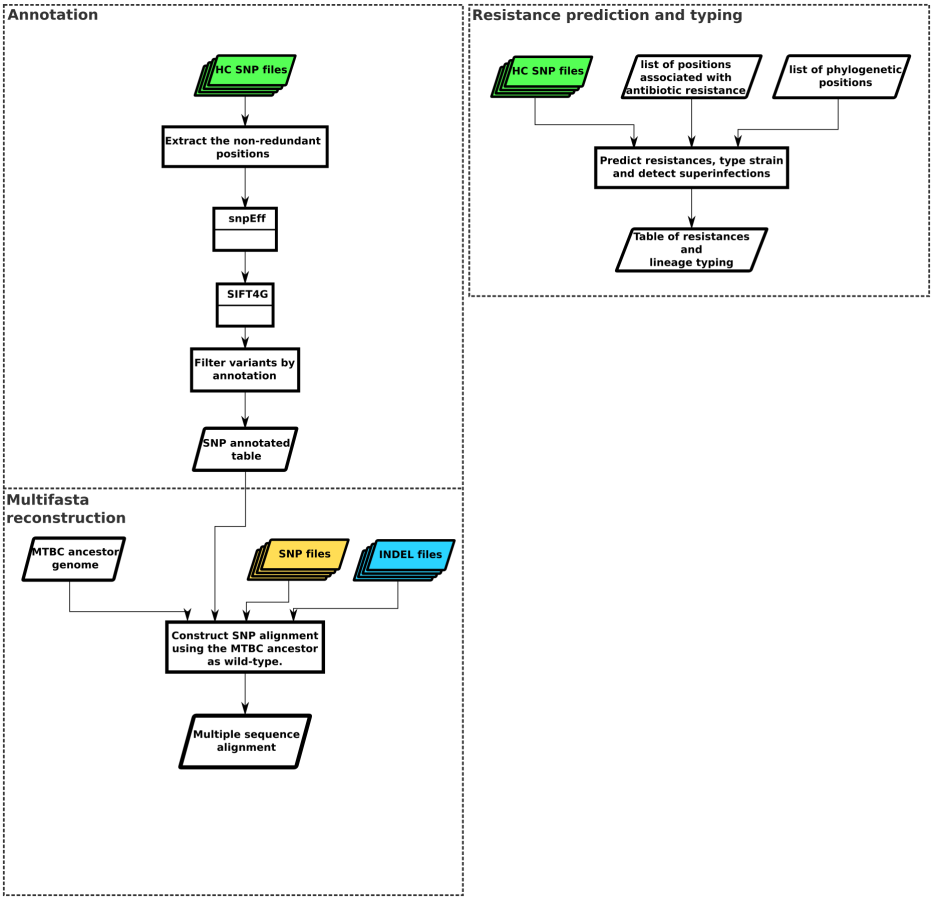


Figure 3.1: Schematic view of the analysis pipelines built to manage the raw data derived from whole-genome sequencing.

sorted alignment files in BAM format from the BWA-mem output. Picard tools [89] was used to remove potential duplicated reads formed during the sequencing process.

Variant calling

From each BAM file, a mpileup file was derived using Samtools. This file was scanned with VarScan [90] to extract single variants and indels. Initially, a Single Nucleotide Position (SNP) was called if variants were supported by 3 reads at least and found in at least 5% of the reads covering the positions. The called SNPs were kept in a SNP file. Similarly, we required a minimum coverage of 10 reads and at least 10% of the reads in this position to call an indel. These variants were stored in the INDEL file. Positions not covered by any read (i.e. regions present in the reference genome but not in our genomic sample) are not reported in the mpileup file and are not detected by VarScan. So, we included also these positions in the INDEL file as well as regions covered by 3 or fewer reads (regions mapped with few reads, in a genomic sample with mean depth $>$ 50, could not be discarded as indels).

Variant filtering

Almost every single variation found in the alignment is called as a potential SNP in the SNP file. However, with these relaxed criteria many of the variants found are potentially false, introduced by errors or inaccuracies in some part of the process. To consider only highly confident variants, a stricter filtering is needed. So the SNP file was filtered with VarScan. We kept variants that were supported by at least 20 reads, called in both strands and present in 90% of the reads covering the corresponding position. In addition, we did not take into account SNPs that were close to indel regions (10 bp upstream and downstream) or in regions with a high accumulation of variants (more than 3 variants in a 10 bp window). These filtered SNPs were stored in a high confident (HC) SNP file.

Annotation

From all the HC SNP files, we created a non-redundant list of variants containing all the SNPs found in the set of samples. The snpEff program [91] was used to create a custom annotation database based on the most likely ancestral positions of the MTBC and using the H37Rv annotation. This snpEff database was extensively used in the present thesis to annotate SNPs lists, both those derived from the pipeline (including the HC SNP files) and those from any other list of SNPs. SIFT4G [92] was used to predict the potential impact of each nonsynonymous mutation in the coding regions. SIFT4G assigns a score for each SNP based on evolutionary information. It predicts the tolerability of a nucleotide substitution based on the abundance of substitutions in this position in related sequences. Next, variants that were predicted to fall in phages, repeated regions or PE/PPE genes were filtered out from the list of variants. These regions are difficult to map with short reads (i.e., Illumina reads) and tend to accumulate many false positive SNPs. The remaining variants along with their annotation were stored in a SNP annotated table.

Multifasta reconstruction

For the subsequent analyses (phylogeny reconstruction, distance calculus, evolutionary traces, etc.) we needed a multiple sequence alignment. For each sample, we used the MTBC ancestor genome to define the wild-type base. Over this structure, we introduced the deletions present in the INDEL file for each sample as well as the variants present in the SNP file that were also present in the SNP annotated table. With this approach we were introducing only HC SNPs (those that are present in the SNP annotated table) or likely true SNPs that did not pass the variant filters in a concrete sample but was observed as a HC SNP in another sample. Finally, we joined all the reconstructed sequences in one single multifasta file.

Resistance prediction and typing

With the information contained in the HC SNP files we can type the samples without constructing a phylogeny. We compared the HC SNP files with a list of phylogenetic SNPs [93, 34] and with a list of known resistance-conferring variants [94, 95]. With these information we constructed a table of samples with their potential phylogenetic assignment and resistance profile.

3.3 Scripting and statistical analyses

All the scripting work was performed using several programming languages, mainly Bash but also Perl, Python and R. Specially useful were the GNU parallel package [96], used for parallelizing some Bash code (i.e. the multifasta reconstruction process in the pipeline described above), and BioPerl [97], adopted for working with sequence data with the Perl language. The biological network visualizations and plotting were performed using Cytoscape [98]. Gene Set Enrichment (GSE) analyses were performed with the BiNGO plugin [99]. This Cytoscape plugin was widely used in the current thesis to study the enrichment in certain functional categories of a specific group of genes, in comparison with the total number of functional categories found in the complete annotation. The tool uses a hypergeometric test (sampling without replacement) and the BH correction for multiple testing comparisons [100]. Almost all the statistical analyses performed in this thesis were implemented using the R statistical language [101], and the RStudio program [102]. R was used not only for the statistical work but also for daily scripting tasks involving tables and numeric data. The following R packages were extensively used:

- **Bioconductor** [103]: An open source project that aims to maintain tools for the analysis of genomic data in R. It was used mainly for microarray data analysis.
- **doParallel** [104] and **foreach** [105]: This two packages combined allow to run loops in parallel. They were extremely useful when working in the

servers as they permit to select the number of nodes/threads in which the code is going to be executed.

- **ggplot2** [106]: Provides a complete set of tools to create colorful and elegant plots. It is helpful to represent a wide range of data types in a graphical manner.
- **DEseq2** [107]: Contains many functions (as well as a complete manual) to handle RNA-seq data and perform several statistical analyses with them.
- **igraph** [108]: Designed to work with graph structures (nodes and edges) and to calculate several specific statistics associated with this type of data.
- **seqinr** [109]: Used to manage sequence data (nucleotides and proteins) in the R environment.

Genomic determinants of speciation and spread

4.1 Introduction

The increasing availability of population genomics data has allowed an improved understanding of genotypic and ecological differentiation among closely related bacteria [110]. While a species concept *sensu stricto* cannot be applied to bacteria [111] models exist to understand how species can emerge in natural populations. Depending on the evolutionary forces involved, models range from differentiation driven by natural selection and adaptation to different ecological niches (Ecological Species Concept) to differentiation as a result of restricted gene flow that reinforces isolation (Biological Species Concept). In reality, most natural populations show a combination of both processes with certain overlap between habitats (Overlapping habitats model, [112]). The study of natural populations and models shows that the emergence of new species is more common among bacterial groups sharing, partially or totally, their habitat, a process also known as sympatric speciation [112]. Processes of bacterial differentiation are often expected to leave measurable genetic signatures in extant genomes including “speciation islands” (regions of high divergence between the nascent species) [113, 112]. These genetic signatures hold clues about the key genomic determinants responsible for ecological differentiation of nascent species from a common genetic pool. However, how

these models apply to professional pathogens, particularly those characterized by an obligate association with their host species, has been little explored.

As stated in the General Introduction, the most closely-related bacteria that fall outside the MTBC include isolates known as *Mycobacterium canettii* (MCAN). MCAN strains differ from MTBC isolates by tens of thousands of SNPs [27, 114]. MCAN strains have been isolated from the Horn of Africa, predominantly from children and often in association with extrapulmonary tuberculosis [115]. It is assumed that MCAN represents an opportunistic pathogen with an unidentified environmental reservoir [28] as opposed to the obligate MTBC pathogen. Genomic comparisons have identified gene-content differences between MTBC, MCAN and other mycobacteria [27, 114, 116] as well as genetic differences in virulence-related loci [117, 118].

Two pieces of evidence suggest that MTBC and MCAN evolved from a common genetic pool in Africa. The high ANI between the MTBC and MCAN suggests incomplete or recent speciation. In addition, most reports suggest lack of on-going recombination between MCAN and MTBC and within the MTBC [119, 120] suggesting complete separation (but see [121]). The second piece of evidence comes from phylogeographic and genetic diversity analyses which identified the origin of the tuberculosis bacilli in Africa [36, 35], the likely place of origin of MCAN [122, 27]. Taken together, the data suggests that ancestral MTBC and MCAN strains at least shared partially the same niche and genetic pool.

Our understanding about the population genomic events mediating the divergence of the ancestor of the MTBC from a common ancestral pool with MCAN is far from complete. In this chapter, the availability of genome sequences from thousands of MTBC clinical strains, as well as of close relatives like MCAN, enables us not only to identify molecular signatures of MTBC speciation events, but also to reveal known and new targets for biomedical research.

4.2 Results

We first analyzed the differentiation between MTBC and MCAN by searching for any hallmark of on-going recombination between and within these groups of strains. Previous reports have suggested that there might be limited but significant recombination among MTBC strains [121, 123] while others failed to identify measurable recombination events [124]. To revisit this question we used a large collection of MTBC genomes. To maximize the chances of identifying potential ongoing recombination events within the MTBC, we screened a data set of complete genome sequences of strains from global sources [93](n=1,591). These genomes are representative of the known geographic and genetic diversity of the MTBC (Supplementary Figure 10.1). Among these genomes, we identified all the variant positions and, more specifically, potential homoplastic sites, i.e., polymorphic sites showing signs of convergent evolution. A total of 96,143 variant positions were called in the 1,591 strains. Homoplasy can arise as a consequence of recombination but it may be caused by other processes, such as positive selection, sequence gaps contributing to homoplastic counts or mapping/calling errors. For example, known drug-resistance positions use to accumulate lots of homoplasies as they are well known instances of convergent evolution [125, 126]. So, to increase the likelihood for homoplastic positions to be due to recombination events, we filtered out known drug-resistance positions (n=48), non-biallelic positions (n=1,076), potential mapping errors identified by generating synthetic reads around each SNP position (n=239). In total we excluded as likely arising from other signals 1,363 positions out of the initial 96,143 positions (1,42%).

As a result, a total of 2,360 core homoplastic sites were identified across the 1,591 strains analyzed (2.5% of all variable sites). Homoplastic sites did not significantly accumulate in any region of the genome, suggesting absence of recombination hotspots (Figure 4.1A). To get a more detailed view and detect small recombination events, we looked for regions with two or more consecutive homoplastic variants (allowing one non-homoplastic variant

between them) co-occurring in the same phylogenetically unrelated strains. We detected only 2 cases in which two variant positions were homoplastic, consecutive and shared phylogenetic congruence (found in the same unrelated strains). The two regions accounted for 4 convergent variants (4.1) and affected strains from different MTBC lineages. Variants in positions 2195896 and 2195899 fell in the primary regulatory region of *mazE5* [127]. Variants in positions 2,641,161 and 2,641,163 fell in the intergenic region of *glyS* and Rv2358. Although we cannot completely discard the possibility that these represent recombination events, it is more likely that the two regions have been under positive selection, a mechanism known to lead to the accumulation of homoplastic accumulation in the MTBC [126]. In summary, this large-scale variant-by-variant analysis could not identify significant ongoing recombination between any of the 1,591 MTBC strains analyzed.

Position	A	G	C	T	Homoplastic steps	Genomic region	Nuc. change	AA. change
2195896	0	4	1587	0	3	Rv1994c	39G>C	K13N
2195899	1587	0	0	4	3	Rv1994c	36T>A	D12E
2641161	0	2	0	1589	2	IG_Rv2357c-Rv2358		
2641163	0	1589	2	0	2	IG_Rv2357c-Rv2358		

Table 4.1: Variants identified as homoplastic and phylogenetically convergent

Due to the low diversity within the MTBC, we also followed alternative approaches to identify recombination events with a high statistical confidence. Using an additional method, we evaluated linkage disequilibrium (LD) as a function of the distance between the 94,780 core variant positions. R^2 has been used to show on-going recombination at very short distances (less than 50 bp.) [121]. In our much larger dataset, R^2 values were also slightly higher at shorter distances, which is compatible with recombination involving larger fragment sizes. However, a close examination reveals that the peak at short distances is misleading, as it is driven by only six points out of more than 11,000 comparisons (Figure 4.1B). In addition, R^2 values are known to fail to reach the

theoretical maximum of 1 when variants compared have very different frequencies [128]. This is likely the case for MTBC, in which there is a strong skew of the site frequency spectrum towards low frequency values (Figure 4.1C) [40]. Thus, as an alternative we calculated D'. In this data set, as expected for a mostly clonal organism, LD measured by D' remained at its maximum value, even when focusing on distant variant positions more than 5 Kb apart, suggesting very little or no ongoing recombination (Figure 4.1B).

To further validate these findings, we ran Gubbins with the same data set and validated them with RDP4 (see methods for details). Gubbins detects the accumulation of a higher than expected number of variants in addition to homoplastic sites as a hallmark of possible recombination. We partitioned the 1,591 strain dataset into the different lineages and screened for possible tracks of recombination. Gubbins reported potential recombining regions characterized by a higher than average number of SNPs. However, none of the RDP4 methods confirmed any of them. Thus, those events maybe real but cannot be confirmed by alternative approaches.

Having established that recombination has little impact on the overall MTBC genetic diversity, we compared a representative data set of MTBC genomes [36] (n = 219) with 7 MCAN genomes to identify and quantify eventual ongoing recombination within MCAN and between MCAN and the MTBC. Of the 93,922 polymorphic sites identified, 22,718 were biallelic homoplasies (24.2%). The genomic distribution of variant positions and homoplasies in the MCAN group showed a landscape different to the MTBC group (Figure 4.2B). A total of 22,464 (98.9%) of those homoplasies were found only among MCAN strains, representing almost half of the variability within this group (22,464/52,392 biallelic sites, 42.9%) which points to recombination as a main source of variability in MCAN. This is consistent with previous reports(Supply et al. 2013). This profile is in sharp contrast with the flat homoplastic profile for the MTBC described above (Figure 4.1A).

To test for ongoing recombination between MCAN and extant MTBC, we identified homoplasies involving both groups. From the 93,922 total variants,

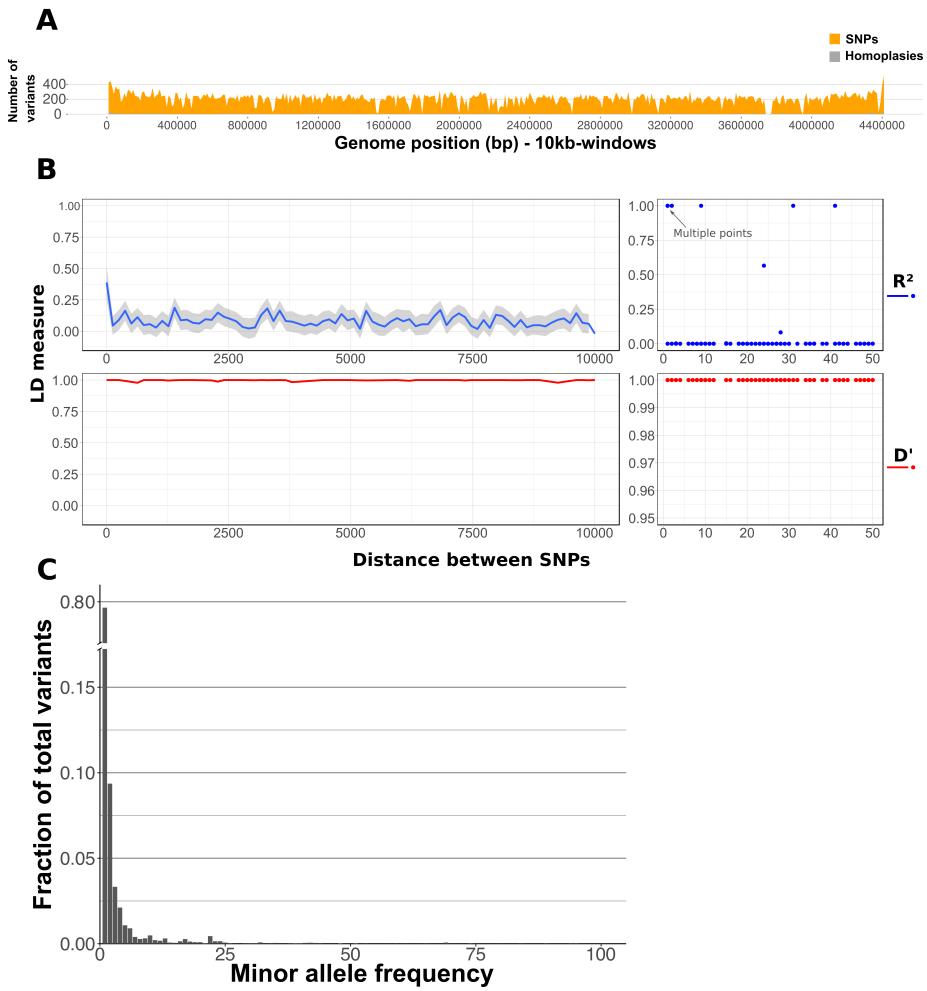


Figure 4.1: No ongoing recombination within the MTBC A) Number of homoplasies (grey) as a function of the total number of variants detected (orange) in the MTBC dataset (n=1,591) B) Linkage disequilibrium as a function of genetic distance detected in the 1,591 strains. C) Site frequency spectrum of MTBC strains using the core variant positions.

7,934 involved MCAN and MTBC strains. We found 234 biallelic homoplasies involving extant MTBC and MCAN strains, thus compatible with ongoing recombination but also with independent diversification. The vast majority of

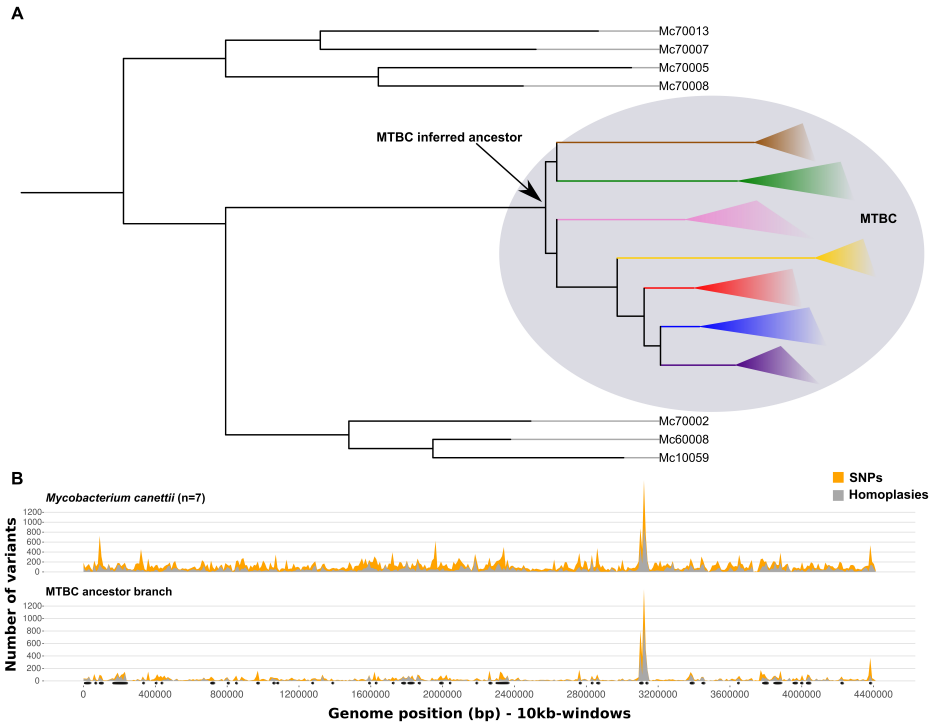


Figure 4.2: Genome-wide variant profiles vary between *M. canettii*, *M. tuberculosis* and the MTBC ancestor A) Schematic view of the phylogenetic relationships between the MCAN groups and the MTBC. In the Supplementary Figure 10.2 a Maximum Likelihood phylogeny of the MCAN group including the MTBC ancestor can be found. B) Number of homoplasies (grey) as a function of the total number of variants detected (orange) in the MCAN dataset and in the branch leading to the MTBC most recent common ancestor. Black dots indicate recombination events detected in the branch leading to the most recent common ancestor of the MTBC.

homoplasies detected (97%) mapped to the branch leading to the MTBC clade (thus fixed within the MTBC but variable within the MCAN group). These results indicate that measurable recombination events were common between MCAN and the ancestral branch of the MTBC, but are unlikely during the subsequent diversification of the MTBC. Consistently, a Gubbins analysis in this dataset did not identify recombination events involving current MTBC and MCAN strains. All the potential recombinant segments mapped on branches involving only MCAN strains or involving MCAN and the common branch of all the MTBC

strains (see External Data1).

Sympatric and stepwise emergence of the MTBC ancestor

Our results show that recombination with closely related mycobacteria occurred during the emergence of the common ancestor of the MTBC. To gain a better insight on how it occurred we reasoned that instead of comparing MCAN strains against extant MTBC strain we should compare against a reconstructed most common ancestor of the MTBC (see Chapter 3). This strategy allowed us to focus on those changes specifically happening in the ancestral branch of the MTBC (see Figure 4.2A and Supplementary Figure 10.2). As described by others, the phylogeny suggests a specific clone of the MCAN group diverged and resulted in the MTBC [129, 122]. To do so we extracted all the variant positions that were homoplastic between the MTBC ancestor and any of the MCAN strains. That is, equal nucleotide changes occurring in the same genomic positions that appeared independently in the branch leading to the MTBC ancestor and in any other branch of the phylogeny (7,700 positions). The SNPs mapping to the branch leading to the MTBC ancestor genome showed a similar homoplasy profile to that of the MCAN strains (Figure 4.2B), suggesting that there were not hard barriers to gene flow between ancestral MTBC and MCAN ancestral strains, thus supporting a model of sympatric speciation. Notably, both MCAN and the MTBC ancestor shared a peak around the CRISPR region, highlighting the dynamic nature of this region possibly as a result of common phage infections.

A Gubbins analysis including MCAN genomes and the most likely common ancestor of the MTBC was performed. Gubbins identified 70 recombination events between the MTBC ancestor and MCAN strains. 5 of these fragments were filtered out due to a high accumulation of gaps (see Methods). So, we kept a total of 65 recombination events mapping to the branch leading to the MTBC (External Data 2). Mapping of variants into the phylogeny revealed that those regions were coincident with a high number of homoplastic variants between MCAN and the MTBC (Suplementary Figure 10.3). To explore

whether these fragments reflected real recombination, we performed a phylogenetic congruence test. First, a likelihood mapping analysis was performed for each fragment (Supplementary Figure 10.4). Fragments 13 and 14 had not enough phylogenetic signal to resolve a reliable phylogeny. The variants found in these regions were present only in the MTBC ancestor branch, so recombination with other organism not present in our dataset is likely to have occurred. Later, the topologies of the trees of the remaining fragments were compared with that of the tree derived from the non-recombinant alignment (whole genome alignment subtracting the recombinant regions). These analyses revealed significant incongruence for all the 63 fragments compared to the non-recombinant phylogeny (Shimodaira-Hasegawa test; p-value¡0.05, External Data 3). The analysis identified consecutive fragments with similar topologies implying not only that the event involved similar donor/strains but also that they likely are part of a larger, unique event (Figure 4.3). This is the case for the genes in fragments 40, 41, 42. The genes involved are almost consecutive and only separated by PE/PPE genes that are not analyzed in this study. The fact that they share a common phylogenetic story indicates that they belong to a unique recombination event involving almost 28 Kb. A similar pattern can be observed for fragments 9-12 in which the fragments are not only consecutive in the genome but they also share a common phylogenetic story. In addition, events falling apart in the genome maybe also share a common phylogenetic story. For example regions 6, 63, 62 are more than 3.8 Mb apart on the genome but they share the same phylogenetic topology. The genes involved are part of the same regulon, KtsR [130], suggesting that selection may have played a role in fixing those independent recombination events. Thus, both Gubbins and phylogenetic approaches indicated that these 65 regions are likely recombinant regions.

To test whether speciation of the MTBC ancestor occurred in one single episode or in multiple episodes over time, we analyzed the relative age of divergence of the recombination fragments from the MCAN closest clade using BEAST. Results show that the MTBC ancestor differentiated from MCAN

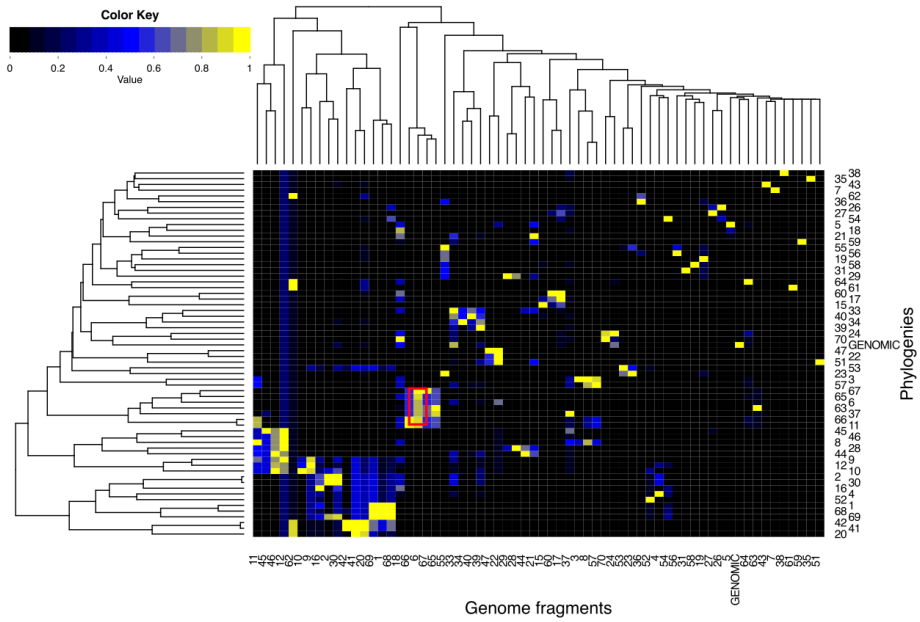
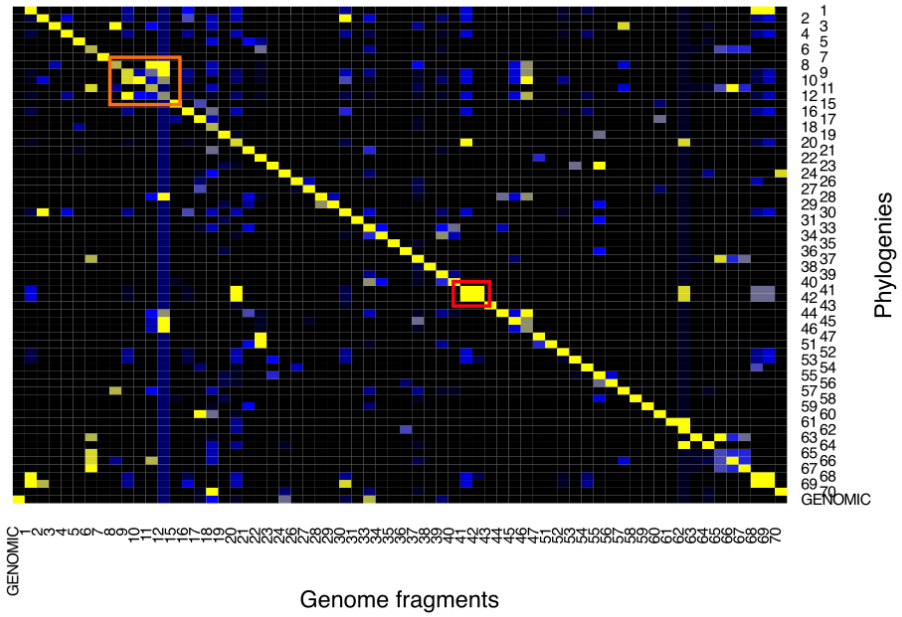


Figure 4.3: Phylogenetic incongruence test. Each fragment alignment was compared against all recombinant fragments phylogenies and the non-recombinant genomic phylogenetic topology. Dark blue indicates strong incongruence and yellow no evidence to reject the topology. In the left side we show a double clustering of fragments and phylogenies in which each row corresponds to a phylogeny and each column to a fragment. In the right plot fragments are organized following their position in the genome.



Fragments 13 and 14 were not included in the analysis as they did not have enough phylogenetic signal to reconstruct a reliable phylogeny. Fragments 40, 41, 42 are marked with a red square in the right panel. They share a common phylogenetic story and are correlatives suggesting that they belong to a unique recombination event. A similar pattern can be observed for fragments 9-12 (orange box, right panel). Regions 6, 63, 62 also share the same phylogenetic topology although not being consecutive (left panel, red square). The genes involved are part of the KtsR regulon.

sequentially (Supplementary Figure 10.5). The estimated ages show large HPD intervals, as expected from the low number of variant positions per fragment. Although the distribution of tMRCA for the fragments represents a continuum, the analysis suggests a separation between “recent” recombination events and “ancient” events, closer to the time of divergence from the MCAN group (Figure 4.4B). The large HPD intervals preclude any firm conclusion but the results suggest that some regions in the MTBC ancestral genome were restricted to gene flow earlier than others (Figure 4.4A).

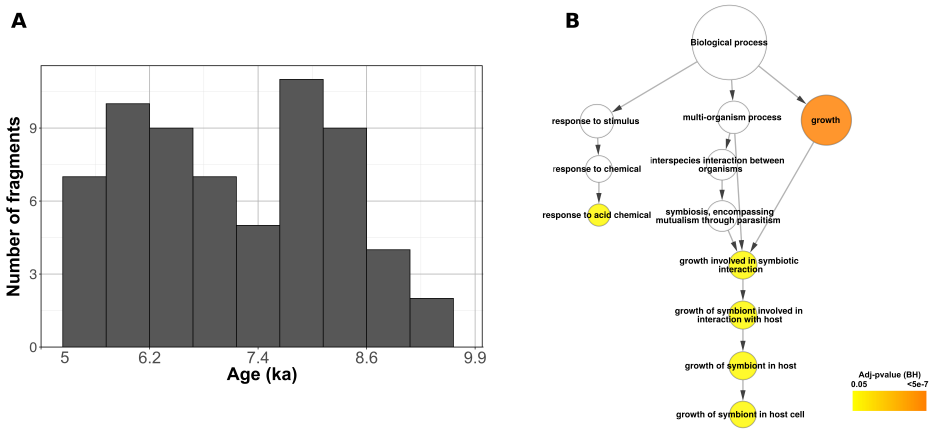


Figure 4.4: Past recombination between *M. canettii* strains and the MTBC ancestor A) Histogram distribution of the recombination fragments ages using the 5ka scenario [82]. A more detailed view can be found in the Supplementary Figure 10.5, with the confidence intervals plotted. B) Gene Ontology terms overrepresented in the coding regions contained in the recombinant fragments

If recombination played a major role in shaping the MTBC ancestral genome with regards to pathogenesis, we would expect some functions related to the interaction with the host to be affected. Indeed, we observed an enrichment in experimentally confirmed essential genes in the regions involved in recombination events, suggesting that recombination targeted important cell functions (Chi-square test; p-value < 0.01). An enrichment analysis of Gene Ontology terms for the genes contained in these regions identified functions related with growth, and most specifically with the category “growth involved in

symbiotic interactions inside a host cell" as significantly overrepresented (Binomial test; adj. p-value < 0.05) (Figure 4.4C). This category can be interpreted as genes involved in a strong association between the pathogen and the host. Remarkably, most of the genes involved have been implicated in virulence using animal models of infection (see Discussion).

The recombination profile shown in Figure 4.2 suggests that the MTBC ancestor recombined with MCAN ancestral strains and, thus, they shared a common niche. A sympatric model of speciation predicts that some parts of the genome will be involved in adaptation to a new niche [113]. The hallmark trace would be the accumulation of variants differentiating the emerging species, at the genome-wide level or in a few loci, as a consequence of reduced recombination between both groups. We identified all the variant positions that appear in the ancestral branch of the MTBC but remain unchanged in all the MCAN strains, the so called divergent variants (divSNPs). These divSNPs are new alleles unique to the MTBC ancestor and not present in any of the MCAN strains. The distribution of divSNPs per gene revealed that only few of them accumulated divSNPs in the branch leading to the ancestor while most genes did not (Figure 4.5A). This pattern is compatible with population differentiation models in which the overlap between emerging species is high [112]. The genome-wide landscape of divergent variants (n = 5,688, Figure 4.5B) revealed that a total of 120 genes harbored more divergent variants than expected by chance (see Methods)(Figure 4.5B).

However, bacterial genomes are highly dynamic and different processes can contribute to the genetic makeup of extant species. Consequently, not all the detected regions necessarily result from pure divergence by accumulation of substitutions. To ascertain the evolutionary origin of the 120 genes containing a high number of divergent variants, we checked whether the abnormal accumulation was due to: (i) horizontal gene transfer with other mycobacteria; (ii) recombination with MCAN strains that were not present in our dataset; or (iii) other evolutionary processes such as mutation combined with natural selection and/or genetic drift, thus representing genes that genuinely

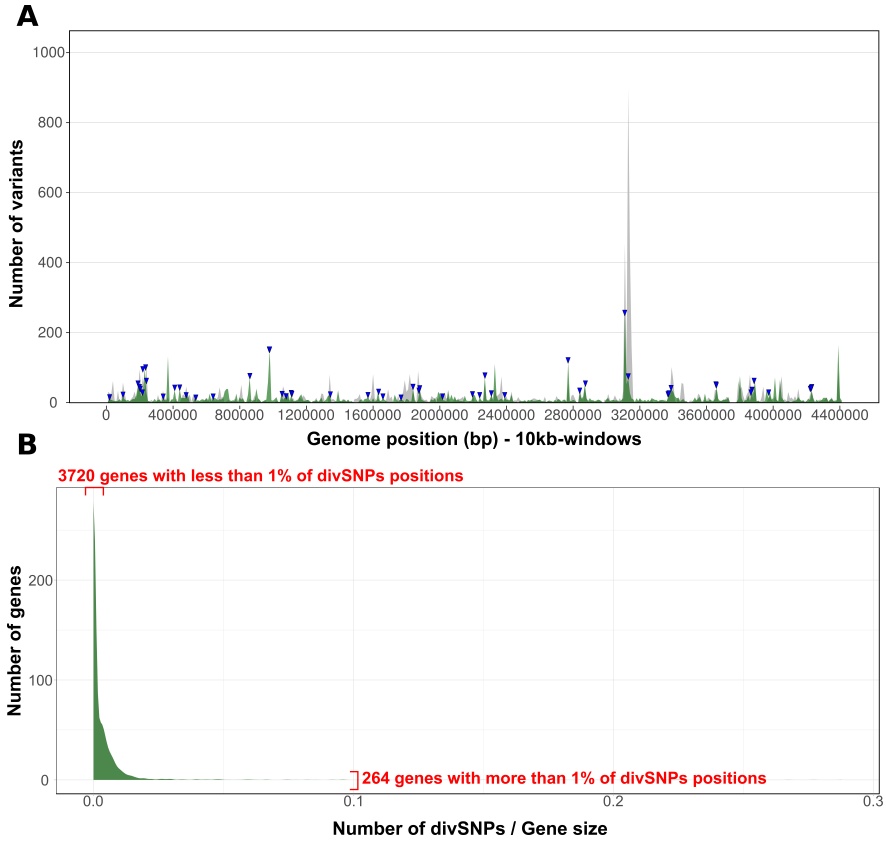


Figure 4.5: Divergent positions between the MTBC ancestor and the *M. canettii clade*. A) Average of divSNPs per 10 kb positions (green) as compared to the average of homoplastic variants (gray). Blue arrows above the distribution are genes that significantly accumulate more divSNPs. B) Accumulation of divSNPs per gene, corrected by gene length. A small number of genes accumulate a high amount of divSNPs while most of the genes have a low number of variants or even none. This pattern resembles those of high habitat overlap derived from Overlapping Habitat Models [112].

have accumulated divergence during the speciation process. To check for horizontal gene transfer events, we downloaded from Refseq and GenBank a set of 155 complete genomes from distinct mycobacterial species. We looked for orthologues of the 120 genes accumulating divSNPs between the MTBC ancestor and the rest of the mycobacterial species. For each gene, we reconstructed a Maximum Likelihood (ML) phylogenetic tree and each of these

phylogenies was compared to a ML reference built from the concatenated core mycobacterial gene set. Phylogenies for 53 of these genes placed MTBC within the MCAN clade, which is compatible with the accumulation of variants by mutation. Sixty-seven of the phylogenies were not topologically congruent with the reference tree. For all these genes, a BLAST search was performed against the NCBI database. In 54 cases the BLAST search gave a best hit with *M. canettii* and in one case no hits were returned. The most plausible explanation for this alternative topology is that recombination with other MCAN strains not included in our data set had occurred. On the other hand, in 12 cases the BLAST search showed a best hit with other mycobacteria, more specifically with *M. chimaera, M. kansasii, M. sp. 3/86Rv*, and *M. shinjukuense*. Interestingly, the consecutive genes from Rv2798c to Rv2803 followed this pattern, giving a better hit with *M. shinjukuense* than with MCAN. The *mazF9* and *mazE9* genes are in this region and were previously reported as a genomic island related with virulence and pathogenesis [131]. Finally, one gene, Rv2804c, returned no results for the BLAST search.

Thus, a total of 53 genes in the MTBC ancestral genome were highly divergent with respect to MCAN due to substitution events (Supplementary Table 10.1). While the genome-wide analysis identified divSNPs that might result from genetic drift or hitch-hiking events associated with selection on other loci, their accumulation in only 53 genes suggests that those regions might have played an important role during the process of niche differentiation. In agreement, those 53 genes are significantly more functionally conserved than the rest of the genome (dN/dS = 0.154 vs genome average dN/dS = 0.279, chi-squared p-value $\leq$ 0.001). This result suggests that, despite the increased divergence from the MCAN strains, those 53 regions have been evolving under purifying selection. Alternatively, the accumulation of divergent variants could also represent hotspot regions for mutation. None of the genes showed a similar pattern of mutation accumulation in other MCAN (no overlap between the divSNPs probabilities distributions for these 53 genes and the rest of the genomes, t-test p-value $<$ 0.05).

Regions under positive selection after the transition to obligate pathogen

Having established that some divSNPs accumulate in genes under purifying selection, we screened for positive selection patterns to identify additional genes relevant in the transition from a newly emerged pathogen to a globally established pathogen. We first revisited the evolution of antigenic proteins. These regions are recognized by the immune system and most of them are hyperconserved within the MTBC [87, 65]. Interestingly, and in agreement with previous data from MCAN genomic analyses [27], the dN/dS calculated in the branch leading to the ancestor showed a very similar pattern, with essential genes being more conserved than non-essential ones and T-cell epitopes being hyperconserved (Figure 4.6). Only nine divSNPs (5 synonymous and 4 nonsynonymous) were found in T-cell epitope regions, which is significantly less than expected by chance (Poisson distribution, p-value < 0.001).

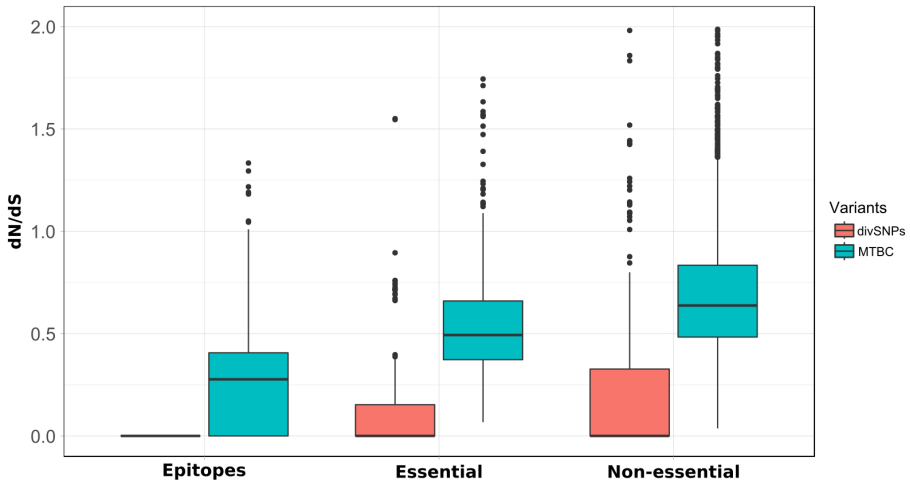


Figure 4.6: The dN/dS ratio distribution for the epitopes regions, essential and nonessential genes. The dN/dS ratios distribution match previous results. Values calculated in the MTBC ancestor branch meet the same pattern as those calculated from modern strains. T-cell epitopes are in both cases hyper-conserved.

Thus, antigenic regions do not show an altered pattern or intensity of selective pressure. We then explored what other regions of the genome changed significantly in selective pressure by comparing the MTBC ancestor dN/dS and the actual dN/dS in extant populations using our global reference dataset of 4,598 MTBC strains. We calculated a dN/dS for all the genes with at least one synonymous and one nonsynonymous mutation for each of the two sets (divSNPs versus within MTBC SNPs). Due to the low number of divSNPs in individual genes, only 499 genes were evaluated. Consequently, although additional genes to those shown in the ensuing analyses may have changed the selection pattern or intensity, they cannot be evaluated properly (External Data 4). We were particularly interested in those genes with a drastic change from purifying (dN/dS $<$ 1) to diversifying or positive selection (dN/dS $>$ 1) or vice versa.

Most of the genes evaluated did not show any sign of changing selective pressure or pattern. However, when looking at the dN/dS variation data, 14 genes appeared as outliers (as defined by Tukey's method [132](Figure 4.7A). Genes Rv1244 (lpqZ), Rv3910, Rv0166 (*fadD5*), Rv0874c, Rv1152, Rv1678, Rv1951c, Rv2584c (*apt*), Rv3026c, Rv3276c (*purK*), Rv3370, Rv3759c (*proX*) and Rv3900c were under a stronger negative selective pressure following speciation. Many of them are annotated [133] as hypothetical conserved proteins. On the other hand, only one gene changed to evolve under positive selection after divergence from the MTBC ancestor Rv0758, also known as *phoR*. Notably, PhoR forms part of the PhoP/PhoR virulence regulation system [134]. In the branch leading to the MTBC ancestor, this gene was as conserved at the amino acid level, as other essential genes (Chi-square test; p-value = 0.4721), but when we looked within the extant MTBC diversity, the gene was significantly less conserved at the amino acid level than essential genes (Chi-square test; p-value $<$ 0.001).

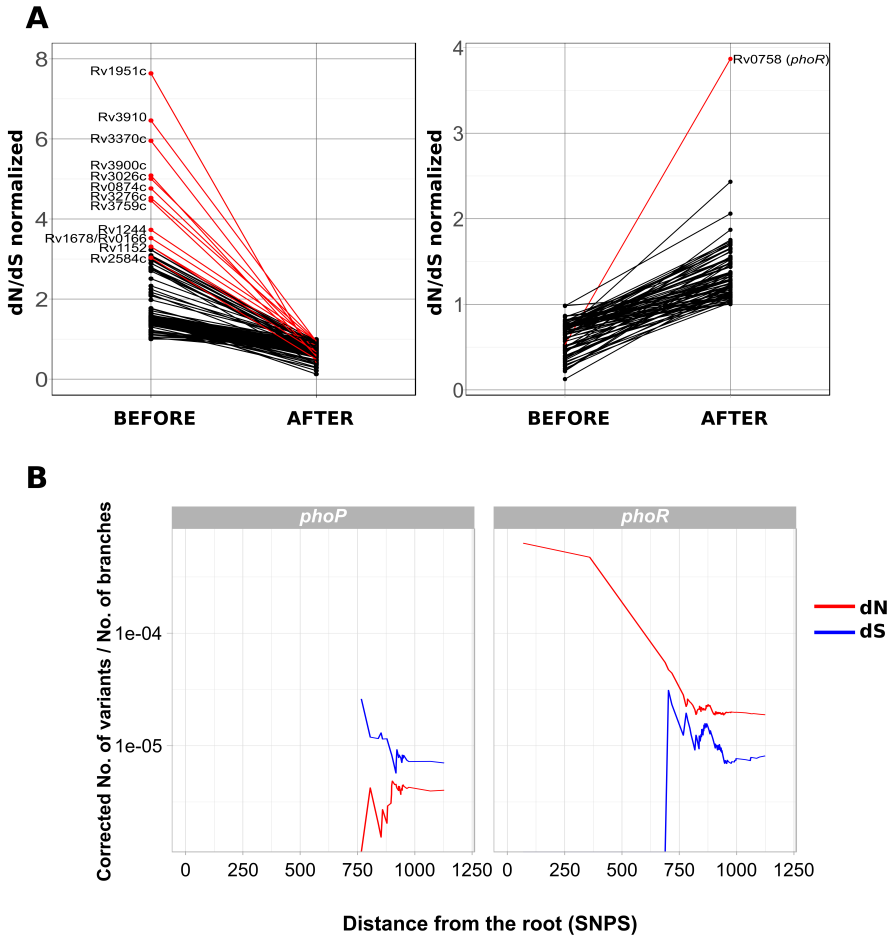


Figure 4.7: Genes with differential selective pressures across the MTBC speciation stages. A) Genes changing selective pressure in the branch of the MTBC ancestor as compared to extant MTBC strains. Red lines mark those genes being outliers of the dN/dS variation distribution. B) *phoR* and *phoP* show different selective pressure dynamics. In both cases the accumulation of nonsynonymous (dN) or synonymous (dS) mutations through time is measure as the distance to the most common ancestor of the MTBC. The dN and dS values have been corrected by the number of branches in the phylogeny at each time point.

Positive selection on *phoR* linked to ongoing selective pressures

Given the known central role of PhoPR in MTBC virulence, we focused our attention on the new mutations found in *phoR*. We observed a total of 193 nonsynonymous mutations and 31 synonymous mutations in *phoR* (Figure 4.8). The average dN/dS for this gene was well above 1 (dN/dS = 2.37), suggesting the action of positive selection.

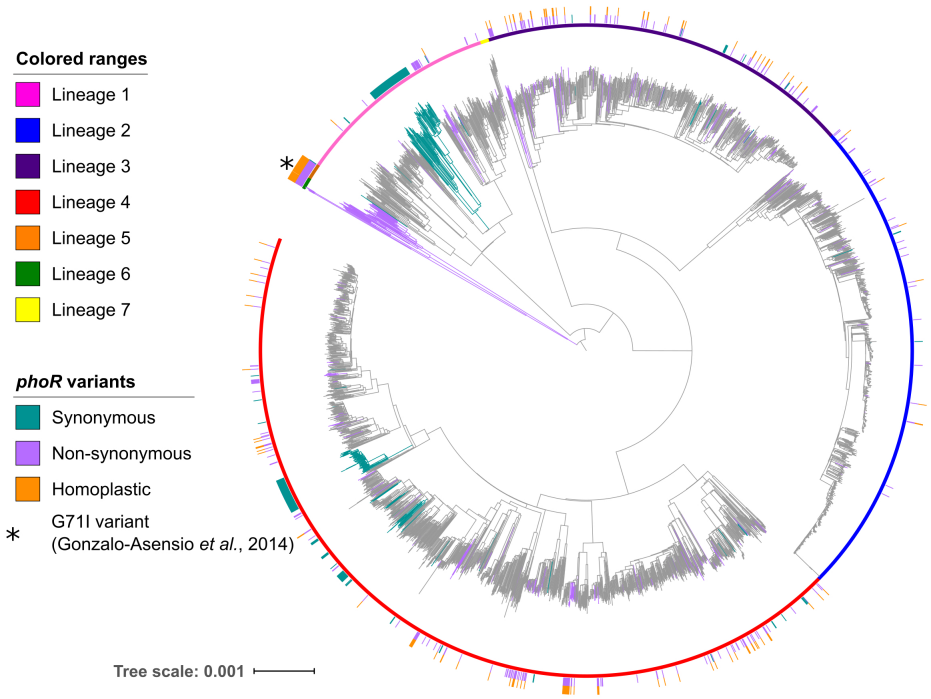


Figure 4.8: *phoR* mutations are phylogeny-wide. Genome-based phylogeny calculated from a total of 4,595 clinical samples obtained from different sources. The synonymous and nonsynonymous variants found in *phoR* are mapped to the corresponding branch. Variants in internal branches affect complete clades which are colored in the phylogeny. Homoplasies are marked in the outer circle of the phylogeny. The star marks the G71I PhoR variant common to L5 and 6 previously reported by Gonzalo-Asensio *et al.*[135]

Furthermore, a plot of the dN and dS values over time reveals that the overall dN/dS remained high along the evolutionary history of the MTBC (Figure 4.7B) corroborating that this gene has likely been under pervasive positive selection. Codon-based maximum likelihood tests of positive selection normally are not suited for intraspecies comparisons. However, in the case of *phoR* the tests identified a higher dN/dS than expected by chance and at least two codons with strong evidence to be under positive selection (Table 4.2). Additional evidence for the action of positive selection on this gene derives from nonsynonymous mutations, among which we found 34 homoplastic variants, which are strong predictors of positive selection in MTBC (External Data 5). Nonsynonymous mutations significantly accumulated in the sensor domain (Chi-square test, p-value < 0.01), further supporting the hypothesis that they could be involved in the fine-tuning of the PhoR sensitive function to the changing environment during infection (Figure 4.9B).

Site	α	β	β - α	**Prob[$\alpha > \beta$]**	**Prob[$\alpha > \beta$]**
71	0.64	11.264	10.624	0.007	0.984
355	2.151	10.592	8.441	0.047	0.915

α = Mean posterior synonymous substitution rate at a site.
β = Mean posterior nonsynonymous substitution rate at a site.
Prob[$\alpha > \beta$] = Posterior probability of negative selection at a site.
Prob[$\alpha > \beta$] = Posterior probability of a positive selection at a site.

Table 4.2: Codons with strong evidence of being under positive selection as detected by FUBAR.

All the mutations identified in our analysis were found in human clinical isolates and mapped to relatively recent branches in the MTBC phylogeny (Figure 4.8). Thus, we reasoned that most mutations were associated with recent selective pressures as opposed to the previously reported mutations found in *Mycobacterium africanum* L5 and L6, and the animal-adapted clade [135] that map to deep branches in the phylogeny (Figure 4.8). To get insights in this hypothesis, we tested whether novel *phoR* mutations are also arising in clinical settings during infection and recent transmission events. We used a

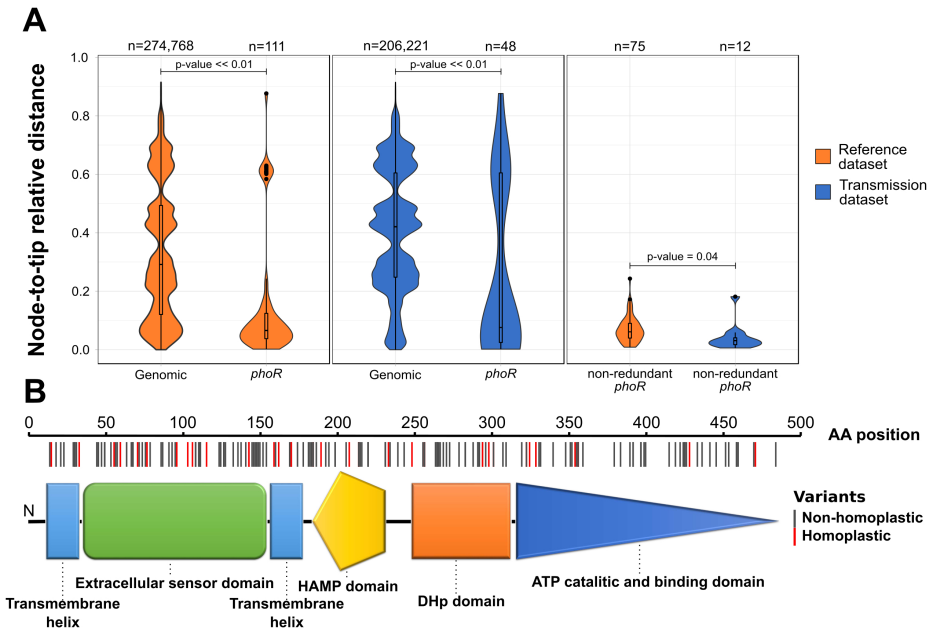


Figure 4.9: Characteristics of the *phoR* nonsynonymous mutations.A) Relative ages distribution of the *phoR* variants in the reference dataset from Coll *et al.*[93] (left pannel) and the transmission dataset [136] (middle pannel) in comparison with the rest of the genome variants. In the right pannel, the relative age of the *phoR* variants exclusive from each of the two datasets were compared. B) Schematic view of PhoR with the amino acid changes found across the 4,595 samples dataset marked on it. Amino acid changes are significantly more abundant in the sensor domain (p-value $<$ 0,01).

population-based data set from Malawi [136] where more than 70% of the strains were collected during fifteen years and their genomes sequenced (n = 1,187). We found 13 mutations (12 nonsynonymous and 1 synonymous) in *phoR* exclusive of the Malawi data set and a *phoR* dN/dS value of 3.93. Moreover, the mean relative age of the nonsynonymous *phoR* variants were significantly younger than that of other nonsynonymous variants in both datasets (Welch's t-test, p-value $\ll$ 0.01) and the *phoR* variants from the Malawi dataset were more recent than those *phoR* mutations from the reference dataset (Welch's t-test, p-value = 0.04)(Figure 4.9A). From the 12

nonsynonymous mutations in the Malawi data set, 8 were markers of recent transmission clusters. Moreover, *phoR* mutations in the Malawi data set involved larger transmission clusters than other mutations (permutations test, p-value < 0.001).

Taken together, there is strong evidence for positive selection acting on *phoR* steaming from higher than expected dN/dS values in the reference dataset, presence of homoplastic variants and new nonsynonymous mutations linked to larger transmission clusters. Thus our data indicates that (i) *phoR* mutations have been selected since the establishment of the MTBC as an obligate pathogen (Figure 4.7); and (ii) novel *phoR* mutations are selected during infection and propagates during human to human transmission in current epidemiological settings (Figure 4.9).

4.3 Discussion

We present evidence that the MTBC ancestor transitioned to an obligate pathogenic lifestyle from a common genetic pool including the ancestors of extant MCAN strains. Specifically, we found common patterns of genome-wide recombination in the branch leading to the MTBC ancestor and the extant MCAN strains. The high recombination rate between MCAN strains, including the MTBC ancestor, stands in sharp contrast to the strictly clonal population structure of extant MTBC strains. By analyzing events leading to the transition from a recombinogenic to a clonal organism, we have also been able to identify genomic regions under different selective pressures. The comparison between selective pressures before and after becoming an obligate pathogen also allow us to propose PhoR as an important player in the past evolutionary history of the MTBC as well as in current clinical settings.

Population genomics data has led to the development and testing of different models of how different genetic clusters of the same species can arise in sympatry [110, 112, 137]. In the case of *Vibrio cholerae*, an appropriate combination of certain virulence-associated variants, ecological opportunity

and additional virulence factors mediated the successful transition of a particular clone from an environmental to a pathogenic lifestyle [138]. Other known cases such as pathogenic *Salmonella* [139] or *Yersinia* species [140] may have followed a similar trajectory. The MTBC represents an extreme case of clonal emergence associated to its obligate pathogenic lifestyle. Here, we have shown that, despite the high ANI between MCAN and the MTBC, there is complete genomic isolation between these organisms. There is experimental evidence that genetic exchange among MCAN strains occurs easily but not between MCAN and the MTBC [118]. We have shown that there is no measurable ongoing recombination among the MTBC strains based on our analysis of 1,591 genomes and in agreement with other recent reports [141, 119]. It is important to note that, due to the low divergence within the MTBC, most methods to detect recombination are limited. Hence, we cannot completely exclude the possibility that we might have missed some recombination events. It was previously suggested that recombination (or gene conversion) could be affecting PE/PPE genes disproportionally [142]. Unfortunately, short reads cannot be properly mapped to these regions so our approach does not allow testing this possibility. However, if recombination does occur in the MTBC, it seems to have a minor impact on the overall genetic diversity of the MTBC. Recombination in natural populations depends both on the capacity of chromosomal DNA exchange between the two groups involved and on the ecological opportunity. The mechanisms, if any, by which the MTBC bacilli lost their capacity to recombine while the ancestral genetic pool showed very similar recombination patterns to MCAN strains, remains to be elucidated. Ecological opportunity may also influence on the lack of opportunities for exchange between MTBC strains. Despite the occurrence of super-infections, the bacilli occupy mainly an intracellular lifestyle, thereby reducing the opportunities for genetic exchange.

We can only speculate about how the transition from a likely environmental or opportunistic pathogen to an obligate pathogen occurred, but our analysis has identified a series of non-random evolutionary events. Notably, these

events involve core pathogenesis genes. We have identified highly divergent regions in the MTBC ancestor compared to MCAN. The pattern of SNP accumulation suggests that those regions were important in the transition to a closer association with the host. In addition, recombination events mapping to the branch leading to the MTBC ancestor affected essential genes as well as genic regions known to be involved in host-pathogen interaction. The *mymA* operon (Rv3083-Rv3089) is related to the production of mycolic acids and its disruption leads to an aberrant cell-wall structure. Importantly, knock-out studies [143] have shown that this operon is essential for growth in macrophages and the spleen of infected mice. Furthermore, the deletion of genes in this operon leads to a higher TNF-alpha production, highlighting their role on regulating host-pathogen interactions [144]. The other major operon identified in our analysis is the *mce1* operon [145]. *mce1* knock-out mutants are hypervirulent in a mouse model of infection and lose the capacity of a proper pro-inflammatory cytokine production that is needed for the establishment of the infection [146] and granuloma [145]. How these processes are mediated by *mce1* is still not clear, pointing at this gene as a priority target for biomedical research.

Our analysis identified one gene, *phoR*, which is under positive selection in extant MTBC strains although it was under purifying selection in the MTBC ancestor. PhoR is the sensor component of the PhoPR two-component system, which plays a major role in MTBC pathogenesis [147, 148]. Previous experimental data show that 1) PhoPR is a major virulence determinant in MTBC [134]; 2) that deep phylogenetic branching mutations in PhoPR were involved in the adaptation of the pathogen to different mammalian hosts [135] and that there is at least one case in which natural overexpression of PhoPR in a *Mycobacterium bovis* clinical isolate was linked to a highly transmissible and virulent phenotype in humans [147]. In fact, mutations affecting the whole animal clade in *phoR* have been proposed to fine-tune MTBC virulence across different animal host species. We find alternative amino acid changes in the same codon experimentally tested by Gonzalo-Asensio *et al.* (2014), thus

changes in this codon could have been selected multiple times in unrelated human isolates. Based on these findings, we speculate that recent *phoR* mutations help to fine-tune the immunogenicity of the pathogen during infection, allowing it to manipulate the human host responses and increase the chances of transmission. However, we still need to understand the stimuli and the molecular pathways that are at the basis of the selective pressures driving the evolution of *phoR*. Given that PhoPR is involved in membrane composition [149], mutations in this regulator might also be involved in susceptibility to some antibiotics. However, antibiotic selection is an unlikely explanation for the oldest mutations in PhoPR, as they likely predate antibiotic usage.

Based on our findings, a model can be proposed in which recombination, together with the acquisition of new genetic material [114, 150], generated a favorable genetic background for the MTBC ancestor to occupy or increase its association with mammalian hosts. We see this emergence only once in the MTBC, perhaps because the right combination of multiple, fortuitous genetic events and the particular ecological conditions has occurred only once. More provocative is the idea that MTBC might just be part of a spectrum of association to the host occupied by the different MCAN-MTBC groups. The fact that the so-called Clone A of MCAN strains are more common in the clinic may suggest differences in ecological niches within the MCAN group [122]. In agreement, previous publications [122, 27] and our own analysis (Supplementary Figure 10.2) have identified Clone A strains as the closest MCAN evolutionary group to MTBC.

In the MTBC, the strong and obligate association with new host(s) was accompanied by new selective pressures. In accordance, we identified genes in the MTBC genome highly diverging from MCAN and evolving under purifying selection, suggesting that they have become essential following MTBC's transition to an obligate pathogenic life-style. In the final stages of adaptation, positive selection on genes such as *phoR* and others [151, 152, 153] likely led to a narrowing of the host-range and later still to a further fine-tuning during the spread of the bacteria within the new host species.

4.4 Materials and methods

Datasets used

***Mycobacterium canettii* dataset.** The *M. canettii* dataset is composed by seven draft genomes downloaded from GenBank (CIPT 140010059, NC_015848.1; CIPT 140060008, NC_019950.1; CIPT 140070008, NC_019965.1; CIPT 140070002, NZ_CAOL00000000.1; CIPT 140070005, NZ_CAOM00000000.1; CIPT 140070013, NZ_CAON00000000.1 and CIPT 140070007, NZ_CAOO00000000.1).

MTBC datasets. We have downloaded all the available genomes from the studies of Coll *et al.* 2014 [93],Walker *et al.* 2015 [154], Guerra-Assunção *et al.* 2015 [136] and Comas *et al.* 2015 [35]. The total number of sequences originally downloaded were 7,977 genomes. For the dN/dS calculations and *phoR* variants screening, we used all the downloaded genomes, with the objective of incrementing the robustness of the measures and the number of variants per gene. We identified all clusters at a maximum distance of 15 fixed SNPs (common threshold in MTB epidemiology), removed samples potentially coinfected with more than one strain, and then randomly select just one representative from each cluster. Thus, the final number of genomes for these analyses were 4,595. The rest of the analyses were performed in smaller subsets of samples, due to computational limitations or the specific features of each dataset. A 1,591 sequences subset from the Coll *et al.* 2014 samples was used for the recombination analyses within the MTBC, as they include global representatives of the MTBC diversity. A smaller subset of these, which included 219 sequences corresponding also to global representatives, was used for Gubbins because it was not computationally feasible to run the program with more strains. Finally, genomes from the Guerra-Assunção *et al.* 2015 dataset, which includes samples taken over a 15-year period in a high transmission setting (thus enriched in transmission clusters) was used for the *phoR* transmission analysis (n=1,187). Information about all the strains used in this study (including its accession numbers) can be found in External Data 6.

Phylogenetic inference and parsimony mapping of SNPs

In the subset of 1,591 strains of the MTBC dataset, we identified 140,239 variants by applying the pipeline defined in Chapter 2. As we wanted to identify nucleotide variants due to recombination events, a stricter filtering was applied to remove putative recombination signal due to polymorphisms introduced by other causes. To avoid false positives, we also removed positions in which a variant was called in at least one strain but also with a gap in at least another strain. Variants related to antibiotic resistance were obtained from PhyResSe [94] and were removed from the analysis. Also, non-biallelic variants were removed from the analysis. To identify variants resulting from mapping errors we generated fragments of 50 bp. downstream, upstream and midstream of the variant positions in the reference genome. With these fragments, we performed a BLAST search over the reference genome to check whether they mapped to other regions. Variants identified in reads that mapped to more than one region of the reference genome (query coverage per HSP over 98% and percentage of identical matches between the query and the reference genome of 98%) were removed from the analysis.

The remaining variants (94,780) were used to infer a phylogenetic tree using RAxML [155] with the GTRCATI (GTR + optimization of substitution rates + optimization of site-specific evolutionary rates) model of evolution. Variants were mapped to the phylogeny using the Mesquite suite [156]. Homoplastic variants were identified based on parsimony criteria. Using these homoplastic variant positions, we looked for consecutive homoplastic variants (allowing at least one variant between them). The detected variants were mapped on the phylogeny using Mesquite to look for coincident phylogenetic patterns.

Linkage-disequilibrium calculation

Using the filtered variant positions (94,780), we used the PLINK software [157] to calculate the linkage-disequilibrium statistics D' and R^2. To estimate these values, we took into account variants with a minimum frequency of 0.01 and

used a sliding window of 10 Kb. To plot the D' and R^2 pattern by variant distance, we calculated average D' and R^2 values for 50 bp. windows.

Multiple alignment of *M. canettii* and MTBC

Seven *M. canettii* draft genomes were aligned to each other and to the ancestor of MTBC using progressiveMauve [158]. The segmented alignment obtained in XMFA format was converted to a plain FASTA format using the MTBC ancestor as reordering reference with a custom Perl script. Positions with gaps in the reference sequence were removed from the final alignment, so the resulting aligned genomes had the same size than the reconstructed MTBC ancestor (4,411,532 Mb). The MTBC pseudogenomes reconstructed from mapping to the MTBC ancestor from the different datasets described above were concatenated to the *M. canettii* alignment obtained in the previous step for further analyses.

From these alignments, homoplastic variants were identified using both, parsimony and maximum-likelihood approaches [159]. Both approaches agreed in identifying the same homoplastic variants.

Recombination analyses and phylogenetic evaluation

Besides SNPs, linkage-disequilibrium analysis and Gubbins, RDP4 [160] was used to detect recombination signal in the MTBC dataset. To mark the regions reported by Gubbins as potentially recombinant we required at least three of the methods implemented in RDP4 to agree in showing a significant signal.

Recombination was evaluated in the alignment containing 219 strains from MTBC and 7 *M. canettii* and in the one containing the MTBC ancestor and 7 *M. canettii*. Firstly, repetitive regions (i.e. PPE/PGRS) were masked from both alignments and, secondly, recombination events were inferred using Gubbins [161], which identifies clusters of high SNP density as markers.

Gubbins identified 70 potential recombinant regions in the alignment containing the 7 MCAN strains and the MTBC ancestor. Four of these regions

were obviated because they fell in regions deleted in several *M. canettii* strains. One more region was removed from the analysis because it was extremely short (41 bp.) and we did not obtain reliable results in the subsequent analyses.

For the remaining 65 fragments a phylogeny was calculated using RAxML [155] and applying the GTRCATI model. Also, a reference phylogeny was calculated with the same method using the complete genomes after subtracting these 65 regions. This reference phylogeny had the same topology as the one obtained from the complete genomes. To test for phylogenetic incongruence between the putative recombination fragments and the genome phylogeny, we applied the Shimodaira-Hasegawa and Expected Likelihood Weight tests implemented in TREE-PUZZLE [162].

Dating analyses

To infer the age of the 65 recombinant fragments we first reasoned that most of the mutations found were contributed by recombination and not by mutation once the fragment had been integrated in the genome. Thus, before dating the fragments we first removed all the homoplastic variants with other MCAN strain found in the fragments. The final alignments for the 65 fragments consisted of only those variants accumulated after the recombination event. We then used the non-recombinant part of the genome to infer a substitution rate assuming two different dating scenarios published for the tMRCA [36, 82]. We ran BEAST for each fragment pre-specifying monophyletic groups and substitution rate based on the non-recombinant genome phylogenetic reconstruction. We used an uncorrelated log-normal distribution for the substitution rate in all cases and a skyline model for population size changes. We ran several chains of up to 10E6 generations sampling every 1E3 generations to ensure independent convergence of the parameters. Convergence was assessed using Tracer [163]. For both evolutionary scenarios, the results obtained were largely congruent and proportional to the age limit imposed for the MTBC ancestor. The 5ka scenario [82] was selected for plotting the ages in Figures 4.4 and Supplementary Figure 10.5, as there is now more evidence for this timeframe.

divSNP analysis

From the MCAN and MTBC ancestor alignment, we extracted those positions having one variant in all the *M. canettii* strains and another variant in the MTB ancestor. The divSNP frequency by nucleotide was calculated by dividing the total number of divSNPs (5,688) by the total number of bases in the alignment. Next, the expected abundance of divSNPs for each gene was calculated by multiplying the nucleotide divSNP frequency by the number of nucleotides in each gene. From the expected and observed divSNP abundances, we used a Poisson distribution to calculate the probability of having the observed divSNPs by chance for each gene. We selected genes having a pFDR $\leq$ 0.01 using the q-value from Storey method [164].

Complete mycobacterial genomes for reference strains [21](External Data 7) were downloaded from RefSeq and GenBank. The orthologous genes were obtained from the amino acid sequences and using the Proteinortho tool [165]. A gene was considered as orthologous based on reciprocal best hits in BLAST. BLAST analysis required a minimum identity of 25%, a query coverage of 50% and a maximum e-value of 1E-05. The orthologous genes were aligned using Clustal-omega [166] and the phylogenies were constructed using RAxML and applying the PROTCATIAUTO model. The reference phylogeny was constructed using only the core genome (proteins having orthologous in all the mycobacterial genomes downloaded) with RAxML using the same options as above. The reference and alternative phylogenies calculated with the orthologous genes for the divSNPs enriched genes were manually inspected to check for congruence.

dN/dS analysis

The potential synonymous and nonsynonymous substitution sites for each region were calculated using the SNAP tool [167]. The dN/dS ratio for each region was calculated using equation 4.1.

$$\frac{\text{Nonsynonymous variants}/\text{Nonsynonymous sites}}{\text{Synonymous variants}/\text{Synonymous sites}} \tag{4.1}$$

The dN/dS for the MTBC ancestor was calculated using the divSNPs while the dN/dS for the MTBC were calculated using 208,238 variants detected in coding regions from the 4,595 strains in the MTBC global data set. To look for a robust comparison between both ratios, only genes having at least 1 synonymous and 1 nonsynonymous variants were taken into account. To compare the dN/dS ratios, both were normalized by the genomic dN/dS for each taxon (0.24 for the MTBC ancestor and 0.59 for the MTBC). The difference between the dN/dS ratio was calculated by subtracting the MTBC dN/dS to that of the MTBC ancestor. The genes that account for the largest differences in the dN/dS were identified as outliers (equation 4.2) of the differences distribution [132].

$$\begin{aligned}(Q2 - 1.5 \times IQR)\\(Q3 + 1.5 \times IQR)\end{aligned} \tag{4.2}$$

phoR positive selection analysis

Positive selection on *phoR* was tested using FUBAR [168] and BUSTED [169]. FUBAR was run with 5 MCMC chains of length 10,000,000. 1,000,000 states were used as burn-in and a Dirichlet prior of 0.5. BUSTED was run with default parameters. To study the potential effect of *phoR* mutations on transmission efficacy we used the data set from Guerra-Assunção *et al.* [136]. We identified SNPs in branches leading either to leaves or to transmission clusters. Transmission clusters were categorized in large, medium or small according to the number of isolates in the cluster (large = over 75th percentile, medium=between 25th and 75th percentile, small = under 25th percentile). Each gene was scored to check for accumulation of mutations in branches leading to large transmission clusters according to equation 4.3.

$$\text{Score} = \text{Large clusters} * 3 + \text{Medium clusters} \quad (4.3)$$

Genes with high mutation rates have a higher number of polymorphisms that could lead to a larger score by chance. To test the probability of obtaining the observed score by chance, a permutation test was carried out 10,000 times. Each of the identified SNP was randomly reassigned to the same branches and the score was recalculated for each gene. The expected score distribution for each gene was compared to the observed score to calculate the probability. This test was performed for transmission events defined at 10 SNPs. The ages for the variant positions were calculated as node-to-tip distances. These distances were relativized to the maximum root-to-tip distance to obtain a relative age value in the [0 - 1] range. In order to have a common framework, a phylogeny was constructed including all the samples from the transmission and the reference data sets. The phylogeny was constructed using RAxML and applying the GTRCATI model. For each variant position we first identified the node in which the variant appeared. The node-to-tip distance was calculated afterwards for each node using the geiger package [170]. Distances were normalized to obtain a relative distance. Later, all the nonsynonymous variants except the phoR polymorphisms were used as a reference set. The nonsynonymous phoR variants to be compared were categorized in two groups, those exclusive to the reference dataset [93] and those derived from the transmission data set [136].

PhoR domains and structure representation

The PhoR domains structure was inferred by using PFAM [171] and SMART [172].

The work described in the present chapter has been published as a Research Article: Chiner-Oms Á., Sánchez-Busó L., Corander J., Gagneux S., Harris S., Young D., González-Candelas F., Comas I. Genomic determinants of speciation and spread of the *Mycobacterium tuberculosis* complex. **Science Advances**. In press.

Impact of the global genetic diversity on the bacterial biological networks.

5.1 Introduction

In the last years, relevant studies based on TB biological networks analysis have been published. Overexpression experiments and chromatin-immunoprecipitation sequencing (ChIP-Seq) data have been used to produce a detailed map of the interactions and regulatory logics of more than 200 transcription factors (TFs) [173, 174]. In addition, networks of protein-protein interactions (PPI or interactome) have been derived using both, experimental and computational approaches [175, 176, 177, 178]. The enormous amount of data generated is publicly available and can be used to study the interactions of the bacteria in several ways [179, 180]. For example, computational models mimicking the regulatory behaviour of the bacteria have been derived from these networks and were used to predict expression changes under different conditions [181].

However, these networks have been derived based on H37Rv, a clinical reference strain. Little attention has been paid to the fact that H37Rv is a clinical strain used in laboratories for decades and that in many aspects it does not represent the whole species. Therefore, natural perturbations in the

biological networks inferred in H37Rv, introduced by naturally occurring mutations in clinical strains, will likely change the models architecture and the predictions derived from them.

The phenotypic role of mutations defining lineages has been extensively studied and some of them are clearly linked to transcriptional differences between the MTBC lineages [182, 183, 184]. It is also clear that one single mutation affecting regulatory processes can impact dramatically on the virulence of the pathogen [135, 185]. In fact, a novel live vaccine, attenuated by carrying a deletion in the key regulator PhoP, is currently in phase 2A of clinical trials [186]. In the general introduction it has been stated the implications of the bacterial genetic diversity for the epidemiology, host immune response and disease progression of the TB disease. As a result of this diversity, novel diagnostics, vaccines and treatments may be compromised by failing to account for the circulating diversity as recently described for several diagnostics tests based on the detection of the protein of Mpt64 [187]. Thus, we are completely blind on whether the topology of the regulatory network and the computational models derived from H37Rv can be extrapolated to other strains of the MTBC and on how the regulatory modulations are affected by the existing bacterial diversity. In the case of PPI networks, there is data from human cells suggesting that the network topology is important for the distribution of synonymous and nonsynonymous substitutions [188]. The identification of key functional nodes, necessary for maintaining the interactome structure and functionality could be of great interest as new biomedical targets.

In this chapter, we derive new gene expression models by pooling existing H37Rv data and explore their predictive power on genome-wide expression patterns when natural variations (mutations) found in clinical strains are considered. We show how different experimental setups can affect the inferred models of gene expression and regulation and how far we are from predicting, only from transcriptomic data, the impact of genetic polymorphisms at a genome-wide expression level. In addition, we study the impact of the global diversity of the MTBC over the topology of the PPI network. We identify

network nodes that are key to maintain the interactome structure and function.

5.2 Results

Building and validation of gene expression models based on strain H37Rv, lineage 4

By taking advantage of recently published experimental datasets testing the regulatory influences of known TFs, we defined gene expression models for the laboratory reference strain H37Rv. The datasets included transcription factors overexpression experiments (TFOE) for $\sim$200 TFs ($\sim$700 tiling microarray experimental tests) [173]. Our aim was to model, for each gene, the level of expression and the resulting changes therein as a function of varying the expression of each TF. We built the models using a linear regression approach as described previously by Galagan *et al.* [181].

Using a backward stepwise algorithm (Figure 5.1, see Materials and methods for details), we generated 3,960 putative gene expression models. When, in addition, we required evidence of physical interaction from ChIP-seq data, the number of initial models was reduced to 755. Therefore, our putative gene models accounted for 98.3% of the coding capacity of the genome when physical interaction was not required, and only 19.24% of it when we used the ChIP-Seq data. Secondly, we cross-validated all the models in the two data sets and then compared them with random models to discard spurious results. Following this approach, we discarded 2,744 models for the TFOE and retained 1,216 (30.8%). For the case of ChIP-Seq data, only 29 models were retained (3.74% of the initial models) (Figure 5.2A). The models derived from TFOEs alone included a larger number of TFs (regressors) per model, as expected due to the larger number of regulatory events incorporated. On the contrary, the models derived from the combination of TFOEs and ChIP-Seq data had fewer TFs influencing the expression, as they only include those TFs physically bound to the gene (Figure 5.2). In summary, our approach shows the

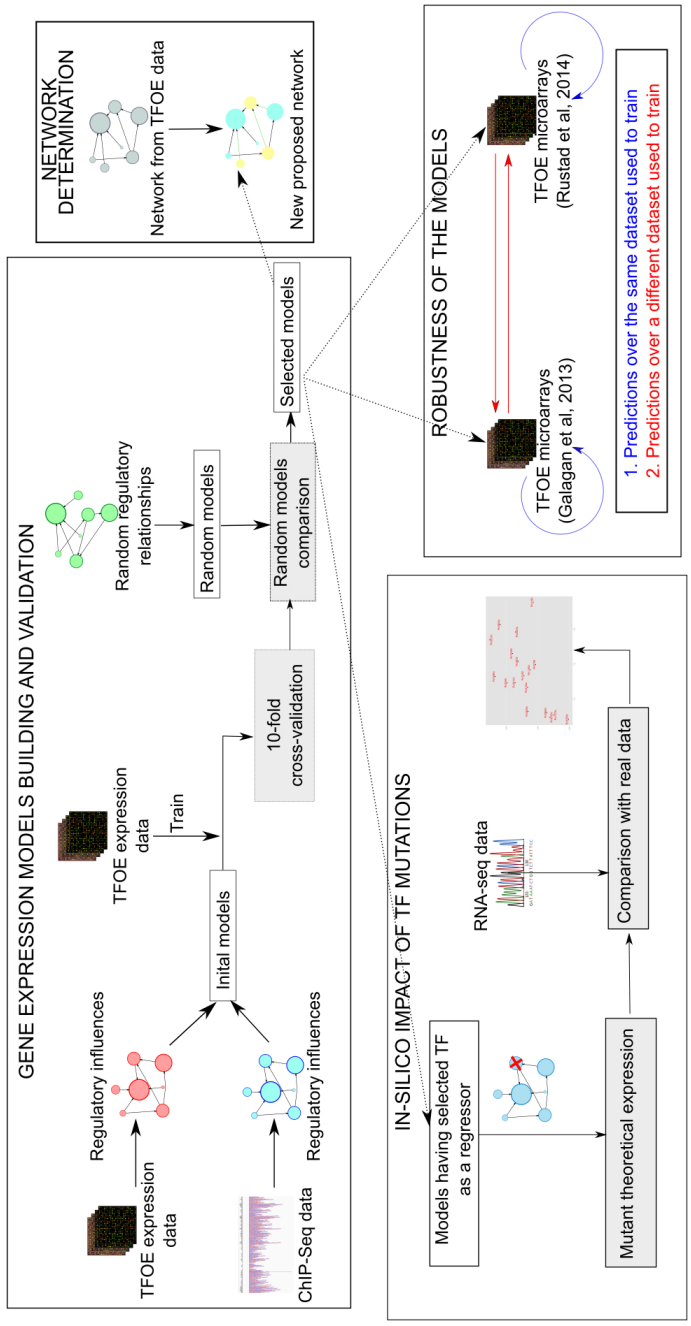


Figure 5.1: Workflow to build and validate gene expression computational models. Initial models were derived from the regulatory relationships derived from the TFOE and ChIP-Seq data. They were trained with the data set from Rustad *et al.* [173] These initial models went through a 10-fold cross-validation process, to discard low accuracy models. The remaining ones were compared with random models. Those showing better performance than the random models were selected as final models. These models were used to (i) cross-check the robustness of the models, (ii) derive a new regulatory network, and (iii) predict the impact of a TF deletion in silico.

relevance of performing sequential statistical validations of expression models derived from experimental data.

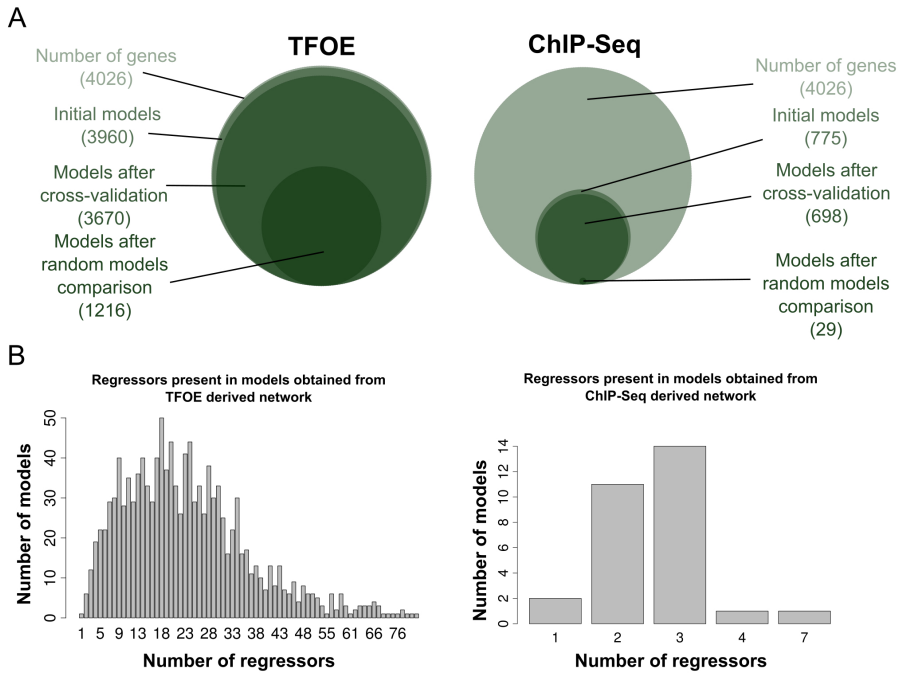


Figure 5.2: Gene expression computational models. A) Overview of the results obtained during the building process and refining steps of the computational models derived from the TFOE data (left) and the ChIP-Seq data (right). B) Distribution of the number of TFs affecting the target gene on each network model.

There is a limited agreement between the values predicted by the TFOE-derived models and the observed values (average of the Pearson's correlation coefficients = 0.71) (Figure 5.3A) despite being trained by the same dataset. To evaluate how robust the predictions were to experimental noise, we compared them with the expression values obtained in a previous, analogous TFOE experiment [181]. To compare the predictive power of the models across data sets we used the housekeeping gene Rv0001 (*dnaA*) as a reference for the expression values. We measured the average fold-change in expression

values between *dnaA* and the remaining genes across all the samples. When we compared the fold-change with the observed Rustad dataset, we found a correlation coefficient of 0.98 (p-value $<$ 0.01) (Figure 5.3B). When we compared the predicted fold-change with the Galagan dataset, the correlation coefficient was 0.97 (p-value $<$ 0.01). As expected, in the first case the correlation was almost perfect, because the models are making predictions over the same data set used to calculate their regressors' coefficients. However, predictions over a similar but different data set were less accurate but still a high predictive power was achieved.

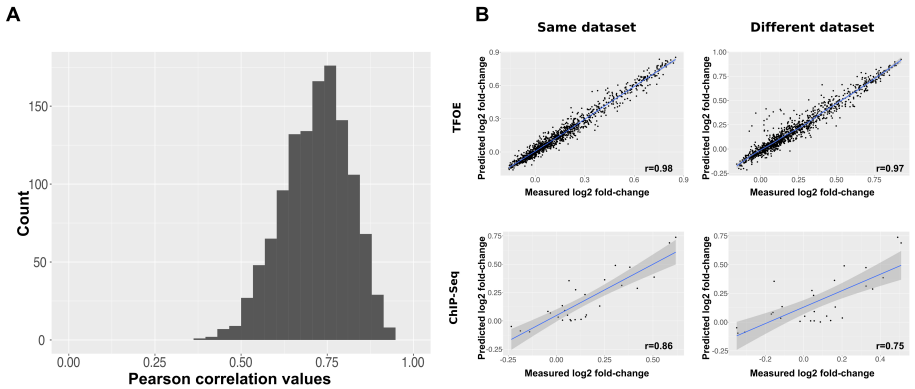


Figure 5.3: Comparison of models performance over different data sets. A) Distribution of the Pearson correlation values obtained when comparing each gene expression measure with its predicted expression. B) There is one data point for each model. Each dot in a plot is a measure of the gene expression fold-change between *dnaA* and the gene represented by this model. The y-axis corresponds to the predicted fold-change in gene expression while the x-axis corresponds to the measured fold-change. The upper row refers to the models derived from the TFOE data set. The lower row contains the models derived from the ChIP-Seq data set. The values in the left column were calculated when training and predictions were performed with the same data set. The values in the right column were calculated when different data sets were used for training and predicting.

Having established that gene expression models can predict gene expression trends in TFOEs experiments, we tried to predict absolute expression values in the Galagan data set [181]. We correctly predicted the

expression for only 128 genes (10.52% of TFOEs-derived gene models, pFDR $\leq$ 0.01). In fact, the comparison of average expression values for each gene between the two datasets (Galagan vs Rustad) revealed that only in 18 cases the mean expression values for the same gene were not different (pFDR $<$ 0.01, see Supplementary Figure 10.6 for more information). Taken together these results show that experimental noise across laboratories has a large influence on the results for analogous experiments, at least for the prediction of absolute quantitative expression levels.

Regulatory network based on statistically validated interactions

The 1,216 expression models obtained from the TFOE data set included 11,253 regulatory relationships. These relationships are the ones selected after applying the backward step-wise method in the building process of the models (see Material and methods for details). Although all of them led to a lower Bayesian Information Criterion in their respective models, most of these relationships are based on a weak regulatory signal. To select the strongest links between TFs and gene regulation influence, we kept those leading to a significant change in gene expression (two-fold change) according to TFOE data [173]. We built a new regulatory network with these subsets of regulatory relationships. The new network comprised 3,396 regulatory events across 1,102 genes (37.15% of the events and 38.76% of the genes from the network proposed by Rustad *et al.* (2014)). The distribution of the in-degree parameter of the network (Figure 5.4) revealed that most genes are regulated by an intermediate number of factors whereas a minority is regulated by a large or small number of them. On the other hand, the distribution of the out-degree parameter followed the expected power-law distribution [189], with most TFs regulating a small amount of genes and a few genes affecting the regulation of many (see Table 5.1 for more details).

In agreement with the original networks of Rustad *et al.* and Galagan *et al.*, in this new regulatory network Rv0023 and Rv0081 are the TFs that regulate

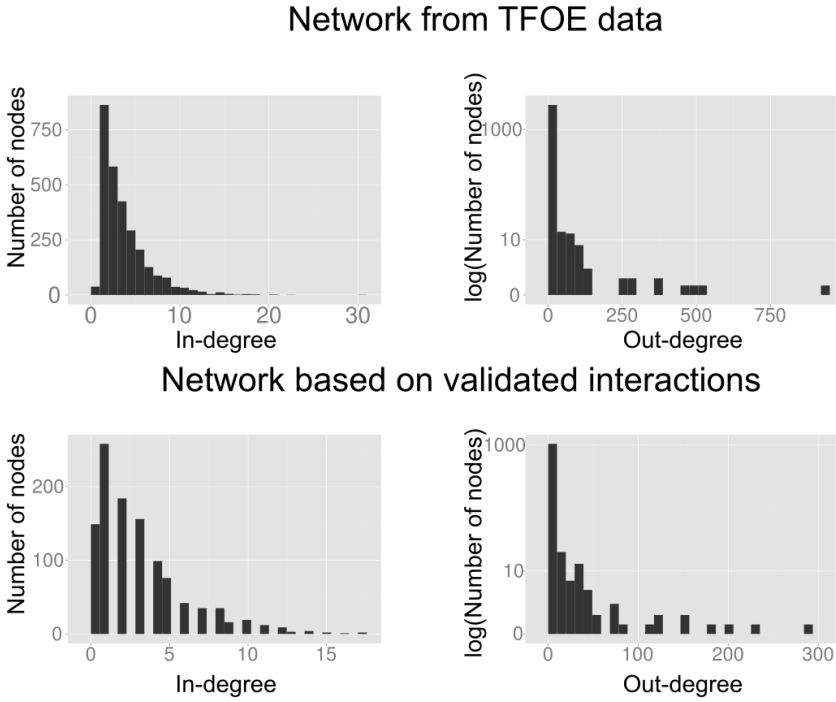


Figure 5.4: Validated gene expression network. Comparison of the out- and in-degree distributions between the network derived from TFOE data and the network derived in this study.

Network	TFOE derived network	Network based on validated interactions
Number of nodes	2,843	1,102
Number of edges	9,142	3,396
Clustreing coefficient	0.065	0.004
Network diameter	12	3
Shortest paths	114,080	4,352
Characteristic path length	3,996	1,239
Mean closeness	1.35E-07	8.27E-07
Mean radiality	3.95	6.54
Mean betweenness	120.2	0.945
Mean eccentricity	5	6.15

Table 5.1: Comparison of mean centralities and descriptive statistics between the TFOE derived network and the new proposed one.

the largest number of genes (672 and 627, respectively). Thus, these genes are regulatory hubs of *M. tuberculosis*. On the other hand, Rv3202c is the gene with the largest number of TFs influencing its expression, as it is indirectly regulated by 26 TFs. This gene has ATPase and helicase activities [133]. The regulatory subnetwork of Rv3202c is related to regulatory DNA and RNA processes as well as to response to external stimuli, transport and secretion. The new network derived can be found in the External Data 8 file.

Transcription factors are not universally conserved in the MTBC

Once gene expression models and a regulatory network for H37Rv were available, we tried to predict the phenotypic effect of natural genetic variation observed in circulating clinical strains. For this, we first examined the degree of conservation of the studied TFs across the MTBC. Previous studies have identified mutations in the genes that code for the PhoPR system in MTBC strains that had important effects on the pathogen's virulence [135], so that not only SNPs in the regulatory regions of the TF but also those located in the coding region could lead to differences in TF activity. Thus, we focused our analyses on mutations falling in regulatory regions but also on those coding mutations that might impair the normal function of the TF.

Using the 219 strains representatives of the global diversity previously used in Chapter 4 [36], we identified a total of 28 transcription factors (TFs), among those present in the TFOE data [173], that are missing or likely dysfunctional (as defined in the Materials and methods section) in one or more clinical strains, including 4 affecting complete lineages of the MTBC (Figure 5.5). Some of these transcription factors are missing in complete lineages and sublineages as they are in known RDs used as phylogenetic markers [30] (all the deletions detected shown in External Data 9A). For example, Rv1994c and Rv2478c are in RD743 and RD715 and they affect the entire L5 [190]. Those lineages represent up to 50% of the tuberculosis cases in West Africa [191]. We have also identified single point mutations disrupting the normal functioning

of some TFs (Figure 5.5 and External Data 9B). This is the case of *sirR* (Rv2788). An early stop codon mutation was found in all the strains of L1. In the proposed regulatory network, Rv2788 regulates 22 genes (Figure 5.6B) and, accordingly, 16 of those genes were expressed differentially in L1 strains with respect to H37Rv using RNAseq data [182]. In our estimates (Figure 5.6A), L1 accounts for roughly 18% of the strains causing active tuberculosis cases each year (almost 1.9 million cases/year). In light of the existing variation in TFs and other regulatory elements among clinical MTBC strains, it is very important to take the circulating diversity when building comprehensive regulatory networks, as these may differ among strains with different variants.

Next, in order to identify the main biological processes involved, we analyzed the relative abundance of Gene Ontology (GO) terms in the regulatory subnetworks for each affected TF. Most of the TFs identified as missing in clinical strains have an important role, with a direct or indirect regulatory influence in up to 210 genes. The GO analysis showed that a wide range of processes are significantly overrepresented among affected TFs, including specific metabolic, regulation, pathogenicity and response to external stimuli pathways (External Data 9). Some deletions affecting TFs appear in single strains, such as one affecting Rv1994c in a strain of L2 or Rv1776c in a strain of L3. A deletion of gene Rv1985c, a known antigen, was also found in a group of strains belonging to L1. It is also remarkable that a stop-codon gain mutation was found in Rv0465c (also known as ramB) in one strain of L4. RamB is related to the glyoxylate cycle in the pathogen and it has been proposed to play an important role in the adaptive response of the bacteria to different host environments during infection [192]. Moreover, the regulatory subnetwork of ramB is involved in several processes such as regulation of RNA biosynthesis, response to hypoxia or interaction with the host.

We also identified 117 SNPs located in the regulatory regions of 44 TFs (Figure 5.5, External Data 9C). Most of these SNPs affect primary or alternative transcription start sites (TSS), as defined previously [127]; two of them correspond to antisense TSS and two more were internal TSS.

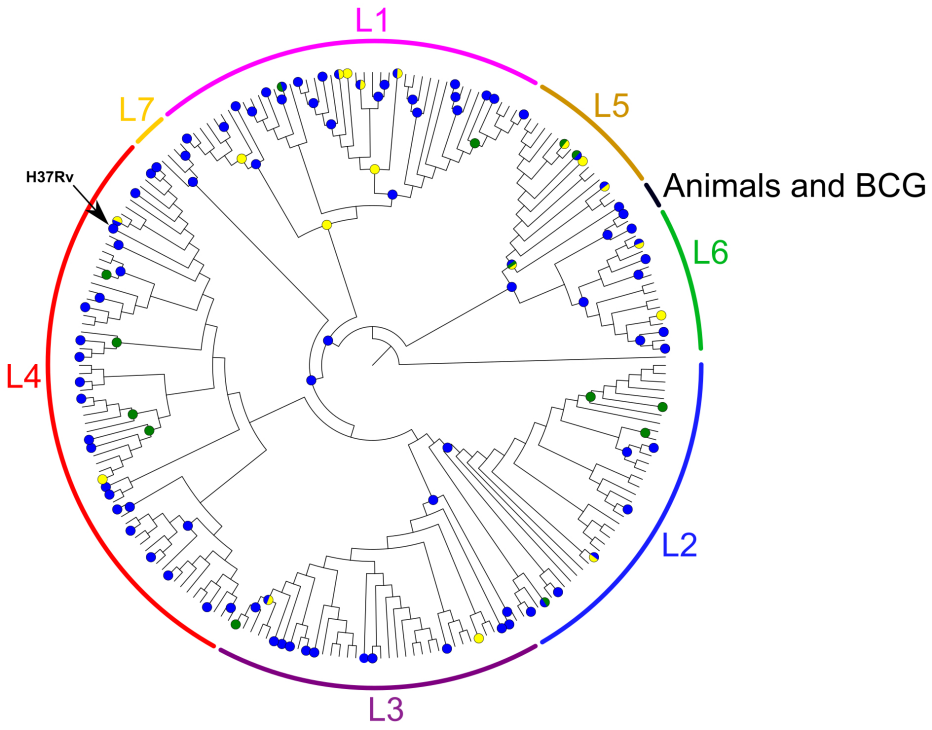


Figure 5.5: MTBC phylogeny comprising the seven major lineages. The figure represents the number of TFs missing or potentially affected in their regulatory functions in one or more clinical strains from an MTBC reference dataset (n=219 strains). Mutations affecting a TF are mapped to the corresponding internal/external node of the phylogeny and highlighted in colour. Green colour indicates total or partial deletions of a TF, yellow indicates stop-codon gain or loss, and blue indicates a SNP in the regulatory region of a TF.

Seventy-four of these SNPs affect one single strain, with the remaining 43 affecting more than one strain. Interestingly, only a few of them affect complete lineages, such as T89200G, which impacts the master regulator Rv0081 in modern L2, 3, 4 and 7 (76% of the circulating strains), or C422745T, which impacts Rv0353 in all lineages except 5 and 6. Rv0081 regulates 188 genes (including *tcrR*, which also regulates 26 genes) (Figure 5.6B). Hence, a SNP potentially affecting Rv0081 regulation could have an important effect on the regulatory network of the bacteria [135]. Besides, we found one homoplastic

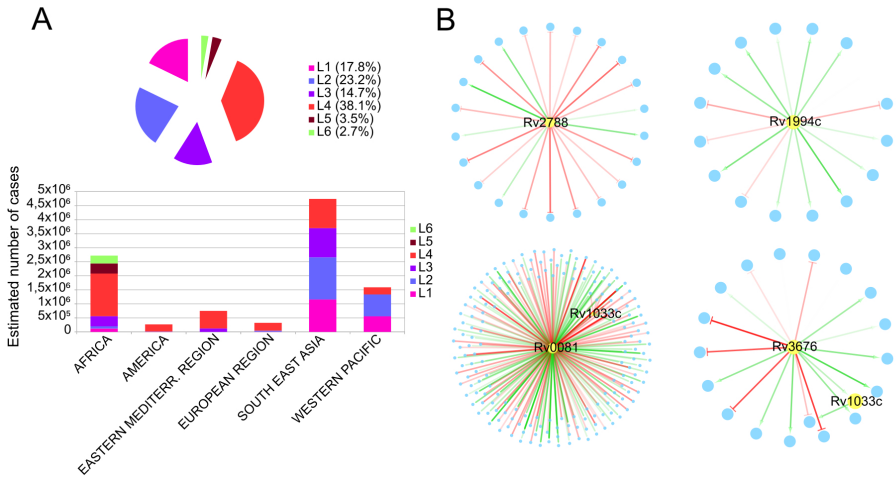


Figure 5.6: Global incidence of the different lineages and representative examples of mutations affecting complete lineages. A) Pie chart showing the estimated number of annual tuberculosis cases attributed to each lineage and a barplot showing the incidence of the different lineages by region. L7 is not shown due to its low incidence in global terms. The data related to the disease incidence by region come from the WHO [1] and the lineage abundance for each region from a previous work [30]. B) Examples of regulatory subnetworks of transcription factors affected by mutations in one or more lineages. From upper-left to lower-right: regulatory subnetwork of Rv2788 (early stop-codon in all L1 strains); regulatory subnetwork of Rv1994c (deleted in all L5 strains); regulatory sub-network of Rv0081 (SNP in regulatory region found in all the strains screened from L2,3,4 and 7) and regulatory sub-network of Rv3676 (SNP in regulatory region in all the strains from L3). Only TF (yellow nodes) were labeled. Green edges indicate positive regulations whereas red edges indicate negative regulation. The intensity of the edges is related to the influence of the TFon the gene (the darker the edge, the higher the regulatory effect).

SNP (C2965900T, which affects Rv2642) that has emerged independently in strains of three different lineages. It has been shown previously that some of the SNPs screened affect the expression of their corresponding TF. For example, SNP G3500149A has been reported to be involved in the regulation of TF Rv3133c in Beijing strains (L2), as it creates a TANNNT box leading to the overexpression of the DosR regulon [182, 193]. The External Data 10 includes a detailed view of the phylogeny with all the variants marked on it.

In-silico expression prediction of genetic backgrounds observed in a clinical and in a vaccine strain

To explore how well the H37Rv-based expression models and the validated network predicted the impact of the genetic background in the transcriptional landscape of the bacteria, we selected a L1 strain (T83) from the comparative genomics analysis. For T83 there is publicly available expression dataset [182] and we have identified a deletion in TF Rv1985c and an early stop-codon in TF Rv2788. Rv1985c or Rv2788 are present in 169 gene models. By reducing the expression of Rv1985c and Rv2788 to its minimum level, we created gene models mimicking the T83 genetic background. With these modifications, we were able to predict that 148 genes will have a significant change in their absolute expression value (pFDR $<$ 0.05). To formally compare with experimental data, we used the RNA-seq data sets from H37Rv and T83. Of those 148 genes only 71 changed in the same direction as determined in RNA-seq dataset irrespective of the absolute expression value. Moreover, out of the 148 genes only 64 showed differential gene expression in RNA-seq experiments with the same strain and with no correlation between predicted and observed values (Pearson correlation coefficient = 0.08, p-value = 0.48) (Figure 5.7A). Although conclusions from a single strain are necessarily provisional, it is also true that mutations in T83 are present in several strains of L1. Thus, from the limited data available we speculate that gene expression models based on H37Rv and derived from TFOE are not likely to predict accurately enough the transcriptional landscape of the MTBC lineages.

We reasoned that, given that it is not possible to predict expression changes in different genetic backgrounds, gene expression models might still be valid for experiments using the H37Rv background. As an example, we selected PhoP for several reasons: (i) it is one of the main regulators in the MTBC [185, 181]; (ii) it is the main gene deleted in a vaccine candidate that is already in clinical trial phase 2A [186]; (iii) there are large datasets available on the expression changes in knock-out strains using two different approaches, microarray [134] and RNAseq [194]; and (iv) there is strong evidence that mutations in the

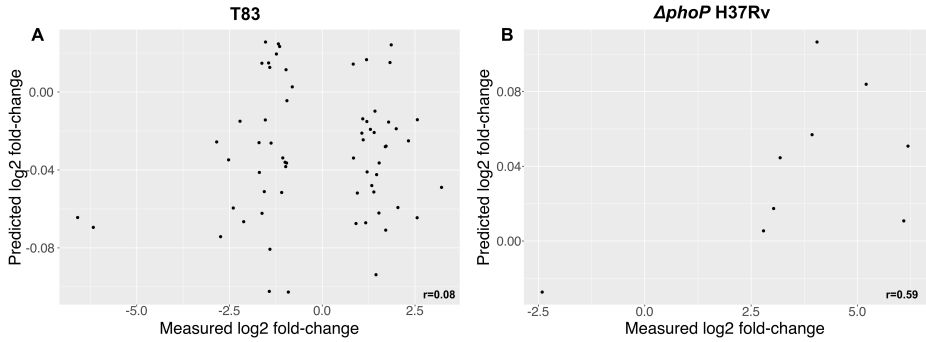


Figure 5.7: Comparison between experimental and predicted fold-changes. A) The x-axis corresponds to the measured log2 fold-change in gene expression between H37Rv and the T83 strain in Rose *et al.* (2013)[182]. The y-axis corresponds to the predicted fold-changes calculated with the predictive models obtained in this work. B) The x-axis corresponds to the measured log2 fold-change in gene expression between the wild-type strain and the $\triangle$*phoP* strain in Solans *et al.* (2014) [194]. The y-axis corresponds to the predicted fold-changes calculated with the predictive models.

PhoPR regulatory regions impact fitness of clinical strains in the human host (see Chapter 4 and [135]).

From the TFOE, we identified 218 models in which *phoP* (Rv0757) is present as a regressor. We lowered the expression value of *phoP* in the models to the minimum, thus simulating that the gene is knocked-out. Comparing the simulated knock-out with the wild-type models, we detected 188 genes with a statistically significant difference in expression (pFDR $<$ 0.05). Very little overlap was found between the 188 genes predicted and those found experimentally to be impacted by a knock-out mutant. Of the 188 predicted genes, Gonzalo-Asensio et al. (2008) described only 10 in microarray experiments. We also contrasted our predictions with an RNA-seq data set of a *phoP* knockout H37Rv strain [194]. We first compared whether the predicted expression for the 188 genes followed the same direction as the ones from the RNA-seq data set. In 96 cases the predictions agreed with the experimental values but in 92 cases the predictions failed. Cohen's kappa test showed a slight agreement between the real and the predicted values (kappa = 0.05). Next, we compared the 188 predictive models showing differential expression

with the genes showing differential expression in the data set (adjusted p-value < 0.05) and we found only 9 coincidences. For these 9 genes, we observed no statistically significant correlation (Pearson correlation coefficient = 0.59, p-value = 0.09) (Figure 5.7B) between the predicted and measured gene expression fold-change in the mutant.

We tested whether the lack of correspondence between our predictions and the experimental data might be due to the former being obtained from TFOE whereas the later were defined after analyzing a knock-out strain. Figure 5.8 shows a graphical comparison between the ChIP-Seq coverage of the over-expressed, the knock-out mutant, and the wild-type strains. Using the wild-type coverage vs the *phoP* mutant coverage as a negative control, we were able to infer the binding sites of PhoP in the H37Rv strain [194]. By comparing these results with the binding sites inferred from the TFOE strains, we observed differences in several genes (the 9 genes showing differences in ChIP-Seq coverage are highlighted in Figure 5.8A). Details for two of these genes are shown in Figure 5.8B and more examples are included in Supplementary Figure 10.7. For example, for the Rv0789c gene there is no evidence of PhoP regulation when the mutant and the wild-type are compared. However, a peak appears in the overexpressed strain and strong regulatory evidence has been reported [173]. In total, from 139 genes predicted to be regulated by PhoP from the TFOE data and 51 from the mutant data, only 16 genes overlapped. Thus, different methodologies to test the function of a gene (overexpression versus deletion) partially account for the limited predictive power of the H37Rv gene expression models even when the mutant is derived from a H37Rv background.

Shaping of genetic diversity by protein-protein interaction networks

Hence, we noticed that the bacterial regulatory network was not conserved and that this fact has implications on predictions derived from H37Rv experiments. Next, we wondered about the conservation degree of other biological networks

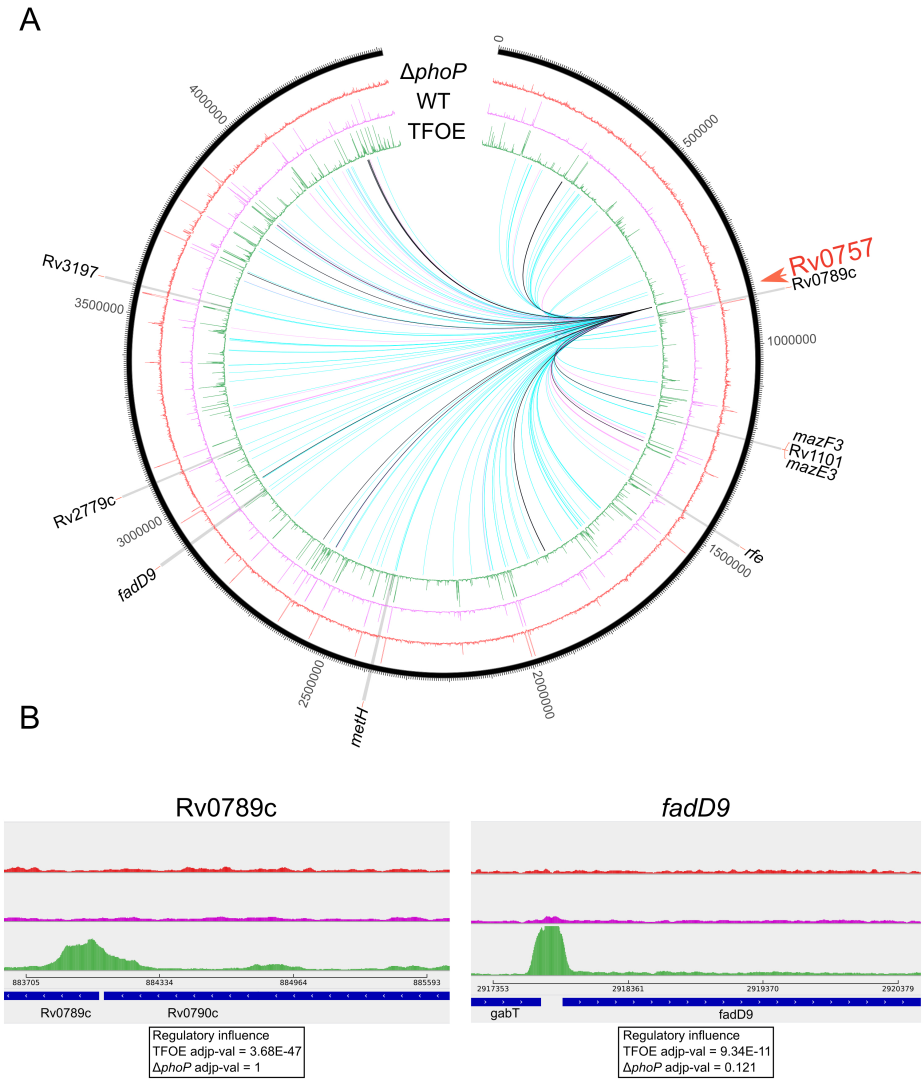


Figure 5.8: ChIP-Seq coverage comparison and regulatory influences between the *phoP* knock-out, wild-type and *phoP* overexpressed strains. A) Circle representation of the H37Rv reference genome. From outside to inside: ChIP-Seq coverage of *phoP* knock-out mutant, wild-type and TFOE. The inner links represent the regulatory influence of *phoP* derived from TFOE (blue), mutant strain (purple) and their overlap (black). B) Detail of two selected genes with regulatory influences derived from TFOE that do not match the evidence from WT and the mutant strain.

derived from the same clinical reference strain. Beyond regulation, cellular functionalities rely on proteins, which are the ultimate product of the cell. Proteins interact to perform a whole catalog of biological functions. Until now, several *M. tuberculosis* PPI networks have been derived [175, 176, 177, 178]. These networks try to model how proteins interact to achieve full functionality but, as in the case of the regulatory network, they were all derived by using the H37Rv strain. So, the MTBC genetic diversity was not taken into account when constructing these network models.

We were not able to select a representative PPI network to analyze from all the published ones, because all of them share a minimum overlap. Thus, we decided to download the protein-protein interactions contained in the STRING database [178] because they are curated, updated and derived from experimental data, scientific literature, computational algorithms and other databases. Moreover, STRING provides a confidence score (range 0-1) for each protein-protein interaction. So, we have downloaded all the interactions in the database having a confidence score > 0.7 and created a PPI network to be used in posterior analyses. This network had 3,272 nodes and 44,784 edges connecting them.

Essential proteins tend to occupy central positions

Essential proteins of *M. tuberculosis* have been characterized previously in *in-vivo* [195] and *in-vitro* [196] assays. Almost all of these proteins are present in the PPI network. As these proteins are required for the pathogen survival, they could have an important role in the biological network function and structure. In graph theory, the relative importance of a node is determined by its centrality values. These attributes summarize structural characteristics of the nodes and are related to their position in the network. The more central the node the more important it is in terms of network stability and communication between nodes. It has been stated that, in biological networks, nodes with higher centrality values have more relevance for the network's biological function [189].

We have observed statistical differences (Welch t-test, p-value < 0.01) in the distribution of centrality values between essential and non-essential nodes (Figure 5.9A). To explore the level of association between the centrality of a node and their essentiality, we have constructed a logistic regression model (see the Materials and methods section). This model determines the probability of a protein of being essential according to a combination of centrality measures. We applied the model to each protein in the network. When splitting the predictions by the protein's essentiality classification, it seems clear that the probability of being essential is higher for the essential proteins than for the non-essential ones (Figure 5.9B).

This result means that essential proteins have a characteristic set of centrality measures that differ from the non-essential proteins. If we check the coefficients of the regression model, that is, the weight of each centrality value over the model predictions (Table 5.2), we see that the high probability of being essential would be determined by high values of degree, closeness and eigenvector centralities while low values of eccentricity and radiality [189, 108]. These values are those that define central proteins. So that, it seems that essential proteins tend to occupy central positions in the PPI network.

Centrality	Degree	Closeness	Eigenvector	Eccentricity	Radiality
Value	1.6E-03	5.50E+04	1.32E+00	-2.19E-01	-4.15E-01

Table 5.2: Coefficients of the computational predictive model derived from the centrality values of the nodes.

Central proteins of the interactome accumulate less mutations impairing gene function

To test how mutations present in the *M. tuberculosis* natural diversity affect the PPI network we screened the MTBC genomic dataset described in Chapter 4. From the 4,595 strains we obtained a total of 235,254 SNPs. We kept the SNPs that affect coding genes (200,033, 85.02% of the total SNPs) as we wanted to map them onto the PPI network. We used the SIFT4G score calculated for

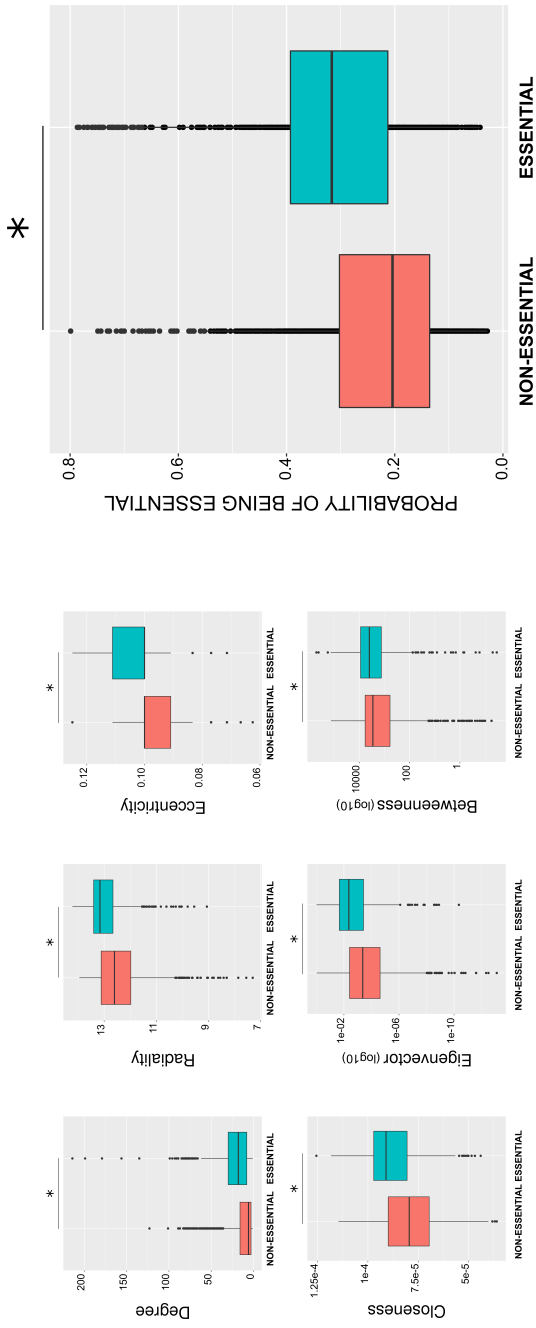


Figure 5.9: Centrality values splitted between essential and non-essential genes. A) There are statistical differences between the distribution of the centrality values for essential and non-essential genes. B) A computational model predicts that the probability of a gene of being essential is higher in the essential genes, due to its centrality values.

each mutation (see General Methods for details) to classify the SNPs as those having high-impact in gene function (HI, SIFT score $\leq$ 0.15) and those having low-impact (LI, SIFT score $>$ 0.15). For the total amount of SNPs in coding regions we obtained 75,545 as HI and 124,488 as LI. With this information, we set a protein as 'Impacted' if the number of HI variants found across all the samples was higher than the number of LI variants. We did not take into account proteins belonging to PE/PPE family, phages and those having repetitive regions as they are prone to false positive SNPs. So that, we had 280 proteins of the network that are impacted versus 2,992 that are not. To analyze the effect of the mutations in terms of PPI structure and function, we first created a clusterized version of the interactome. We defined communities (clusters of proteins) which are densely connected subgraphs. All the nodes belonging to a community are more connected between them than with nodes from other communities. These communities include proteins that are related in similar biological processes [197]. Later, we calculated an impact value for each community of proteins. The community impact value was derived from the number of impacted proteins in the community (those having a higher number of SNPs potentially affecting gene function). We observed a significant correlation between some centrality measures and the impact value of a community (Figure 5.10A)

So, communities with a higher impact value (those having a high percentage of impacted proteins) tended be periferic while communities placed in the core of the network accumulated less impacted proteins. Finally, instead of putting all the variants together, we repeated the same analysis for each lineage. We noticed that the community impact patterns of the different lineages clustered almost according to its phylogenetic relationships (Figure 5.10B). So, close lineages (in phylogenetic terms) had similar communities affected, probably because of sharing common phylogenetic variants. It is remarkable however, that all the patterns were different, and we can observe that the protein communities are not equally impacted in all the lineages.

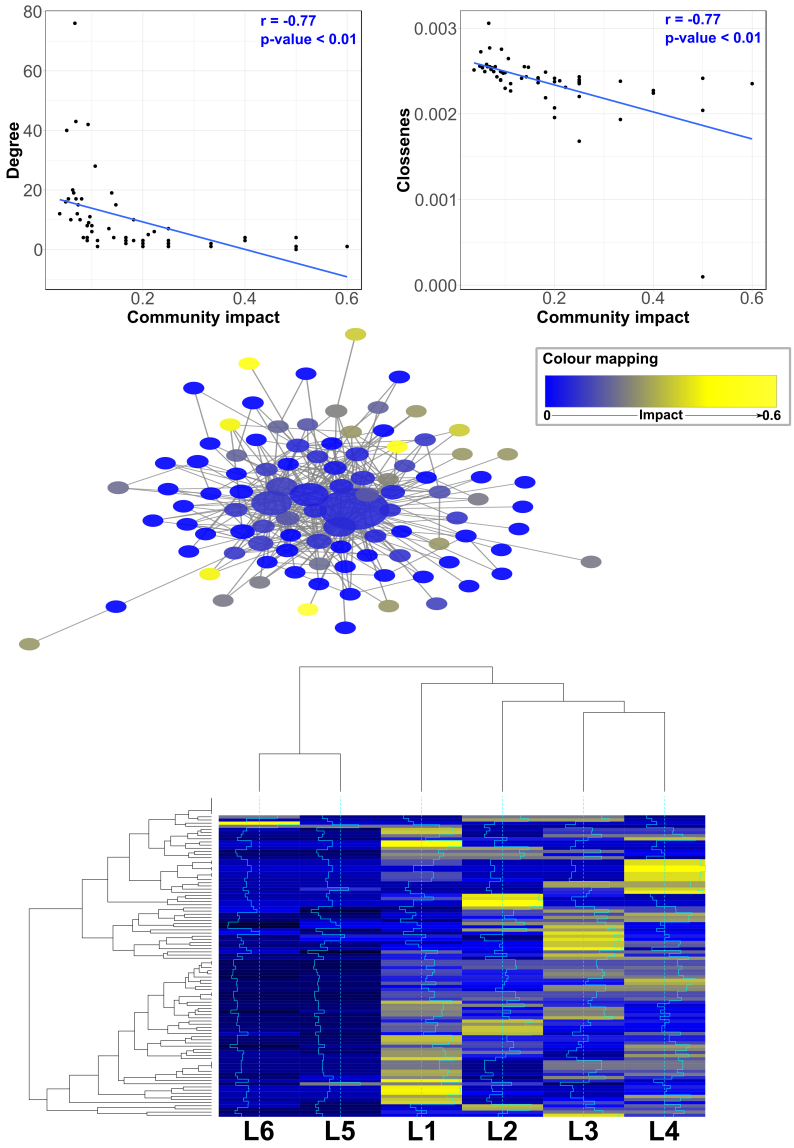


Figure 5.10: Impact of the mutations that potentially affect gene function on the PPI network. A) The most impacted protein communities of the network are those that are more periferic, while the central ones accumulate less HI mutations. B) The distinct lineages have different protein communities impacted. The clustering derived from the communities impact value is highly congruent with the phylogenetic topology of the MTBC.

5.3 Discussion

We have shown that the genetic diversity naturally present in the MTBC has impact on the biological networks. Until now, *in-silico* modeling of these networks has not taken into account this diversity, and only the clinical reference strain H37Rv has been used to derive these approximations. Overall, our results suggest that using a reference strain for generating complex network models only grabs a minor part of the true phenotypic variation, thus leading to inaccurate predictions when using these models.

Regarding the regulation of gene expression, we have tested the predictive power of state-of-the-art *M. tuberculosis* regulatory networks and expression models when the system is disturbed by (i) experimental noise, (ii) mutations associated to a clinical strain with a different genetic background to that of the training data set, and (iii) a knock-out mutation in the key regulator PhoP in the reference strain used for the training data set. For the genetic background and single mutations predictions our results show very little overlap between the genes predicted to be significantly impacted and those determined experimentally.

One striking result is that gene expression models are not statistically different from random generated models in 66.87% of the cases. This result suggests that subtle impacts of TFs in expression maybe missing even for a data set that comprises the construction of 206 TFOEs strains and 698 microarray for the analysis. In this analysis, only the TFs having a strong signal were taken into account. Background noise introduced by the experimental system prevented us from incorporating TFs with more subtle effects. Thus, we cannot discard the possibility that the sum of weak effects may account for some of the expression differences. For the remaining models there is a moderate correlation, with predicted absolute quantitative expression ($r = 0.71$) and a high correlation ($r = 0.98$) to predicted fold-changes using *dnaA* to normalize. When applied to an analogous data set, we found a good agreement with fold-change data ($r = 0.93$) but almost none when we tried to

predict absolute expression values (128 out of 1216). Thus, although the models grab the intensity of the interaction they are not able of quantifying the changes.

The limitations of the models, even when applied to the same data used to train them, can be explained by two different, non-mutually excluding alternatives.

Firstly, our results show that the statistical validation of gene expression models is essential to remove methodology-dependent effects that may or may not correspond to actual biological differences and contrasted biological effects. Only 1,216 gene models derived from the TFOE dataset and 29 from the ChIP-Seq were significantly different from random-generated gene expression models. In addition, predictions with an alternative data set, generated in a different laboratory but following the same protocols, show that absolute quantification of gene expression is not possible with the current models. This suggests that, in order to understand the impact of different perturbations in the system such as genetic mutations, the noise introduced by the experimental setting must be taken into account, especially in genes with a low expression level [198].

Secondly, the regulatory network inferred is highly dependent on the experimental methodology. Overexpression of transcription factors is a common, widely used technique to identify regulatory influences but it can fail in making accurate predictions when an increase in gene expression has no physiological effect or, on the contrary, it can overestimate the regulatory effect due to a loss of specificity [199]. A recent work with *Mycoplasma pneumoniae* demonstrated that the overexpression of regulatory molecules (asRNAs in this case) leads to an overestimation of the regulatory effect of these molecules [200]. In addition, the ChIP-Seq technique might introduce false positives when overexpressing the TF [201]. Besides, other studies [202] have shown that there are many sRNAs in bacteria that regulate gene expression. Their effect is not reflected in this type of networks and analyses, as we only look for the amount of mRNA produced by a gene. Finally, the amount of mRNA does not

always agree with the amount of translated protein [203, 204]. However, new experimental techniques, such as CRISPRi, are being tested in *M. tuberculosis* and other organisms to characterize and modify gene expression [205].

Our results show that the different TFs tested in H37Rv are not universally conserved. Some of those mutations (deletions and single point mutations) are present in complete lineages and in up to 76% of the circulating strains. Using comparative genomic data we have predicted the transcriptional landscape of a L1 strain. We found 64 out of 148 matches between the genes predicted to be impacted and those found in an RNA-seq experiment [182]. Strain T83 belongs to L1 and its genetic distance to H37Rv, the strain used to build the models, is more than 1,800 SNPs [36]. Thus, other genetic differences besides those found in TFs between this strain and the one used to infer the regulatory influences will certainly impact the genome-wide transcriptional landscape of T83. For example, we have mapped two SNPs in the regulatory region of TF Rv0353 in T83. Current models do not take into account the potential influence of these SNPs nor that of other regulatory layers that possibly differ between lineages. In addition, we have previously shown that specific SNPs of L1 alter the expression levels of sense and antisense transcripts by means of new TSSs [182]. Our predictions on a PhoP knock-out H37Rv strain were also poor. The regulatory influence was predicted correctly only for 96 out of 188 genes.

Modulation of the regulatory processes does not always agree with changes in translation rates, as post-transcriptional mechanisms are known to have an effect over the translation of the final protein product [206, 207]. So that, it could be the case that the impact of the genetic diversity over the regulatory network could be balanced by post-transcriptional mechanisms in the different MTBC genetic backgrounds. This fact perhaps could lead to convergent translation patterns despite the heterogeneity found in the regulatory network due to genetic variants. However, nonsynonymous mutations that potentially affect protein function could have an impact in the way that proteins interact (in a post-traductional stage). In this way, we have seen that the interactome structure is highly related with the biological function of the proteins. The most

important proteins, those essential for bacterial survival, are located in the central region of the interactome, being key hubs for communication and involved in multiple processes. The proteins tend to form clusters with its related neighbours and these structures prevent the accumulation of mutations that affect the gene function in the most important regions of the interactome. This type of structure confers a high robustness to random mutations that can alter the network structure and function [189]. However, it is susceptible to be disrupted with directed "attacks" over specific key nodes. We could use the central proteins of the network structure as potential biomedical targets as a direct "attack" on them could affect multiple essential functions of the pathogen physiology.

Additionally, we have checked that different lineages have different nodes impacted. Again, the PPI network used was derived from H37Rv. Similarly to what happens with the regulatory network, some of the network edges (interactions between proteins) could be different as there are genetic differences naturally present in MTBC strains. Will the network map change in the different lineages? Could the proteins affected by HI mutations be bypassed thus maintaining the network functionality? If this is the case, we should need a different network representation for each lineage, to look for central proteins/communities that could be potential biomedical targets.

5.4 Materials and methods

Datasets and techniques used

The main microarray expression data sets were obtained from Rustad *et al.* [173] (GEO accession number GSE59086). The ChIP-Seq data were obtained from Minsch *et al.* [174]. The TFOE-derived network used to compare with the TFOE network generated in this work was obtained from Rustad *et al.* [173]. The *phoP* mutant data were obtained from Solans *et al.* [194] (GEO accession number GSE54241). The RNA-seq data from L1 were obtained from Rose *et*

al.[182] (EBI ENA accession number ERP002122). The H37Rv RNA-seq data were obtained from Arnvig *et al.*[202].

Gene expression models construction

The regulatory relationships used for model construction were obtained from Rustad *et al.*[173] data set. We selected all the regulatory interactions with adjusted p-value$\leq$0.01 (Benjamini-Hochberg) regardless the fold-change in the expression values. In consequence, all the statistically significant regulatory influences (even the weak ones) were taken into account. From Minsch *et al.* (2015), we selected the physical bindings that demonstrated a regulatory effect on the level of gene expression of the target. To select the TFs to be incorporated into gene-expression models we compared two different strategies. Firstly, for each gene expression model we selected those TFs that had a large influence on the expression response (Moderated t-test, adjusted p-value$<$0.01 and fold-change$>$2). Alternatively, we selected as relevant TFs all the genes showing regulatory influence (Moderated t-test, adjusted p-value$<$0.01) without taking into account the intensity of the effect. The latter models were trimmed a posteriori following a backward stepwise methodology. In both cases, we evaluated whether the predicted values departed from random estimates by generating at least 20 random models incorporating some of the 200 TFs. To identify which of the two strategies predicted better the observed expression values in the Rustad *et al.* data set and to evaluate the strength of the prediction, we computed the correlation between predicted and real values of gene expression. We obtained an average correlation of 0.32 for the a priori approach and 0.71 for the a posteriori approach. As the differences in correlation estimates between both strategies were statistically significant (t-test, p-value$<$0.01), we selected the latter method to construct the final models. The following process was performed for each target gene in the TFOE- and the ChIP-Seq-derived models:

1. All the TFs affecting the gene were selected as regressors for the model. A TF is said to affect a gene if there is statistical evidence that the expression

of the gene (moderated t-test adjusted p-valueleq0.01) changes when the TF is over-expressed following the procedure described in Rustad *et al.* [173]. In addition, the RNA polymerase alpha chain gene, *rpoA* (Rv3457c), and the sigma factor gene, *sigA* (Rv2703), were included as normalization factors. Interactions between TFs were also taken into account. The model structure [181] was based on equation 5.1.

$$y = a + \sum_{i=1}^{T} b_i x_i + \sum_{i=1}^{T} \sum_{j=i+1}^{T} c_{ij} x_i x_j + d x_{sigA} + e x_{rpoA} + \varepsilon \tag{5.1}$$

where y is the target gene expression, x_i are the expression values of the selected TFs (from $i = 1$ to T), a, b, d and e are the linear coefficients in the regression model, c are the interaction coefficients, and ε is the error term.

2. A linear regression model with all the TFs selected as regressors and based on the previous structure was constructed. Next, the model was parameterized using microarray data from Rustad *et al.* [173]. This data set consists of 698 tiling microarrays, so that the model was fitted using 698 data points corresponding to expression values of the strains present in the data set.

3. The Bayesian Information Criterion (BIC) and the Akaike Information Criterion (AIC) associated to the model were calculated. To limit the overfitting error we used the BIC in the TFOE-derived models because it penalizes models with a large number of regressors [208]. In turn, the AIC was used when calculating models from the ChIP-Seq data set given the low number of regressors involved.

4. We sequentially eliminated from the model regressors whose removal led to the largest decrease in the BIC/AIC. So, we deleted from the model those terms whose removal had a minor or null contribution on the model's performance. In biological terms, we filtered out the TFs that led to a minor or weak regulatory response in the target gene, in comparison with the rest of the TFs. The remaining TFs were retained and we returned to step 2. In case we did not observe a decrease in the BIC/AIC after the removal of any regressor,

we considered that model as optimal for the corresponding gene.

5. A Fisher's F-test was performed to check the null hypothesis that the retained regressors do not have predictive power [208]. P-values were adjusted to multiple testing by Benjamini and Hochberg's false discovery rate (FDR) [100] and all models with adjusted p-value$\geq$0.05 were rejected.

Cross-validation of the models

We checked the initial models obtained above in a 10-fold cross-validation. For each gene:

1. The optimal model selected was parameterized using a random subset of the 90% TFOE data set as a training-set.

2. Next, the remaining 10% of the data set was used as a test-set to make predictions. A Fisher's F-test was performed to check differences between residuals of the training set and the test set. Also, the Root Mean Squared Error (RMSE) [209] was obtained when predicting over the test set.

3. Steps 1 and 2 were repeated 10 times (10-fold cross-validation) We retained those models that showed no difference between predictions over the training set and the test set, by comparing the average adjusted p-value of the F-test over the 10 iterations ($\alpha\geq$0.05). In some cases, we could not find differences between residuals but the squared error was high. In consequence, we also rejected models with RMSE$>$Q3+1.5 * IQR, as they were considered outliers of the RMSE distribution [132].

Comparisons to random models

We considered each TF of the data sets as a potential regulator. For each gene, we listed all the TFs that do not have a regulatory influence on it. From this list, we created 100 random subsets of TFs. The number of elements in each subset was equal to the number of real factors with regulatory influence on the corresponding gene. With this random subset of TFs, we followed the

steps described above to create the random models. Also, the 10-fold cross-validation was performed for each random model. For each model, a Welch's t-test was performed to compare the distribution of p-values from the 10-fold cross-validation of the real model versus the random ones. P-values from Welch's tests were adjusted by Storey's method [164]. Tests showing a pFDR$\leq$0.01 were accepted as having a better fit than random models. Also, the RMSE distributions of random models were tested versus the real ones by means of a Welch's t-test, correcting the p-values with Storey's method. Tests showing a pFDR$\leq$0.05 were accepted.

Comparison to a different TFOEs dataset

To check the ability of the models to perform accurate predictions over different datasets, we used a TFOE dataset similar to the one used to built the models but derived only from 50 TFs construct strains [181]. As gene expression values can have very different ranges it is not advisable to compare the numeric values directly. Predictive models are uncertain when predicting in a range of values different to the ones used to train them [208]. So, to compare both data sets we used the expression value of dnaA to normalize both the predicted values from the models and the absolute expression values from the experiments by Galagan *et al.* [181] and Rustad *et al.*[173]). Pearson's correlation was calculated between predicted and observed fold-changes.

Evolutionary conservation of TFs within the MTBC

We have analyzed the 219 representative strains of the complex described in Chapter 4. A custom script was used to search for TFs with at least 25% of their gene length deleted. A manual inspection was performed for the detected TF deletions to filter false positives and mapping errors. Also, single nucleotide polymorphisms (SNPs) leading to stop-codon gains or losses, and point mutations affecting TF regulatory regions [127] were extracted from this dataset. The terminology used to classify the TSSs follows that of Cortes *et*

al.[127].

We defined the regulatory sub-network associated to each TF as that defined by the one-step distance nodes to the TFOE network. We studied for each sub-network the enrichment in certain functional categories as defined by the GO classification [210] by using the tools described in the Chapter 3.

RNA-seq analysis

Expression data from H37Rv and L1 strains were obtained from Rose *et al.*[182]. The differential expression analysis was performed using the DESeq2 package [107]. Differentially expressed genes were those with an adjusted p-value of 0.05. Differential expression analyses were performed over H37Rv versus all L1 strains to test the differences in those genes regulated by Rv2788 (*sirR*). In addition, we specifically analyzed differentially expressed genes between strains T83 (L1) and H37Rv (L4).

Predicting the impact of genetic polymorphisms in the regulatory network

To predict the impact of the genetic background on the transcriptional landscape of the *M. tuberculosis* complex we chose a L1 strain, namely T83. We have identified two genetic mutations in L1 which likely have a major impact on the functionality of TFs. One of the mutations corresponds to a deletion affecting TF Rv1985c whereas the other is in an early stop codon in TF Rv2788. To simulate a transcriptional landscape for L1, the expression values of both TFs were set to the minimum value found in the training dataset because standard regression models are only valid to make predictions in the same data range used to parameterize the model [208]. Gene expression values predicted for T83 were compared to those obtained from H37Rv expression models. We performed a Welch's t-test over the expression values. P-values were adjusted by Storey's method. A pFDR$<$0.05 was considered for accepting the difference between models as significant. For the genes that

showed differential expression, we calculated the log2 fold-change between H37Rv and T83. We compared these values with those obtained from the RNA-seq analysis. For a qualitative approach, we checked whether changes in expression values had the same sign (positive for induction and negative for repression). We constructed a 22 matrix with the predicted effect vs the measured effect and a Cohen's kappa test was performed over this matrix to check the agreement between predicted and real data. To test the quantitative accuracy of the models we selected those genes that showed differential expression in the RNA-seq data set (adjusted p-value$<$0.05) and in the predicted expression to calculate Pearson's correlation. Similarly, to make predictions on a *phoP* mutant in a H37Rv background we set its expression to the minimum value found in the training data set for this TF. The analysis was performed following the steps described above. To analyze how accurately the models reflect fold-changes in experimental data, we used an RNA-seq data from a *phoP* mutant and H37Rv [194] as explained in the previous section. We compared the log2-based fold-changes between the predictive models and experimental data comparisons. To compare the ChIP-Seq coverages in the different cases, we obtained raw data from the wild-type strain and the *phoP* mutant from Solans *et al.* [194]. We also downloaded ChIP-Seq data from the overexpression experiment of *phoP* (Rv0757_B167) from the MTB network portal [174]. The circular diagram was constructed with the Circos tool [211] and the values from the regulatory influence were extracted from the TFOE data set.

PPI network construction and analysis

Every interaction involving two proteins of the *M. tuberculosis* H37Rv strain, and having a confidence score $>$ 0.7, was retrieved from the STRING database [178]. Centrality values were calculated by using the igraph package from R [108]. A logistic regression model was constructed, using a generalized linear model with the distribution function specified as binomial. The independent variable of the model was the probability of being essential, and the predictors

were the centrality values calculated for each protein. Mutation affecting coding regions were extracted from the MTBC global dataset (n=4,595) defined in Chapter 4. The potential impact over the gene function of each mutation was inferred by using SIFT4G [92]. This tool works by predicting the potential impact of an amino-acid substitution over the gene function, by comparing the conservation of these gene in a group of closely related species. We established a threshold of 0.15 in the SIFT score to define a mutation as potentially affecting gene function. The SIFT manual establish a threshold of 0.05 to define a mutations as impacting gene function. However, and due to the low genetic variability found in the MTBC, we have tried to be more conservative and we have penalized more the fact of finding a nonsynonymous mutation. The clusters of proteins (communities) were defined by using the walktrap algorithm [212] which defines dense connected clusters of nodes by performing multiple random "walks" across the network edges. For each community an impact value was calculated. The community impact value was defined as the number of impacted proteins in the community corrected by the number of proteins in this specific community. A protein was set as impacted if the number of variants that potentially affect gene function (SIFT value $<$ 0.15) was higher than the number of variants that potentially do not affect gene function (SIFT value $>$ 0.15).

Part of the work described in the present chapter has been published as:

Chiner-Oms Á., González-Candelas F., Comas I. Gene expression models based on a reference laboratory strain are poor predictors of *Mycobacterium tuberculosis* complex transcriptional diversity. **Scientific Reports**:8(1),3813. DOI:10.1038/s41598-018-22237-5

The roles of mutation and methylation on transcriptional heterogeneity

6.1 Introduction

We have previously stated that, despite the low diversity found in the MTBC, there are biological differences between lineages which result in different phenotypic characteristics. For example, there are multiple examples of the association of MTBC lineages with specific populations [213, 214] and in some settings this association could be linked to differential transmission efficacy depending on the host population [30, 47]. Apart from transmission, the progression from latent infection to active disease differs among the different MTBC members [42]. Moreover, there are differences in growth rates for different MTBC strains in different *in-vitro* and *in-vivo* conditions [38]. Some of these phenotypic characteristics seem to depend on transcriptional differences, illustrated by the gene expression differences reported in MTBC strains grown *in-vitro* and *in-cellula* [184, 152].

In Chapter 5, we have shown that regulatory layers in the MTBC are multiple and that expression models based on H37Rv do not adjust the introducing perturbations. Also, we have shown that the MTBC regulatory networks vary across strains and lineages, with several transcription factors

carrying mutations that potentially impair regulatory function (External Data 9). For example, sequence variants that affect coding regions of a signaling cascade [147] or create new transcriptional start sites (TSS) [182] result in major gene expression changes, especially if they affect regulatory hubs. Some of these new TSS were previously reported to be favoured by a genome-wide mutational bias in the MTBC towards AT genetic changes [215, 183]. However we still miss the link between individual variants, underlying population processes and and phylogenetically-wide transcriptional diversity.

In addition, there is now substantial evidence in bacteria that DNA methylation can affect transcription, and that changes in methylation patterns can modify transcription patterns over long or short timescales [216]. The development of Single Molecule Real Time (SMRT) sequencing now allows direct detection of this DNA methylation [217]. Previous work has shown that DNA modifications induced by the MamA methyltransferase affect transcription of several genes in H37Rv [218]. Moreover, other regulatory mechanisms such as non-coding RNAs are known to be present among the MTBC members, although their roles have not yet been clearly defined [219]. All these factors, together with previously published transcriptional data [152, 184] suggest that the MTBC has a rich transcriptional diversity.

Considering that the genetic heterogeneity among lineages could lead to differences in gene expression regulatory mechanisms, it is of particular interest to characterize the gene expression signatures of each MTBC lineage. As phenotypic traits are affected by the expression of specific genes and the distinct MTBC members show heterogeneous phenotypes, those differences could be related to specific gene expression patterns for each MTBC clade. However, gene expression studies to date have been based on microarray technology [184] or have used RNA-seq but addressing only single strains, some phylogenetic groups or compared to distant mycobacteria [183, 182]. Therefore, there is a lack of transcriptomic studies based on recent technology (RNA-seq) and taking into account the whole MTBC diversity.

In this chapter, we studied the transcriptomic signatures of different MTBC

members, from RNA-seq data, and identify differentially expressed genes using a novel phylogeny-based approach. In addition, we have revisited the mutational biases observed in MTBC populations and quantified the direct impact of individual genetic changes on transcriptional patterns. We extend our analyses to the impact of individual variants in methylation patterns and assessed its role in the regulation of *in-vitro* gene expression. We demonstrate the hypothesis, previously suggested [183, 182], that the universal genome-wide mutational bias on MTBC leads to a higher phenotypic plasticity at the transcriptome level.

6.2 Results

RNA-seq data and analysis

We selected 19 strains from clinical samples which are representative of the MTBC global diversity (Supplementary Figure 10.8). Each lineage (L1-6) was represented by at least 3 strains. Two replicates per strain were grown in standard 7H9 medium with the addition of 30 mM pyruvate to account for strains with potential pyruvate kinase mutations (Supplementary Table 10.2). Cells were harvested for DNA and RNA extraction at an OD_{600} between 0.5 and 0.7.

From the RNA extracted, ribosomal RNA was depleted by using a Ribo-Zero Magnetic Kit. After that, sequencing libraries were prepared using the TrueSeq stranded Illumina protocol and sequenced on an Illumina HiSeq 2500 platform. RNA-seq analysis was performed using a custom analysis pipeline (see Materials and methods for details). From the DNA extracted, long-read sequencing was performed on the PacBio RSII platform. In addition to transcriptome and long-read sequencing, short-read sequences for the selected strains were obtained from a previous publication [220].

Global transcriptomic patterns

As a control, we first checked the agreement between sample replicates. We calculated the pairwise Pearson correlation between each pair of replicates. An almost perfect correlation (range 0.9996 - 0.9999) was achieved between each pair of replicates derived from the same strain. So, after evaluating the minimum level of variability between both replicates, the coverage data from the two biological replicates of each sample were merged for the subsequent analyses.

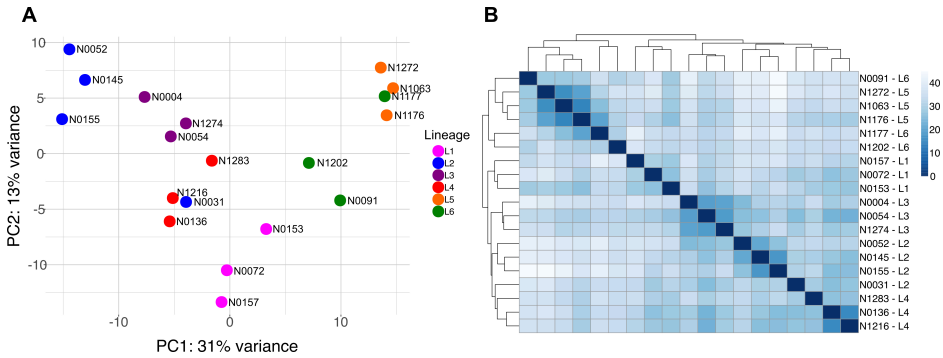


Figure 6.1: Global transcriptomic profiles of the samples A) The PCA plot shows that samples belonging to the same phylogenetic clade tend to group closely, except for two cases. B) A cluster analysis reinforces the trend derived from the PCA, with almost all the samples belonging to the same lineage clustering together.

Next, we surveyed the transcriptomic profile of the whole MTBC. A principal component analysis (PCA) was performed with the gene expression profiles of all the samples. Samples belonging to the same phylogenetic lineage group closely in the PCA (Figure 6.1A). *M. africanum* (L5 and L6) and *M. tuberculosis* (L1-4) samples split along the first component (31% of the variance), grouping according to their phylogenetic clade. In addition, strains belonging to L1 were found between the MAF group and the modern lineages (L2, L3 and L4). As a further step, we performed an unsupervised hierarchical clustering (Euclidean distance, clustering method complete). The results agreed with the observations derived from the PCA, with the samples clustering according to

their phylogenetic relationships (Figure 6.1B). However, there were two exceptions. N0031 belongs to L2 but its transcriptomic signature was different from other L2 strains. It was previously reported that N0031 belongs to a basal branch of L2 in which the *dosR* regulon is differentially expressed in comparison with L2 Beijing strains. Thus, its transcriptomic profile is expected to differ from others strains belonging to the same lineage [182]. On the other hand, N1177 belongs to L6 but it clustered with L5 samples. After the initial analysis, we realized that N1177 harbours a mutation in the *rpoB* gene (D435Y) that confers resistance to rifampicin. As this mutation affects the RNA-polymerase and genome-wide transcriptional, levels it is not surprising that it does not cluster together with the other L6 strains. Therefore, for subsequent analyses we removed N1177 as it may not be representative of the common L6 transcriptional profile.

As the RNA-seq profiles were congruent with the topology of the MTBC phylogeny, we investigated whether the number of differentially expressed genes between different clades was related to the genetic distance between them. We performed a Phylogenetically aware Differentially Expressed Genes (PDEG) analysis (see Materials and methods for details) to infer the number of differentially expressed genes on each of the main branches of the phylogeny (Supplementary Table 10.3 and External Data 11A). The results were highly variable, with a maximum of 42 PDEG genes in the branch leading to L6 and a minimum of 7 in the common branch of the modern lineages (Figure 6.2A). We observed a reasonable trend in the data with the number of PDEG genes varying according to the genetic distance between groups (Pearson's correlation value 0.57, p-value = 0.04). This suggests that the differences in the transcriptomic profiles between each group were accumulated gradually as the MTBC lineages diverged. However, there were two branches that break slightly away from this trend. The split between *M. tuberculosis* and the two *M. africanum* lineages was defined by a short genetic distance but by a high number of PDEG genes while in the branch leading to the modern lineages we found the opposite situation. A complete list of the PDEG genes detected in

each of the main branches can be found in External Data 11A.

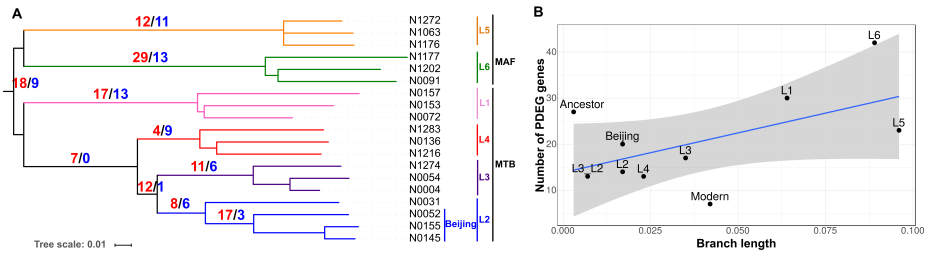


Figure 6.2: Gene expression changes across the MTBC phylogeny. A) Number of genes differentially expressed (red up, blue down) in each of the main branches of the MTBC phylogeny. The phylogeny was constructed using Illumina sequencing data, the Maximum-Likelihood algorithm and a bootstrapping of 1,000 replicates. B) Number of PDEG genes in each of the main MTBC branches plotted against the genetic distances.

Differential expression between phylogenetic clades

We performed an enrichment analysis of Gene Ontology functions for the up- and down-regulated sets of genes for each of the branches analyzed above. This analysis highlights the relative abundance of specific biological functions in a set of genes in comparison to the rest of the genome. Diverse biological functions appear as up- and down-regulated in each of the branches (Table 6.1 and External Data 11B), most of them related to host interactions and metabolic processes. Several studies have shown that strains belonging to different MTBC lineages show phenotypic differences when growing *in-vitro* and *in-vivo* [41, 42, 43]. In this context it is not surprising that biological functions related with host environment, nutrient uptake and metabolic processes are differentially modulated across the MTBC.

The deepest phylogenetic split split in the MTBC phylogeny is between MAF and MTB (Figure 6.2A). As stated above, the short genetic distance between both groups contrasts with a high number of differentially expressed genes. 18 genes were significantly upregulated and 9 were significantly downregulated between both groups (BH adjusted p-value < 0.05, fold-change > 1.5).

Branch	Up-regulated genes	Down-regulated genes
MTB-MAF	Response to iron starvation, siderophore metabolic process	
L6	Response to copper ion	Reactive nitrogen species metabolic process
L5	Reactive nitrogen species metabolic process	Response to host immune response and oxygen-containing compounds
L1		Growth of symbiont in host cell and response to acid chemical
L4	Oxalate metabolic process	Molybdopterin cofactor biosynthetic process
L3 & L2 common branch	Response to hypoxia and sulfolipid biosynthetic process	
L3		Response to heat
L2	Phosphorelay signal transduction system and regulation of fatty acid metabolic process	
Beijing	Response to hypoxia and signalling	Phospholipid catabolic process

Table 6.1: Gene Ontology enrichment analysis. Main GO functions enriched in the up- and down-regulated genes for each branch of the PDGE analysis. More detail can be found in the External Data 11.

Almost all of the mbt operon genes are upregulated in the MTB clade (*mbtI, mbtC, mbtH, mbtE, mbtG, mbtD, mbtB* and *mbtF*). These genes code for the siderophore (mycobactin) system that is necessary for iron acquisition in iron-limited environments (i.e. macrophages)[221]. Genes *ctpG* and *ctpC* were also overexpressed and are involved in metal cation transport [222]. However, the *mbtJ* gene was not upregulated but its antisense transcript was highly overexpressed in MTB suggesting a differential regulation between MAF and MTB. On the other hand, gene Rv0216, which is known to be essential for bacterial survival during infection, is overexpressed in MAF [223]. It is also interesting that ncRNA-mcr16, a non-coding RNA located in the *fadB* locus, appears as differentially down-regulated in MTB in comparison with MAF. FadB is involved in mycolic acid synthesis [224] and the PDEG of this ncRNA could be potentially involved in transcriptional regulation of FadB.

Although MAF lineages are geographically and genetically related, studies show that there are phenotypic and genetic differences between both clades [225, 226]. There is a high genetic distance (Figure 6.2A) between both lineages. Consequently, many PDEG genes appear in the branches that lead to current strains in both groups. The VapBC3 and VapBC5 toxin-antitoxin systems are upregulated in the L6 clade. Toxin-antitoxin systems have been proposed to play a role in response to stress. Specifically, VapBC3 and VapBC5 are up-regulated in moderately low pH conditions (i.e., the phagosome) [227]. As the pathogen is able to reside in the acidic phagosome during infection [228], the basal up-regulation of this system in L6 could be related to adaptation to the host environment during disease progression. We also found upregulated genes related to the copper ion response (*lpqS*, Rv0967, Rv2642 and Rv2963) in this lineage. Copper ions affect bacteria during the infectious process and the management of high levels of this substrate is required for full virulence in animal infections [229]. With respect to L5, the most upregulated gene is *acyP*. AcyP is an acylphosphatase involved in the pathway of pyruvate metabolism [230]. Also *nirB* and *nirD*, which are known to play a role during dormancy [231], were upregulated. On the other hand, an

important number of genes involved in parasitic functions such as virulence, persistence and macrophage infection are downregulated in L5 (External Data 11). For example, *icl, fadB2, tgs2* and *mmpL12* are significantly underexpressed with a fold-change range [$\sim$2 - 3.4].

L1 belongs to the so-called ancient lineages with *M. africanum*. However, it is genetically closer to the modern lineages than to the *M. africanum* strains. One of the most upregulated genes in the L1 clade is *virS*, which encodes a transcriptional regulator essential for the transcription of the virulence-related *mymA* operon under acidic conditions [232]. The *mymA* operon is known to be required for growth in macrophages and spleen. This upregulation of *virS* seems to have no effect on *mymA* regulation in this condition, as previously reported [182]. It is also interesting that *mpt63*, which encodes an epitope recognized by the immune system [233], had a strong antisense signal in L1 strains.

The modern lineages form a monophyletic clade which is $\sim$300 SNPs distant from the common ancestor of all the MTB strains. In this branch only 8 genes are upregulated. From these genes, 3 of them seem to form an operon (*nrdE, nrdI* and *nrdH*). NrdE is an essential protein involved in DNA replication and its transcriptional levels vary according to oxygen level [234]. Interestingly, these genes are significantly down-regulated in the L1 strains.

Regarding each of the single modern lineages, the L4 branch had only 4 genes upregulated. From them, Rv2159 and Rv2160A form part of an operon previously identified as being overexpressed in *M. tuberculosis* H37Rv strain compared to *M. bovis* strains due to the loss of a transcriptional repressor [235]. In contrast, genes involved in molybdopterin cofactor biosynthesis (*moaC* and *moaX*) [236] were downregulated in this branch as well as part of the *mce2* operon (*mce2C, mce2D, lprL* and *mce2F*). *mce2* mutant strains showed an attenuated phenotype in a mouse model of infection, with less pro-inflammatory cytokine recruitment and lower mortality rates in comparison with the wild-type [237].

In the branch leading to L3, 11 genes were upregulated while 6 were

repressed. Surprisingly, the most upregulated gene in L3 strains was *oxyR*. This gene is related to detoxification of ROS, contributing to the survival of the bacterium in the host, and also related to isoniazid resistance [238, 239]. It was reported previously that *oxyR* is inactivated in H37Rv, BCG, *M. africanum* and *M. microtti* [238]. Intriguingly, we have found that in L3, this gene had a 3-fold increase in expression compared to L2. The *ahpC* and *ahpD* loci upstream of *oxyR* were also overexpressed in L3 strains.

For L2, we have studied 4 representative strains. N0031 which belongs to a basal branch of this lineage and N0052, N0145 and N0155 which belong to the Beijing clade. As we noted in the global transcriptomic analysis, the N0031 transcriptomic profile was markedly different to those of the Beijing group. The DosR/DosS system was overexpressed in Beijing strains as previously reported [182] as well as the genes regulated by them. The DosR regulon is related to virulence and response to hypoxia [240]. In contrast, *plcD*, a gene related with extrathoracic progression of the disease and pathogenesis, was strongly repressed (fold-change = -7.7).

Non-random processes lead to higher transcriptional plasticity

Next we tried to link specific transcriptional changes to genetic variants and to underlying mutational biases. It has been reported previously that mutations can create new Pribnow boxes (TANNNT motifs) which are recognised by sigma factor A, SigA, and lead to the overexpression of nearby genes [182, 127, 183]. To test the influence of such mutations, we scanned all the single nucleotide variants across the 19 samples that either create or disrupt TANNNT motifs. We found 603 variants that created new Pribnow boxes in at least one strain and 81 that disrupted existing boxes (External Data 12). We investigated whether the observed impact on the Pribnow boxes resulted from stochastic mechanisms (i.e. genetic drift) or from non-random processes (i.e. selection). By comparing the number of expected versus observed occurrences (see Materials and methods), we have obtained a probability of 0.006 for the

observed number of disrupted boxes by chance and a probability of 2.5E-54 for the observed number of new boxes by random processes. So, it seems that non-random processes are acting to modulate the number of Pribnow boxes in the MTBC.

In addition, we randomly introduced all the mutations observed in the genome and repeated the process 1,000 times (Figure 6.3A). We obtained a probability of 0 for having at least the same number of observed new boxes (n=683) and a probability of 0.015 of having at least the same number of disrupted boxes (n = 81). Hence, it is unlikely that stochastic processes have been responsible for the observed appearance of Pribnow boxes across the MTBC. In addition when we looked at other sigma factors' -10 consensus sequences such as SigE (cGTT), SigG (CGANCA) and SigJ (CGTCCT) [241], no difference in the permutation was identified supporting the hypothesis that new SigA boxes are maintained by selection and not drift.

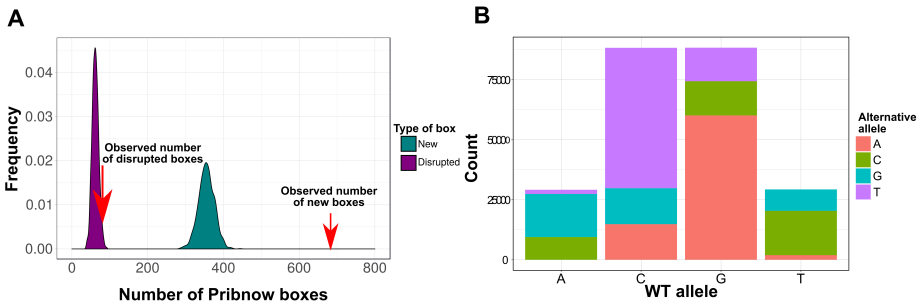


Figure 6.3: Non-random processes impact the emergence and disruption of Pribnow boxes. A) Distribution of new (green) and disrupted (purple) Pribnow boxes in 1,000 random simulations. Red arrows mark the observed value for each type of event in our dataset. B) Distribution of the alternative alleles from 235,254 mutations (obtained from 4,595 clinical samples of the MTBC). There is a clear bias favouring new T and A alleles.

We also noted that there was a remarkable difference between the number of new versus disrupted Pribnow boxes (ratio = 7.88). To get insights into the mechanism behind this figure, we randomly reordered all the mutations observed in our dataset by maintaining the alternative alleles but reshuffling the

genomic positions. After that, we searched for new/disrupted Pribnow boxes in these 'reordered' mutations. A Fisher-exact test showed that there was no difference between real and reordered mutations in terms of new/disrupted boxes ratio (p-value = 0.39). Thus, the higher ratio observed between both type of events is independent of the genomic context in which the new allele appears. It seems that these differences are caused by the type of substitution (TA alternative alleles could create TANNNT motifs, while mutation of wild-type TA bases disrupts them). It is known that there is a bias towards TA substitutions in bacteria [215]. Hence, this could be the cause of the notable difference between new acquisition and loss of TANNNT motifs. Using the global dataset described in Chapter 4, (n=4,595), we checked the alternative alleles derived from single nucleotide mutations and we observed that this pattern was also present across the MTBC (Figure 6.3B). Thus, the mutational signature of the MTBC facilitates the appearance of new Pribnow boxes which, ultimately, supplies the bacteria with a higher transcriptional plasticity.

We tested the potential impact on gene expression of these new and disrupted Pribnow boxes. (see Materials and methods, External Data 12). We took into account only those mutations affecting the clades defined previously in the PDEG as the analysis of individual strains could lead to inconsistent results due to the lack of statistical power. We identified a trend in which new boxes increased the transcription of nearby genes and the disruption of boxes decreased gene expression (Figure 6.4A).

Interestingly, 57 genes identified above as PDEG seem to be differentially expressed by means of a new or disrupted Pribnow box (External Data 11A), meaning that $\sim$26% of the transcriptional variability found across the MTBC main clades was due to single point mutations. Despite the genetic context in which the Pribnow box appears (Figure 6.4C, External Data 12) there is always an increase in transcription rates of nearby genes. Transcription is found to be induced in the sense or antisense direction, depending on the strand in which the mutation appeared, creating a complex regulatory scenario (Figure 6.4C). For example, a mutation previously reported in L3 upstream of the *ahpC* gene

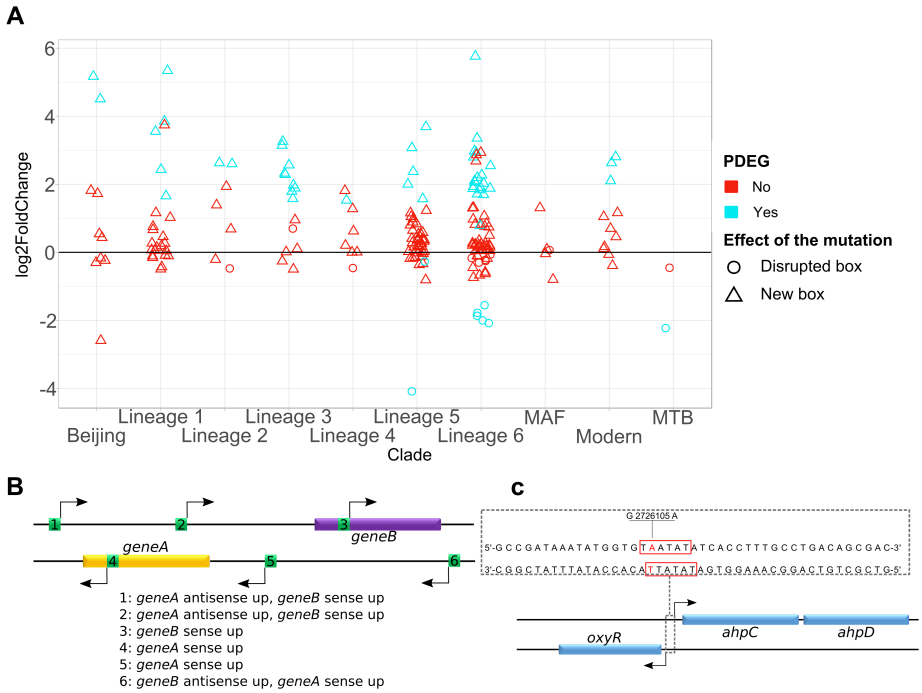


Figure 6.4: Impact of natural mutations in the appearance and disruption of Pribnow boxes. A) Effect of the new/disrupted Pribnow boxes over the expression of nearby genes. New boxes tend to upregulate gene expression while disrupted boxes tend to down regulate transcription. B) The G2726105A mutation, common to all L3 strains, create two new Pribnow boxes in the intergenic region of *oxyR* and *ahpC*. These new boxes are the potential explanation for the observed upregulation of *oxyR*, *ahpC* and *ahpD* in the L3 strains. C) New Pribnow boxes can increase sense and/or antisense expression, depending on the genomic context on which the mutation appears.

creates a new TANNNT motif in both the forward and reverse strands (Figure 6.4B). This new box could be the cause of the observed overexpression of *oxyR, ahpC* and *ahpD* in all L3 strains reported in the section above (External Data 11A and 12).

New boxes enhancing the transcription of complete genes or operons (i.e. those falling in intergenic regions) are more probable to be functionally relevant. An interesting case is the upregulation of the ribonucleotide reductase genes *nrdE*, *nrdI* and *nrdH* (Rv3051c-Rv3053c operon) in the modern lineages

(External Data 11A). The expression of these genes are regulated by the transcription factor NrdR [242]. It was previously hypothesized [243] that a new TSS created by the G3415332A variant present in all the modern lineages could potentially lead to NrdR-independent expression of the operon. Our results support this prediction, as we observe an upregulation of these genes in the strains that harbour this mutation. Other interesting case is the *narG* operon. In our analysis, we detected that the intergenic region located upstream the *narG* gene accumulates 2 variants that create 4 different TANNNT motifs (External Data 12). The C1287112T variant is found in the common branch of the modern lineages and seem to upregulate the transcription of the *narG* operon in the modern lineages compared with L6 and and L1 (Supplementary Figure 10.9). NarG is a nitrate reductase known to be related with the switch from dormancy to active state [244]. L5 had also the NarG operon upregulated but by means of an unidentified mechanism. The other variant found was C1287068T, which creates a new TANNNT motif in N0153 (L1) and increases the expression of the operon in N0153 with respect to the other L1 strains (Supplementary Figure 10.9A). To gain a global understanding of the abundance of these variants, we checked their distribution in the 4,595 strain dataset (Chapter 4). In addition to the variants reported above, we found 2 more variants, C1287081T and G1287182T, that also created a new Pribnow box. Interestingly, all the variants that generate new Pribnow boxes upstream were found to be highly homoplastic (Supplementary Figure 10.9A) suggesting the action of positive selection. Most of our analyses can only be focused in the deeper branches of the phylogeny. However, the data from *narG* suggests that new Pribnow boxes fuelled by genome-wide mutational biases is a mechanism under selection to access transcriptional diversity in clinical strains. One more obvious case was the previously reported G3500149A SNP present in Beijing strains which leads to the overexpression of the *dosR* regulon [182].

Differential methylation patterns across the MTBC

From our RNA-seq analysis it is clear that there are marked differences in gene expression between the main MTBC strain groups. Several mechanisms are known to impact gene expression in addition to sequence changes. Recent studies have shown that DNA methylation can have an effect on gene expression in bacteria [216]. To test the potential transcriptional effect of methylation in the MTBC we tried to link differential methylation (DM) patterns between samples in our dataset with differences in gene expression. To do this, each sample was sequenced using PacBio technology and analyzed with the SMRT Analysis Software to identify methyltransferase recognition motifs (see Materials and methods for details). Consistent with previous reports, we identified three main methylated motifs in almost all the samples [245, 246, 247]. The frequency of methylated sequences among these motifs was near 100% in almost all the samples. In some of the strains however, the sequences recognized by the methyltransferase were not methylated (frequency of methylated motifs 0%), suggesting that the methyltransferase that recognises this pattern is inactive (Table 6.2).

The three main methyltransferases that recognise these motifs are MamA, MamB and HdsM/HsdS.1/HsdS. Interestingly, in two cases (N0052 and N0136) we observed that only a fraction of the motifs recognised by MamA were methylated (20% in N0052 and 56% in N0136). The sequences recognised by MamB and HdsM/HsdS.1/HsdS in N0052 and N0091, respectively, were also partially methylated along the genome ($\sim$70% of the sequences), suggesting that the activity of those methyltransferases was reduced, but not eliminated.

We wanted to identify the genetic variants that could be responsible for these functional differences. We therefore analysed the methyltransferase coding genes in the strains lacking methylation of one or more of the three motifs. From that, we identified several nonsynonymous SNPs potentially involved in the methyltransferase inactivation (or partial inactivation) (Table 6.2). Some of these variants have already been reported [246] while others are

		mamA			*hdsM/hsdS.1/hsdS*			*mamB*		
	Strain	**Variant**	**Phenotype**	**Methyl freq.**	**Variant**	**Phenotype**	**Methyl freq.**	**Variant**	**Phenotype**	**Methyl freq.**
L1	N0072			0.97			0.91			0.96
	N0153			0.96			0.92			0.96
	N0157	W136R	Inactive	0			0.93	D59G	Inactive	0
L2	N0031			0.98			0.93			0.97
	N0052	E270A	Partially methylated	0.19			0.93		Partially methylated	0.7
	N0145	E270A	Inactive	0			0.92			0.97
	N0155	E270A	Inactive	0			0.88			0.98
L3	N0004			0.98	G173D(*hsdM*) L119R(*hsdS*)	Inactive	0			0.97
	N0054			0.98	G173D(*hsdM*) L119R(*hsdS*)	Inactive	0			0.98
	N1274			0.98	G173D(*hsdM*) L119R(*hsdS*)	Inactive	0			0.98
L4	N0136	G152S	Partially methylated	0.56	P306L (*hsdM*)	Inactive	0			0.98
	N1216			0.97	P306L(*hsdM*)	Inactive	0			0.98
	N1283			0.99			0.96			0.99
L5	N1063			0.97			0.97			0.98
	N1176			0.94			0.93			0.96
	N1272			0.97			0.98			0.98
L6	N0091			0.96	E481A (*hsdM*)	Partially methylated	0.68			0.96
	N1177			0.97	T393A (*hsdM*)	Inactive	0			0.96
	N1202			0.96			0.9			0.97

Table 6.2: Methylation profile of the 19 MTBC samples. Summary of the methylation report derived from the PacBio sequencing for each of the three main methyltransferases.

novel. Curiously, *mamA* in N0052 carries the same mutation as *mamA* in N0145 and N0155, although the activity in N0052 was only partially lost, compared to full loss of activity in the latter two, suggesting other genetic variants outside the gene may be having an effect. We expanded our analysis to the bigger dataset of 4,595 strains to get a global picture of the methyltransferase conservation degree. Some of these variants located deep in the phylogeny affected complete lineages while others were more recent and affected only a subset of strains (Figure 6.5). For example, of the W136R and G152S mutations that were found in MamA inactive strains, G152S was found in a subset of L4.3.3 strains and a small clade of L1.1.2 (it is homoplastic) while W136R affected a subset of L1.2.1 samples. Interestingly, a new *mamB* variant (D59G) was found also in these strains potentially linked to MamB inactivation. On the other hand, a T393A variant was found to affect *hsdM* in a subset of L6 strains potentially affecting the methyltransferase activity in those strains.

To gain a wider perspective on the main MTBC methyltransferase diversity, we analyzed all the variants present in these genes in the larger dataset. In all the methyltransferase coding genes, we found nonsynonymous variants (External Data 13), most of them potentially compromising gene functionality. The dN/dS values for *mamA* (0.75) and *mamB* (0.72) were slightly higher than the mean dN/dS value for non-essential genes (0.66 [87]). In contrast, HdsM/HsdS.1/HsdS has a different pattern, with the genes that encode for the specificity units *hsdS* (0.73) and *hsdS.1* (0.69) having similar values than *mamA* and *mamB*, and the gene that encode for the methyltransferase unit *hsdM* (0.5) showing a value similar to that of the essential genes (0.53 [87]).

The impact of DM on gene expression is subtle and lineage independent

DM in regulatory regions has been reported as potentially affecting gene expression in H37Rv [218]. We wanted to check if DM naturally present in our strains could be linked to differential gene expression. To do that, we looked for SigA recognition motifs (TANNNT / GNNANNNT [183]) in gene promoter

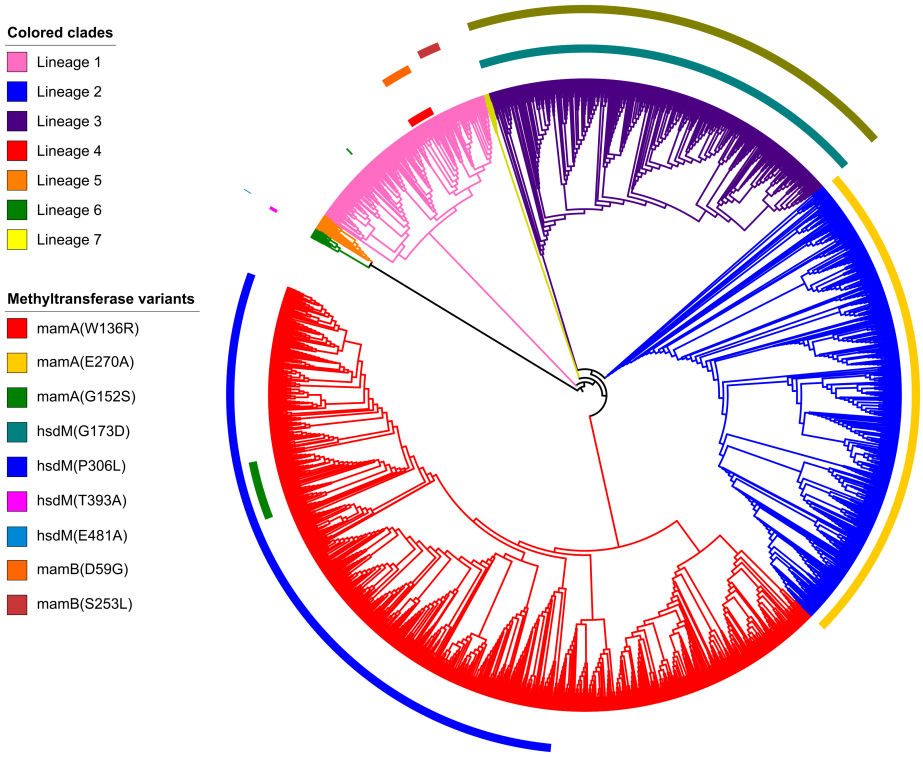


Figure 6.5: Methyltransferase activity of the MTBC. Distribution of the characterized mutations that potentially impair methyltransferase function on the global dataset.

regions (-50 bp upstream the TSS previously defined [127]) that overlap with methyltransferase recognition motifs. We managed to identify SigA recognition motifs for 13 genes overlapping with the MamA motif, 24 with the HdsM/HsdS.1/HsdS motif and with the MamB motif (Figure 6.2B). To account for differential gene expression due to DM and not for other evolutionary reasons, we compared gene expression values in strains which belonged to the same lineage but in which the specific methylase was either active or inactive. This was the case for MamA in L1 and L2, HdsM/HsdS.1/HsdS in L4 and L6, and MamB in L4 (Table 6.2).

First, we compared the expression of the 13 genes identified in both

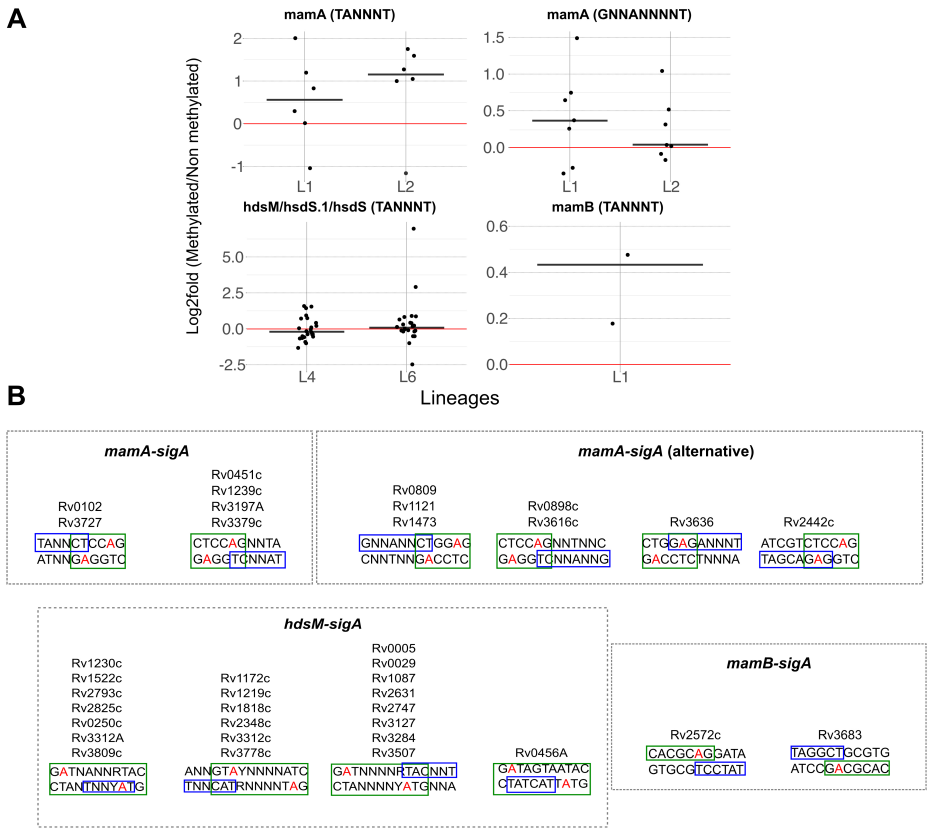


Figure 6.6: Impact of DM on gene expression. A) Gene expression differences between with SigA recognition motifs differentially methylated by each of the three methyltransferases. Red line maks a 0 fold-change in gene expression (no differences). The expression of each gene was tested in both situations, methylated and non-methylated strains, in independent lineages (when possible). B) Different overlapping patterns found between SigA recognition motifs and the methylated motifs. The red adenines in the motif are the methylated ones.

situations (MamA activated or inactivated) in L1 and L2 strains. We observed that almost all genes increase their expression values in the methylated strains in both lineages, matching previous observations in H37Rv [218] (Figure 6.2A). However, we identified some exceptions in which the gene expression behaved differently in each lineage. This was the case for example for Rv3272. In L1, this gene showed a lower expression in methylated strains (N0072 and N0153)

than in the non-methylated strain (N0157). Rv3272 is regulated by the transcription factor Rv0022c which had an early-stop codon in the N0157 strains [248]. This type of polymorphism could be the cause of the discordant results. For MamB, we only found 2 genes where SigA and MamB motifs overlapped. Even so, for these two genes we observed the same effect as in MamA DM strains.

However, for HdsM/HsdS.1/HsdS we did not observe this pattern of changes in gene expression. The overlap between SigA recognition motifs and HdsM/HsdS.1/HsdS motifs in the regulatory regions seemed to have no impact on the gene expression. In some cases the genes increased their expression in the non-methylated strains while some others behave in the opposite manner. Moreover, this behaviour was not congruent in L4 and L6 as half of the genes showed the same regulatory response in both lineages while the other half behave differently in each lineage. So, MamB and MamA methylation over SigA motifs seems to cause a similar effect independently of the strain genetic background while HdsM/HsdS.1/HsdS seems to have no effect.

We searched for other sigma factors apart from SigA that could potentially have an overlap between their recognition motifs and methyltransferase recognition motifs. We found that the SigB recognition motif (NNGNNG) could overlap, so we applied the same analysis as for SigA. However, DM seemed to have no effect on SigB regulated genes. It is known that SigB plays a role during stress response [249] but it is dispensable for growth. As the RNA-seq samples were collected during exponential growth (applying no stress), DM over SigB influenced genes could show little or no differential expression.

A different mechanism for HdsM/HsdS.1/HsdS gene expression regulation

Although we were not able to link HsdM DM in the SigA sites with regulation of gene expression, the natural variability of HsdM found in the MTBC suggests that there may be some biological relevance associated with this protein. We

therefore created an HsdM mutant by deleting the *hsdM* gene in a N1283 background (L4).

We performed a transcriptomic analysis comparing the Δ*hsdM* strain and the wild-type. An initial analysis showed differences between the strains, as the transcriptomes split into two groups in a PCA analysis (Figure 6.7A). In the Differential Expression (DE) analysis, we observed that these differences were mainly driven by a small number of genes (BH adj-pvalue $<$ 0.05 and log2fold-change $>$ 1, External Data 14, Figure 6.7B). In N1283-Δ*hsdM*, several genes were increased in expression in comparison with the wild-type. First, *hsdS.1* expression was increased in the mutant, suggesting that its regulation is linked to *hsdM* (which is found upstream in the H37Rv genomic context). In addition, a set of 7 consecutive genes (Rv0081-Rv0087), potentially forming an operon, were found to have increased expression. Interestingly, Rv0081 is a transcriptional hub involved in the regulation of multiple genes [181, 248, 173], including the hyc-family genes, which have homology to so-called EHR (energy-converting hydrogenases related) complexes. Evolutionarily, EHR proteins stand between complex 1 and NiFe-hydrogenases [250] and their functions have yet to be determined. The EHR complex of the MTBC is with high certainty not a functional hydrogenase, because *M. tuberculosis* lacks the cluster of assembly genes needed to mature NiFe centers and insert them in to the protein [251]. In contrast, Rv1813c, Rv0080, Rv3131 (all hypothetical proteins) were decreased in expression in the mutant, as well as *ctpJ*. Thus, HsdM methylation has an effect on gene expression, but the mechanism seems to be different to that of MamA and MamB, as the genes reported above did not have any overlap between SigA and HsdM motifs. Moreover, we found no bases methylated by HsdM near these genes, suggesting an indirect effect of HsdM DM on gene expression.

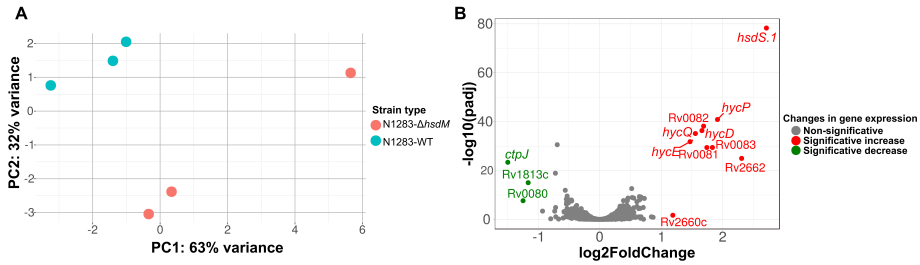


Figure 6.7: Gene expression differences due to an *hsdM* deletion. A) Overall transcriptomic profiles of the wild-type versus the Δ*hsdM* strains. B) Volcano-plot of the gene expression differences of the wild-type versus de mutant strains. A small numbers of genes showed significative differential expression (3 down-regulated and 10 up-regulated).

6.3 Discussion

Our results show that the different MTBC clades have their own transcriptomic signature. Each main lineage showed a transcriptomic landscape, clearly separated from the rest of the lineages. However, mutations impacting key regulators can blur these separations and alter the transcriptomic profile of the strains carrying these mutations. An example was strain N1177, which carries a single mutation in *rpoB* gene conferring rifampicin resistance which modified the transcriptional levels of multiple genes.

Our phylogeny-based approach allowed us to identify gene expression changes that took place during the evolution of the MTBC. We observed that, as the MTBC diverged into the different lineages, the expression of specific genes varied. There was a moderate correlation between increasing genetic distance and increasing gene expression differences. Modification of gene expression could be a rapid mechanism of physiological adaptation to a new environment, without the need to substantially change the genome. This seems to have been the case when MAF and MTB split from a common ancestor. In a relatively short genetic distance, many genes changed their expression. We propose that a sudden environmental change (possibly a change in host population) rapidly selected nascent phylogenetic groups that behaved

differentially in terms of gene expression, or that standing variation in regulation allowed the ancestor to differentially specialise in different environments. In accordance with this, enrichment in genes involved in metal homeostasis may be related to different concentrations of ions in different host populations or animals [252]. For example, *ctpA* has been found to be under positive selection in ancient strains of animal origins infecting humans [82]. This could be related to the point at which the basal MTBC clades diverged, following the human population migrations [36, 82]. The opposite situation seems to have occurred in the modern lineages. They seem to have not accumulated large transcriptomic differences although a large genetic distance separates them from their common ancestor.

The GO analysis showed that the main functions affected by the expression differences were those related to nutrient uptake and the macrophage environmental conditions, despite the genes involved being different in each clade. This reinforces the idea that the different lineages have adapted to different hosts, not only at the gene sequence level, but also by altering the expression levels of specific genes. Thus, the variability in gene expression patterns found in the different clades could be related to the adaptation to specific host populations (i.e. different environments).

The analyses of the genetic bases of expression differences between phylogenetic clades reveals an interplay of natural selection and mutational processes. Up to 26% of the core expression differences between lineages were due to single point mutation creating new Pribnow boxes in gene regulatory regions. The number of new Pribnow boxes are more than expected by chance and thus selection probably played a role fixing expression differences. Importantly, the underlying AT mutational bias across the genome has been a source of expression diversity through random generation of new Pribnow boxes, as previously theorized [183]. The reason why selection is apparent for SigA motifs but not for other sigma factors remains unclear but at least two non-mutually exclusive explanations are possible. On the one hand, SNPs impacting SigA recognition motifs can have an impact across

environmental conditions while SNPs for other sigma factors only will be relevant for specific conditions. They may happen but are more difficult to detect in our analyses. On the other hand, SigA motifs are enriched in AT bases and thus it is not surprising that new SigA motifs are generated at a faster pace leaving more room for selection to act.

Our results also show that methylation seems to play a minimal role in shaping *in-vitro* gene expression. We have not been able to detect a regulatory impact for the main methyltransferases, except for a subtle effect on few genes having overlapping SigA and MamA/MamB recognition motifs, consistent with previous reports [218]. This could be due to our inability to identify genes that are actually influenced by the methyltransferases, as the Δ*hsdM* strain shows differential expression in genes that we had not previously identified as potentially influenced by HsdM. MamA/MamB methylation motifs do overlap with SigA recognition motifs, affecting the transcription mediated by SigA, however, this seems to not be the case for HsdM. It is intriguing that the GC content in MamA/MamB recognition motifs is higher than in HsdM motifs. As the transcription machinery needs to open the DNA strands, it is possible methylation could have some synergistic effect with the high content of GC bases on initiation of transcription, reducing the potential effect of the methylation in the AT-rich HsdM motifs. Whatever the mechanism, the analysis of the *hsdM* mutant has shown that HsdM-mediated methylation does have an effect on gene expression of several genes, although the exact mechanism remains to be elucidated.

It seems clear from our results that there has been a convergence of the methylation patterns in the different phylogenetic groups of the MTBC, instead of a lineage-specific pattern as proposed previously [246]. Equivalent phenotypes (non-methylation of specific motifs) appear to be produced by different genetic variations. For example, W136R mutations in a subset of L1 strains seem to have the same effect as E270A in a subset of L2 strains, impairing MamA activity. Similarly, G173D and/or L119R mutations in HsdM and HsdS seem to inactivate HdsM/HsdS.1/HsdS in L3, which is the same

phenotype we found in a sublineage of L4 potentially due to the effect of P306L in HsdM. Moreover, one of the variants characterized was found to be homoplastic. These convergent change could suggest that methylation in the MTBC still plays a biological role, although not necessarily related to gene expression regulation.

In summary, in this chapter we have carried out a comprehensive comparison of transcriptomes and DNA-methylomes of nineteen clinical isolates representative of the global phylogenetic spectrum of human *Mycobacterium tuberculosis* complex. Patterns of differential transcription between lineages reflected constitutive expression of genes that are normally regulated in response to environmental cues, as a result of mutations that introduce novel TANNNT Pribnow boxes and mutations that impair the function of transcriptional repressors. The role of methylation is more elusive but it is clear from pattern of inactivating mutations that methylases are not conserved across the MTBC. Isolated from the opportunity to generate diversity by horizontal gene transfer, transcriptional adaptation may allow *M. tuberculosis* isolates to optimise their infectivity and transmission in subtly differing environments provided by different human host populations.

6.4 Materials and methods

The culture, DNA/RNA extraction and sequencing processes were performed by external collaborators in different institutes. More concretely, strain culture and DNA/RNA isolation were performed by Michael Berney at the Albert Einstein College of Medicine, in New York, USA. The complete sequencing process was performed by Christine Boinett and Julian Parkhill at the Sanger Institute, in Hinxton, UK. The accesion numbers for the newly generated data can be found in External Data 15.

The rest of the research comprising the transcriptome, methylome and genome analysis was performed by the author of the thesis. Part of this research took place during a short stay of three months in the London School

of Hygiene and Tropical Medicine, in London, UK, under the supervision of Teresa Cortés.

Culture conditions

All cultures were grown in ink wells containing 10 ml Middlebrook 7H9 OADC medium supplemented with 30 mM sodium pyruvate to account for pyruvate kinase mutations in L5 and L6. Cultures were grown on orbital shakers at 80 rpms at 37ºC. For each strain, two biological replicates were generated.

RNA isolation

For RNA extraction cultures were grown to OD_{600} of 0.5 - 0.7. Ten ml aliquots were spun down and immediately processed with TRIZOL reagent according to manufacturer protocols. Cells were harvested from exponential cultures and RNA extracted using the Direct-zol™ RNA Kit from Zymo according to manufacturer's instructions. From the RNA extracted, ribosomal RNA was depleted by using a Ribo-Zero Magnetic Kit. After that, sequencing libraries were prepared using the TrueSeq stranded Illumina protocol and sequenced on an Illumina HiSeq 2500 platform.

DNA isolation

For DNA extraction cultures were harvest between OD_{600} of 0.5 - 0.7 by spinning down 5 ml culture and immediately starting DNA extraction by CTAB method[253].

RNA-seq pipeline

Fastq files qualities were assessed using FastQC [254]. Trimmomatic, a program that uses a dynamic trimming approach [255], was used to remove bases from the start and the end of the reads when its quality was below 20. Reads were mapped to the H37Rv reference strain [64] using BWA-mem

algorithm [86]. Potential duplicates were removed by using the MarkDuplicates option from the Picard tools package [89]. Bedtools [256] was used to calculate the read coverage for each genomic feature. To precissely report the coding and non-coding coverage, each read was classified according to the strand from which it was initially derived.

Transcriptomic analysis

The transcriptomic analysis was performed using the R statistical language [101], specifically the DESeq2 package [107]. The input data was the count table containing the coverage information for each feature for all the samples. The PCA and the hierarchical clustering were performed by previously normalizing the count data across samples and scaling it into a log2 scale, by using the rlog function from the DESeq2 package. For the analysis of Phylogenetically aware Differentially Expressed Genes (PDEG), we performed a two-step process. First, we identified all the genes having differential gene expression (adjusted BH p-value < 0.05 and log2fold-change > 1.5) between each pair of phylogenetic groups with a common origin (for example L5 and L6, MAF and MTB, etc). Therefore, we identified the genes changing their expression between these groups. This information however, is not enough to assign the expression change to one group or the other. We can not know if an increase in gene expression for one gene is due to an up-regulation in one group or to a down-regulation in the other. To resolve this, for each gene identified as differentially expressed, we compared its expression value in each of the two groups against the rest of the MTBC samples. This analysis allowed us to identify the group in which the change in gene expression took place and the direction of this change. Finally, we assigned all the changes in a group to the tree branch common to this clade. For this part of the analysis, sample N1177 (L6) was excluded, as the *rpoB* mutation alters its transcriptomic signature and it is therefore not representative of the L6 transcriptomic signature. Genes with deletions in each of the groups were not taken into account in the pairwise comparisons, as they result in false positive signals.

These genes were identified by mapping long-reads obtained from PacBio sequencing against the H37Rv reference genome, and assessing the genomic coverage. PE/PPE, phages and repetitive genes have not been taken into account in any of the analyses, as their sequenced reads are prone to map erroneously. The enrichment analysis in GO functions was performed using the BiNGO tool [99]. BiNGO identifies the most abundant functions in a subset of genes, compared to all functions present in a complete genome using a hypergeometric test (sampling without replacement).

FASTQ mapping and variant calling from the Illumina data

For each of the analyzed strains, we downloaded the publicly available genomic data from a previous work [220] (Supplementary Table 7). The variant calling was performed as described in the Chapter 3.

Creation and disruption of Pribnow boxes

Using the MTBC ancestor genome as a template, we introduced all the mutations found in the dataset, and look for new/disappeared TANNNT motifs with the fuzznuc tool included in the EMBOSS program [257]. By doing this, we have obtained the number of affected Pribnow boxes in our dataset.

To calculate the probability of appearance or disruption of Pribnow boxes we have first scanned the MTBC ancestor genome looking for the 'ancestral' TANNNT motifs, or for motifs that could result in TANNNT motifs by introducing one single mutation (VANNNT, TANNNV, TBNNNT). In parallel, from the observed number of variants in the MTBC dataset, we calculated the probability of a non-A (B), non-T(V), A and T mutations. After that, we calculate the expected disruption of boxes by inferring the probability of non-A or non-T mutations to fall in the 1st, 2nd or 6th position of the 'ancestral' motifs. The expected generation of new boxes was calculated by inferring the probability of A and T mutations to fall in the corresponding VANNNT, TANNNV and TBNNNT motifs. In a last step, we have used a Poisson distribution to calculate the

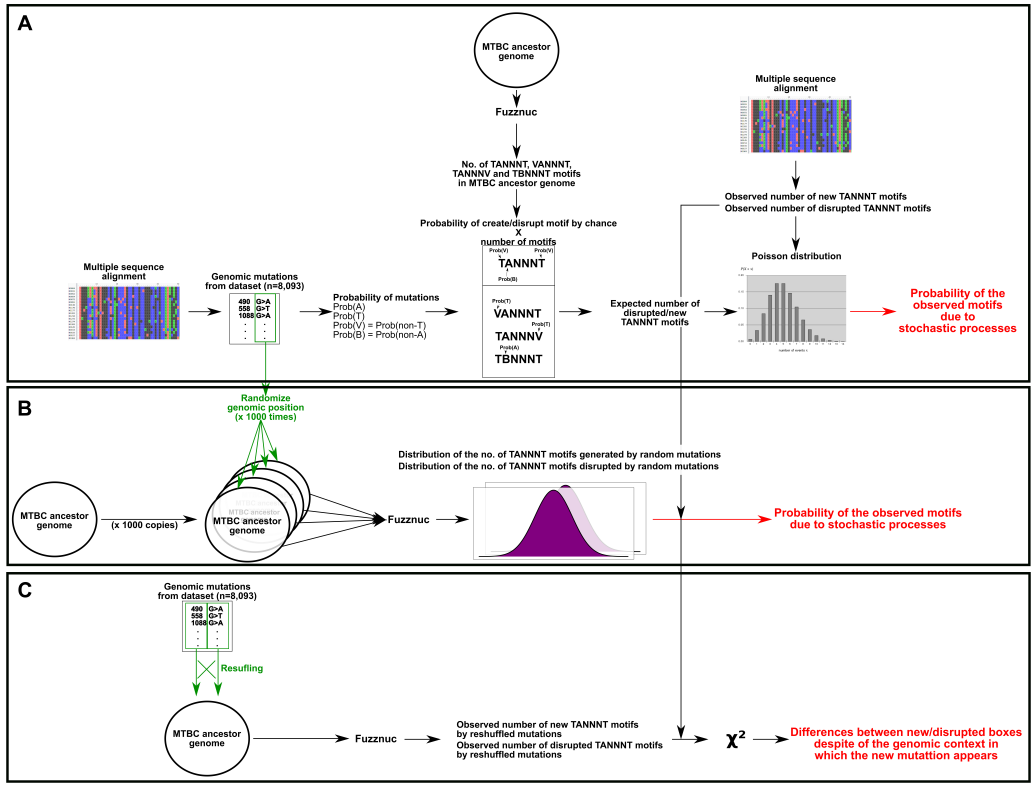


Figure 6.8: Statistical analysis performed to evaluate if non-random process are influencing the observed number of new and disrupted Pribnow boxes. A) Probability of the observed creation/disruption of Pribnow boxes, given the probability of each genomic position to accumulate variants. B) Permutation test to assess the number of new/disrupted boxes expected by random mutations impacting the MTBC genome C) Chi-squared distribution of expected vs observed ratio of new/disrupted boxes, when the mutations in the dataset are reshuffled.

probability of the expected versus the observed number of disruptions in our dataset (Figure 6.8A).

A random permutation test was performed by keeping the alternative alleles for the 8,093 SNPs found in the global dataset, but randomly assigning a new genomic position in which those SNPs appear. Later, we scanned for new/disappeared TANNNT motifs with the fuzznuc tool. This process was repeated 1,000 times. With the number of new/disrupted boxes in this 1,000 simulations we calculated a cumulative empirical distribution of expected Pribnow boxes affected by random mutations. Later, we compared these distributions with the number of boxes affected by the real variants (Figure 6.8B).

Finally, we have reshuffled the 8,093 mutations so the alternative alleles were randomly assigned to genomic positions that initially harbour other variants. Again, we impacted the MTBC ancestor genome with these variants and assessed the number of new/disrupted Pribnow boxes by using fuzznuc. The results obtained were confronted against the number of new/disrupted boxes with the real 8,093 variants in a chi-squared test (Figure 6.8C).

Genomic DNA isolation for PacBio Sequencing

DNA was prepared and sequenced on the Pacific Biosciences RSII machine as described previously [258]. Briefly, we used template preparation kit version 3.0, polymerase binding P6 version 2, and sequencing reagents version 4.0 (C4). Data were captured using 3-h movies. Each sample replicate was sequenced on two to four chips to get enough genome coverage for the detection of the methylated patterns.

HsdM mutant construction

The gene *hsdM* was deleted in sample N1283 by specialized transduction as described previously [259]. Transductants were recovered on 7H10 OADC plates containing hygromycin (75 g/mL). Mutations were confirmed by

three-primer PCR using primers *hsdM* L, *hsdM* R, and Universal uptag, listed in Table 6.3.

Primer	Secuence
hsdM L	CTCTGGTCAACGCAATGT
hsdM R	ATCTGAGACTCCTCCATTCC
Universal uptag	GATGTCTCACTGAGGTCTCT

Table 6.3: Number of Phylogenetically Aware Differentially Expressed genes and the branch length for each of the main clades in the MTBC phylogeny. The phylogeny was calculated by using the ML algorithm, with 1,000 iteration bootstraping.

Methylation analysis and variant search

DNA isolated from the cultured samples were used for PacBio sequencing. The SMRT Analysis Software from PacBio [260] was used to detect methylation patterns in the PacBio sequencing data. Sequencing reads from both biological replicates per strain were merged to assess a higher sequencing depth. The Modification_and_Motif_Analysis protocol was used as defined in the SMRT manual. This protocol detects the Interpulse Duration (IPD) to classify one base as methylated. After that, it looks for over-represented methylation motifs in the genome. For those strains lacking at least one of the three main methylated motifs present in the rest of the dataset, we looked for nonsynonymous variants affecting the methyltransferases. These variants, that potentially affect the methyltransferase function, were also scanned in the 4,595 strains dataset representative of the MTBC global diversity described in Chapter 4. The potential effect of the nonsynonymous variants over gene functionality was assessed by using the SIFT4G tool [92]. dN/dS values for the methyltransferases were calculated as explained in Chapter 4 methods.

RNA-seq analysis linked to differential methylation

We used fuzznuc from the EMBOSS package [257] to identify genes whose conserved -10 TANNNT motif overlapped with identified methylated motifs. The

potential effect of methylation over the expression of these genes was assessed by comparing the expression values in strains with a similar genetic background (same lineage), but having differential methyltransferase activity.

The work described in the present chapter is currently under second review:
Chiner-Oms Á., Berney M., Boinett C., González-Candelas F., Young D., Gagneux S., Jacobs W.R., Parkhill J., Cortes T., Comas I. Genome-wide mutational biases fuel transcriptional diversity in the *Mycobacterium tuberculosis* complex . **Nature Communications**. Under review.

General discussion

Infectious diseases rank among the top 10 causes of death worldwide, being a global health concern for the WHO. Among them, tuberculosis is the leading cause by a single infectious agent. As a result, multiple international initiatives involving public and private funds have been proposed in the last decades to stop TB pandemic. However, an annual declining rate of 2% has been achieved in 2017, not enough yet to accomplish the WHO objective of eradicating the disease by 2035. Innovative solutions to this problem depend on gaining more knowledge about the complex biological mechanisms of the pathogen, the host and their interaction. From the pathogen point of view, we need to gain more insights into transmission dynamics, persistence, virulence (in all its alternative meanings) and drug-resistance development, as these are the main bacterial determinants of the disease outcome.

These characteristics are highly heterogeneous in pathogens with complex population structures. For example, *Salmonella enterica* is a pathogenic bacteria, with a population structure formed by several subspecies [261]. It has been shown that recombination exists within and between these subspecies [262]. The *S. enterica* genetic lineage defines the host range for the different subspecies, as well as the symptoms of the salmonellosis infection [263]. Another example is *Legionella pneumophila*, causative agent of Legionnaires' disease, which has a mosaic population structure [264, 265]. Genetic differences between the distinct *L. pneumophila* clades have impact in

pathogenicity and intracellular growth rate [266].

In the case of TB, it has been shown that heterogeneities disease outcome are also related with the genetic background of the bacterial strain [42, 38], despite its (almost) clonal population structure. So, genetic diversity of the MTBC is gaining momentum as a major point to evaluate in TB research, in contrast to what happened until very recently. For example, systems biology approaches rarely accommodate information about natural polymorphisms in the systems studied. In addition, even for model organisms, this type of approach has been rarely applied to more than one genetic lineage [267]. This is also true in the application of systems biology to TB research. As we have seen in Chapter 5, computational models derived from a single strain do not make reliable predictions in MTBC strains from different genetic backgrounds. We focus our research in human MTBC lineages but these differences are even more pronounced when human- and animal-adapted strains are compared or to the BCG (animal-like) vaccine [152, 268, 235]. We have also seen how mutations found across the whole human MTBC impact different modules of the regulatory and PPI networks depending on the bacterial clade. These mutations could result in a high impact on gene functionality, sometimes impairing it (i.e., early stop-codons). Despite the existing variability, most of the analyses and predictions performed until now have been based on network models derived from clinical reference strains used in the laboratory, mainly H37RV (L4), CDC1551 (L4), Erdman (L4), HN878 (L2 - Beijing), AF2122/97 (*M. bovis*). It is true that, despite the limitations of current predictive models, we cannot preclude that significant results can be obtained from their application [79]. Nevertheless, we are losing the details of the complete picture. For example, we do not know how resilient to perturbations these networks are. Does the topology of these biological networks differ between different strains, in order to bypass dysfunctional nodes while maintaining their main performances? Or, on the contrary, is the topology maintained but existing nodes gain new functions to supply the impaired ones?

So, it is apparent that at least lineage- and condition-specific network

models will be necessary to generate more accurate predictions across the *Mycobacterium tuberculosis* complex. We are aware that meeting the above conditions is a major experimental and computational accomplishment. For example, the regulatory network and the gene expression data used in Chapter 5 were derived from almost 200 TFs genetic constructs and $\sim$700 microarray experiments and used only strain H37Rv. Generating comprehensive models for all major human- and animal-adapted lineages of the *M. tuberculosis* complex will represent a challenge in the years to come.

The "limited" genetic diversity of the MTBC responds to several factors. First, the bacteria has a low mutation rate in comparison with other pathogens [269]. This is probably influenced by its low growth rate (doubling time of 24 h. in culture) and dormancy times, in which the bacteria can reside for years within granulomas with a reduced physiological activity. Second, as it has been shown in Chapter 4, there is neither ongoing recombination among MTBC strains nor modern acquisition of genetic material from other bacteria. This last point is specially relevant for understanding the development of drug-resistance in the MTBC. In many pathogenic bacteria, drug-resistance mutations spread in bacterial populations through recombination and horizontal gene transfer. This is the case of *Neisseria gonorrhoeae* [270], for example, in which a multi-resistant lineage has established globally and its resistant determinants have disseminated in the bacterial population through homologous recombination. In contrast, drug-resistance mutations in the MTBC are all chromosomic and do not spread through genetic interchange (they appear independently in different bacterial strains). This fact has important implications in terms of surveillance and epidemiology. Recombinant pathogens are able to share its drug-resistant determinants with other members of the population, so the control and eradication of a drug-resistant strain do not guarantee the stagnation of the resistant phenotypes. In the case of the MTBC, the spread of drug-resistant phenotypes in the bacterial population is through clonal inheritance, so the identification and treatment of these strains is determinant to cut transmission of specific drug-resistant phenotypes.

WGS is gaining momentum as a key methodology to tackle the TB epidemic. The analysis of bacterial genome sequences allow us to identify drug-resistance mutations even when they are not fixed in the population, ie. at a low frequency. In a recent work [271], our research group described a patient which carried a persistent infection over nine years, with recurrent relapses. In a first TB episode, the patient was initially diagnosed with a susceptible strain by routine means of different drug susceptibility tests, and it ended up developing several drug-resistances through time. The authors performed a retrospective study using WGS and found that uncommon drug-resistance mutations were present in a relapse episode 4 years later. These rare variants are missed by routine clinical diagnostic methods. In addition, they identified resistant variants related with MDR, indicating that the patient developed an MDR infection during time. With this information, the patient treatment was changed to finally be cured. In another study [272], the authors identified a patient which was infected by a strain with a complex resistant profile. Bedaquiline and delamanid, a recently approved anti-tubercular drugs at that time, were added to the treatment at different points. By using WGS, the authors realised that, after several relapses, the infectious strain had acquired resistance to these new drugs. The patient was finally cured after undergoing surgery, but the study shows that inadequate treatment of MDR-TB and XDR-TB cases could lead to the development of new resistant mutations, even for newly developed drugs. These studies highlight the importance of using WGS approaches as a first line tool for diagnosis in clinical settings, at least in high income countries. In fact, WGS is becoming the *de facto* standard for surveillance and monitoring of epidemic outbreaks [273, 274]. Specifically, in the last years this methodology has been applied to large-scale surveys of pathogenic bacteria such as *Listeria monocytogenes*, *Staphylococcus aureus* and *Salmonella* species [261].

In addition to the significative advantage that WGS entails in applied and clinical contexts, the amount of information derived from this technique has also been used to greatly improve the quality of basic research. Almost all the

results derived in the present thesis were obtained from WGS data, downloaded from public databases or generated by collaborators for other projects. We have been able to study genomic polymorphisms at the single nucleotide level in thousands of genomes which, in the end, have allowed us to perform robust statistical analyses. An illustrative example of this is the linkage-disequilibrium analysis performed in Chapter 4. While comparing the co-occurrence of independent alleles in thousand of genomes, in a dataset of 98,780 variant positions, we can affirm that our results are statistically well supported. Thus, we can be very confident about the conclusions derived from them. Another example is the important discovery of *phoR* as an ongoing key player in the MTBC evolution. Again, the low genetic diversity of the MTBC made it difficult to perform selection tests on specific genes. So, the management of thousands of strains has allowed us to perform these tests and to detect the traces that positive selection has left on the *phoR* genetic sequence.

Besides DNA, high-throughput sequencing techniques can be applied to RNA to gain insight into the organism physiology that we cannot derive solely from the genomic sequence. Several studies comprising different organisms have reported that small genomic changes can have a strong impact on gene expression patterns [275, 276, 277, 278]. So, species with low genomic variability among its members can show wide transcriptional fluctuations which, ultimately, could have an effect on the phenotype and physiology of these organisms at multiple levels. In Chapter 6, we have made use of high-throughput sequencing to define the gene expression patterns of the main clades of the MTBC. We have observed that the different clades have a variety of genes showing changes in their expression values through time. These fluctuations affect several biological functions, almost all of them related to the interplay between host and pathogen. It makes complete sense as changes in gene expression are a mechanism of adaptive plasticity to environmental changes.

The variability in gene expression patterns found in the different clades

could be related to the fact that many MTBC lineages and sublineages are better adapted to specific host populations (i.e., different environments). Nevertheless, studies in *Escherichia coli* have shown that populations evolving in different environments and with diverse gene expression patterns show similar phenotypes [279]. So, it is possible to reach phenotypic convergence even when having different transcriptomic patterns in independently evolved clades. We are aware that an important limitation of the studies presented in this thesis is that we have not used *in-vivo* or *in-vitro* infection models. Although that does not detract from our conclusions, future *in-vivo* and *in-vitro* experiments could be necessary to identify the genetic determinants that lead to this host-specificity, and to link gene expression with phenotypic characteristics.

We have shown different mechanisms for which the transcriptional diversity of the MTBC have been generated. As we have shown and discussed in Chapters 5 and 6, mutations affecting TFs and regulatory regions have heterogeneous impact on gene expression regulation. For example, the D435Y mutation in *rpoB* gene in N1177 alters its transcriptomic profile to the extent that its pattern is different to the other strains tested from the same phylogenetic clade. And this impact could potentially affect the phenotypic outcome. In addition, there is now substantial evidence in bacteria that DNA methylation can affect transcription, and that changes in methylation patterns can modify transcription patterns over long or short timescales [280]. The use of SMRT sequencing in Chapter 6 has allowed us the direct detection of this DNA methylation [217]. Regarding *M. tuberculosis*, it has been published previously that modifications induced by the MamA methyltransferase affect the regulation of several genes, in an H37Rv background [218]. We have scaled this analysis to study the potential regulatory effect of the three main methyltransferases in the major MTBC lineages. As a result, we have not been able to detect a regulatory role for the main methyltransferases, except for a subtle effect on a few genes. This could respond to three main reasons: (i) phase variation mechanisms act when the organism needs a quick response to

rapid changing environments; hence,to see a regulatory effect we might need to apply several stresses, (ii) methylation could be an ancestral phase variation mechanism in the MTBC an it has lost its regulatory role in current strains, and (iii) we have not tackled the proper approach to identify genes influenced by the methyltransferases, as the $\triangle$*hsdm* strain shows differential expression in genes that we have not determined previously as influenced by HsdM. Thus, understanding gene expression regulation in an holistic manner is complex and difficult to fulfill even with high-throughput datasets. Moreover, transcriptional levels do not correlate perfectly with translation rates, as post-transcriptional regulatory events can occur before obtaining the final protein [203, 204]. Future approaches in this field should generate and integrate multiomics data in order to capture the major regulatory layers [281]. The challenge will be to integrate all of them in a manner that can inform each other [282] and to accommodate and predict the role of existing MTBC genetic diversity[283].

What is clear from our analyses in Chapters 4 and 6 is that, from a physiological point of view, the MTBC members are adapted to the host environments. This is not a surprise, as it is well known that the MTBC has evolved in parallel with its host population for millennia [82, 36, 35]. Previous reports [65] and our data in Chapter 4 have shown that purifying selection is acting in certain parts of the genome, specifically in the epitopes regions. Given that the bacteria need to be recognized by the immune system to complete their infectious cycle, and that this recognition involved the epitopes, it has been proposed that the hyper-conservation of the epitopes is advantageous for the MTBC [87] in contrast to what occurs with other pathogens [284, 285].

Despite purifying selection in epitopes, there are reports describing the action of other types of selection over specific genes in current MTBC strains. As shown in Chapter 4, several methods agree in concluding that positive selection is shaping the evolution of the *phoR* gene in past and current settings. However, this is not the only case, as other studies have identified more genes subjected to the action of this type of selection. For example, early studies in *Mycobacterium marinum* identified a *ppe38* knockout strain unable to secrete

ESX5 dependent substrates involved in virulence [286]. This gene is part of a larger family, difficult to study with current short-read sequencing techniques, with some of its members linked to virulence and host-pathogen crosstalk. More recent studies using whole genome comparisons corroborated that strains across the MTBC and particularly the Beijing clade had naturally occurring insertions/deletions occurring in this gene [151, 287]. The mutant strains with *ppe38* inactivated had no capacity to secrete a large number of PE_PPE members leading to a higher virulence [151]. As Beijing strains are thought to be hypervirulent [288] this led to the hypothesis that *ppe38* is under selection in natural populations. With this idea in mind, Ates *et al.* identified strains across the world with likely *ppe38*-inactivating mutations [151]. First, the authors showed that the situation in *M. marinum ppe38*-deleted strains was paralleled by *M. tuberculosis* strains when different parts of the *ppe38*-related region were knocked out. Second, the authors showed that clinical strains of MTBC with putative inactivating mutations did not secrete the ESX-5-associated proteins. Third, they showed that virulent strains of the Beijing family had inactivation of *pp38* and thus impaired secretion of the PE_PPE protein cluster. These strains were also highly virulent in a mice model of infection, showing again that natural inactivation of *ppe38* leads to highly virulent strains *in vitro* and *in vivo*. Furthermore, the inactivating mutations are shared by a cluster of Beijing strains called “modern”, which is particularly successful across the globe. Another study that analyzed a large number of strains collected from patients over 2.5 years in Vietnam found that a gene belonging to the *esx* family, *esxW*, is under positive selection [153]. In this case, the authors identified a mutation common to all strains of a particular sublineage of Beijing strains. The mutation is also homoplastic, and was identified in other unrelated phylogenetic clades including L1 and L3 strains. Homoplasy in *M. tuberculosis* is very rare: only around 1-2% of all mutations appear more than once in the phylogeny of the MTBC [126]. Detecting positive selection in *M. tuberculosis* depends critically on the existence of these convergence episodes of homoplasy. At the same time, the number of homoplasies of a character largely depend on the strength of the selective

force. Thus, those residues evolving under strong selective forces are easier to spot than those masked by other phenomena or when selection is weak. Thus, it is not surprising that most residues described under positive selection are linked to drug resistance, arguably the strongest selective force that *M. tuberculosis* encounters. In comparison, it is much more difficult to find variants positively associated to virulence as this trait is not easily defined which, in turn, may cause spurious associations.

The results achieved in the two studies referenced above have been obtained by applying WGS to thousands of samples. As stated before in this discussion, the use of WGS data is improving the quality and the extent of TB research. Nevertheless, it is important to note that, after genome sequencing, subsequent analyses with WGS data must be performed with bioinformatic tools. So, strong computing skills are needed to analyze all the data generated in an accurate and proper manner. In this thesis, computing had a preeminent role as all the analyses and processes described here have been performed with bioinformatic tools and programming languages. For example, the WGS analysis pipeline described in Chapter 3 has been essential for managing and analyzing MTBC genomes. An initial pipeline developed by Iñaki Comas [36] was at the core of the current pipeline. Over this initial basis, several modules, such as duplicate read removal, refinement of indel detection, strict variant filtering (near indels and high density regions), and the resistance and typing prediction were added during this thesis. In addition, some processes such as annotation, fastq trimming, and multifasta reconstruction were updated to improve their performance and/or computing time. The current pipeline, including some extended functionalities developed by other members of our research group and not described in this thesis, is used in our daily analyses. In fact, routine analyses of tuberculosis samples derived from all the Comunitat Valenciana hospitals are performed in our laboratory, based on WGS and the application of our analysis pipeline. Recently, a proficiency study [unpublished data] performed at Dick Van Soolingen group compared different analysis pipelines from different Mycobacterial National Reference Laboratories. Our

pipeline performed among the best ones, as we implemented stricter criteria regarding initial sample filtering and variant calling. So, false positive variants are not likely to be included in our final results. The application of our pipeline has allowed us to construct a database of clinical samples representative of the global MTBC diversity, composed by ∼4,600 strains. Sharing this database has allowed us to establish collaborations with other laboratories [289, 290] and to participate in projects developed by other members of our research group. In addition to the analysis pipeline, most of the methodology described in the Materials and methods sections has been specifically developed for each of the projects included in this thesis. For example, the ascertainment that mutations in *phoR* were related to transmission and the evolution of dN/dS through time in Chapter 4 were methods newly devised for this project. Another example is the PDEG analysis explained in Chapter 6, which has allowed us to track expression changes of specific genes along the MTBC evolution. We are convinced that these methodologies can be exported to other organisms and projects because they are not TB-specific. Thus, we think that the contribution of this thesis to the scientific field is not limited to the results and conclusions shown here, but it also includes the methodology developed over this time.

In summary, we have studied the evolution of the MTBC in current and past environments, by taking advantage of new available technologies. As the eminent scientist Douglas Young stated in an opinion article before his retirement [291], the biology research field is moving forward, and the classical reductionist mode of investigation 'gene-by-gene' is now getting complemented with holistic approaches. These methods will allow us to draw a general picture of the problem and to fill the gaps in our knowledge about the disease. With this in mind, we have tried to contribute to the TB research field by highlighting the importance of the MTBC genetic diversity as a characteristic to be taken into account in order to fight this deadly pathogen.

Conclusions

- There is no measurable recombination in current MTBC strains, neither among the MTBC strains nor between them and *M. canettii*. However, recombination between the *M. canettii* clade and the ancestor of the MTBC group occurred in the past.
- The MTBC and *M. canettii* evolved in sympatry from a common genetic pool in Africa. Recombination and selective pressures on specific genes, most of them involved in pathogenic functions, have driven the divergence between both groups of bacteria.
- The *phoR* gene has been subjected to pervasive positive selection since the divergence between *M. canettii* and the MTBC until nowadays. It has been involved in host environment adaptation and transmission.
- The MTBC regulatory and PPI networks are not conserved, and have many of their nodes impacted by potential dysfunctional mutations.
- Computational predictive models derived from the clinical reference strain H37Rv show inaccurate and poor predictions.
- Essential proteins tend to occupy central nodes in the protein-rotein interaction network. Moreover, the structure of the interactome prevents the accumulation of dysfunctional mutations in the core of the network, maintaining its functionality.

- The MTBC transcriptional profiles are diverse, with each main clade having a well defined transcriptional space. However, point mutations can alter the transcriptomic profile of single strains, making them to separate clearly from their phylogenetic relatives.
- The main functional categories affected by expression differences in the MTBC clades are those related to nutrient uptake and to the macrophage environmental conditions. This reinforces the idea that the different lineages have adapted to different hosts, not only at the genetic level, but also at the transcriptome level.
- Point mutations can affect significantly the regulation of specific genes. Creation and disruption of Pribnow boxes results in patterns of differential gene expression.
- There is evolutionary convergence of the methylation patterns in the different phylogenetic groups as identical phenotypes (methyltransferase inactivations) are presumably produced by different genetic variations. However, there is not a clear impact of differential methylation on gene expression regulation in *in-vitro* conditions.

Bibliography

[1] World Health Organization. Global tuberculosis report 2018, 2018.

[2] Frieden TR, Fujiwara PI, Washko RM and Hamburg MA. Tuberculosis in new york city — turning the tide. *N Engl J Med*, 1995: **333**(4):229–233.

[3] Barry CE 3rd, Boshoff HI, Dartois V, Dick T, Ehrt S, Flynn J, Schnappinger D, Wilkinson RJ and Young D. The spectrum of latent tuberculosis: rethinking the biology and intervention strategies. *Nat Rev Microbiol*, 2009: **7**(12):845–855.

[4] Esmail H, Barry CE 3rd, Young DB and Wilkinson RJ. The ongoing challenge of latent tuberculosis. *Philos Trans R Soc Lond B Biol Sci*, 2014: **369**(1645):20130437.

[5] Lin PL and Flynn JL. The end of the binary era: revisiting the spectrum of tuberculosis. *J Immunol*, 2018: **201**(9):2541–2548.

[6] Pai M, Behr MA, Dowdy D, Dheda K, Divangahi M, Boehme CC, Ginsberg A, Swaminathan S, Spigelman M, Getahun H *et al.* Tuberculosis. *Nat Rev Dis Primers*, 2016: **2**:16076.

[7] World Health Organization. *Companion handbook to the WHO guidelines for the programmatic management of drug-resistant tuberculosis*. World Health Organization, Geneva, 2014.

[8] World Health Organization. *Guidelines on the management of latent tuberculosis infection*. World Health Organization, 2015.

[9] Dheda K, Barry CE and Maartens G. Tuberculosis. *Lancet*, 2016: **387**(10024):1211–1226.

[10] Testing & diagnosis — TB — CDC. `https://www.cdc.gov/tb/topic/testing/default.htm`, 2018. Accessed: 2019-1-25.

[11] Walzl G, McNerney R, du Plessis N, Bates M, McHugh TD, Chegou NN and Zumla A. Tuberculosis: advances and challenges in development of new diagnostics and biomarkers. *Lancet Infect Dis*, 2018: **18**(7):e199–e210.

[12] Doyle RM, Burgess C, Williams R, Gorton R, Booth H, Brown J, Bryant JM, Chan J, Creer D, Holdstock J *et al.* Direct whole-genome sequencing of sputum

accurately identifies drug-resistant *Mycobacterium tuberculosis* faster than MGIT culture sequencing. *J Clin Microbiol*, 2018: **56**(8):e00666–18.

[13] Nour-Neamatollahi A, Siadat SD, Yari S, Tasbiti AH, Ebrahimzadeh N, Vaziri F, Fateh A, Ghazanfari M, Abdolrahimi F, Pourazar S *et al.* A new diagnostic tool for rapid and accurate detection of *Mycobacterium tuberculosis*. *Saudi J Biol Sci*, 2018: **25**(3):418–425.

[14] World Health Organization. Global strategy and targets for tuberculosis prevention, care and control after 2015. Technical Report EB134/12, 2013.

[15] Dye C, Glaziou P, Floyd K and Raviglione M. Prospects for tuberculosis elimination. *Annu Rev Public Health*, 2013: **34**:271–286.

[16] World Health Organization. The EndTB strategy. Online, 2015.

[17] Koch R. Die aetiologie der tuberkulose. *Mittbeihgen aus dem Kaiserlichen Gesundbeitsamte*, 1844: **2**:1–88.

[18] de Jong BC, Antonio M and Gagneux S. *Mycobacterium africanum*—review of an important cause of human tuberculosis in west africa. *PLoS Negl Trop Dis*, 2010: **4**(9):e744.

[19] Brites D, Loiseau C, Menardo F, Borrell S, Boniotti MB, Warren R, Dippenaar A, Parsons SDC, Beisel C, Behr MA *et al.* A new phylogenetic framework for the animal-adapted *Mycobacterium tuberculosis* complex. *Front Microbiol*, 2018: **9**.

[20] Gagneux S. Ecology and evolution of *Mycobacterium tuberculosis*. *Nat Rev Microbiol*, 2018: **16**(4):202–213.

[21] Fedrizzi T, Meehan CJ, Grottola A, Giacobazzi E, Fregni Serpini G, Tagliazucchi S, Fabio A, Bettua C, Bertorelli R, De Sanctis V *et al.* Genomic characterization of nontuberculous mycobacteria. *Sci Rep*, 2017: **7**:45258.

[22] Sidders B and Stoker NG. Mycobacteria: Biology. In John Wiley & Sons, Ltd, ed., *Encyclopedia of Life Sciences*, volume 99, page 3684. John Wiley & Sons, Ltd, Chichester, UK, 2001: .

[23] Tortoli E, Fedrizzi T, Meehan CJ, Trovato A, Grottola A, Giacobazzi E, Serpini GF, Tagliazucchi S, Fabio A, Bettua C *et al.* The new phylogeny of the genus *Mycobacterium*: the old and the news. *Infect Genet Evol*, 2017: **56**:19–25.

[24] Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P and Tiedje JM. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *Int J Syst Evol Microbiol*, 2007: **57**(Pt 1):81–91.

[25] Gupta RS, Lo B and Son J. Phylogenomics and comparative genomic studies robustly support division of the genus *Mycobacterium* into an emended genus Mycobacterium and four novel genera. *Front Microbiol*, 2018: **9**.

[26] Brosch R, Gordon SV, Marmiesse M, Brodin P, Buchrieser C, Eiglmeier K, Garnier T, Gutierrez C, Hewinson G, Kremer K *et al.* A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. *Proc Natl Acad Sci U S A*, 2002: **99**(6):3684–3689.

[27] Supply P, Marceau M, Mangenot S, Roche D, Rouanet C, Khanna V, Majlessi L, Criscuolo A, Tap J, Pawlik A *et al.* Genomic analysis of smooth tubercle bacilli provides insights into ancestry and pathoadaptation of *Mycobacterium tuberculosis*. *Nat Genet*, 2013: **45**(2):172–179.

[28] Koeck JL, L Koeck J, Fabre M, Simon F, Daffé M, Garnotel É, Matan AB, Gérôme P, J Bernatas J, Buisson Y *et al.* Clinical characteristics of the smooth tubercle bacilli '*Mycobacterium canettii*' infection suggest the existence of an environmental reservoir. *Clin Microbiol Infect*, 2011: **17**(7):1013–1019.

[29] Riojas MA, McGough KJ, Rider-Riojas CJ, Rastogi N and Hazbón MH. Phylogenomic analysis of the species of the *Mycobacterium tuberculosis* complex demonstrates that *Mycobacterium africanum*, *Mycobacterium bovis*, *Mycobacterium caprae*, *Mycobacterium microti* and *Mycobacterium pinnipedii* are later heterotypic synonyms of *Mycobacterium tuberculosis*. *Int J Syst Evol Microbiol*, 2018: **68**(1):324–332.

[30] Gagneux S, DeRiemer K, Van T, Kato-Maeda M, de Jong BC, Narayanan S, Nicol M, Niemann S, Kremer K, Gutierrez MC *et al.* Variable host-pathogen compatibility in *Mycobacterium tuberculosis*. *Proc Natl Acad Sci U S A*, 2006: **103**(8):2869–2873.

[31] Comas I, Homolka S, Niemann S and Gagneux S. Genotyping of genetically monomorphic bacteria: DNA sequencing in *Mycobacterium tuberculosis* highlights the limitations of current methodologies. *PLoS One*, 2009: **4**(11):e7815.

[32] Behr MA, Wilson MA, Gill WP, Salamon H, Schoolnik GK, Rane S and Small PM. Comparative genomics of BCG vaccines by whole-genome DNA microarray. *Science*, 1999: **284**(5419):1520–1523.

[33] Brosch R, Gordon SV, Garnier T, Eiglmeier K, Frigui W, Valenti P, Dos Santos S, Duthoy S, Lacroix C, Garcia-Pelayo C *et al.* Genome plasticity of BCG and impact on vaccine efficacy. *Proc Natl Acad Sci U S A*, 2007: **104**(13):5596–5601.

[34] Stucki D, Brites D, Jeljeli L, Coscolla M, Liu Q, Trauner A, Fenner L, Rutaihwa L, Borrell S, Luo T *et al.* *Mycobacterium tuberculosis* lineage 4 comprises globally distributed and geographically restricted sublineages. *Nat Genet*, 2016: **48**(12):1535–1543.

[35] Comas I, Hailu E, Kiros T, Bekele S, Mekonnen W, Gumi B, Tschopp R, Ameni G, Hewinson RG, Robertson BD *et al.* Population genomics of *Mycobacterium tuberculosis* in ethiopia contradicts the virgin soil hypothesis for human tuberculosis in Sub-Saharan africa. *Curr Biol*, 2015: **25**(24):3260–3266.

[36] Comas I, Coscolla M, Luo T, Borrell S, Holt KE, Kato-Maeda M, Parkhill J, Malla B, Berg S, Thwaites G *et al.* Out-of-Africa migration and neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans. *Nat Genet*, 2013: **45**(10):1176–1182.

[37] Brynildsrud OB, Pepperell CS, Suffys P, Grandjean L, Monteserin J, Debech N, Bohlin J, Alfsnes K, H Pettersson JO, Kirkeleite I *et al.* Global expansion of *Mycobacterium tuberculosis* lineage 4 shaped by colonial migration and local adaptation. *Science Advances*, 2018: **4**(10):eaat5869.

[38] Coscolla M and Gagneux S. Consequences of genomic diversity in *Mycobacterium tuberculosis*. *Semin Immunol*, 2014: **26**(6):431–444.

[39] Goig GA, Blanco S, Garcia-Basteiro A and Comas I. Pervasive contaminations in sequencing experiments are a major source of false genetic variability: *Mycobacterium tuberculosis* meta-analysis, 2018.

[40] Hershberg R, Lipatov M, Small PM, Sheffer H, Niemann S, Homolka S, Roach JC, Kremer K, Petrov DA, Feldman MW *et al.* High functional diversity in *Mycobacterium tuberculosis* driven by genetic drift and human demography. *PLoS Biol*, 2008: **6**(12):e311.

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

[42] de Jong BC, Hill PC, Aiken A, Awine T, Antonio M, Adetifa IM, Jackson-Sillah DJ, Fox A, DeRiemer K, Gagneux S *et al.* Progression to active tuberculosis, but not transmission, varies by *Mycobacterium tuberculosis* lineage in The Gambia. *J Infect Dis*, 2008: **198**(7):1037–1043.

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

[44] Di Pietrantonio T, Correa JA, Orlova M, Behr MA and Schurr E. Joint effects of host genetic background and mycobacterial pathogen on susceptibility to infection. *Infect Immun*, 2011: **79**(6):2372–2378.

[45] Berg S and Smith NH. Why doesn't bovine tuberculosis transmit between humans? *Trends Microbiol*, 2014: **22**(10):552–553.

[46] Reed MB, Pichler VK, McIntosh F, Mattia A, Fallow A, Masala S, Domenech P, Zwerling A, Thibert L, Menzies D *et al.* Major *Mycobacterium tuberculosis* lineages associate with patient country of origin. *J Clin Microbiol*, 2009: **47**(4):1119–1128.

[47] Fenner L, Egger M, Bodmer T, Furrer H, Ballif M, Battegay M, Helbling P, Fehr J, Gsponer T, Rieder HL *et al.* HIV infection disrupts the sympatric host-pathogen relationship in human tuberculosis. *PLoS Genet*, 2013: **9**(3):e1003318.

[48] Vandal OH, Nathan CF and Ehrt S. Acid resistance in *Mycobacterium tuberculosis*. *J Bacteriol*, 2009: **191**(15):4714–4721.

[49] Walpole GFW, Grinstein S and Westman J. The role of lipids in host-pathogen interactions. *IUBMB Life*, 2018: **70**(5):384–392.

[50] Delogu G, Sali M and Fadda G. The biology of *Mycobacterium tuberculosis* infection. *Mediterr J Hematol Infect Dis*, 2013: **5**(1):e2013070.

[51] Flynn JL, Chan J and Lin PL. Macrophages and control of granulomatous inflammation in tuberculosis. *Mucosal Immunol*, 2011: **4**(3):271–278.

[52] Getahun H, Matteelli A, Chaisson RE and Raviglione M. Latent *Mycobacterium tuberculosis* infection. *N Engl J Med*, 2015: **372**(22):2127–2135.

[53] Rosser A, Stover C, Pareek M and Mukamolova GV. Resuscitation-promoting factors are important determinants of the pathophysiology in *Mycobacterium tuberculosis* infection. *Crit Rev Microbiol*, 2017: **43**(5):621–630.

[54] Abel L, El-Baghdadi J, Bousfiha AA, Casanova JL and Schurr E. Human genetics of tuberculosis: a long and winding road. *Philos Trans R Soc Lond B Biol Sci*, 2014: **369**(1645).

[55] Khan N, Vidyarthi A, Nadeem S, Negi S, Nair G and Agrewala JN. Alteration in the gut microbiota provokes susceptibility to tuberculosis. *Front Immunol*, 2016: **7**.

[56] Bates MN, Khalakdina A, Pai M, Chang L, Lessa F and Smith KR. Risk of tuberculosis from exposure to tobacco smoke: a systematic review and meta-analysis. *Arch Intern Med*, 2007: **167**(4):335–342.

[57] Cegielski JP and McMurray DN. The relationship between malnutrition and tuberculosis: evidence from studies in humans and experimental animals. *Int J Tuberc Lung Dis*, 2004: **8**(3):286–298.

[58] Kim JH, Park JS, Cho YJ, Yoon HI, Song JH, Lee CT and Lee JH. Low serum 25-hydroxyvitamin D level: an independent risk factor for tuberculosis? *Clin Nutr*, 2014: **33**(6):1081–1086.

[59] Bastos HN, Osório NS, Gagneux S, Comas I and Saraiva M. The troika host–pathogen–extrinsic factors in tuberculosis: modulating inflammation and clinical outcomes. *Front Immunol*, 2017: **8**.

[60] Ong CWM, Elkington PT and Friedland JS. Tuberculosis, pulmonary cavitation, and matrix metalloproteinases. *Am J Respir Crit Care Med*, 2014: **190**(1):9–18.

[61] Saini D, Hopkins GW, Seay SA, Chen CJ, Perley CC, Click EM and Frothingham R. Ultra-low dose of *Mycobacterium tuberculosis* aerosol creates partial infection in mice. *Tuberculosis*, 2012: **92**(2):160.

[62] Dean GS, Rhodes SG, Coad M, Whelan AO, Cockle PJ, Clifford DJ, Glyn Hewinson R and Martin Vordermeier H. Minimum infective dose of *Mycobacterium bovis* in cattle. *Infect Immun*, 2005: **73**(10):6467.

[63] Stutz MD, Clark MP, Doerflinger M and Pellegrini M. *Mycobacterium tuberculosis*: rewiring host cell signaling to promote infection. *J Leukoc Biol*, 2017: **103**(2):259–268.

[64] Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry CE III *et al.* Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature*, 1998: **393**(6685):537.

[65] Coscolla M, Copin R, Sutherland J, Gehre F, de Jong B, Owolabi O, Mbayo G, Giardina F, Ernst JD and Gagneux S. *M. tuberculosis* T cell epitope analysis reveals paucity of antigenic variation and identifies rare variable TB antigens. *Cell Host Microbe*, 2015: **18**(5):538.

[66] Scott AN, Menzies D, Tannenbaum TN, Thibert L, Kozak R, Joseph L, Schwartzman K and Behr MA. Sensitivities and specificities of spoligotyping and mycobacterial interspersed repetitive unit-variable-number tandem repeat typing

methods for studying molecular epidemiology of tuberculosis. *J Clin Microbiol*, 2005: **43**(1):89–94.

[67] Niemann S and Supply P. Diversity and evolution of *Mycobacterium tuberculosis*: moving to whole-genome-based approaches. *Cold Spring Harb Perspect Med*, 2014: **4**(12):a021188–a021188.

[68] Schürch AC and van Soolingen D. DNA fingerprinting of *Mycobacterium tuberculosis*: from phage typing to whole-genome sequencing. *Infect Genet Evol*, 2012: **12**(4):602–609.

[69] Wyllie DH, Davidson JA, Grace Smith E, Rathod P, Crook DW, Peto TEA, Robinson E, Walker T and Campbell C. A quantitative evaluation of MIRU-VNTR typing against whole-genome sequencing for identifying *Mycobacterium tuberculosis* transmission: a prospective observational cohort study. *EBioMedicine*, 2018: **34**:122–130.

[70] Satta G, Lipman M, Smith GP, Arnold C, Kon OM and McHugh TD. *Mycobacterium tuberculosis* and whole-genome sequencing: how close are we to unleashing its full potential? *Clin Microbiol Infect*, 2018: **24**(6):604–609.

[71] CRyPTIC Consortium and the 100,000 Genomes Project, Allix-Béguec C, Arandjelovic I, Bi L, Beckert P, Bonnet M, Bradley P, Cabibbe AM, Cancino-Muñoz I, Caulfield MJ *et al.* Prediction of susceptibility to first-line tuberculosis drugs by DNA sequencing. *N Engl J Med*, 2018: **379**(15):1403–1415.

[72] Abdurakhmonov IY. Bioinformatics: basics, development, and future. In *Bioinformatics - ppdated features and applications*. 2016: .

[73] Zou J, Zheng MW, Li G and Su ZG. Advanced systems biology methods in drug discovery and translational biomedicine. *Biomed Res Int*, 2013: **2013**:742835.

[74] Iyengar R, Zhao S, Chung SW, Mager DE and Gallo JM. Merging systems biology with pharmacodynamics. *Sci Transl Med*, 2012: **4**(126):126ps7.

[75] Duncan Ayers PJD. Systems medicine: the application of systems biology approaches for modern medical research and drug development. *Mol Biol Int*, 2015: **2015**.

[76] Dutta NK, Bandyopadhyay N, Veeramani B, Lamichhane G, Karakousis PC and Bader JS. Systems biology-based identification of *Mycobacterium tuberculosis* persistence genes in mouse lungs. *MBio*, 2014: **5**(1).

[77] Pienaar E, Cilfone NA, Lin PL, Dartois V, Mattila JT, Russell Butler J, Flynn JL, Kirschner DE and Linderman JJ. A computational tool integrating host immunity with antibiotic dynamics to study tuberculosis treatment. *J Theor Biol*, 2015: **367**:166–179.

[78] Lalande L, Bourguignon L, Maire P and Goutelle S. Mathematical modeling and systems pharmacology of tuberculosis: Isoniazid as a case study. *J Theor Biol*, 2016: **399**:43–52.

[79] Peterson EJR, Ma S, Sherman DR and Baliga NS. Network analysis identifies Rv0324 and Rv0880 as regulators of bedaquiline tolerance in *Mycobacterium tuberculosis*. *Nat Microbiol*, 2016: **1**(8):16078.

[80] Sambarey A, Devaprasad A, Mohan A, Ahmed A, Nayak S, Swaminathan S, D'Souza G, Jesuraj A, Dhar C, Babu S *et al.* Unbiased identification of blood-based biomarkers for pulmonary tuberculosis by modeling and mining molecular interaction networks. *EBioMedicine*, 2017: **15**:112–126.

[81] Farrell D, Chubb AJ, Rue-Albrecht K, Malone K, Pirson C, Jones G, Gordon SV and Vordermeier M. Integrated computational prediction and experimental validation identifies promiscuous T cell epitopes in the proteome of *Mycobacterium bovis*. *Microbial Genomics*, 2016: **2**(8).

[82] Bos KI, Harkins KM, Herbig A, Coscolla M, Weber N, Comas I, Forrest SA, Bryant JM, Harris SR, Schuenemann VJ *et al.* Pre-Columbian mycobacterial genomes reveal seals as a source of new world human tuberculosis. *Nature*, 2014: **514**(7523):494–497.

[83] optimuscoprime. autoadapt. `https://github.com/optimuscoprime/autoadapt`.

[84] Schmieder R and Edwards R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, 2011: **27**(6):863–864.

[85] Chen S, Zhou Y, Chen Y and Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, 2018: **34**(17):i884–i890.

[86] Li H and Durbin R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics*, 2010: **26**(5):589–595.

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

[88] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R and 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and SAMtools. *Bioinformatics*, 2009: **25**(16):2078–2079.

[89] Picard tools - by broad institute. `http://broadinstitute.github.io/picard/`. Accessed: 2018-12-13.

[90] Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L and Wilson RK. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res*, 2012: **22**(3):568–576.

[91] Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X and Ruden DM. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly*, 2012: **6**(2):80–92.

[92] Vaser R, Adusumalli S, Leng SN, Sikic M and Ng PC. SIFT missense predictions for genomes. *Nat Protoc*, 2016: **11**(1):1–9.

[93] Coll F, McNerney R, Guerra-Assunção JA, Glynn JR, Perdigão J, Viveiros M, Portugal I, Pain A, Martin N and Clark TG. A robust SNP barcode for typing *Mycobacterium tuberculosis* complex strains. *Nat Commun*, 2014: **5**:4812.

[94] Feuerriegel S, Schleusener V, Beckert P, Kohl TA, Miotto P, Cirillo DM, Cabibbe AM, Niemann S and Fellenberg K. PhyResSE: a web tool delineating *Mycobacterium tuberculosis* antibiotic resistance and lineage from whole-genome sequencing data. *J Clin Microbiol*, 2015: **53**(6):1908–1914.

[95] Ezewudo M, Borens A, Chiner-Oms Á, Miotto P, Chindelevitch L, Starks AM, Hanna D, Liwski R, Zignol M, Gilpin C *et al.* Integrating standardized whole genome sequence analysis with a global *Mycobacterium tuberculosis* antibiotic resistance knowledgebase. *Sci Rep*, 2018: **8**(1):15382.

[96] Tange O. GNU parallel: The command-line power tool — USENIX. *The USENIX Magazine*, 2011: **36**:42–47.

[97] Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JGR, Korf I, Lapp H *et al.* The bioperl toolkit: Perl modules for the life sciences. *Genome Res*, 2002: **12**(10):1611–1618.

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

[99] Maere S, Heymans K and Kuiper M. BiNGO: a cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, 2005: **21**(16):3448–3449.

[100] Benjamini Y and Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J R Stat Soc Series B Stat Methodol*, 1995: **57**(1):289–300.

[101] R Core Team. R: The R project for statistical computing, 2004.

[102] RStudio Team. RStudio: Integrated development for R, 2015.

[103] Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, Bravo HC, Davis S, Gatto L, Girke T *et al.* Orchestrating high-throughput genomic analysis with bioconductor. *Nat Methods*, 2015: **12**(2):115–121.

[104] Microsoft Corporation and Steve Weston. doparallel: foreach parallel adaptor for the 'parallel' package, 2018.

[105] Microsoft Corporation and Steve Weston. foreach: Provides foreach looping construct for R, 2017.

[106] Wickham H. *ggplot2: elegant graphics for data analysis*. Springer, 2016.

[107] Love MI, Huber W and Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*, 2014: **15**(12):550.

[108] Csardi G and Nepusz T. The igraph software package for complex network research. *InterJournal*, 2006: **Complex Systems**:1695.

[109] Charif D and Lobry JR. SeqinR 1.0-2: A contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In *Biological and Medical Physics, Biomedical Engineering*, pages 207–232. 2007: .

[110] Vos M. A species concept for bacteria based on adaptive divergence. *Trends Microbiol*, 2011: **19**(1):1–7.

[111] Shapiro BJ. What microbial population genomics has taught us about speciation. In M Polz And O, ed., *Population Genomics*, pages 31–47. Springer, 2018: .

[112] Marttinen P and Hanage WP. Speciation trajectories in recombining bacterial species. *PLoS Comput Biol*, 2017: **13**(7):e1005640.

[113] Shapiro BJ and Polz MF. Microbial speciation. *Cold Spring Harb Perspect Biol*, 2015: **7**(10):a018143.

[114] Veyrier F, Pletzer D, Turenne C and Behr MA. Phylogenetic detection of horizontal gene transfer during the step-wise genesis of *Mycobacterium tuberculosis*. *BMC Evol Biol*, 2009: **9**(1):196.

[115] Aboubaker Osman D, Osman DA, Bouzid F, Canaan S and Drancourt M. Smooth tubercle bacilli: neglected opportunistic tropical pathogens. *Frontiers in Public Health*, 2016: **3**.

[116] Levillain F, Poquet Y, Mallet L, Mazères S, Marceau M, Brosch R, Bange FC, Supply P, Magalon A and Neyrolles O. Horizontal acquisition of a hypoxia-responsive molybdenum cofactor biosynthesis pathway contributed to *Mycobacterium tuberculosis* pathoadaptation. *PLoS Pathog*, 2017: **13**(11):e1006752.

[117] Brennan PJ. Bacterial evolution: emergence of virulence in TB. *Nat Microbiol*, 2016: **1**:15031.

[118] Boritsch EC, Frigui W, Cascioferro A, Malaga W, Etienne G, Laval F, Pawlik A, Le Chevalier F, Orgeur M, Ma L *et al.* *pks5*-recombination-mediated surface remodelling in *Mycobacterium tuberculosis* emergence. *Nature Microbiology*, 2016: **1**(2):15019.

[119] Boritsch EC, Khanna V, Pawlik A, Honoré N, Navas VH, Ma L, Bouchier C, Seemann T, Supply P, Stinear TP *et al.* Key experimental evidence of chromosomal DNA transfer among selected tuberculosis-causing mycobacteria. *Proceedings of the National Academy of Sciences*, 2016: **113**(35):9876–9881.

[120] Gagneux S and Small PM. Global phylogeography of *Mycobacterium tuberculosis* and implications for tuberculosis product development. *Lancet Infect Dis*, 2007: **7**(5):328–337.

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

[122] Blouin Y, Cazajous G, Dehan C, Soler C, Vong R, Hassan MO, Hauck Y, Boulais C, Andriamanantena D, Martinaud C *et al.* Progenitor "*Mycobacterium canettii*" clone responsible for lymph node tuberculosis epidemic, djibouti. *Emerg Infect Dis*, 2014: **20**(1):21–28.

[123] Liu X, Gutacker MM, Musser JM and X Fu Y. Evidence for recombination in *Mycobacterium tuberculosis*. *J Bacteriol*, 2006: **188**(23):8169–8177.

[124] Pepperell CS, Casto AM, Kitchen A, Granka JM, Cornejo OE, Holmes EC, Birren B, Galagan J and Feldman MW. The role of selection in shaping diversity of natural *M. tuberculosis* populations. *PLoS Pathog*, 2013: **9**(8):e1003543.

[125] Farhat MR, Jesse Shapiro B, Kieser KJ, Sultana R, Jacobson KR, Victor TC, Warren RM, Streicher EM, Calver A, Sloutsky A *et al.* Genomic analysis identifies targets of convergent positive selection in drug-resistant *Mycobacterium tuberculosis*. *Nat Genet*, 2013: **45**(10):1183–1189.

[126] Mortimer TD, Weber AM and Pepperell CS. Signatures of selection at drug resistance loci in *Mycobacterium tuberculosis*. *mSystems*, 2018: **3**(1).

[127] Cortes T, Schubert OT, Rose G, Arnvig KB, Comas I, Aebersold R and Young DB. Genome-wide mapping of transcriptional start sites defines an extensive leaderless transcriptome in *Mycobacterium tuberculosis*. *Cell Rep*, 2013: **5**(4):1121–1131.

[128] Hedrick P and Kumar S. Mutation and linkage disequilibrium in human mtDNA. *Eur J Hum Genet*, 2001: **9**(12):969–972.

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

[130] Kendall SL, Withers M, Soffair CN, Moreland NJ, Gurcha S, Sidders B, Frita R, Ten Bokum A, Besra GS, Lott JS *et al.* A highly conserved transcriptional repressor controls a large regulon involved in lipid degradation in *Mycobacterium smegmatis* and *Mycobacterium tuberculosis*. *Mol Microbiol*, 2007: **65**(3):684–699.

[131] Becq J, Gutierrez MC, Rosas-Magallanes V, Rauzier J, Gicquel B, Neyrolles O and Deschavanne P. Contribution of horizontally acquired genomic islands to the evolution of the tubercle bacilli. *Mol Biol Evol*, 2007: **24**(8):1861–1871.

[132] Tukey JW. *Exploratory data analysis*. Pearson College Division, 1977.

[133] Lew JM, Kapopoulou A, Jones LM and Cole ST. TubercuList – 10 years after. *Kekkaku*, 2011: **91**(1):1–7.

[134] Gonzalo-Asensio J, Mostowy S, Harders-Westerveen J, Huygen K, Hernández-Pando R, Thole J, Behr M, Gicquel B and Martín C. PhoP: a missing piece in the intricate puzzle of *Mycobacterium tuberculosis* virulence. *PLoS One*, 2008: **3**(10):e3496.

[135] Gonzalo-Asensio J, Malaga W, Pawlik A, Astarie-Dequeker C, Passemar C, Moreau F, Laval F, Daffe M, Martin C, Brosch R *et al.* Evolutionary history of tuberculosis shaped by conserved mutations in the PhoPR virulence regulator. *Proceedings of the National Academy of Sciences*, 2014: **111**(31):11491–11496.

[136] Guerra-Assunção JA, Crampin AC, Houben R, Mzembe T, Mallard K, Coll F, Khan P, Banda L, Chiwaya A, Pereira RPA *et al.* Large-scale whole genome sequencing of *M. tuberculosis* provides insights into transmission in a high prevalence area. *Elife*, 2015: **4**.

[137] Shapiro BJ, Friedman J, Cordero OX, Preheim SP, Timberlake SC, Szabó G, Polz MF and Alm EJ. Population genomics of early events in the ecological differentiation of bacteria. *Science*, 2012: **336**(6077):48–51.

[138] Shapiro BJ, Levade I, Kovacikova G, Taylor RK and Almagro-Moreno S. Origins of pandemic *Vibrio cholerae* from environmental gene pools. *Nat Microbiol*, 2016: **2**:16240.

[139] Baumler A and Fang FC. Host specificity of bacterial pathogens. *Cold Spring Harb Perspect Med*, 2013: **3**(12):a010041–a010041.

[140] McNally A, Thomson NR, Reuter S and Wren BW. 'add, stir and reduce': *Yersinia* spp. as model bacteria for pathogen evolution. *Nat Rev Microbiol*, 2016: **14**(3):177–190.

[141] Mortimer TD and Pepperell CS. Genomic signatures of distributive conjugal transfer among mycobacteria. *Genome Biol Evol*, 2014: **6**(9):2489–2500.

[142] Phelan JE, Coll F, Bergval I, Anthony RM, Warren R, Sampson SL, Gey van Pittius NC, Glynn JR, Crampin AC, Alves A *et al.* Recombination in pe/ppe genes contributes to genetic variation in *Mycobacterium tuberculosis* lineages. *BMC Genomics*, 2016: **17**:151.

[143] Cheruvu M, Plikaytis BB and Shinnick TM. The acid-induced operon Rv3083-Rv3089 is required for growth of *Mycobacterium tuberculosis* in macrophages. *Tuberculosis*, 2007: **87**(1):12–20.

[144] Olsen A, Chen Y, Ji Q, Zhu G, De Silva AD, Vilchèze C, Weisbrod T, Li W, Xu J, Larsen M *et al.* Targeting *Mycobacterium tuberculosis* tumor necrosis factor alpha-downregulating genes for the development of antituberculous vaccines. *MBio*, 2016: **7**(3).

[145] Casali N, White AM and Riley LW. Regulation of the *Mycobacterium tuberculosis mce1* operon. *J Bacteriol*, 2006: **188**(2):441–449.

[146] Shimono N, Morici L, Casali N, Cantrell S, Sidders B, Ehrt S and Riley LW. Hypervirulent mutant of *Mycobacterium tuberculosis* resulting from disruption of the *mce1* operon. *Proceedings of the National Academy of Sciences*, 2003: **100**(26):15918–15923.

[147] Broset E, Martín C and Gonzalo-Asensio J. Evolutionary landscape of the *Mycobacterium tuberculosis* complex from the viewpoint of PhoPR: implications for virulence regulation and application to vaccine development. *MBio*, 2015: **6**(5):e01289–15.

[148] Soto CY, Menendez MC, Perez E, Samper S, Gomez AB, Garcia MJ and Martin C. IS6110 mediates increased transcription of the *phoP* virulence gene in a multidrug-resistant clinical isolate responsible for tuberculosis outbreaks. *J Clin Microbiol*, 2004: **42**(1):212–219.

[149] Walters SB, Dubnau E, Kolesnikova I, Laval F, Daffe M and Smith I. The *Mycobacterium tuberculosis* PhoPR two-component system regulates genes essential for virulence and complex lipid biosynthesis. *Mol Microbiol*, 2006: **60**(2):312–330.

[150] Orgeur M and Brosch R. Evolution of virulence in the *Mycobacterium tuberculosis* complex. *Curr Opin Microbiol*, 2018: **41**:68–75.

[151] Ates LS, Dippenaar A, Ummels R, Piersma SR, van der Woude AD, van der Kuij K, Le Chevalier F, Mata-Espinosa D, Barrios-Payán J, Marquina-Castillo B *et al.* Mutations in *ppe38* block PE_PGRS secretion and increase virulence of *Mycobacterium tuberculosis*. *Nature Microbiology*, 2018: **3**(2):181.

[152] Malone KM, Rue-Albrecht K, Magee DA, Conlon K, Schubert OT, Nalpas NC, Browne JA, Smyth A, Gormley E, Aebersold R *et al.* Comparative 'omics analyses differentiate *Mycobacterium tuberculosis* and *Mycobacterium bovis* and reveal distinct macrophage responses to infection with the human and bovine tubercle bacilli. *Microb Genom*, 2018: .

[153] Holt KE, McAdam P, Thai PVK, Thuong NTT, Ha DTM, Lan NN, Lan NH, Nhu NTQ, Hai HT, Ha VTN *et al.* Frequent transmission of the *Mycobacterium tuberculosis* beijing lineage and positive selection for the EsxW beijing variant in vietnam. *Nat Genet*, 2018: **50**(6):849.

[154] Walker TM, Kohl TA, Omar SV, Hedge J, Del Ojo Elias C, Bradley P, Iqbal Z, Feuerriegel S, Niehaus KE, Wilson DJ *et al.* Whole-genome sequencing for prediction of *Mycobacterium tuberculosis* drug susceptibility and resistance: a retrospective cohort study. *Lancet Infect Dis*, 2015: **15**(10):1193–1202.

[155] Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 2006: **22**(21):2688–2690.

[156] Maddison WP and Maddison DR. Mesquite: a modular system for evolutionary analyses, 2017.

[157] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*, 2007: **81**(3):559–575.

[158] Darling AE, Mau B and Perna NT. progressivemauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One*, 2010: **5**(6):e11147.

[159] Ashkenazy H, Penn O, Doron-Faigenboim A, Cohen O, Cannarozzi G, Zomer O and Pupko T. FastML: a web server for probabilistic reconstruction of ancestral sequences. *Nucleic Acids Res*, 2012: **40**(Web Server issue):W580–4.

[160] Martin DP, Murrell B, Golden M, Khoosal A and Muhire B. RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus Evol*, 2015: **1**(1):vev003.

[161] Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, Parkhill J and Harris SR. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res*, 2015: **43**(3):e15.

[162] Schmidt HA, Strimmer K, Vingron M and von Haeseler A. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics*, 2002: **18**(3):502–504.

[163] Rambaut A, Suchard M, Xie D and Drummond. Tracer, 2016.

[164] Storey JD. The positive false discovery rate: a bayesian interpretation and the q -value. *Ann Stat*, 2003: **31**(6):2013–2035.

[165] Lechner M, Findeiss S, Steiner L, Marz M, Stadler PF and Prohaska SJ. Proteinortho: detection of (co-)orthologs in large-scale analysis. *BMC Bioinformatics*, 2011: **12**:124.

[166] Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Mol Syst Biol*, 2011: **7**:539.

[167] Ota T and Nei M. Variance and covariances of the numbers of synonymous and nonsynonymous substitutions per site. *Mol Biol Evol*, 1994: **11**(4):613–619.

[168] Murrell B, Moola S, Mabona A, Weighill T, Sheward D, Kosakovsky Pond SL and Scheffler K. FUBAR: a fast, unconstrained bayesian approximation for inferring selection. *Mol Biol Evol*, 2013: **30**(5):1196–1205.

[169] Murrell B, Weaver S, Smith MD, Wertheim JO, Murrell S, Aylward A, Eren K, Pollner T, Martin DP, Smith DM *et al.* Gene-wide identification of episodic selection. *Mol Biol Evol*, 2015: **32**(5):1365–1371.

[170] Pennell MW, Eastman JM, Slater GJ, Brown JW, Uyeda JC, FitzJohn RG, Alfaro ME and Harmon LJ. geiger v2.0: an expanded suite of methods for fitting macroevolutionary models to phylogenetic trees. *Bioinformatics*, 2014: **30**(15):2216–2218.

[171] Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A *et al.* The pfam protein families database: towards a more sustainable future. *Nucleic Acids Res*, 2016: **44**(D1):D279–85.

[172] Letunic I and Bork P. 20 years of the SMART protein domain annotation resource. *Nucleic Acids Res*, 2018: **46**(D1):D493–D496.

[173] Rustad TR, Minch KJ, Ma S, Winkler JK, Hobbs S, Hickey M, Brabant W, Turkarslan S, Price ND, Baliga NS *et al.* Mapping and manipulating the *Mycobacterium tuberculosis* transcriptome using a transcription factor overexpression-derived regulatory network. *Genome Biol*, 2014: **15**(11):502.

[174] Minch KJ, Rustad TR, Peterson EJR, Winkler J, Reiss DJ, Ma S, Hickey M, Brabant W, Morrison B, Turkarslan S *et al.* The DNA-binding network of *Mycobacterium tuberculosis*. *Nat Commun*, 2015: **6**:5829.

[175] Wang Y, Cui T, Zhang C, Yang M, Huang Y, Li W, Zhang L, Gao C, He Y, Li Y *et al.* Global protein-protein interaction network in the human pathogen *Mycobacterium tuberculosis* H37Rv. *J Proteome Res*, 2010: **9**(12):6665–6677.

[176] Hegde SR, Rajasingh H, Das C, Mande SS and Mande SC. Understanding communication signals during mycobacterial latency through predicted genome-wide protein interactions and boolean modeling. *PLoS One*, 2012: **7**(3):e33893.

[177] Liu ZP, Wang J, Qiu YQ, Leung RKK, Zhang XS, Tsui SKW and Chen L. Inferring a protein interaction map of *Mycobacterium tuberculosis* based on sequences and interologs. *BMC Bioinformatics*, 2012: **13 Suppl 7**:S6.

[178] Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP *et al.* STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res*, 2015: **43**(Database issue):D447–52.

[179] Turkarslan S, Peterson EJR, Rustad TR, Minch KJ, Reiss DJ, Morrison R, Ma S, Price ND, Sherman DR and Baliga NS. A comprehensive map of genome-wide gene regulation in *Mycobacterium tuberculosis*. *Sci Data*, 2015: **2**:150010.

[180] Mei S. In silico enhancing *M. tuberculosis* protein interaction networks in STRING to predict drug-resistance pathways and pharmacological risks. *J Proteome Res*, 2018: **17**(5):1749–1760.

[181] Galagan JE, Minch K, Peterson M, Lyubetskaya A, Azizi E, Sweet L, Gomes A, Rustad T, Dolganov G, Glotova I *et al.* The *Mycobacterium tuberculosis* regulatory network and hypoxia. *Nature*, 2013: **499**(7457):178–183.

[182] Rose G, Cortes T, Comas I, Coscolla M, Gagneux S and Young DB. Mapping of genotype-phenotype diversity among clinical isolates of *Mycobacterium tuberculosis* by sequence-based transcriptional profiling. *Genome Biol Evol*, 2013: **5**(10):1849.

[183] Dinan AM, Tong P, Lohan AJ, Conlon KM, Miranda-CasoLuengo AA, Malone KM, Gordon SV and Loftus BJ. Relaxed selection drives a noisy noncoding transcriptome in members of the *Mycobacterium tuberculosis* complex. *MBio*, 2014: **5**(4):e01169–14.

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

[185] Pérez E, Samper S, Bordas Y, Guilhot C, Gicquel B and Martín C. An essential role for *phoP* in *Mycobacterium tuberculosis* virulence. *Mol Microbiol*, 2001: **41**(1):179–187.

[186] Biofabri SL. Dose-defining safety and immunogenicity study of MTBVAC in south african neonates. Study report of clinical trial, online, 2018.

[187] Ofori-Anyinam B, Kanuteh F, Agbla SC, Adetifa I, Okoi C, Dolganov G, Schoolnik G, Secka O, Antonio M, de Jong BC *et al.* Impact of the *Mycobacterium africanum* West Africa 2 lineage on TB diagnostics in West Africa: decreased sensitivity of rapid identification tests in The Gambia. *PLoS Negl Trop Dis*, 2016: **10**(7):e0004801.

[188] Garcia-Alonso L, Jiménez-Almazán J, Carbonell-Caballero J, Vela-Boza A, Santoyo-López J, Antiñolo G and Dopazo J. The role of the interactome in the maintenance of deleterious variability in human populations. *Mol Syst Biol*, 2014: **10**:752.

[189] Junker BH and Schreiber F. *Analysis of biological networks*. John Wiley & Sons, 2011.

[190] Mostowy S, Onipede A, Gagneux S, Niemann S, Kremer K, Desmond EP, Kato-Maeda M and Behr M. Genomic analysis distinguishes *Mycobacterium africanum*. *J Clin Microbiol*, 2004: **42**(8):3594–3599.

[191] Gehre F, Kumar S, Kendall L, Ejo M, Secka O, Ofori-Anyinam B, Abatih E, Antonio M, Berkvens D and de Jong BC. A mycobacterial perspective on tuberculosis in West Africa: Significant geographical variation of *M. africanum* and other *M. tuberculosis* complex lineages. *PLoS Negl Trop Dis*, 2016: **10**(3):e0004408.

[192] Micklinghoff JC, Breitinger KJ, Schmidt M, Geffers R, Eikmanns BJ and Bange FC. Role of the transcriptional regulator RamB (Rv0465c) in the control of the glyoxylate cycle in *Mycobacterium tuberculosis*. *J Bacteriol*, 2009: **191**(23):7260–7269.

[193] Domenech P, Zou J, Averback A, Syed N, Curtis D, Donato S and Reed MB. Unique regulation of the DosR regulon in the beijing lineage of *Mycobacterium tuberculosis*. *J Bacteriol*, 2017: **199**(2).

[194] Solans L, Gonzalo-Asensio J, Sala C, Benjak A, Uplekar S, Rougemont J, Guilhot C, Malaga W, Martín C and Cole ST. The PhoP-dependent ncRNA Mcr7 modulates the TAT secretion system in *Mycobacterium tuberculosis*. *PLoS Pathog*, 2014: **10**(5):e1004183.

[195] Sassetti CM and Rubin EJ. Genetic requirements for mycobacterial survival during infection. *Proc Natl Acad Sci U S A*, 2003: **100**(22):12989–12994.

[196] Sassetti CM, Boyd DH and Rubin EJ. Genes required for mycobacterial growth defined by high density mutagenesis. *Mol Microbiol*, 2003: **48**(1):77–84.

[197] Pizzuti C and Rombo SE. Algorithms and tools for protein–protein interaction networks clustering, with a special focus on population-based stochastic methods. *Bioinformatics*, 2014: **30**(10):1343–1352.

[198] Malone JH and Oliver B. Microarrays, deep sequencing and the true measure of the transcriptome. *BMC Biol*, 2011: **9**:34.

[199] Blais A and Dynlacht BD. Constructing transcriptional regulatory networks. *Genes Dev*, 2005: **19**(13):1499–1511.

[200] Lloréns-Rico V, Cano J, Kamminga T, Gil R, Latorre A, Chen WH, Bork P, Glass JI, Serrano L and Lluch-Senar M. Bacterial antisense RNAs are mainly the product of transcriptional noise. *Sci Adv*, 2016: **2**(3):e1501363.

[201] Park D, Lee Y, Bhupindersingh G and Iyer VR. Widespread misinterpretable ChIP-seq bias in yeast. *PLoS One*, 2013: **8**(12):e83506.

[202] Arnvig KB, Comas I, Thomson NR, Houghton J, Boshoff HI, Croucher NJ, Rose G, Perkins TT, Parkhill J, Dougan G *et al.* Sequence-based analysis uncovers an abundance of non-coding RNA in the total transcriptome of *Mycobacterium tuberculosis*. *PLoS Pathog*, 2011: **7**(11):e1002342.

[203] Maier T, Güell M and Serrano L. Correlation of mRNA and protein in complex biological samples. *FEBS Lett*, 2009: **583**(24):3966–3973.

[204] Vogel C and Marcotte EM. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat Rev Genet*, 2012: **13**(4):227–232.

[205] Singh AK, Carette X, Potluri LP, Sharp JD, Xu R, Prisic S and Husson RN. Investigating essential gene function in *Mycobacterium tuberculosis* using an efficient CRISPR interference system. *Nucleic Acids Res*, 2016: **44**(18):e143.

[206] Zhang Y, Xiao Z, Zou Q, Fang J, Wang Q, Yang X and Gao N. Ribosome profiling reveals genome-wide cellular translational regulation upon heat stress in *Escherichia coli*. *Genomics Proteomics Bioinformatics*, 2017: **15**(5):324–330.

[207] Eastman G, Smircich P and Sotelo-Silveira JR. Following ribosome footprints to understand translation at a genome wide level. *Comput Struct Biotechnol J*, 2018: **16**:167–176.

[208] Faraway JJ. *Linear models with R*. CRC Press, 2016.

[209] Chai T and Draxler RR. Root mean square error (RMSE) or mean absolute error (MAE)? – arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, 2014: **7**(3):1247–1250.

[210] The Gene Ontology Consortium. Gene ontology consortium: going forward. *Nucleic Acids Res*, 2014: **43**(D1):D1049–D1056.

[211] Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ and Marra MA. Circos: an information aesthetic for comparative genomics. *Genome Res*, 2009: **19**(9):1639–1645.

[212] Pons P and Latapy M. Computing communities in large networks using random walks. *J Graph Algorithms Appl*, 2006: **10**(2):191–218.

[213] Palittapongarnpim P, Ajawatanawong P, Viratyosin W, Smittipat N, Disratthakit A, Mahasirimongkol S, Yanai H, Yamada N, Nedsuwan S, Imasanguan W *et al.* Evidence for host-bacterial co-evolution via genome sequence analysis of 480 thai *Mycobacterium tuberculosis* lineage 1 isolates. *Sci Rep*, 2018: **8**(1):11597.

[214] Hirsh AE, Tsolaki AG, DeRiemer K, Feldman MW and Small PM. Stable association between strains of *Mycobacterium tuberculosis* and their human host populations. *Proc Natl Acad Sci U S A*, 2004: **101**(14):4871–4876.

[215] Hershberg R and Petrov DA. Evidence that mutation is universally biased towards AT in bacteria. *PLoS Genet*, 2010: **6**(9):e1001115.

[216] Adhikari S and Curtis PD. DNA methyltransferases and epigenetic regulation in bacteria. *FEMS Microbiol Rev*, 2016: **40**(5):575–591.

[217] Davis BM, Chao MC and Waldor MK. Entering the era of bacterial epigenomics with single molecule real time DNA sequencing. *Curr Opin Microbiol*, 2013: **16**(2):192.

[218] Shell SS, Prestwich EG, Baek SH, Shah RR, Sassetti CM, Dedon PC and Fortune SM. DNA methylation impacts gene expression and ensures hypoxic survival of *Mycobacterium tuberculosis*. *PLoS Pathog*, 2013: **9**(7):e1003419.

[219] Arnvig K and Young D. Non-coding RNA and its potential role in *Mycobacterium tuberculosis* pathogenesis. *RNA Biol*, 2012: **9**(4):427–436.

[220] Borrell S, Trauner A, Brites D, Rigouts L, Loiseau C, Coscolla M, Niemann S, De Jong B, Yeboah-Manu D, Kato-Maeda M *et al.* Reference set of *Mycobacterium tuberculosis* clinical strains: a tool for research and product development, 2018.

[221] Chavadi SS, Stirrett KL, Edupuganti UR, Vergnolle O, Sadhanandan G, Marchiano E, Martin C, Qiu WG, Soll CE and Quadri LEN. Mutational and phylogenetic analyses of the mycobacterial *mbt* gene cluster. *J Bacteriol*, 2011: **193**(21):5905.

[222] Botella H, Peyron P, Levillain F, Poincloux R, Poquet Y, Brandli I, Wang C, Tailleux L, Tilleul S, Charrière GM *et al.* Mycobacterial P1-Type ATPases mediate resistance to zinc poisoning in human macrophages. *Cell Host Microbe*, 2011: **10**(3):248.

[223] Castell A, Johansson P, Unge T, Alwyn Jones T and Bäckbro K. Rv0216, a conserved hypothetical protein from *Mycobacterium tuberculosis* that is essential for bacterial survival during infection, has a double hotdog fold. *Protein Sci*, 2005: **14**(7):1850.

[224] Takayama K, Wang C and Besra GS. Pathway to synthesis and processing of mycolic acids in *Mycobacterium tuberculosis*. *Clin Microbiol Rev*, 2005: **18**(1):81.

[225] Ates LS, Dippenaar A, Sayes F, Pawlik A, Bouchier C, Ma L, Warren RM, Sougakoff W, Majlessi L, van Heijst JWJ *et al.* Unexpected genomic and phenotypic diversity of *mycobacterium africanum* lineage 5 affects drug resistance, protein secretion, and immunogenicity. *Genome Biol Evol*, 2018: **10**(8):1858.

[226] Otchere ID, Coscollá M, Sánchez-Busó L, Asante-Poku A, Brites D, Loiseau C, Meehan C, Osei-Wusu S, Forson A, Laryea C *et al.* Comparative genomics of *Mycobacterium africanum* lineage 5 and lineage 6 from Ghana suggests distinct ecological niches. *Sci Rep*, 2018: **8**(1):11269.

[227] Gupta A, Venkataraman B, Vasudevan M and Bankar KG. Publisher correction: Co-expression network analysis of toxin-antitoxin loci in *Mycobacterium tuberculosis* reveals key modulators of cellular stress. *Sci Rep*, 2018: **8**(1):7554.

[228] Vandal OH, Nathan CF and Ehrt S. Acid resistance in *Mycobacterium tuberculosis*. *J Bacteriol*, 2009: **191**(15):4714–4721.

[229] Heran Darwin K. *Mycobacterium tuberculosis* and copper: a newly appreciated defense against an old foe? *J Biol Chem*, 2015: **290**(31):18962.

[230] Keating LA, Wheeler PR, Mansoor H, Inwald JK, Dale J, Glyn Hewinson R and Gordon SV. The pyruvate requirement of some members of the *Mycobacterium tuberculosis* complex is due to an inactive pyruvate kinase: implications for in vivo growth. *Mol Microbiol*, 2005: **56**(1):163–174.

[231] Akhtar S, Khan A, Sohaskey CD, Jagannath C and Sarkar D. Nitrite reductase NirBD is induced and plays an important role during in vitro dormancy of *Mycobacterium tuberculosis*. *J Bacteriol*, 2013: **195**(20):4592–4599.

[232] Singh A, Gupta R, Vishwakarma RA, Narayanan PR, Paramasivan CN, Ramanathan VD and Tyagi AK. Requirement of the *mymA* operon for appropriate cell wall ultrastructure and persistence of *Mycobacterium tuberculosis* in the spleens of guinea pigs. *J Bacteriol*, 2005: **187**(12):4173–4186.

[233] Yruela I, Contreras-Moreira B, Magalhães C, Osório NS and Gonzalo-Asensio J. *Mycobacterium tuberculosis* complex exhibits lineage-specific variations affecting protein ductility and epitope recognition. *Genome Biol Evol*, 2016: **8**(12):3751–3764.

[234] Dawes SS, Warner DF, Tsenova L, Timm J, McKinney JD, Kaplan G, Rubin H and Mizrahi V. Ribonucleotide reduction in *Mycobacterium tuberculosis*: Function and expression of genes encoding class Ib and class II ribonucleotide reductases. *Infect Immun*, 2003: **71**(11):6124.

[235] Golby P, Hatch KA, Bacon J, Cooney R, Riley P, Allnutt J, Hinds J, Nunez J, Marsh PD, Hewinson RG *et al.* Comparative transcriptomics reveals key gene expression differences between the human and bovine pathogens of the *Mycobacterium tuberculosis* complex. *Microbiology*, 2007: **153**(Pt 10):3323–3336.

[236] Williams MJ, Kana BD and Mizrahi V. Functional analysis of molybdopterin biosynthesis in mycobacteria identifies a fused molybdopterin synthase in *Mycobacterium tuberculosis*. *J Bacteriol*, 2011: **193**(1):98–106.

[237] Marjanovic O, Miyata T, Goodridge A, Kendall LV and Riley LW. Mce2 operon mutant strain of *Mycobacterium tuberculosis* is attenuated in C57BL/6 mice. *Tuberculosis*, 2010: **90**(1):50–56.

[238] Deretic V, Philipp W, Dhandayuthapani S, Mudd MH, Curcic R, Garbe T, Heym B, Via LE and Cole ST. *Mycobacterium tuberculosis* is a natural mutant with an inactivated oxidative-stress regulatory gene: implications for sensitivity to isoniazid. *Mol Microbiol*, 1995: **17**(5):889–900.

[239] Jena L, Waghmare P, Kashikar S, Kumar S and Harinath BC. Computational approach to understanding the mechanism of action of isoniazid, an anti-TB drug. *Int J Mycobacteriol*, 2014: **3**(4):276–282.

[240] Bretl DJ, He H, Demetriadou C, White MJ, Penoske RM, Salzman NH and Zahrt TC. MprA and DosR coregulate a *Mycobacterium tuberculosis* virulence operon encoding Rv1813c and Rv1812c. *Infect Immun*, 2012: **80**(9):3018.

[241] Chauhan R, Ravi J, Datta P, Chen T, Schnappinger D, Bassler KE, Balázsi G and Gennaro ML. Reconstruction and topological characterization of the sigma factor regulatory network of *Mycobacterium tuberculosis*. *Nat Commun*, 2016: **7**:11062.

[242] Rodionov DA and Gelfand MS. Identification of a bacterial regulatory system for ribonucleotide reductases by phylogenetic profiling. *Trends Genet*, 2005: **21**(7):385–389.

[243] Young DB, Comas I and de Carvalho LPS. Phylogenetic analysis of vitamin b12-related metabolism in *Mycobacterium tuberculosis*. *Frontiers in Molecular Biosciences*, 2015: **2**.

[244] Iona E, Pardini M, Mustazzolu A, Piccaro G, Nisini R, Fattorini L and Giannoni F. *Mycobacterium tuberculosis* gene expression at different stages of hypoxia-induced dormancy and upon resuscitation. *J Microbiol*, 2016: **54**(8):565–572.

[245] Zhu L, Zhong J, Jia X, Liu G, Kang Y, Dong M, Zhang X, Li Q, Yue L, Li C *et al.* Precision methylome characterization of *Mycobacterium tuberculosis* complex (MTBC) using PacBio single-molecule real-time (SMRT) technology. *Nucleic Acids Res*, 2016: **44**(2):730–743.

[246] Phelan J, de Sessions PF, Tientcheu L, Perdigao J, Machado D, Hasan R, Hasan Z, Bergval IL, Anthony R, McNerney R *et al.* Methylation in *Mycobacterium tuberculosis* is lineage specific with associated mutations present globally. *Sci Rep*, 2018: **8**(1):160.

[247] Roberts RJ, Vincze T, Posfai J and Macelis D. REBASE: restriction enzymes and methyltransferases. *Nucleic Acids Res*, 2003: **31**(1):418–420.

[248] Chiner-Oms Á, González-Candelas F and Comas I. Gene expression models based on a reference laboratory strain are poor predictors of *Mycobacterium tuberculosis* complex transcriptional diversity. *Sci Rep*, 2018: **8**(1):3813.

[249] Sachdeva P, Misra R, Tyagi AK and Singh Y. The sigma factors of *Mycobacterium tuberculosis*: regulation of the regulators. *FEBS J*, 2010: **277**(3):605–626.

[250] Marreiros BC, Batista AP, Duarte AMS and Pereira MM. A missing link between complex I and group 4 membrane-bound [NiFe] hydrogenases. *Biochim Biophys Acta*, 2013: **1827**(2):198–209.

[251] Cook GM, Hards K, Vilchèze C, Hartman T and Berney M. Energetics of respiration and oxidative phosphorylation in mycobacteria. *Microbiol Spectr*, 2014: **2**(3).

[252] Botella H, Stadthagen G, Lugo-Villarino G, de Chastellier C and Neyrolles O. Metallobiology of host-pathogen interactions: an intoxicating new insight. *Trends Microbiol*, 2012: **20**(3):106–112.

[253] Larsen MH, Biermann K, Tandberg S, and Jacobs H Jr. Genetic manipulation of *Mycobacterium tuberculosis*. *Curr Protoc Microbiol*, 2007: **6**(1):10A.2.1–10A.2.21.

[254] Andrews S. FastQC: A quality control tool for high throughput sequence data.

[255] Bolger AM, Lohse M and Usadel B. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, 2014: **30**(15):2114.

[256] Quinlan AR and Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 2010: **26**(6):841–842.

[257] Rice P, Longden I and Bleasby A. EMBOSS: the european molecular biology open software suite. *Trends Genet*, 2000: **16**(6):276–277.

[258] Berney M, Berney-Meyer L, Wong KW, Chen B, Chen M, Kim J, Wang J, Harris D, Parkhill J, Chan J *et al.* Essential roles of methionine and s-adenosylmethionine in the autarkic lifestyle of *Mycobacterium tuberculosis*. *Proc Natl Acad Sci U S A*, 2015: **112**(32):10008–10013.

[259] Jain P, Hsu T, Arai M, Biermann K, Thaler DS, Nguyen A, González PA, Tufariello JM, Kriakov J, Chen B *et al.* Specialized transduction designed for precise high-throughput unmarked deletions in *Mycobacterium tuberculosis*. *MBio*, 2014: **5**(3):e01245–14.

[260] SMRT analysis software - PacBio. https://www.pacb.com/products-and-services/analytical-software/smrt-analysis/. Accessed: 2018-12-14.

[261] Laing CR, Whiteside MD and Gannon VPJ. Pan-genome analyses of the species *Salmonella enterica*, and identification of genomic markers predictive for species, subspecies, and serovar. *Front Microbiol*, 2017: **8**.

[262] Didelot X, Bowden R, Street T, Golubchik T, Spencer C, McVean G, Sangal V, Anjum MF, Achtman M, Falush D *et al.* Recombination and population structure in *Salmonella enterica*. *PLoS Genet*, 2011: **7**(7):e1002191.

[263] Achtman M, Wain J, Weill FX, Nair S, Zhou Z, Sangal V, Krauland MG, Hale JL, Harbottle H, Uesbeck A *et al.* Multilocus sequence typing as a replacement for serotyping in *Salmonella enterica*. *PLoS Pathog*, 2012: **8**(6):e1002776.

[264] Sánchez-Busó L, Comas I, Jorques G and González-Candelas F. Recombination drives genome evolution in outbreak-related *Legionella pneumophila* isolates. *Nat Genet*, 2014: **46**(11):1205.

[265] Gomez-Valero L, Rusniok C and Buchrieser C. *Legionella pneumophila*: population genetics, phylogeny and genomics. *Infect Genet Evol*, 2009: **9**(5):727–739.

[266] Qin T, Zhang W, Liu W, Zhou H, Ren H, Shao Z, Lan R and Xu J. Population structure and minimum core genome typing of *Legionella pneumophila*. *Sci Rep*, 2016: **6**:21356.

[267] Gasch AP, Payseur BA and Pool JE. The power of natural variation for model organism biology. *Trends Genet*, 2016: **32**(3):147–154.

[268] Rehren G, Walters S, Fontan P, Smith I and Zárraga AM. Differential gene expression between *Mycobacterium bovis* and *Mycobacterium tuberculosis*. *Tuberculosis*, 2007: **87**(4):347.

[269] Eldholm V and Balloux F. Antimicrobial resistance in *Mycobacterium tuberculosis*: the odd one out. *Trends Microbiol*, 2016: **24**(8):637–648.

[270] Sánchez-Busó L, Golparian D, Corander J, Grad YH, Ohnishi M, Flemming R, Parkhill J, Bentley SD, Unemo M and Harris SR. Antimicrobial exposure in sexual networks drives divergent evolution in modern gonococci, 2018.

[271] Cancino-Muñoz I, Moreno-Molina M, Furió V, Goig GA, Torres-Puente M, Chiner-Oms A, Villamayor LM, Sanz F, Guna-Serrano MR and Comas I. Cryptic resistance mutations associated to misdiagnoses of multidrug-resistant tuberculosis. *J Infect Dis*, 2019: **In press**.

[272] Bloemberg GV, Gagneux S and Böttger EC. Acquired resistance to bedaquiline and delamanid in therapy for tuberculosis. *N Engl J Med*, 2015: **373**(20):1986.

[273] Deng X, den Bakker HC and Hendriksen RS. Genomic epidemiology: Whole-genome-sequencing-powered surveillance and outbreak investigation of foodborne bacterial pathogens. *Annu Rev Food Sci Technol*, 2016: **7**:353–374.

[274] Franz E, Gras LM and Dallman T. Significance of whole genome sequencing for surveillance, source attribution and microbial risk assessment of foodborne pathogens. *Current Opinion in Food Science*, 2016: **8**:74–79.

[275] Kim S, Cho H, Lee D and Webster MJ. Association between SNPs and gene expression in multiple regions of the human brain. *Transl Psychiatry*, 2012: **2**(5):e113.

[276] Shastry BS. SNPs: impact on gene function and phenotype. In A K, ed., *Single Nucleotide Polymorphisms*, volume 578 of *Methods in Molecular Biology™*, pages 3–22. Humana Press, Totowa, NJ, 2009: .

[277] Robledo D, Rubiolo JA, Cabaleiro S, Martínez P and Bouza C. Differential gene expression and SNP association between fast- and slow-growing turbot (*Scophthalmus maximus*). *Sci Rep*, 2017: **7**.

[278] Hammarlöf DL, Kröger C, Owen SV, Canals R, Lacharme-Lora L, Wenner N, Schager AE, Wells TJ, Henderson IR, Wigley P *et al.* Role of a single noncoding nucleotide in the evolution of an epidemic african clade of *Salmonella*. *Proc Natl Acad Sci U S A*, 2018: **115**(11):E2614–E2623.

[279] Fong SS, Joyce AR and Palsson BØ. Parallel adaptive evolution cultures of *Escherichia coli* lead to convergent growth phenotypes with different gene expression states. *Genome Res*, 2005: **15**(10):1365.

[280] Adhikari S and Curtis PD. DNA methyltransferases and epigenetic regulation in bacteria. *FEMS Microbiology reviews*, 2016: **40**(5):575–591.

[281] Monk JM, Koza A, Campodonico MA, Machado D, Seoane JM, Palsson BO, Herrgård MJ and Feist AM. Multi-omics quantification of species variation of *Escherichia coli* links molecular features with strain phenotypes. *Cell Syst*, 2016: **3**(3):238–251.e12.

[282] Ma S, Minch KJ, Rustad TR, Hobbs S, Zhou SL, Sherman DR and Price ND. Integrated modeling of gene regulatory and metabolic networks in *Mycobacterium tuberculosis*. *PLoS Comput Biol*, 2015: **11**(11):e1004543.

[283] Comas I and Gagneux S. A role for systems epidemiology in tuberculosis research. *Trends Microbiol*, 2011: **19**(10):492–500.

[284] Allen TM, Altfeld M, Yu XG, O'Sullivan KM, Lichterfeld M, Le Gall S, John M, Mothe BR, Lee PK, Kalife ET *et al.* Selection, transmission, and reversion of an antigen-processing cytotoxic T-lymphocyte escape mutation in human immunodeficiency virus type 1 infection. *J Virol*, 2004: **78**(13):7069–7078.

[285] Finlay BB and McFadden G. Anti-immunology: evasion of the host immune system by bacterial and viral pathogens. *Cell*, 2006: **124**(4):767–782.

[286] Dong D, Wang D, Li M, Wang H, Yu J, Wang C, Liu J and Gao Q. PPE38 modulates the innate immune response and is required for *Mycobacterium marinum* virulence. *Infect Immun*, 2012: **80**(1):43–54.

[287] McEvoy CRE, van Helden PD, Warren RM and Gey van Pittius NC. Evidence for a rapid rate of molecular evolution at the hypervariable and immunogenic *Mycobacterium tuberculosis* PPE38 gene region. *BMC Evol Biol*, 2009: **9**:237.

[288] Merker M, Blin C, Mona S, Duforet-Frebourg N, Lecher S, Willery E, Blum MGB, Rüsch-Gerdes S, Mokrousov I, Aleksic E *et al.* Evolutionary history and global spread of the *Mycobacterium tuberculosis* Beijing lineage. *Nat Genet*, 2015: **47**(3):242–249.

[289] Abascal E, Pérez-Lago L, Martínez-Lirola M, Chiner-Oms Á, Herranz M, Chaoui I, Comas I, El Messaoudi MD, Cárdenas JAG, Santantón S *et al.* Whole genome sequencing-based analysis of tuberculosis (TB) in migrants: rapid tools for cross-border surveillance and to distinguish between recent transmission in the host country and new importations. *Euro Surveill*, 2019: **24**(4).

[290] Herranz M, Pole I, Ozere I, Chiner-Oms Á, Martínez-Lirola M, Pérez-García F, Gijón P, Serrano MJR, Romero LC, Cuevas O *et al. Mycobacterium tuberculosis* acquires limited genetic diversity in prolonged infections, reactivations and transmissions involving multiple hosts. *Front Microbiol*, 2018: **8**.

[291] Four decades of tuberculosis. *The biomedical dcientist magazine of the IBMS*, 2017: pages 17–18.

Supplementary material

10.1 Tables

Gene	Gene phylogenycongruent with reference	Nearest species (BLAST)
Rv0007	Yes	
Rv0083	Yes	
Rv0150c	No	*M. canettii*
Rv0153c	No	*M. canettii*
Rv0154c	Yes	
Rv0161	Yes	
Rv0166	Yes	
Rv0178	Yes	
Rv0180c	Yes	
Rv0194	Yes	
Rv0197	Yes	
Rv0276	Yes	
Rv0338c	Yes	
Rv0356c	No	*M. canettii*
Rv0357c	Yes	
Rv0389	Yes	
Rv0393	No	*M. shinjukuense*
Rv0399c	No	*M. canettii*

Rv0405	No	*M. canettii*
Rv0438c	Yes	
Rv0524	No	*M. canettii*
Rv0538	Yes	
Rv0609	No	*M. canettii*
Rv0609A	No	*M. canettii*
Rv0620	No	*M. canettii*
Rv0630c	No	*M. canettii*
Rv0673	No	*M. canettii*
Rv0698	No	*M. canettii*
Rv0756c	Yes	
Rv0870c	No	*M. canettii*
Rv0871	No	*M. canettii*
Rv0873	Yes	
Rv0875c	Yes	
Rv0914c	No	*M. canettii*
Rv0936	Yes	
Rv0957	Yes	
Rv0971c	No	*M. canettii*
Rv0983	Yes	
Rv0987	Yes	
Rv1185c	No	*M. canettii*
Rv1187	No	*M. canettii*
Rv1192	Yes	
Rv1244	No	*M. canettii*
Rv1385	Yes	
Rv1443c	No	*M. canettii*
Rv1449c	Yes	
Rv1462	Yes	
Rv1554	Yes	
Rv1629	Yes	
Rv1631	No	*M. canettii*

Rv1657	Yes	
Rv1658	Yes	
Rv1736c	No	*M. canettii*
Rv1775	Yes	
Rv1817	No	*M. sp 3/86Rv*
Rv1899c	No	*M. canettii*
Rv1937	Yes	
Rv1951c	No	*M. canettii*
Rv1963c	No	*M. canettii*
Rv1990c	No	*M. chimaera*
Rv1992c	Yes	
Rv2015c	Yes	
Rv2016	No	*M. sp 3/86Rv*
Rv2017	No	*M. canettii*
Rv2021c	No	*M. canettii*
Rv2022c	No	*M. sp 3/86Rv*
Rv2048c	Yes	
Rv2064	No	*M. canettii*
Rv2125	Yes	
Rv2160A	No	*M. canettii*
Rv2463	No	*M. canettii*
Rv2464c	Yes	
Rv2465c	No	*M. sp. 3/86Rv*
Rv2515c	Yes	
Rv2528c	No	*M. canettii*
Rv2541	No	*M. canettii*
Rv2542	No	*M. shinjukuense*
Rv2552c	Yes	
Rv2774c	No	*M. canettii*
Rv2798c	No	*M. shinjukuense*
Rv2799	No	*M. shinjukuense*
Rv2800	No	*M. shinjukuense*

Rv2802c	No	*M. shinjukuense*
Rv2803	No	*M. shinjukuense*
Rv2804c	No	
Rv2806	Yes	
Rv2829c	No	*M. shinjukuense*
Rv2833c	Yes	
Rv3009c	Yes	
Rv3014c	Yes	
Rv3027c	Yes	
Rv3037c	No	*M. canettii*
Rv3273	Yes	
Rv3275c	Yes	
Rv3339c	No	*M. canettii*
Rv3384c	No	*M. shinjukuense*
Rv3385c	No	*M. shinjukuense*
Rv3423c	No	*M. canettii*
Rv3447c	Yes	
Rv3451	Yes	
Rv3464	No	*M. canettii*
Rv3465	No	*M. canettii*
Rv3466	No	*M. canettii*
Rv3467	No	*M. canettii*
Rv3468c	Yes	
Rv3522	No	*M. canettii*
Rv3534c	Yes	
Rv3559c	No	*M. canettii*
Rv3561	No	*M. canettii*
Rv3589	No	*M. sp 3/86Rv*
Rv3591c	No	*M. canettii*
Rv3593	No	*M. canettii*
Rv3777	Yes	
Rv3782	Yes	

Rv3785	Yes	
Rv3896c	No	*M. canettii*
Rv3897c	No	*M. canettii*
Rv3899c	No	*M. canettii*
Rv3900c	No	*M. kansasii*
Rv3901c	No	*M. canettii*

Table 10.1: Results of the phylogenetic comparison of genes having a significant accumulation of divSNPs

Strain	Lineage	Genotype
N0157	L1	W.T.
N0072	L1	W.T.
N0153	L1	W.T.
N0145	L2	W.T.
N0052	L2	W.T.
N0031	L2	W.T.
N0155	L2	W.T.
N0004	L3	G1816338A syn
N1274	L3	C1816370T A61V
N0054	L3	W.T.
N1216	L4	W.T.
N0136	L4	W.T.
N1283	L4	W.T.
N1063	L5	G1816848T E220D
N1272	L5	G1816848T E220D
N1176	L5	G1816848T E220D
N0091	L6	C1816587G syn; G1816848T E220D
N1202	L6	C1816587G syn; G1816848T E220D
N1177	L6	C1816587G syn; G1816848T E220D

Table 10.2: Mutations found in the pyruvate kinase gene *pykA* (Rv1617) in the cultured strains

Branch	Branch length	PDEG genes
L5	0.096	23
L6	0.089	42
Ancestor	0.003	27
L1	0.064	30
Modern	0.042	7
L4	0.023	13
L3&L2	0.007	13
L3	0.035	17
L2	0.017	14
Beijing	0.017	20

Table 10.3: Number of Phylogenetically Aware Differentially Expressed genes and the branch length for each of the main clades in the MTBC phylogeny. The phylogeny was calculated by using the ML algorithm, with 1,000 iteration bootstraping.

10.2 Figures

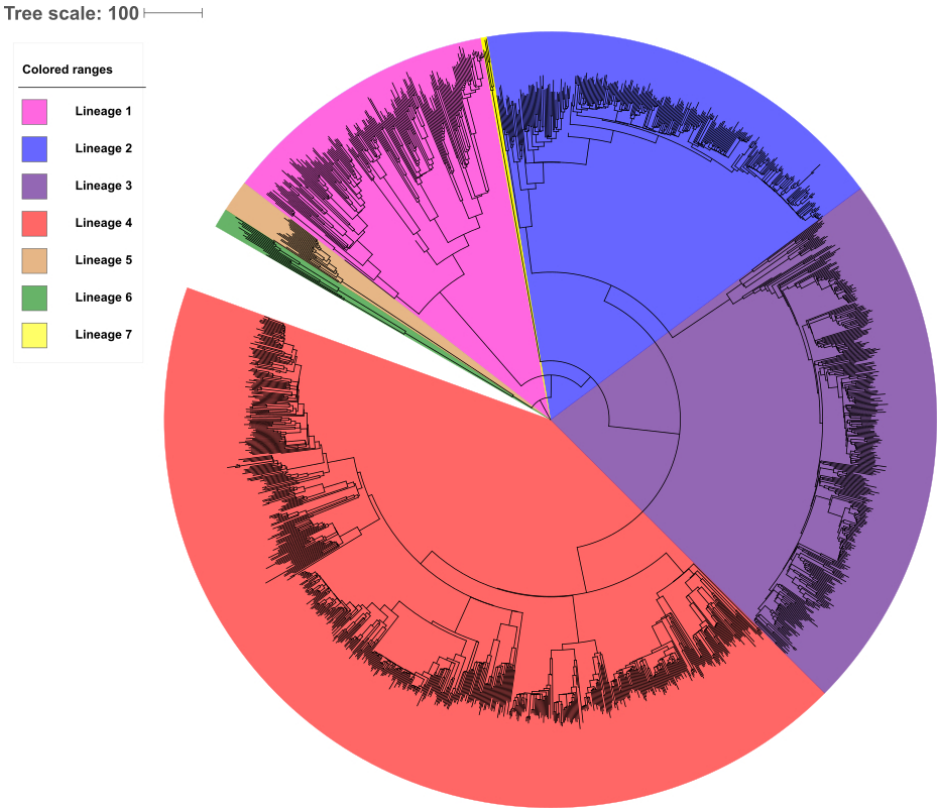


Figure 10.1: MTBC global phylogeny. The dataset used in the recombination analyses of Chapter 4 comprises 1,591 strains representatives of the global MTBC diversity. The phylogeny was constructed using both, ML and Neighbor-Joining methods. Congruent trees were obtained in both approaches. Tree scale in the figure refers to the number of genomic variants.

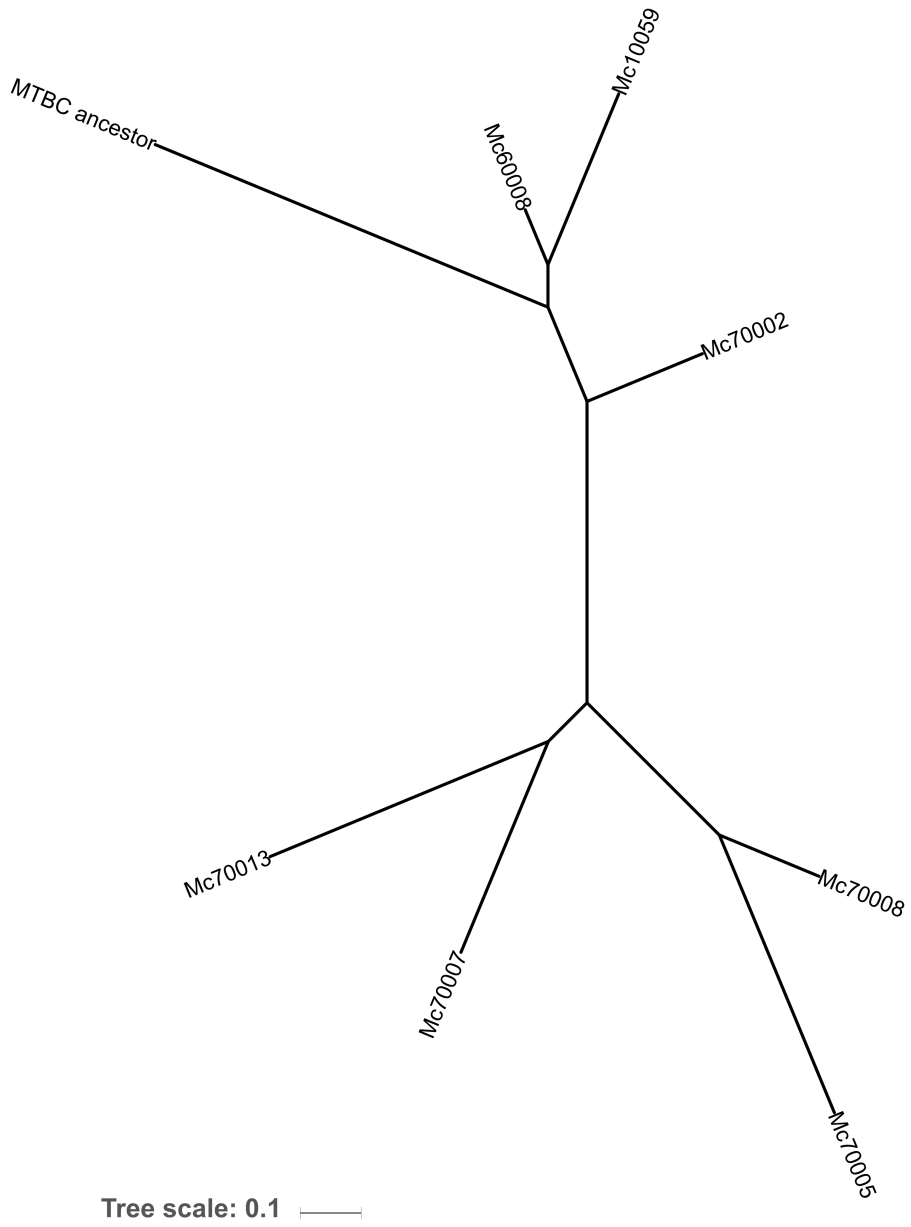


Figure 10.2: ML phylogeny of the MCAN group, including the MTBC most likely inferred ancestor.

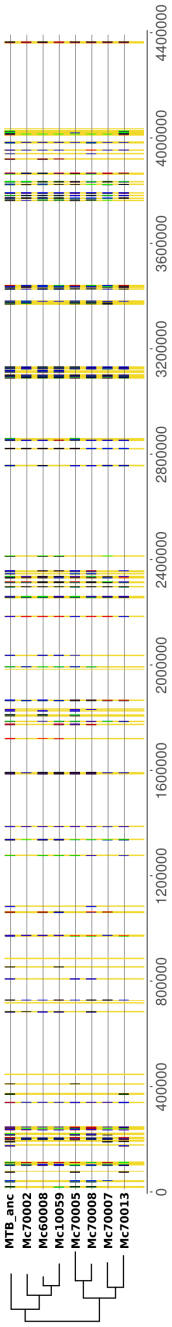


Figure 10.3: Homoplastic variants between MCAN and the MTBC ancestor. Homoplastic variant positions mapping to the branch of the most recent common ancestor of the MTBC coincide with events detected by Gubbins (highlighted in yellow) and show correlated phylogenetic patterns within each event.

Figure 10.4: Likelihood mapping analysis performed with TREEPUZZLE. For almost all of the fragments the phylogenetic signal was enough to perform a phylogenetic reconstruction for each fragment. Only fragments 13 and 14 had not enough phylogenetic signal to reconstruct a reliable phylogeny. The variants present in these regions were found only in the MTBC ancestor branch. This is congruent with a potential recombination with other organisms not present in our dataset.

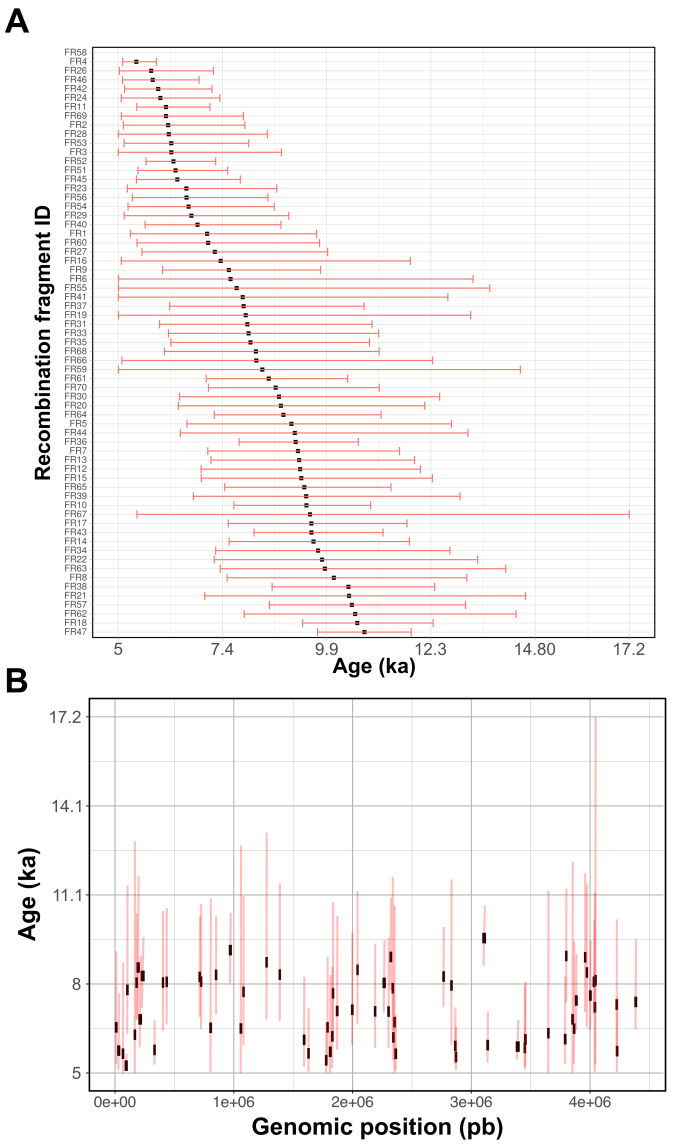


Figure 10.5: Recombination fragments ages derived from BEAST. A) Ages of the recombinant fragments (x-axis), sorted by age. B) Ages of the recombination fragments (y-axis) sorted by its genomic position (x-axis). In both panels, the red error bars represent the 95% highest probability density (HPD).

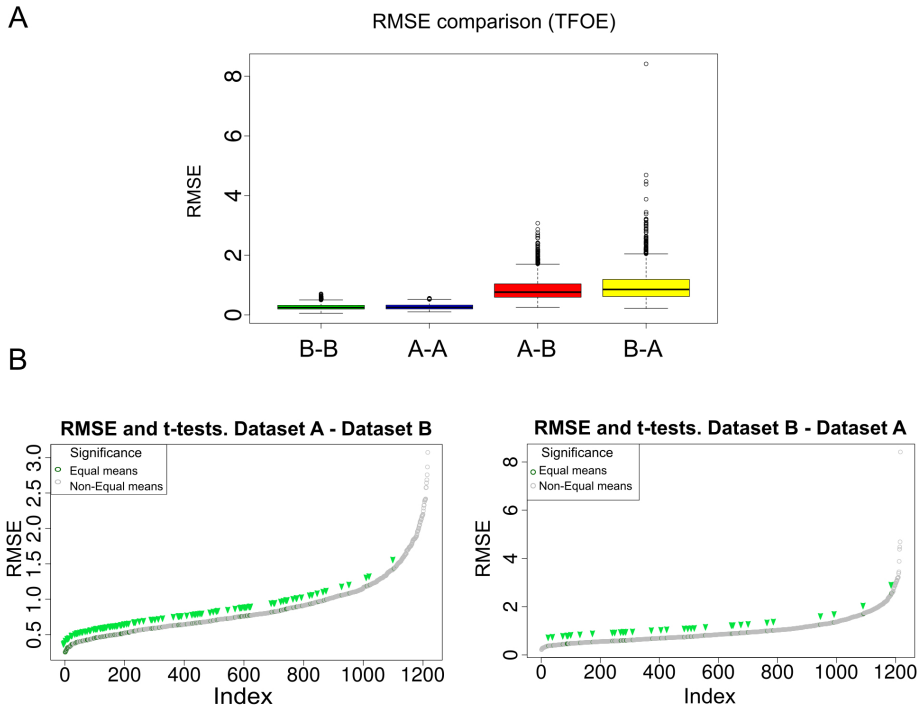
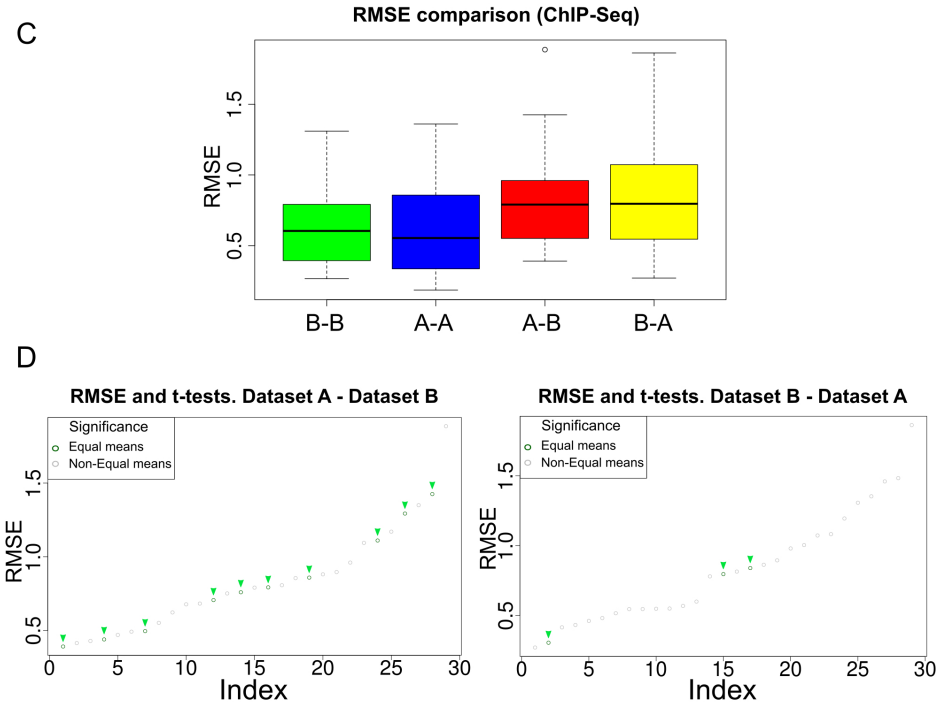


Figure 10.6: Evaluation of accuracy and comparison of model's behavior between different datasets. The main goal of the models is to make predictions under different conditions and with several data sources. Therefore, apart from training the models with the same dataset used to calculate the models and the regulatory networks we trained them with the other analogous dataset. In this figure, dataset A is the one obtained from Rustad *et al.*, and dataset B from Galagan *et al.* A) Root mean squared error comparison (RMSE) for the models obtained from TFOE data. Values when training and testing with dataset B (green), training and testing with dataset A (blue), training with dataset A and testing with dataset B (red) and training with dataset Band testing with dataset A (yellow). B) Plot showing RMSE values for TFOE derived models. Index refers to the list of models sorted by RMSE. The green arrows mark those models having no differences between predicted and measured mean expression. The left plot shows the case of training with dataset A and testing with dataset B while the right plot shows the reverse case. In the left plot, 128 genes show no differences between real and predicted values in terms of equaliy of means while in the right plot 33 genes show no statistical differences (pFDR $\leq$ 0.01).



C) RMSE comparison for the models obtained from ChIP-Seq data. Values when training and testing with dataset B (green), training and testing with dataset A (blue), training with dataset A and testing with dataset B (red) and training with dataset B and testing with dataset A (yellow). D) Plot showing RMSE values for ChIP-Seq derived models. Index refers to the list of models sorted by RMSE. The green arrows mark those models having no differences between predicted and measured mean expression. The left plot shows the case of training with dataset B and testing with dataset A while the right plot shows the reverse case. In the left plot, 10 genes show no differences between real and predicted values in terms of equality of means while in the right plot only 3 genes show no statistical differences (pFDR $\leq$ 0.01).

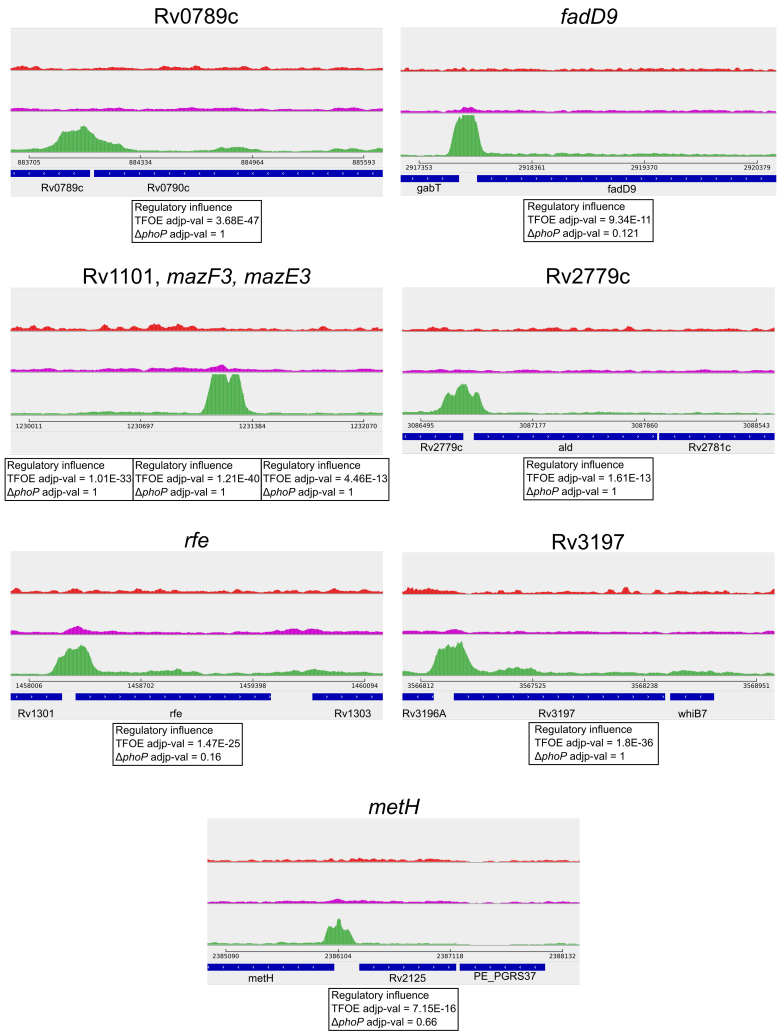


Figure 10.7: Detailed information about 9 selected genes showing no agreement in their *phoP* regulatory signal depending on the experiment used to detected it. Detail of the 9 genes selected for which regulation is affected in *phoP* overexpression experiments but not in *phoP* knockout experiments. The red track corresponds to the level of PhoP binding in the regulatory region of the genes in a knockout strain. The magenta track corresponds to the level of binding in the wild type strain. The green track corresponds to level of binding in the *phoP* overexpressed strain. The impact on downstream transcriptional levels are shown as published previously for the TFOE data and for the *phoP* knockout data.

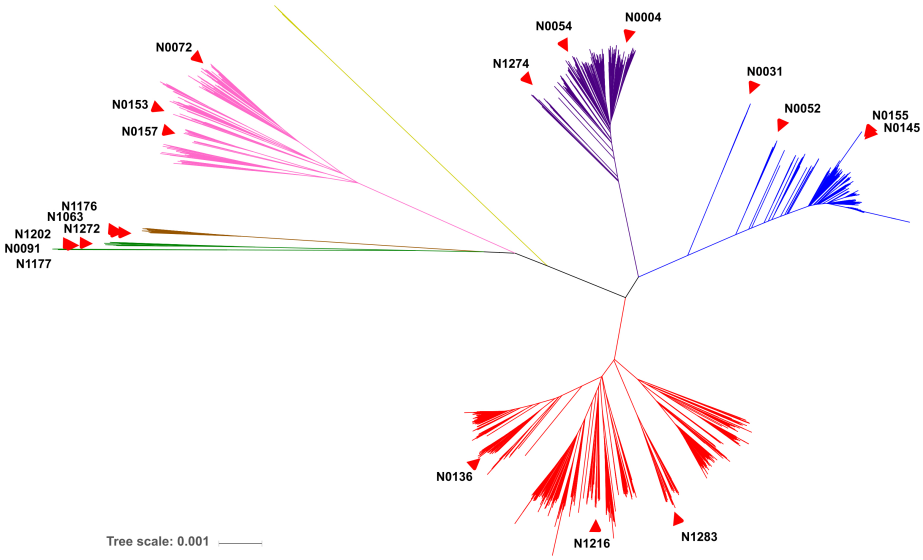


Figure 10.8: ML phylogeny, constructed with 4,595 strains representative of the global MTBC diversity. Red marks point to the 19 strains used for the main analyses in Chapter 6.

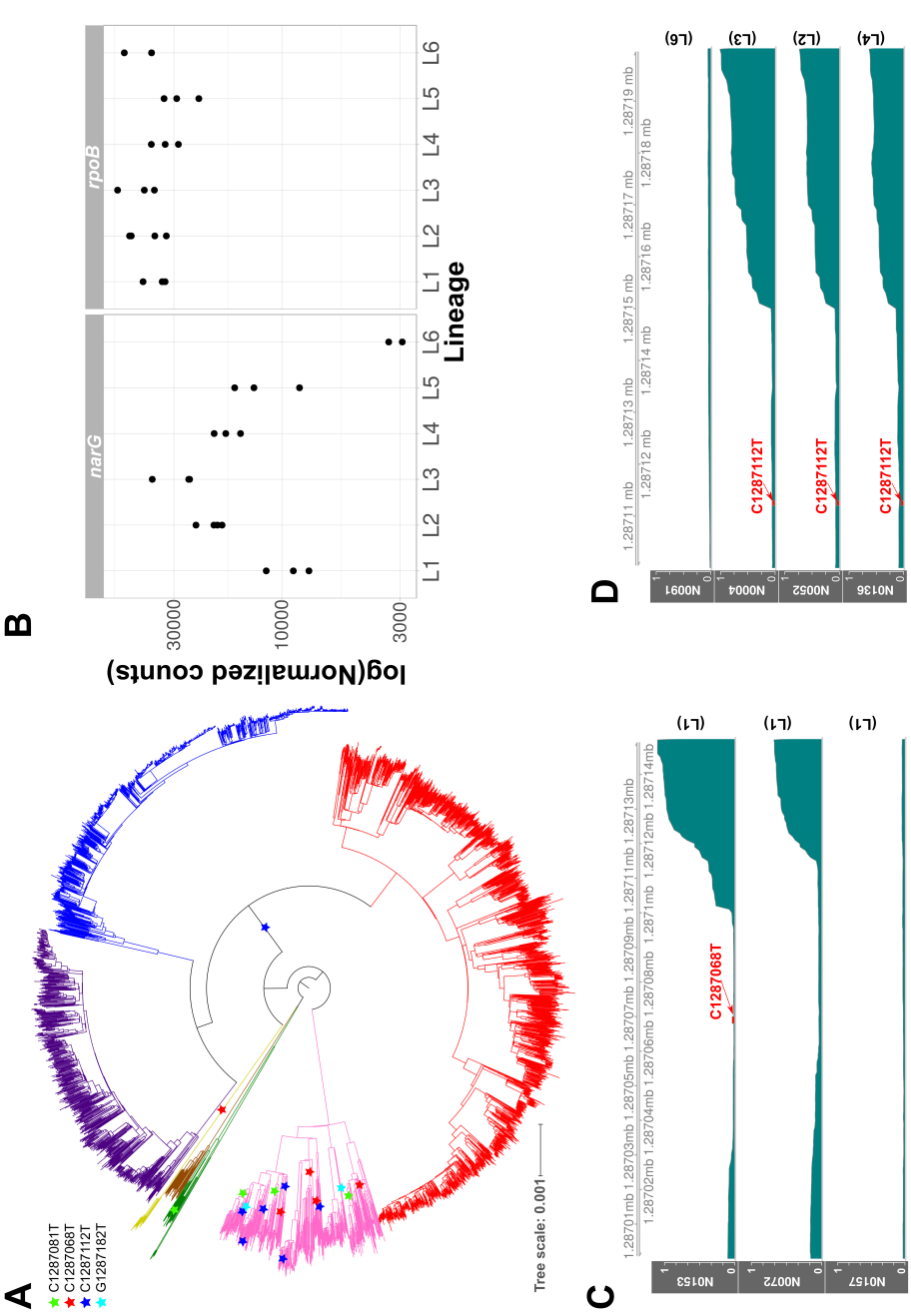


Figure 10.9: Overexpression of the *narG* operon due mutations found in its promoter region. A) The *narG* promoter mutations that create new Pribnow boxes are highly homoplastic. The phylogeny plotted was constructed by using the ML method, with the 4,595 strains dataset. B) Comparative between the expression values (y-axis, log of the normalized counts) of the *narG* gene and the housekeeping gene *rpoB* for the different lineages (x-axis). C) The new Pribnow box present in the N0153 strains upregulates the expression of *narG* in contrast to the other L1 strains. D) The new Pribnow box in the common branch of the modern lineages upregulates the transcription of *narG* in contrast to L6 strains.

10.3 External Data

Some of the supporting data generated is not suitable to be printed. It can be found in the following link:

http://tgu.ibv.csic.es/wp-content/uploads/2019/03/External-Data.zip.

External Data 1. Gubbins report for the MTBC-MCAN alignment. Each color block is a potential recombination region identified by Gubbins represented by its genomic position. Red blocks mark recombination events mapping in internal nodes, while blue blocks mark recombination events in terminal branches. All the recombinant blocks identified in the MTBC clade were common to all the strains and no recombinant traits were found in terminal branches. In contrast, the MCAN clade has a higher abundance of recombination events detected either in internal nodes or in terminal branches.

External Data 2.Potential recombination fragments detected between the MTBC ancestor and MCAN. The inferred dating for the 5ka and 70ka scenarios are included, as well as the genes contained in these sections.

External Data 3. Values for the Shimodaira-Hasegawa test (p-SH) between the recombinant fragments, the non-recombinant genomic fragment and the phylogenies derived from them.

External Data 4. Analysis of dN/dS variation between the MTBC ancestor and the MTBC. Only genes with at least 1 variant positions in each category were taken into account.

External Data 5. Variants found in the *phoR* gene.

External Data 6. Accession numbers and description of the strains analyzed in Chapter 5.

External Data 7. Accession numbers of the mycobacterial genomes used to construct the reference phylogeny.

External Data 8. Regulatory network obtained from statistically validated interactions. Each row correspond to an edge of the network. The edges are

directed the way TF $\rightarrow$ Target. The coefficient is a measure of the regulatory influence the TF has over the target gene. The sign of the interacions refers to a positive interaction (+) or to a negative interaction (-).

External Data 9. Mutations potentially affecting different TFs of the regulatory network Each row of the table corresponds to a mutation. Each row number agree with the number set in the External Data 10. Strain names came from Comas *et al.* 2013. The regulatory influence in the network refers to the number of genes regulated by the mutated TF in network derived from H37Rv network. The column principal GO terms overrepressented on the subnetwork came from the Gene Set Enrichment Analysis performed over the genes regulated by the TF A) Transcription factors missing in MTBC main lineages or sublineages. The RD column reffers to the Region of Difference (if any) associated to this TF's deletion. B) Transcription factors likely dysfunctional due to SNPs provoking stop-codons gain or loss. C) SNPs present in TF's regulatory regions.

External Data 10. Mutations affecting TFs in the MTBC phylogeny comprising the seven major lineages. The figure represents the number of TFs missing or potentially affected in their regulatory functions in one or more clinical strains from an MTBC reference dataset (n = 219 strains). The mutations affecting a TF are mapped to the corresponding internal/external node of the phylogeny. Each panel shows the same phylogeny and the mutations affecting a TF are mapped to the corresponding branch in the tree and highlighted in red. Label numbers correspond to entries in External Data 9. The mutations considered are either partial or complete deletions of the TF (A) (External Data 9A), single point mutations leading to gain or loss of stop codons (B) (External Data 9B) and single point mutations affecting the regulatory region of a TF (C) (External Data 9C).

External Data 11. PDEG Genes and Gene Ontology enrichment analysis A) Phylogenetically Aware Differentially Expressed genes for the comparison between the main phylogenetic groups. BaseMean, log2FoldChange and padj (BH) values were calculated with the DEseq2 package. The branch assignation

was obtained while comparing the branch again the rest of the MTBC. B) Gene Ontology enrichment analysis for the up and down-regulated genes for each branch of the PDGE analysis.

External Data 12. SNPs that either create or disrupt Pribnow boxes, and their potential effect over the expression of specific genes.

External Data 13. Nonsynonymous SNPs affecting the three main methyltransferases found in an MTBC dataset (n=4,595) representative of the global MTBC diversity.

External Data 14. Differential expression analysis performed with DEseq2, for N1283-△*hsdM* and N1283 strains

External Data 15. Accession numbers for the samples

10.4 Abbreviations

AIC Akaike Information Criterion

ANI Average Nucleotide Identity

BCG Bacillus Calmette-Guérin

BIC Bayesian Information Criterion

ChIP-Seq Chromatin ImmunoPrecipitation Sequencing

CRISPR Clustered Regularly Interspaced Short Palindromic Repeats

DE Differentially Expressed

DM Differentially Methylated

ENA European Nucleotide Archive

GEO Gene Expression Omnibus

GO Gene Ontology

GTR General Time Reversible

HC High Confident

LD Linkage Disequilibrium

MCAN *Mycobacterium canettii*

MDR Multi-Drug Resistant

MIRU Mycobacterial Interspersed Repetitive Units

ML Maximum Likelihood

MTBC *Mycobacterium tuberculosis* complex

PCA Principal Component Analysis

PDEG Phylogenetically aware Differentially Expressed Genes

FDR False Discovery Rate

PPI Protein-Protein Interaction

RD Region of Difference

RMSE Root Mean Squared Error

ROS Reactive Oxygen Species

SMRT Single Molecule Real Time

SNP Single Nucleotide Position

TB Tuberculosis

TFOE Transcription Factor Overexpression Experiment

TSS Transcription Start Site

VNTR Variable Number Tandem Repeats

WGS Whole-Genome Sequencing

WHO World Health Organization

XDR eXtensively-Drug Resistant

Resumen en castellano

Introducción

La tuberculosis

La tuberculosis es una enfermedad de transmisión aérea que afecta principalmente al sistema respiratorio. Según la OMS, es la principal causa de muerte por un único agente infeccioso, causando 1,6 millones de fallecimientos en 2017. La tuberculosis está especialmente presente en el continente africano, Asia y Sudamérica.

Clásicamente se ha clasificado la enfermedad en dos estadíos clínicos principales, tuberculosis latente, que es asintomática y no transmisible, y tuberculosis activa, que es transmisible y sintomática. Sin embargo, estudios recientes proponen un abanico muy amplio de estadíos clínicos, con la tuberculosis activa mostrando una combinación de síntomas desde leves a muy graves en distintos pacientes. Además, hay pacientes asintomáticos que también pueden estar afectados por la enfermedad a distintos niveles.

Los síntomas incluyen expectoración sanguinolenta con esputo, dolor de pecho, fiebre, sudores nocturnos y pérdida de peso. El tratamiento estándar para pacientes con tuberculosis activa consiste en un mínimo de 6 meses de terapia con 4 antibióticos combinados. Aunque este tratamiento se ha estado utilizando durante décadas, hoy en día hay un gran número de casos que no

responden al mismo. Por una parte, las cepas multidrogo-resistentes son aquellas que no responden, al menos, a dos de los más potentes antibióticos del tratamiento, isoniacida y rifampicina. Más difíciles de tratar son las cepas extremadamente resistentes que, además de los antibióticos anteriormente mencionados, son resistentes a varios antibióticos de segunda línea.

Con respecto a la detección, el diagnóstico preliminar de la tuberculosis se basa en dos tests rápidos pero poco precisos, que indican si el paciente ha tenido contacto con la bacteria en el pasado. Si dan un resultado positivo, al paciente se le somete a radiografías de pecho y, finalmente, a un cultivo a partir de muestras de esputo. Todo este proceso puede durar más de 3 semanas, y suelen obtenerse falsos negativos.

En 2014, la OMS comenzó una estrategia con el objetivo de eliminar la epidemia mundial de tuberculosis para 2035. Esta estrategia se basa en tres pilares principales: mejorar los protocolos de prevención y atención a pacientes, aumentar las acciones políticas y sistemas de soporte e intensificar el apoyo a la investigación. A pesar de las intenciones de la OMS, la incidencia global de la tuberculosis decrece a una tasa del 2% por año, insuficiente para cumplir el objetivo de erradicar la enfermedad en 2035. Para poder llegar a cumplir esa meta son necesarios nuevos métodos diagnósticos, estrategias para interrumpir la transmisión y tratamientos más efectivos. Todas estas mejoras necesitan de una fuerte inversión en investigación básica.

El complejo de *Mycobacterium tuberculosis*

La tuberculosis humana es causada principalmente por las bacteria *Mycobacterium tuberculosis* y *Mycobacterium africanum*, dos linajes bacterianos hermanos. Además, distintas especies de mamíferos pueden ser infectados por micobacterias que les provocan la enfermedad tuberculosa. Todas estas bacterias (*M. tuberculosis, M. africanum* y las micobacterias que infectan animales) forman un grupo monofilético llamado el complejo de *Mycobacterium tuberculosis* (MTBC). El MTBC tiene un único ancestro y sus

miembros comparten un 99% de su genoma, por lo que en realidad forman parte de una única especie pese a tener designaciones taxonómicas diferentes. En la presente tesis, nos hemos referido a las distintas cepas del MTBC por sus nombres clásicos, para facilitar la comprensión y legibilidad.

El MTBC tiene una estructura poblacional clonal, consistente en 7 linajes bacterianos adaptados a infectar humanos y varias cepas específicamente adaptadas para infectar animales. Los linajes 1,2,3,4 y 7 se clasifican como *M. tuberculosis sensu estricto* (MTB) mientras que los linajes 5 y 6 pertenecen a *M. africanum* (MAF). Los distintos linajes están distribuidos de forma heterogénea por el globo. Esto responde al hecho de que la bacteria se diversificó en los distintos linajes y ocupó distintas regiones geográficas siguiendo las migraciones y los cambios de las poblaciones humanas desde el neolítico hasta la actualidad.

Distintos estudios han mostrado que hay diferencias fenotípicas en cuanto a tasas de crecimiento, virulencia y capacidad de transmisión entre los distintos linajes. Esto indica que, pese a que el MTBC es poco diverso genéticamente hablando, la variabilidad genética presente tiene importantes implicaciones epidemiológicas y sobre la respuesta del hospedador a la enfermedad. La evolución paralela que se ha dado entre las poblaciones humanas y la bacteria, ha generado un cierto nivel de especialización del patógeno. Así, distintos linajes tienen preferencia por ciertos grupos de hospedadores en función de su origen filogeográfico.

El MTBC, un patógeno profesional

El proceso infeccioso comienza cuando la bacteria penetra en el cuerpo del hospedador a través del aparato respiratorio, donde es fagocitada por los macrofagos, desatando una respuesta inflamatoria leve y formando una estructura conocida como granuloma. La bacteria puede sobrevivir dentro del granuloma desde días hasta décadas en un estado asintomático llamado latencia. La transición de latencia a estado activo depende de factores

bacterianos, del hospedador y de la interacción entre ambos. Cuando se da esta transición, la bacteria comienza a replicarse dentro del granuloma hasta que este ya no es capaz de contener al patógeno y la bacteria es liberada. En ese punto, la bacteria causa necrosis y cavitación pulmonar, lo que provoca que el hospedador expectore. La bacteria aprovecha este hecho para diseminarse e infectar nuevos individuos.

La bacteria es capaz de interferir las defensas de los macrófagos cuando estos la fagocitan. Puede resistir el entorno ácido del fagosoma e interrumpir el proceso de maduración del mismo. La capacidad del bacilo para alterar el sistema inmunitario del hospedador es una de las razones por las que *M. tuberculosis* es considerado uno de los patógenos humanos más exitosos.

En resumen, el proceso por el que el bacilo provoca infección y se transmite es multifactorial, y todavía no conocemos todos los detalles del mismo. Un paso crucial para combatir al patógeno es entender de forma lo más completa posible sus características biológicas, las del hospedador y el entorno en que ambos interactúan. En la tesis que se presenta, nos hemos centrado en descifrar aspectos biológicos de la bacteria desconocidos hasta ahora y que son relevantes para su capacidad patogénica.

Características genómicas

El MTBC no tiene plásmidos, y su genoma es de 4,5 millones de bases. Tiene un contenido en GC elevado ($\sim$65%) y está muy conservado a nivel funcional. La máxima distancia genética entre cualquier par de cepas del MTBC son 2.500 SNPs. Además de SNPs, hay polimorfismos que afectan a regiones más grandes del genoma. Estos polimorfismos son regiones genómicas que están ausentes en algunas cepas del MTBC en comparación con la cepa de referencia H37Rv.

El impacto de las nuevas herramientas computacionales en la investigación de la tuberculosis

Actualmente, la secuenciación de genomas completos (WGS, por sus siglas en inglés) ha proveído a los investigadores con una gran cantidad de nuevos datos. Esto ha incrementado enormemente la resolución de los análisis que se venían haciendo hasta el momento, en múltiples áreas (evolución, epidemiología, estudio de mecanismos de resistencia, ...). La bioinformática y la biología computacional se han transformado en herramientas imprescindibles manejar los datos derivados de WGS. En consecuencia, los laboratorios y los equipos de trabajo se han transformado para dar cabida al material y los perfiles humanos necesarios para manejar estas nuevas herramientas. En ámbitos como la biología de sistemas, los equipos son multidisciplinares y aglutinan perfiles como matemáticos, biólogos, físicos, etc.

En el campo de la tuberculosis, estas disciplinas han producido importantes resultados. Por ejemplo, en la identificación de genes de persistencia, en el modelado de la dinámica del proceso infeccioso dentro del granuloma al aplicar antibióticos o en la farmacodinámica y farmacocinética de distintas drogas antituberculosas. Dado que la presente tesis se basa en el manejo de datos WGS, las metodologías de biología computacional y bioinformática tienen un papel predominante.

Motivación

En el pasado se han realizado algunos análisis sobre el proceso de diversificación del MTBC desde su ancestro común. Para estos análisis se utilizaron cientos de genomas, lo que contrasta con los miles de genomas que actualmente hay disponibles en las bases de datos. Se pretende usar toda la información disponible para entender en detalle, a nivel genómico y poblacional, el proceso de especiación del MTBC (capítulo 4). Además, pese a que la diversidad genética del MTBC es modesta en comparación con otras bacterias patógenas, tiene una importante implicación en el desenlace de la

enfermedad. Sin embargo, gran parte de las investigaciones publicadas no tienen en cuenta esta diversidad. Por tanto, comprobaremos si las conclusiones derivadas de algunas de estas investigaciones pueden ser generalizadas a todo el MTBC o, por el contrario, están sesgadas (capítulo 5). Por último, se plantea estudiar el papel que distintos mecanismos han jugado a lo largo de la historia evolutiva del MTBC para generar la diversidad fenotípica actual. Para ello, caracterizamos en detalle la diversidad transcripcional del MTBC como una medida de su diversidad fenotípica, y los procesos evolutivos que la han moldeado (capítulo 6).

Objetivos

La presente tesis pretende estudiar la evolución y las características biológicas del complejo de *Mycobacterium tuberculosis* usando herramientas computacionales y las nuevas técnicas de biología de sistemas. Utilizaremos secuencias de genomas completos como fuente de datos principal. Específicamente, perseguimos la consecución de los siguientes objetivos:

- Estudiar los distintos procesos evolutivos que han guiado la evolución del MTBC, desde un reservorio potencialmente ambiental a su nicho ecológico actual como patógeno obligado.
- Describir las modificaciones genéticas principales implicadas en la adaptación del MTBC a distintas especies de hospedadores.
- Auditar la capacidad de predicción de los modelos computacionales desarrollados hasta el momento, para realizar predicciones precisas en cualquier miembro del MTBC.
- Estudiar el impacto de la diversidad genética del MTBC sobre distintas redes biológicas de la bacteria.
- Analizar los perfiles transcriptómicos de distintos miembros del MTBC y los principales procesos evolutivos que han dado lugar a los patrones

regulatorios específicos de cada clado bacteriano.

Determinantes genómicos del proceso de especiación y expansión del MTBC

Introducción

Actualmente hay modelos teóricos que explican los procesos evolutivos por los cuales, a partir de una población natural de bacterias, pueden aparecer nuevas especies. La aparición de nuevas especies es más común en grupos de bacterias que comparten hábitat, en un proceso llamado especiación en simpatría.

Mycobacterium canettii (MCAN) es la bacteria más próxima al MTBC genéticamente hablando y solo se encuentra en una región geográfica pequeña del Cuerno de África. Precisamente en esa región es donde algunos estudios sitúan el origen del MTBC. Pese a no haber encontrado el reservorio natural de MCAN, diversas pistas apuntan a que es una bacteria ambiental que en ciertas circunstancias puede infectar humanos. Su gran parecido genético y el hecho de encontrarse en el mismo lugar donde se cree que se originó el MTBC hacen pensar que el MTBC y MCAN evolucionaron a partir de un mismo grupo de bacterias.

Nuestro entendimiento actual del proceso de especiación entre MCAN y el MTBC no es completo. En el presente capítulo, gracias al análisis de miles de secuencias del MTBC y de muestras de MCAN, identificamos las señales moleculares que dejó aquel proceso de especiación pasado, y revelamos nuevas dianas útiles para la investigación biomédica actual.

Resultados

Como primera aproximación para estudiar el proceso de diferenciación entre el MTBC y MCAN, buscamos señales de procesos de recombinación inter- e

intragrupo en ambas poblaciones actuales de bacterias. Para evaluar esta cuestión utilizamos un conjunto de 1.591 genomas representativos de la diversidad global del MTBC y distintas aproximaciones:

- Identificamos zonas con variantes homoplásicas (posiciones en las que un mismo alelo aparece de forma independiente en distintos linajes de la filogenia), y que compartan historia filogenética. Después de aplicar distintos filtros para asegurarnos de la robustez de los resultados, únicamente obtuvimos dos regiones que acumulan dos mutaciones homoplásicas cada una.
- Se utilizaron los estadísticos D' y R^2 para evaluar potenciales desequilibrios de ligamiento entre todas las posiciones polimórficas identificadas entre los 1.591 genomas (fueran homoplásicas o no). Ambos estadísticos coincidieron en asignar un papel mínimo o inexistente a los procesos de recombinación en la población actual del MTBC.
- Usamos los programas Gubbins y RDP4, sobre el mismo conjunto de datos. De nuevo, ninguno de los dos programas coincidió en identificar ninguna región genómica con fuerte señal recombinatoria.

Por tanto la recombinación, de estar presente, juega un impacto mínimo en la diversidad global del MTBC. A continuación, comparamos una selección de cepas del MTBC (n=219) contra 7 genomas de MCAN para identificar y cuantificar eventos de recombinación recientes entre ambos grupos. Para ello evaluamos el número de variantes homoplásicas entre ambos. De todas las homoplasias detectadas, el 98% estaban presentes en cepas de MCAN, lo que indica que la recombinación juega un papel muy importante en ese grupo. Sin embargo, no parecía haber eventos de recombinación actuales entre ambos grupos.

Especiación en simpatría del ancestor del MTBC

Hicimos un análisis con el programa Gubbins incluyendo los 7 genomas de

MCAN y el genoma del ancestro del MTBC. Identificamos un total de 65 eventos de recombinación entre ambos. Se realizó un test de congruencia filogenética con cada uno de estos fragmentos y comprobamos que, efectivamente, provenían de potenciales eventos de recombinación.

Analizamos la edad relativa de cada uno de los fragmentos con BEAST. Los resultados nos mostraron que el ancestro del MTBC se diferenció del grupo de MCAN de forma secuencial, acumulando eventos de recombinación a lo largo del tiempo. Aunque la poca variabilidad genética presente en los fragmentos impide establecer conclusiones firmes, los resultados sugieren que algunas regiones del genoma del ancestro acumularon eventos de recombinación antes que otras.

Estas regiones recombinantes además, están enriquecidas para genes esenciales (Chi-cuadrado, p-valor $<$ 0,01) lo que indica que la recombinación afectó a funciones celulares importantes. Además de esto, un análisis de enriquecimiento de funciones biológicas (términos de Gene Ontology) identificó funciones relacionadas con 'crecimiento dentro de células del hospedador' como sobrerrepresentadas en estas zonas. Es importante resaltar que muchos de los genes presentes en estas regiones están implicados en procesos de virulencia testados en animales.

Hay modelos teóricos que predicen que, durante el proceso de especiación simpátrica, hay partes concretas del genoma que acumulan mucha variabilidad al estar implicados en la adaptación a un nuevo nicho ecológico. Buscamos este tipo de variantes escaneando aquellas posiciones genómicas que mostraban un alelo en todas las cepas de MCAN y un alelo distinto en el ancestro del MTBC (divSNPs). Tras el análisis, identificamos 120 genes que acumulaban más divSNPs que las esperadas por azar.

Para comprobar a qué se debía la acumulación de divSNPs en estos genes los comparamos con bases de datos públicas y con genomas de otras micobacterias. Concluímos que de esos 120 genes, 53 son altamente divergentes entre ambos grupos debido a eventos de sustitución nucleotídica.

Además, estos están significativamente más conservados que el resto del genoma del MTBC. Esto indica que, después de divergir de MCAN, estos genes evolucionaron bajo presiones de selección purificadora.

Regiones bajo selección positiva después de la transición a patógeno obligado

Después buscar pistas sobre el proceso de diferenciación del MTBC y MCAN, quisimos encontrar genes relevantes para la 'profesionalización' y expansión global del nuevo patógeno. Primero nos centramos en la evolución de las proteínas antigénicas. Vimos como los epítopos de las células T estaban hiperconservados en la rama del ancestro del MTBC, lo que concuerda con otras observaciones anteriores en MCAN.

Buscamos luego regiones del genoma que si mostraran esa alteración. Calculamos el dN/dS en la rama del ancestro (usando los divSNPs) y el dN/dS actual del MTBC. Nos centramos en genes sujetos a selección purificadora antes del ancestro (dN/dS $<$ 1) y selección positiva después (dN/dS $>$ 1), y viceversa. Detectamos 14 genes en los que había un cambio de selección positiva a selección purificadora y 1 con cambio de selección purificadora en la rama del ancestro a selección positiva en las cepas actuales del MTBC. Ese gen es Rv0758, también conocido como *phoR*, y codifica para parte del sistema regulador PhoP/PhoR, implicado en virulencia.

Selección positiva en *phoR* ligada a presiones selectivas actuales

Dado que PhoPR es un sistema implicado en múltiples funciones de virulencia, centramos nuestra atención en las mutaciones encontradas en *phoR*. Ampliamos el número de cepas a analizar hasta 4.595, y encontramos 193 mutaciones no sinónimas y 31 sinónimas, lo que nos da un dN/dS de 2,37. Esto sugiere un fuerte efecto de la selección positiva. De hecho, comprobamos de forma retrospectiva cómo a lo largo del tiempo, desde el ancestro del MTBC hasta la actualidad, *phoR* ha estado sujeto a la acción de la selección positiva. Distintos tests de selección usando máxima verosimilitud nos permiten identificar, al menos, dos codones con una fuerte evidencia de estar bajo

selección positiva. Además, el hecho de encontrar 34 variantes homoplásicas apoya la idea de que el gen se encuentra evolucionando bajo una fuerte presión de selección. Las mutaciones no sinónimas se acumulan de manera estadísticamente significativa en la parte sensora del gen, lo que nos hace pensar que están relacionadas con la adaptación de la función sensora de la proteína a un ambiente cambiante (el hospedador) durante el proceso infeccioso.

Por último, para conocer la relevancia de las mutaciones de *phoR* en ambientes clínicos reales, analizamos las variantes del gen encontradas en un conjunto de cepas obtenidas en Malawi (un país con una alta tasa de transmisión de TB) durante 10 años. Encontramos 14 mutaciones nuevas en *phoR* exclusivas de este grupo de cepas. Además, al evaluar la edad relativa de las mutaciones no sinónimas de *phoR*, vimos que son significativamente más recientes que el resto de mutaciones no sinónimas del genoma. Finalmente, encontramos que mutaciones en *phoR* están significativamente sobrerrepresentadas en grupos de transmisión de mayor tamaño, en comparación con el resto de mutaciones no sinónimas en todo el genoma. Todos estos datos juntos nos indican que nuevas mutaciones en *phoR* están implicadas en la transmisión actual del MTBC en humanos.

Discusión

En este capítulo presentamos evidencia de que el MTBC comparte ancestralidad con el grupo de *M. canettii*, y que el ancestro del MTBC se separó del grupo bacteriano común especiando en simpatría. Durante las primeras etapas del proceso encontramos múltiples eventos de recombinación entre ambos grupos, lo que contrasta con la estructura poblacional actual del MTBC, que es prácticamente clonal. Además, hemos sido capaces de identificar regiones genómicas con diferentes presiones de selección antes y después del establecimiento del MTBC como un patógeno obligado.

Otras bacterias como *Vibrio cholerae* y especies de los géneros *Salmonella*

o *Yersinia* parecen haber especiado también en simpatría. Sin embargo, el MTBC parece un caso extremo de emergencia clonal. Pese a haber aumentado la resolución y contar con más de 1.500 genomas, no hemos encontrado ninguna prueba sobre eventos de recombinación recientes. Este cambio de perfil, de una bacteria altamente recombinogénica (el grupo ancestral del MTBC) a un organismo con herencia prácticamente clonal puede explicarse por dos causas no excluyentes: restricciones ecológicas y restricciones genéticas. Es cierto que no podemos descartar totalmente la presencia de recombinación, ya que la poca diversidad genética presente en las muestras limita la efectividad de las metodologías de detección. Además, hay regiones genómicas que no podemos evaluar con la tecnología de secuenciación de lecturas cortas. Sin embargo, parece claro que la recombinación, de estar presente, tiene un impacto mínimo en la diversidad genética del MTBC.

Los eventos de ancestrales de recombinación descritos implican a genes relevantes para las funciones patogénicas de la bacteria, como por ejemplo los operones *mymA* (esencial para el crecimiento dentro de macrófagos) y *mce1* (necesario desencadenar una la respuesta pro-inflamatoria adecuada para el desarrollo de la infección). Alternativamente, hemos identificado genes concretos que fueron relevantes en el proceso de adaptación del MTBC a su nuevo nicho ecológico.

Nuestros análisis nos han permitido identificar un gen, *phoR*, que está bajo selección positiva en cepas actuales del MTBC. Estudios previos ya han mostrado que:

- PhoPR es uno de los principales factores de virulencia del MTBC.
- Mutaciones antiguas en PhoPR están relacionadas con la adaptación del MTBC a nuevos hospedadores animales.

Hemos encontrado cambios aminoacídicos en codones que ya han sido propuestos como implicados en adaptación a hospedadores animales, en muestras clínicas de humanos. Esto nos lleva a pensar que las mutaciones en

phoR pueden estar implicadas en el ajuste de la respuesta inmunogénica del patógeno durante la infección, permitiéndole manipular la reacción del hospedador e incrementar las posibilidades de transmisión. Sin embargo, aun desconocemos los estímulos concretos que activan *phoR*, y que han sido la base de las presiones de selección de este gen durante su evolución.

En resumen, proponemos que la recombinación, así como la adquisición de nuevo material genético (demostrado en otros estudios), le permitieron al ancestro del MTBC especializarse como patógeno obligado de mamíferos. Esta asociación obligada al hospedador, se vió acompañada de nuevas presiones de selección, que actuaron sobre diversos genes. El hecho de que esos genes hoy en día evolucionen bajo selección purificadora sugiere que son importantes para la adaptación al nicho ecológico actual. Por último, en las últimas fases de expansión y especialización del MTBC a distintos hospedadores, encontramos trazas de selección positiva en distintos genes, *phoR* entre ellos.

Impacto de la diversidad global del MTBC en las redes biológicas de la bacteria

Introducción

En los últimos años se han publicado importantes artículos basados en el estudio de las redes biológicas de *M. tuberculosis*. Por ejemplo, se ha publicado una red de regulación de la bacteria basada en datos de ChIP-Seq y experimentos de sobreexpresión, así como múltiples modelos de redes de interacción de proteínas (redes PPI o interactomas). Algunos de estos modelos han permitido predecir el comportamiento de la bacteria en distintas condiciones ambientales.

Sin embargo, todas estas redes han sido construidas basándose en la cepa de referencia H37Rv (linaje 4). Esta cepa es una cepa clínica de referencia usada durante décadas en laboratorios de todo el mundo, y en muchos aspectos no representa la complejidad del MTBC. Así pues, mutaciones presentes en el

MTBC de forma natural, pueden cambiar la arquitectura de estas redes basadas en H37Rv.

Se sabe que hay mutaciones puntuales, presentes en clados principales del MTBC, que afectan a procesos regulatorios y tienen impacto sobre la virulencia del patógeno. De hecho, actualmente se está testando una vacuna (fase 2 de ensayos clínicos) basada en una deleción del factor de transcripción PhoP. Además, como se comentó en la introducción general, la diversidad genética de la bacteria tiene implicaciones epidemiológicas y para el desarrollo de la enfermedad en el hospedador. Desconocemos el efecto que esta diversidad pueda tener sobre la topología de las redes biológicas de la bacteria y sobre las predicciones que los modelos computacionales calculados en base a H37Rv ofrecen. Tampoco sabemos si estas predicciones pueden ser extrapolables a otras cepas del MTBC.

En este capítulo, construímos nuevos modelos de expresión basados en H37Rv, y comprobaremos su capacidad predictiva al introducir mutaciones presentes en cepas del MTBC. Además, estudiamos el impacto de la diversidad global del MTBC en la topología del interactoma, e identificamos nodos de la red claves para mantener su estructura.

Resultados

Construcción y validación de modelos de expresión génica basados en H37Rv

Basándonos en datos calculados a partir de experimentos de sobreexpresión (TFOE) sobre más de 200 factores de transcripción (TF), generamos modelos computacionales que nos permiten predecir el nivel de expresión de un gen a partir de los niveles de expresión de los TF que lo regulan. Los modelos predictivos se construyeron mediante regresión lineal, utilizando los datos derivados de los experimentos de TFOE. Los modelos calculados inicialmente fueron evaluados en un proceso de validación cruzada y frente a modelos calculados con datos aleatorizados. Esto hizo que, al final

del proceso, tuviéramos 1.216 modelos predictivos que superaran las pruebas de evaluación (30,8% de los modelos iniciales).

Para evaluar la robustez de las predicciones, comprobamos el funcionamiento de los modelos con un conjunto de datos de expresión distinto a los usados para construirlos y entrenarlos, aunque de nuevo basados en H37Rv. Si intentamos predecir el valor de expresión bruto (valor cuantitativo), obtenemos una correlación de 0,71 entre los datos predichos y los reales. Intentamos otra aproximación, usando un gen como referencia (*dnaA*), e intentamos predecir el nivel de cambio de expresión entre este gen y el resto de genes del genoma. En este caso, obtuvimos un coeficiente de correlación de 0,97 entre los datos reales y los datos predichos.

Red de regulación basada en interacciones estadísticamente validadas

Los 1.216 modelos obtenidos incluyen 11.253 relaciones de regulación. Algunas de estas relaciones están basadas en una señal regulatoria muy débil, por lo que seleccionamos únicamente las relaciones con una señal más fuerte (aquellas que, en los experimentos de TFOE, provocan una respuesta regulatoria 2 veces superior o inferior a la expresión inicial). Con estas relaciones creamos una nueva red de regulación, que incluye 1.102 genes y 3.396 relaciones. La distribución del parámetro de grado de la red indica que muchos TF regulan un número pequeño de genes mientras que un número bajo de TF regulan a muchos de ellos, siguiendo una ley de potencia.

En esta nueva red, los nodos Rv0023 y Rv0081 son los TF que regulan un número mayor de genes, lo que los convierte en nodos extremadamente importantes. En contraposición, Rv3202c es el gen que tiene un mayor número de TFs modulando su expresión. Este gen tiene actividad ATPasa y helicasa.

Los factores de transcripción no están globalmente conservados en el MTBC

Una vez calculamos los modelos de expresión y la nueva red de regulación, intentamos predecir el efecto fenotípico de la variación genética natural existente en cepas clínicas. Para ello examinamos el grado de conservación de

los TF estudiados en todo el MTBC. Buscamos mutaciones tanto en las regiones reguladoras de los TF como en las regiones codificantes de los mismos.

Usamos una colección de SNPS obtenidos de una publicación anterior. En esta colección, identificamos mutaciones con efecto potencial sobre la función génica. Por ejemplo, tenemos 15 TF con SNPs que introducen alteraciones en codones de parada. Además, 12 TF están delecionados en algunos linajes y sublinajes, ya que caen en conocidas regiones de diferencia (RD). Para cada una de los subconjuntos de genes regulados por estos factores de transcripción, hacemos un análisis de enriquecimiento de categorías funcionales. Un amplio espectro de funciones están representadas en estas redes como por ejemplo rutas metabólicas, respiración, patogenicidad y respuesta a estímulos externos.

Además de estas mutaciones, identificamos 117 SNPs en las zonas reguladoras de 44 TFs, algunos de ellos afectando a linajes y sublinajes completos. Algunos de estos SNPs, ya reportados en estudios anteriores, tienen efecto sobre la expresión del TF. Otras variantes, pese a no haber sido reportadas, pueden tener efecto potencial sobre la regulación de los TF. Por ejemplo, el cambio T89200G que afecta al gen Rv0081, que es uno de los reguladores más importantes de la bacteria. O la variante C2965900T, que se encuentra en la región reguladora del TF Rv2642 y es homoplásica.

Predicción *in-silico* de los niveles de expresión en cepas con distintos trasfondos genéticos

A continuación, nos interesa comprobar si los modelos construidos son capaces de predecir el impacto del trasfondo genético en el perfil transcripcional de la bacteria. Para ello seleccionamos una cepa del linaje 1 (T83) para la cual hay publicados datos de expresión, y en la que hemos identificado una deleción en el TF Rv1985c y un stop-codon en el TF Rv2788. En los modelos, reducimos al mínimo el nivel de expresión de Rv1985c y Rv2788 (imitando el potencial efecto de ambas mutaciones) y predecimos el efecto sobre la expresión de los genes regulados por ambos TF. Predecimos

diferencias significativas en la expresión de 148 genes. De ellos, solo 64 muestran diferencias de expresión en experimentos de RNA-seq reales, y obteniendo una correlación de 0,08 entre los niveles predichos por nosotros y los datos reales de esta cepa. No podemos obtener conclusiones definitivas a partir del análisis de una sola cepa, pero nuestros resultados parecen indicar que los modelos obtenidos únicamente con datos de H37Rv no pueden aplicarse para hacer predicciones en todo el MTBC.

Razonamos que, si no es posible predecir cambios de expresión en los distintos linajes, tal vez los modelos podrían ser aplicados en experimentos basados en H37Rv. Seleccionamos datos provenientes de un estudio en el que a la cepa H37Rv se le deleciona el TF *phoP*. De nuevo, seleccionamos todos los modelos en que *phoP* está presente como regresor y disminuimos su expresión al mínimo. Predecimos 188 genes en los que su expresión génica está significativamente alterada en el mutante. Al contrastar estos resultados con datos de RNA-seq obtenidos de otro estudio, solo encontramos 9 genes coincidentes. Además, para esos genes tampoco encontramos una correlación estadísticamente significativa entre la expresión predicha y la expresión medida.

En vista de estos resultados, evaluamos si la ausencia de correspondencia entre las predicciones y los datos experimentales se debía a que los modelos predictivos se han generado con datos de sobreexpresión de TFs mientras que los datos con los que intentamos contrastar las predicciones se han obtenido por la deleción de TFs. Comparamos los sitios genómicos a los que el factor PhoP se liga en el genoma en el mutante knock-out, en la cepa original H37Rv y en la cepa con *phoP* sobreexpresado. Vimos que, en los experimentos de sobreexpresión, el factor PhoP se liga a más sitios que en H37Rv, lo que hace pensar en ligamientos inespecíficos. Además, los datos de expresión de los TFOE indican que estos ligamientos inespecíficos generan respuesta transcripcional. Esto nos hace sospechar que el hecho de trabajar con distintas metodologías (sobreexpresión y deleción) puede ser causa de parte de las inexactitudes en las predicciones de los modelos.

Las proteínas esenciales tienden a ocupar lugares centrales en el interactoma

Al darnos cuenta de que la red de regulación no está conservada, nos preguntamos por el nivel de conservación de otras redes biológicas de la bacteria. Las proteínas, que son el producto final de los procesos de regulación, interactúan entre ellas para desarrollar un completo abanico de funciones biológicas. Para los análisis posteriores, elegimos descargarnos las PPI contenidas en la base de datos STRING, eligiendo aquellas con un índice de confianza mayor, para construir una red. En esta red, las proteínas son los nodos, y las interacciones entre ellas aparecen como conexiones aristas.

La rama matemática que estudia teoría de grafos dice que la importancia relativa de un nodo en una red viene dada por sus valores de centralidad. Cuantos más elevados sus valores, más importancia en términos de estabilidad y comunicación de la red. En las redes biológicas se ha propuesto que los nodos centrales son más relevantes para la funcionalidad de la red. Por otro lado, estudios previos han determinado la esencialidad de ciertas proteínas del MTBC para su supervivencia en distintas condiciones. Enfrentando ambos conceptos, encontramos diferencias estadísticamente significativas en la distribución de los valores de centralidad entre las proteínas esenciales y no esenciales. En base a esto, construimos un modelo para determinar la probabilidad de que una proteína sea esencial, en base a sus medidas de centralidad. Al aplicarlo, vemos que las proteínas esenciales tienen una probabilidad mayor de ser esenciales que las no esenciales, en base a sus medidas de centralidad. En concreto, si analizamos el modelo vemos que una alta probabilidad de ser esencial la tienen proteínas con valores altos de centralidad de grado, cercanía y de vector propio y con valores bajos de excentricidad y radialidad. Por tanto, parece claro que las proteínas esenciales tienden a ocupar puestos centrales en el interactoma.

Las proteínas centrales del interactoma tienden a acumular menos mutaciones que afectan a la función génica

Para comprobar el impacto de la diversidad genética presente en el MTBC, extrajimos todas las posiciones variables de las 4.598 cepas clínicas usadas en el capítulo 4 (n=235.254). Calculamos el potencial efecto sobre la función génica de estas mutaciones y las mapeamos sobre el interactoma. 280 de las proteínas presentes en la red tendían a acumular más mutaciones con potencial efecto disruptor de la función génica que de otro tipo.

Posteriormente creamos una versión simplificada del interactoma. Para ello definimos comunidades, que son grupos de proteínas más conectadas entre sí que con proteínas de otros grupos. Estas comunidades incluyen proteínas implicadas en procesos biológicos comunes. Definimos un valor de impacto para cada grupo en base al número de proteínas con potenciales mutaciones disfuncionales en la comunidad. Observamos que hay una correlación significativa entre los valores de centralidad de las comunidades y el valor de impacto, localizándose las comunidades más impactadas en la periferia del interactoma. Si en vez de poner todas las mutaciones juntas las segregamos por linaje, el patrón de comunidades impactadas es similar en los linajes que están filogenéticamente más relacionados.

Discusión

En este capítulo hemos visto cómo la diversidad genética presente de forma natural en el MTBC tiene impacto sobre las redes biológicas del patógeno. Nuestros resultados sugieren que el hecho de utilizar una cepa de referencia (H37Rv en este caso) para generar modelos complejos de redes puede generar conclusiones inexactas y no generalizables a todo el complejo.

Con respecto a la expresión génica, los modelos predictivos derivados de la red basada en H37Rv no difieren de modelos aleatorios en el 66,87% de los casos. Por otro lado, el 33,13% restante, tampoco ofrecen predicciones precisas, ni siquiera al aplicarlos a los mismos datos utilizados para generarlos y entrenarlos. Este hecho puede tener dos causas principales no excluyentes. Por una parte, el ruido introducido por las técnicas de cuantificación ha de ser

tenido en cuenta a la hora de crear estos modelos, principalmente en genes con un nivel de expresión bajo. Variaciones en estos niveles de expresión pueden deberse a procesos estocásticos y ser ruido de fondo. Además, la aplicación de estos modelos sobre datos generados en distintos estudios hace que no sea posible cuantificar de forma absoluta la expresión génica. Por otra parte, la red de regulación inferida es altamente dependiente de la metodología experimental utilizada para generarla. La sobreexpresión de factores de transcripción pueden introducir falsos positivos a la hora de cuantificar la regulación de la expresión debido a la aparición de inespecificidades.

Nuestros resultados muestran que los TFs testados en H37Rv no están conservados en el resto del MTBC. Mutaciones y deleciones afectan al 76% de las cepas circulantes. Las redes de regulación modeladas hasta este momento no tienen en cuenta la influencia potencial de estas variantes, ni tampoco el efecto de otras capas de regulación (ARN no codificante, metilación o modificaciones post-transcripcionales). Todo esto hace que, al aplicar los modelos a otras cepas distintas de H37RV, no obtengamos buenas predicciones.

Además de encontrar mutaciones que afectan a nodos importantes (TFs) de la red de regulación, también encontramos mutaciones impactando la red de interacción de proteínas. Hemos visto cómo las proteínas más importantes para el funcionamiento de la bacteria se sitúan en las zonas centrales del interactoma. Las proteínas tienden a agruparse en base a su función, y estas agrupaciones previenen la acumulación de mutaciones aleatorias con potencial efecto disfuncional en las zonas más centrales del interactoma. Sin embargo, esto hace también que el interactoma sea sensible de ser desestabilizado por 'ataques' dirigidos contra nodos específicos que sean ejes centrales de esta red. Además, hemos visto cómo distintos linajes del MTBC tienen distintos nodos impactados por estas mutaciones. Esto puede provocar que las interacciones entre las proteínas no sean las mismas en todas las cepas, ya que se pueden crear nuevas interacciones entre distintas proteínas para suplir aquellas afectadas por mutaciones disfuncionales. Si esto es así, deberíamos

generar, al menos, un modelo de interactoma para cada uno de los grandes grupos filogenéticos del MTBC, con ánimo de identificar proteínas o grupos de ellas que sean centrales y puedan ser potenciales dianas terapéuticas.

El papel de los mecanismos de mutación y metilación en la heterogeneidad transcripcional del MTBC

Introducción

Como se ha detallado en la introducción general, pese a la baja diversidad genética presente en el MTBC las diferencias entre los distintos linajes se traducen en características fenotípicas distintas. En el capítulo anterior hemos comprobado como hay variantes genéticas que impactan la red de regulación del complejo a distintos niveles filogenéticos, lo que sugiere la presencia de una variabilidad transcripcional elevada en el MTBC.

Dado que las características fenotípicas pueden venir dadas por diferencias en los mecanismos de regulación de la expresión, es de especial interés caracterizar el perfil transcripcional propio de cada uno de los linajes del MTBC. Hasta ahora, los estudios de expresión génica en el MTBC se han basado en el uso de microarrays o de secuenciación de cepas concretas. No hay estudios, a nivel de linaje, utilizando las herramientas de secuenciación del transcriptoma (RNA-seq). Así pues, en este capítulo final se han estudiado los perfiles transcripcionales de los principales grupos filogenéticos del MTBC usando datos de RNA-seq. Además, se ha caracterizado el impacto de dos mecanismos diferentes sobre la plasticidad transcripcional del complejo; mutaciones puntuales y patrones de metilación diferencial, ambos afectando a regiones reguladoras.

Resultados

Patrones generales de transcripción en el MTBC

Se seleccionaron 19 cepas, representantes de los linajes 1 a 6, para las cuales se extrajo y secuenciaron el ADN y el ARN. En un primer paso, se analizó el perfil transcripcional global del complejo a partir de los transcriptomas de cada una de las muestras. Todas las muestras pertenecientes a un mismo linaje filogenético mostraron un perfil transcripcional similar, con dos excepciones. Ambas excepciones debían su alteración del perfil a mutaciones puntuales en reguladores principales (los genes *dosR* y *rpoB*). Al ser congruentes las agrupaciones de los perfiles transcripcionales con la topología de la filogenia, investigamos el número de genes diferencialmente expresados (DE) en cada una de las ramas principales. Observamos un patrón, en el cual el número de genes DE es proporcional a la longitud de la rama. Esto sugiere que los cambios en los perfiles de transcripción se acumularon gradualmente, a medida que los linajes fueron divergiendo. Encontramos dos excepciones. La divergencia entre MTB y MAF viene definida por una distancia genética pequeña, pero comprende un gran número de genes DE. Por otro lado, en la rama común de todos los linajes modernos (linajes 2,3 y 4) encontramos justo la situación contraria.

Expresión diferencial entre principales clados del complejo

Después de analizar los perfiles transcriptómicos generales, quisimos ver las diferencias específicas entre cada uno de los linajes. Encontramos que los genes que codifican para sideróforos (sistemas necesarios para adquisición de hierro en ambientes limitados, como el macrófago) están sobreexpresadas en MTB frente a MAF; y genes relacionados con transporte de cationes metálicos están sobreexpresados en MAF frente a MTB. Con respecto al linaje 6, encontramos genes relacionados con el metabolismo de iones de cobre; sustrato necesario para que la bacteria desarrolle características virulentas en animales. Con respecto al linaje 5, encontramos genes relacionados con el estado metabólico dormante sobreexpresados; mientras que genes

relacionados con funciones de virulencia están reprimidos. En el linaje 1, encontramos genes como *virS*, otro regulador de virulencia, sobreexpresados. Por otro lado, tenemos *mpt63*, un epítopo reconocido por el sistema inmune, con una fuerte señal transcripcional en antisentido. En la rama común del linaje 4 tenemos parte del operón *mce2* reprimido así como genes relacionados con biosíntesis de molybdopterin. Con respecto al linaje 3, el gen más sobreexpresado es *oxyR*. Este hecho es bastante intrigante, ya que *oxyR* es un pseudogen que ha perdido su función en el MTBC. Por último, dentro del linaje 2 tenemos dos grandes grupos. Por una parte el llamado clado proto-Beijing, que es la rama más basal del linaje, y por otro el clado Beijing, que comprende cepas con un impacto epidemiológico muy importante, principalmente en Asia. Encontramos que el sistema DosR/DosS está sobreexpresado en las cepas Beijing con respecto a la rama basal del linaje, lo cual es un hecho ya conocido. Esta sobreexpresión viene dada por una mutación que genera un nuevo sitio de inicio de la transcripción aguas arriba del TF *dosR*. Un análisis de enriquecimiento para cada una de las ramas resaltó que la mayoría de funciones biológicas sobrerrepresentadas en los genes DE están relacionados con interacciones con el hospedador y procesos metabólicos.

Procesos no aleatorios generan gran plasticidad transcripcional en el MTBC

En un siguiente paso, intentamos ligar cambios transcripcionales con mutaciones genéticas concretas. Como adelantamos en el capítulo 5, ciertas mutaciones pueden crear nuevos sitios de inicio de transcripción (motivos TANNNT) que pueden provocar sobreexpresión de genes adyacentes. Decidimos buscar en nuestras cepas, mutaciones que pudieran o bien crear nuevos motivos TANNNT o interrumpir los ya existentes. Encontramos 683 mutaciones que creaban nuevos motivos y 81 que interrumpían motivos ya existentes. Análisis de permutaciones aleatorias y comparaciones frente a distribuciones de probabilidad, indicaron que estos valores eran mayores de los esperados por azar. Además, comprobamos como el elevado ratio de nuevos

motivos frente a interrumpidos es independiente de la posición genética en la que está la mutación. Parece ser que este hecho está relacionado con el elevado sesgo en mutaciones que generan nuevas bases AT.

Mirando los genes que anteriormente habíamos detectado como DE en cada una de las ramas, vemos que los nuevos motivos tienden a incrementar la transcripción de genes adyacentes (es el caso de *oxyR* en linaje 3) mientras que la interrupción de los motivos la disminuyen. Además, el $\sim$26% de los genes detectados como DE están asociados a este tipo de eventos mutacionales.

Patrones diferenciales de metilación

Intentamos testar el efecto de la metilación sobre la regulación génica. Para ello secuenciamos el ADN de las cepas con la plataforma PacBio, y buscamos bases metiladas. Encontramos tres motivos principales metilados en la mayoría de nuestras muestras, correspondientes a tres metiltransferasas identificadas en estudios previos (MamA, MamB y HsdM). En algunas cepas, los motivos reconocidos por cada una de las metiltransferasas no estaban metilados, lo que sugería que en determinados casos estas proteínas estaban inactivas. En estos casos, buscamos mutaciones no sinónimas en los genes codificantes para las metiltranferasas. Encontramos algunas mutaciones previamente caracterizadas, y algunas otras nuevas. Para tener una idea de la distribución global de estas mutaciones, las buscamos en las 4.595 cepas usadas en el capítulo 4. Algunas de las mutaciones las encontramos enraizadas muy profundamente en la filogenia, afectando a linajes completos. Por tanto, parece que gran parte de las cepas del MTBC pueden tener gran parte de su genoma metilado de forma diferencial.

El impacto de la metilación diferencial sobre la expresión génica es sutil e independiente de linaje

Para comprobar el impacto de la metilación diferencial (MD) sobre la regulación génica, buscamos lugares de inicios de la transcripción que estén metilados por alguna de las tres principales metiltransferasas. Encontramos 13

genes metilados por MamA, 24 por HsdM y 2 por MamB. A continuación, comparamos la expresión de cada uno de estos genes en cepas del mismo linaje, que tengan MD. Encontramos que, para los genes afectados por MamA y MamB, la expresión génica es ligeramente superior en las cepas metiladas que en las no metiladas, independientemente del linaje. Sin embargo, la metilación por HsdM no parece afectar a la expresión génica de los 24 genes con MD.

El mecanismo de regulación de HsdM es distinto a MamA y MamB

Para dilucidar el potencial efecto regulador la MD por HsdM, creamos un mutante *knock-out* de este gen, y analizamos su transcriptoma comparándolo con el *wild-type*. Encontramos varios genes con su expresión incrementada en el mutante. Principalmente los genes comprendidos entre Rv0081 a Rv0087. Por otro lado, 3 genes ven su expresión reducida en el mutante. Así pues, parece que la MD por HsdM tiene efecto sobre la regulación de algunos genes, aunque el mecanismo parece ser distinto a MamA y MamB, ya que no se han encontrado bases metiladas cerca de los genes DE.

Discusión

De nuestro análisis se deduce que cada uno de los linajes del MTBC tiene un perfil transcripcional propio. Sin embargo, pequeñas variaciones genéticas pueden tener un gran impacto sobre el perfil transcripcional de las muestras.

Hemos caracterizado genes que han visto su expresión modificada o lo largo de la evolución del complejo. La modificación de los niveles de expresión puede ser un mecanismo de adaptación fisiológica rápido ante un ambiente cambiante. Este parece haber sido el escenario cuando MAF y MTB se separaron de su ancestro común. En una distancia genética pequeña, muchos genes vieron alterada su expresión. Un cambio ambiental brusco relacionado con las poblaciones del hospedador, pudo ser la causa de estos cambios. El análisis funcional refuerza la idea de que los distintos linajes se han adaptado a diferentes hospedadores no solo a nivel de secuencia, si no también

alterando la expresión de distintos genes.

En este capítulo se han estudiado dos procesos que potencialmente tienen efecto sobre la regulación de la expresión. Por un lado, cerca del 26% de la expresión génica diferencial detectada en el MTBC está provocada por la alteración de sitios de inicio de la transcripción reconocidos por SigA. La selección, junto al sesgo mutacional AT, parece haber jugado un importante papel en este hecho. Por el contrario, la metilación parece tener un efecto mínimo en la regulación de la expresión génica *in-vitro*. No hemos sido capaces de detectar un gran impacto sobre la regulación debido a las metiltransferasas principales, excepto por un efecto sutil en unos pocos genes cuyas regiones reguladoras están metiladas por MamA o MamB. Esto puede ser debido a que:

- Para ver un efecto regulador importante debemos aplicar estress
- La metilación pudo tener un efecto regulador en cepas ancestrales, pero ahora no juega ese papel
- No hemos usado la aproximación correcta que ya el mutante Δ*hsdM* no muestra expresión diferencial en genes con el motivo TANNNT metilado de forma diferencial.

Lo que sí se deduce de nuestros resultados es que hay convergencia en los patrones de metilación de los distintos grupos filogenéticos. Fenotipos similares están presumiblemente producidos por variantes genéticas diferentes. Esto, junto con el hecho de que las metiltransferasas hayan estado evolucionando bajo selección purificadora, sugiere que la metilación en el MTBC aún puede estar jugando algún papel biológico.

Discusión general

Para conseguir erradicar la tuberculosis, la enfermedad infecciosa que más muertes humanas ha causado, necesitamos adquirir más conocimiento en

cuanto a los mecanismos de transmisión, persistencia, virulencia y adquisición de resistencias del patógeno. Todas estas características son altamente heterogéneas en patógenos con estructuras poblacionales muy complejas. En el caso de la tuberculosis, diversos autores han demostrado que también hay fluctuaciones en estos factores relacionadas con el perfil filogenético de la cepa bacteriana, pese a tener una estructura poblacional altamente clonal.

Es por esto que la diversidad genética del MTBC debería ser tenida en cuenta en las investigaciones que se realizan con este patógeno. Sin embargo, en muchos casos esto no sucede. Por ejemplo, las técnicas de biología de sistemas suelen trabajar con organismos modelo, y raramente incluyen polimorfismos. En el caso del MTBC y como vimos en el capítulo 4, modelos computacionales calculados con datos de una cepa de referencia no son capaces de realizar predicciones fiables sobre cepas de otros orígenes genéticos. Además, mutaciones encontradas de forma natural en el MTBC impactan distintos módulos de las redes de regulación y PPI, potencialmente afectando a su función. Sin embargo, todas las redes biológicas modeladas hasta ahora en el MTBC están basadas en la cepa de referencia H37Rv. Sería de gran utilidad generar, al menos, redes específicas de linaje y/o en distintas condiciones. Sin embargo, lo cierto es que generar datos a esa escala es técnica y económicamente muy costoso.

La limitada diversidad genética del MTBC puede deberse a distintos factores:

- El MTBC tiene una tasa evolutiva baja, probablemente influenciada por su lenta tasa de crecimiento y los largos tiempos de latencia, en los que la bacteria puede estar años con una actividad biológica reducida.
- Como vimos en el capítulo 4, no hay adquisición de material genético externo ni recombinación en el MTBC.

Este último punto es especialmente relevante para entender la aparición de resistencias a antibióticos en el MTBC. Las mutaciones de resistencia aparecen de forma independiente en la filogenia y se fijan a causa de la

presión de selección que supone el uso de antibióticos. La identificación de las cepas que contienen estas mutaciones es clave para vigilancia epidemiológica. La WGS está ganando cada vez más relevancia para la consecución de este objetivo.

Además, la WGS está aplicándose cada vez con más éxito en contextos clínicos y en investigación básica. Casi todos los resultados que se han presentado en esta tesis provienen de WGS, descargados de bases de datos públicas o generados por colaboradores en otros proyectos. Gracias a esto, hemos sido capaces de estudiar polimorfismos genómicos a nivel de SNP en miles de genomas, lo que nos ha permitido obtener resultados estadísticamente robustos.

Pero no solo hemos usado datos de secuenciación de genomas, si no que también hemos secuenciado transcriptomas completos. Los análisis del capítulo 6 nos han permitido arrojar luz sobre la heterogeneidad transcripcional presente en el MTBC, y los mecanismos por los que se produce esta variabilidad. Hemos logrado identificar como mutaciones en genes clave como *rpoB*, pueden alterar completamente el perfil transcripcional de la bacteria. Por otro lado, mutaciones que crean nuevos sitios de inicio de la transcripción también tienen efecto sobre la regulación de ciertos genes. Además de las modificaciones genéticas, hemos testado modificaciones epigenéticas (metilación de adeninas) para ver si tienen efecto regulador. Sin embargo, en este caso solo hemos sido capaces de encontrar un efecto sutil de la metilación del ADN sobre la expresión de ciertos genes.

Lo que se desprende de todos los análisis que hemos realizado es que los miembros del MTBC están perfectamente adaptados a su hospedador. La selección natural parece estar actuando para mantener la asociación patógeno-hospedador en un nivel máximo. En consecuencia, la mayor parte del genoma del MTBC parece estar bajo selección purificadora, con valores de dN/dS menores que 1. Pese a esta tendencia general, hay genes concretos que evolucionan bajo selección positiva. En el capítulo 4 hemos mostrado como *phoR* está evolucionando bajo selección positiva. No es el único sin

embargo, ya que otros estudios han mostrado como otros genes como *ppe38*, *esxW* o genes relacionados con resistencia a antibióticos están evolucionando bajo este tipo de selección.

Todos estos resultados se han obtenido mediante el uso de datos WGS. Y como ya hicimos notar en la introducción general, para analizar este tipo de datos es necesario el uso de herramientas bioinformáticas. En la presente tesis, la bioinformática ha tenido un papel protagonista. Por ejemplo, el pipeline de análisis que se ha desarrollado para analizar los datos ha sido esencial no solo para el desarrollo de esta tesis, si no para el trabajo diario de muchos de los miembros del grupo. De hecho, una comparativa realizada por un laboratorio independiente entre pipelines de distintos centros de referencia demostró que nuestra metodología estaba entre las que ofrecían resultados más robustos y reproducibles. Además del pipeline, muchos de los métodos usados han sido específicamente desarrollados por el doctorando para los proyectos y análisis que comprende esta tesis. Estamos convencidos de que la metodología desarrollada puede ser exportada a otros organismos y proyectos. Por tanto, la contribución de esta tesis al ámbito científico no se limita solo a los resultados y conclusiones, sino que también incluye la metodología desarrollada durante este tiempo.

En resumen, hemos estudiado la evolución del MTBC en distintos escenarios valiéndonos de las nuevas tecnologías disponibles. Hemos intentado poner en valor la diversidad genética del MTBC, como una característica clave a tener en cuenta si queremos encontrar una solución global para este patógeno mortal.