Genomics and Transcriptomics of the Mycobacterium tuberculosis complex

THÈSE N^{O} 5550 (2012)

PRÉSENTÉE LE 20 NOVEMBRE 2012

À LA FACULTÉ DES SCIENCES DE LA VIE

UNITÉ DU PROF. COLE

PROGRAMME DOCTORAL EN BIOTECHNOLOGIE ET GÉNIE BIOLOGIQUE

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Swapna UPLEKAR

acceptée sur proposition du jury:

Prof. J. McKinney, président du jury
Prof. S. Cole, Dr J. Rougemont, directeurs de thèse
Prof. R. Brosch, rapporteur
Prof. B. Deplancke, rapporteur
Dr K. Harshman, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2012

Table of Contents

1. Abstract

The goal of eliminating tuberculosis (TB) by 2050 depends on the development of improved TB diagnostics, drugs and vaccines. Advances in these areas require a deep understanding of the disease and its causative agent, *Mycobacterium tuberculosis* (*M. tb*). Mycobacterial species that cause TB in humans and other mammalian hosts are grouped within the *M. tb* complex. Development of powerful technologies such as next-generation sequencing and microarrays opened up new avenues for comparative and functional genomics of the *M. tb* complex. Due to the large and increasingly complex datasets generated from these technologies, the bottleneck in biological investigation has shifted from data generation to analysis. The objectives of this thesis were to establish and employ strategies for the analysis, integration, and interpretation of high-throughput sequencing and microarray datasets using a range of bioinformatics and statistical tools.

In the area of comparative genomics, we assessed the genetic diversity in the *M. tb* complex using various methods, such as SNP (single nucleotide polymorphism) genotyping, automated Sanger sequencing and next-generation sequencing. In a study comparing the genomes of the virulent *M. bovis* and *M. bovis* BCG vaccine strains, we identified a set of SNPs that were common to all BCG strains, and could provide novel insights on the molecular basis of BCG attenuation. In another study, we surveyed the genetic variation in the highly immunodominant *esx* gene family among clinical isolates of *M. tb* and identified sequence polymorphisms in known T-cell epitopes on Esx proteins that could affect their immunogenicity. We exploited the power of next-generation sequencing to detect sequence variation among *M. tb* strains that could result in phenotypic differences. By comparing the genomes of drug-resistant mutants with the sensitive wild-type strain we were able to identify the target of the anti-TB drug, pyridomycin. Using a similar approach we identified a mutation that makes *M. tb* strains incapable of producing PDIMs (phthiocerol dimycocerosates), which are cell wall associated lipids involved in *M. tb* virulence.

In the area of functional genomics, we mapped genome-wide binding sites for transcription factors using chromatin immunoprecipitation followed by hybridization to microarrays (ChIP-on-chip) or sequencing (ChIP-seq), and performed transcription profiling by means of high-throughput cDNA sequencing (RNA-seq). We carried out a comprehensive study to characterize the whole transcriptome of *M. tb* in exponential and stationary phases of growth, and understand the genome-wide dynamics of two key components of the transcription machinery, namely, RNA polymerase and NusA. By systematic integration of the ChIP-seq and RNA-seq data, we identified a set of transcription units (TU) in the *M. tb* genome, and mapped their putative promoters. Analysis of RNAP and NusA binding across the promoter

and body of TUs and their correlation with transcription uncovered new functional aspects of the transcriptional complex in *M. tb*. We also exploited the ChIP-on-chip and ChIP-seq technologies to define the regulon of the *M. tb* sigma factor F, and gain a better understanding of the regulatory role of the nucleoid associated protein, EspR.

Altogether, this thesis has improved our knowledge of the evolution, physiology and virulence of the *M. tb* complex. In addition, we have established next generation sequencing as a powerful tool for comparative and functional studies, with potential applications in the clinical setting.

Keywords: *Mycobacterium tuberculosis*, genomics, transcriptomics, ChIP-seq, RNA-seq, ChIP-on-chip, microarrays, next-generation sequencing, tuberculosis, bioinformatics, single nucleotide polymorphisms

Résumé

L'objectif d'éradiquer la tuberculose (TB) pour 2050 dépend du développement de nouveaux tests diagnostiques, d'agents antimycobactériens et de vaccins. Les progrès dans ces domaines nécessitent une bonne compréhension de la maladie et de l'agent infectieux en cause, *Mycobacterium tuberculosis* (*M. tb*). Les espèces de mycobactéries responsables de la tuberculose chez l'homme et chez d'autres mammifères forment le "complexe *M. tb*". Le développement de technologies très performantes telles que les puces à ADN et le séquençage à haut débit a ouvert la voie de la génomique comparative et fonctionnelle du « complexe *M. tb* ». En raison de la masse considérable et de la complexité des données générées par ces nouvelles technologies, l'étape limitante en génonique aujourd'hui n'est plus l'obtention de données mais leur analyse. Les objectifs de cette thèse étaient d'établir et d'utiliser des méthodes d'analyse, d'intégration, et d'interprétation des données des puces à ADN et du séquençage à haut débit en s'appuyant sur un panel d'outils statistiques et bio-informatiques.

Dans le domaine de la génomique comparative, nous avons étudié la diversité génétique du complexe *M.tb* en utilisant diverses méthodes telles que le génotypage de polymorphismes mononucléotidiques (SNP), le séquençage automatisé Sanger et le séquençage de nouvelle génération à haut débit. L'étude comparative des génomes d'une souche de *M. bovis* virulente et de souches de *M. bovis* BCG atténuées nous a permis d'identifier un certain nombre de SNPs communs à toutes les souches BCG, révélant de nouvelles bases moléculaires pouvant être à l'origine de l'atténuation des souches. Dans une autre étude nous avons examiné la variation génétique des gènes de la famille esx, chez des isolats cliniques de *M. tb*. Nous avons identifié des polymorphismes dans des épitopes connus de lymphocytes T qui pourraient affecter l'immunogénicité des protéines Esx. Nous avons également exploité la puissance du séquençage à haut débit pour détecter des variations de séquence entre des souches de *M. tb* dans le but de découvrir des mutations responsables de différences phénotypiques. En comparant les génomes de la souche sauvage *M. tb* H37Rv et d'un mutant résistant à la pyridomycine, nous avons pu identifier la protéine cible de cet antibiotique, la protéine InhA. Par la même approche, nous avons mis en évidence une mutation responsable d'un défaut de production de PDIM, un lipide associé à l'enveloppe de la bactérie et qui contribue à la virulence de *M. tuberculosis*.

Dans le domaine de la génomique fonctionnelle, nous avons cartographié à l'échelle du génome les sites de liaison à l'ADN de facteurs de transcription par la technique d'immunoprécipitation de la chromatine suivie d'hybridation de puces à ADN (ChIP-on-chip) ou de séquencage (ChIP-seq) et nous avons étudié le transcriptome

de *M. tb* par séquencage d'ADNc à haut débit (RNA seq). Nous avons réalisé une étude globale visant à caractériser le transcriptome complet de *M. tb* en phases de croissance exponentielle et stationnaire et à comprendre la dynamique de deux facteurs clés de la machinerie transcriptionnelle à savoir l'ARN polymérase et le facteur de transcription NusA. En intégrant de façon systématique les données obtenues par ChIP-seq et RNA-seq, nous avons identifié une série d'unités transcriptionnelles dans le génome de *M. tb* et localisé leurs promoteurs putatifs. L'analyse des sites de liaison de l'ARN polymérase et de NusA au niveau des promoteurs et des corps principaux des unités de transcription corrélée à l'activité transcriptionnelle a mis en évidence de nouveaux aspects fonctionnels du complexe de transcription. Nous avons également exploité les technologies de ChIP-on-chip et ChIP-seq pour définir le régulon du facteur sigma F et approfondir les connaissances sur le rôle régulateur de EspR, une protéine associée aux nucléoides.

Globalement, les études décrites dans cette thèse ont amélioré nos connaissances sur l'évolution, la physiologie et la virulence du complexe *M. tb*. De plus, nous avons établi le séquençage de nouvelle génération comme un outil performant pour réaliser des études comparatives et fonctionnelles ayant des applications cliniques.

Mots-clés: *Mycobacterium tuberculosis*, génomique , transcriptomicque ChIP-seq, RNA-seq, ChIP-on-chip, microarrays, séquençage à haut débit , tuberculose, bioinformatique, polymorphismes mononucléotidiques

2. Background

2.1 Tuberculosis

Tuberculosis (TB) is a chronic infectious disease that has afflicted humans for thousands of years. *Mycobacterium tuberculosis* (*M. tb*), the bacterium that causes TB is an obligate human pathogen. It is transmitted by aerosols containing infectious *M. tb* released from the lungs of infected individuals upon coughing. TB predominantly affects the lungs, but can occur in any tissue. An infection with *M. tb* can lead to one of three possible outcomes: (Glickman & Jacobs 2001; Young *et al.* 2009)

1) a minority of infected individuals develop rapidly progressive disease, referred to as 'active' primary TB. Active disease is a chronic wasting illness characterized by fever, weight loss and cough (in the case of pulmonary infection) that could lead to death.

2) the majority of infected individuals develop a clinically 'latent' TB infection (LTBI) characterized by an effective acquired immune response and absence of disease symptoms.

3) a small proportion of the 'latently' infected individuals, especially those who are immunosuppressed, may develop 'active' disease at a later stage, which is referred to as 'post-primary' TB.

2.1.1 History of the Disease

Archaeological evidence from Egyptian mummies indicated manifestation of TB as early as 5000 BC (Schaaf & Zumla 2009). TB was designated "The Captain of all the Men of Death" as it continued to claim lives through the ages (Daniel 2006). In the 18th and 19th centuries, TB reached epidemic proportions in Europe and North America and was referred to as "The Great White Plague" (Dubos 1952). On March 24, 1882, the German scientist Robert Koch changed the course of TB history when he identified *Mycobacterium tuberculosis* (*M. tb*) as the causative agent of TB. The discovery of the tubercle bacillus paved the way for great advances in the areas of TB diagnosis, prevention and control.

In 1907 Clemens von Pirquet developed the tuberculin skin test for diagnosing TB and used it to demonstrate latent TB infection in asymptomatic children (Daniel 2006). In 1908, Albert Calmette and Camille Guérin began their efforts to attenuate *Mycobacterium bovis* (a close relative of *M. tb*), which finally led to the development of the BCG (Bacille Calmette-Guérin) vaccine in 1921 (Calmette 1931). BCG remains the most widely used vaccine in the world. In the decades that followed, a therapeutic revolution took place with the discoveries of para-amino salicylic acid (PAS) in 1943 and streptomycin in 1944 (Daniel 2006; Zumla *et al.* 2009). In 1952, Isoniazid became the first oral anti-TB drug, followed by rifampicin in 1963 (Girling *et al.* 1976; Daniel 2006). Short-course chemotherapy regimens developed in the early 1970s proved to be highly efficient in the treatment of TB (Schaaf & Zumla 2009; Zumla *et al.* 2009; Girling *et al.* 1976).

Even before the introduction of BCG and chemotherapy, improvements in the socioeconomic conditions and living standards had led to a decrease in the

death toll from TB in the Western world. As tuberculosis was no longer considered a threat to developed countries funding for tuberculosis research declined. However, at the same time, TB assumed greater prevalence in many poor and developing countries. There was a global resurgence of TB by the late 1980s primarily due to the arrival and spread of HIV infection and the emergence of drug resistant strains of *M. tb*. This combined with the increased immigration from high-incidence countries, the deterioration in the quality of TB control, and intermittent adherence to treatment fuelled the TB epidemic further (Murray 2004; Zumla *et al.* 2009). In the 1990s, it was estimated that one-third of the world's population had been infected with *M. tb* (LTBI) putting these individuals at risk of developing active TB at a later stage (Dye *et al.* 1999). HIV infection was identified as the most powerful risk factor for progression of TB from latency to active disease, along with diabetes, old age, malnutrition, and other factors leading to immunosuppression (Aaron *et al.* 2004; Corbett *et al.* 2006; Lawn & Zumla 2011).

The World Health Organization (WHO) declared TB a global emergency in 1993, and re-established its TB control program by promoting the Directly Observed Therapy Short-Course (DOTS) strategy in 1994. The DOTS strategy relies on efficient case detection, standardized chemotherapy, uninterrupted drug supply, standardized recording and reporting systems and government commitment in order to cope with the high burden of TB (Schaaf & Zumla 2009; Espinal *et al.* 1999). In 2001, the Stop TB Partnership (http://www.stoptb.org) was established with the goal of eliminating TB as a public health problem. Led by the WHO and involving nearly 1000 organizations worldwide, the partnership outlined a global plan to halve TB prevalence and mortality by 2015 and to eliminate TB as a public health

problem by 2050 (Stop TB Partnership, 2006).

Despite the growing number of local, national and international research and TB control initiatives in the past decade TB is still among the top causes of death and morbidity worldwide, especially in the low- and mid-income countries.

2.1.2 Epidemiology of Tuberculosis

TB remains a global pandemic, killing someone approximately every 25 seconds. In 2010 there were an estimated 8.8 million cases of TB and 1.4 million estimated deaths caused by TB. The highest incidence rate was observed in sub-Saharan Africa (Figure 1), reflecting the high prevalence of HIV-infection in this region (Figure 2) (Raviglione *et al.* 2012). Over 80% of the worldwide TB burden is borne by 22 low-income and middle-income countries, with the majority of the estimated cases occurring in Asia (59%) and Africa (26%). The five countries with largest number of incident cases in 2010 were India, China, South Africa, Indonesia, and Pakistan (World Health Organisation, 2011b) (Figure 1). Although it appears that TB rates in North America and most western European countries have reached record lows, the number of cases in some metropolitan areas, such as London, has doubled in the past 10 years (Raviglione *et al.* 2012) (Figure 1). HIV-associated TB presents a formidable challenge, accounting for 1.1 million of the total TB cases in 2010. The African region accounts for the highest proportion (82%) of TB cases co-infected with HIV (Figure 2).

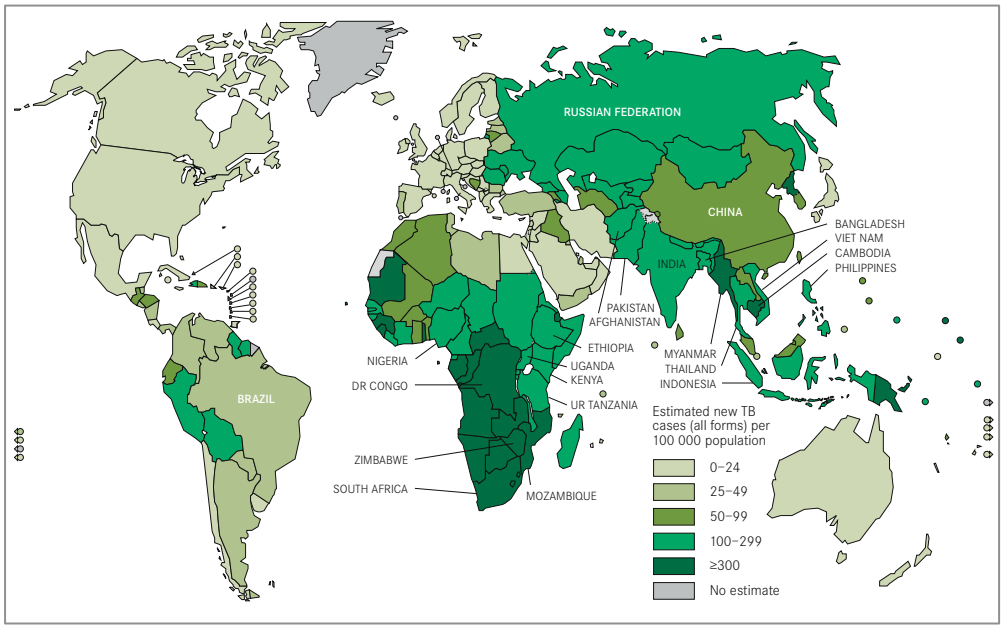


Figure 1. Estimated tuberculosis incidence in 2010 (WHO, 2011b).

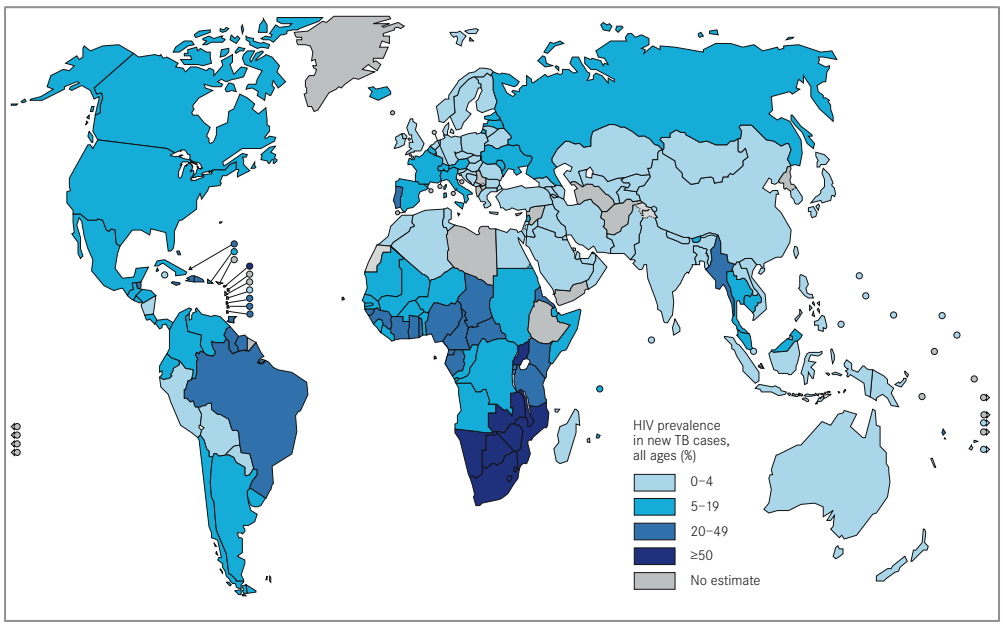


Figure 2. Estimated HIV prevalence in new cases of TB in 2010 (WHO, 2011b).

Another major threat to TB control is the emergence of drug-resistant *M. tb* strains. Drug-resistant TB is widespread and found in all countries surveyed by the WHO. Multidrug-resistant TB (MDR-TB) is caused by bacteria that are resistant to two of the first-line anti-TB drugs (rifampicin and isoniazid). In 2010, the top five countries in terms of total numbers of MDR cases were India, China, the Russian Federation, South Africa, and Bangladesh. Estimates for the incidence of drug-resistant tuberculosis are difficult to obtain due to insufficient laboratory facilities for drug susceptibility testing in most endemic countries (Raviglione *et al.* 2012). Of the expected 250,000 MDR-TB cases in 2009, only 24,511 were reported to have enrolled for treatment (WHO Stop TB Department, 2011d). In 2010, the estimated prevalence of MDR-TB rose to 650,000 cases among which only 46,000 enrolled for treatment. These data reveal the urgent need for improving the diagnosis and management of drug-resistant tuberculosis, especially in resource-poor countries. Extensively drug-resistant TB (XDR-TB) is a serious threat to global health caused by bacterial strains that are resistant to the two first-line drugs (rifampicin and isoniazid) and to second-line drugs

(fluoroquinolones and injectable aminoglycosides) (WHO Stop TB Department, 2011d). Globally, there are an estimated 25,000 annual cases of XDR-TB. A total of 69 countries have reported at least one case of XDR-TB (Migliori *et al.* 2007; N Sarita Shah 2011). The situation is worse in Eastern Europe where 10% of all MDR-TB cases are extensively drug-resistant.

Since 1995, 46 million people have been successfully treated and up to 6.8 million lives have been saved through the Stop TB Strategy (Raviglione *et al.* 2012). To eliminate TB as a public health problem by 2050, the global incidence must fall at a rate of 16% annually over the next four decades (Lönnroth *et al.* 2009). Unfortunately, the rate of decline in TB incidences has been very slow, at less than 1% since 2002, stressing the need for intensifying efforts to combat TB.

2.1.3 Methods for TB Control and Challenges

In spite of substantial advances in TB control, and a steady decline in the global TB incidence in the last decade, the disease still represents an unacceptable burden of human suffering and death in the world. Elimination of TB can only be achieved by intensifying basic and operational research in the three key areas, namely, effective and faster-acting anti-TB drugs, more sensitive and specific diagnostic tests and a universally efficacious vaccine (Cole 2002a; Zumla *et al.* 2009). The Stop TB Partnership and WHO created the TB Research Movement to support and stimulate fundamental research on tuberculosis. The result was a roadmap for global TB research – a list of the main research priorities that need to be addressed to help us get closer to the goal of TB elimination (Stop TB Partnership, 2011c). This section will summarize the important recent developments in global research efforts for TB control (Lawn & Zumla 2011; Raviglione *et al.* 2012) along with the key research priorities identified in the International Roadmap for TB research and selected review publications in areas of diagnostics (Wallis *et al.* 2010), vaccine (Brennan 2012; Kaufmann & Hussey 2010), and drug (Zumla, Hafner, *et al.* 2012a; Sala & Hartkoorn 2011) development.

2.1.3.1 Diagnosis

The last few years have seen substantial progress in the development of new diagnostic tools for TB. In 2007, WHO recommended liquid culture systems for TB diagnosis and for drug-susceptibility testing owing to their increased sensitivity and reduced delays in obtaining results compared to the use of solid media. The following year, molecular line probe assays (LPAs) based on reverse hybridization technology were approved for detection of drug resistance in smear-positive patients (Morgan *et al.* 2005; Ling, Zwerling, *et al.* 2008b). The most promising development in TB diagnostics came in 2010 with the introduction of nucleic acid amplification tests (NAATs) (Ling, Flores, *et al.* 2008a). The GeneXpert MTB/RIF assay is a quick (< 2 hours), automated diagnostic test that can detect *M. tb* and resistance to rifampicin (RIF) (Helb *et al.* 2010). It has been endorsed by WHO and adapted on a wide scale over the last couple of years (Boehme *et al.* 2010).

For the past century, the primary diagnosis of LTBI was carried out using the tuberculin skin test. Unfortunately, this test is unable to distinguish individuals infected with *M. tb* from those exposed to other mycobacteria, including those immunized with *M. bovis* BCG. About a decade ago, interferon gamma (IFN-γ) release assays (IGRA) were developed for diagnosing latent TB infection based on the release of IFN-γ triggered by immunodominant *M.tb* antigens such as ESAT-6 and CFP-10 (Pai *et al.* 2008). QuantiFERON-TB Gold (Mazurek *et al.* 2001) and the T-SPOT.TB (Meier *et al.* 2005) tests are commercially available IGRAs that can detect latent TB but have a reduced specificity for diagnosing active TB.

Despite the recent progress in the field of diagnostics, most of the high-burden, resource-poor countries rely on the inaccurate and relatively

antiquated methods such as direct smear microscopy, chest radiography and tuberculin skin testing. The diagnostics pipeline is still missing a simple, rapid, inexpensive point-of-care test for active tuberculosis that can work in high-burden settings without the need for sophisticated equipment or laboratory infrastructure. The existing diagnostic tests are also restricted in their ability to diagnose latent TB infection, predict progression of disease and, most importantly, to rapidly screen and detect MDR-TB, XDR-TB, and HIV-associated TB. At present, the major research priorities for diagnostics development are:

(1) Identify and evaluate new host- and pathogen-specific biomarkers that can be used to discriminate between active and latent TB from infected samples.

(2) Devise affordable, rapid, simple and multi-functional diagnostic platforms that can be appropriately used in high-burden settings.

(3) Combine new and existing diagnostic tests in order to detect various forms of TB (drug-sensitive, drug-resistant, LTBI) in various population settings and at all health-care levels.

2.1.3.2 Treatment

The current DOTS treatment regimen for drug-sensitive TB lasts for six months and involves administration of four drugs, isoniazid, rifampicin, pyrazinamide and ethambutol for the first two months followed by four months of isoniazid and rifampicin. A key feature of the DOTS regimen is giving the drugs under direct observation to ensure completion of therapy in order to prevent relapse of the disease and minimize the likelihood of spreading drug resistance. Although this makes the therapy highly effective, its lengthy duration represents a challenge, especially in most developing countries where adherence to treatment is often poor. For TB patients infected with MDR-TB strains, or those who are intolerant to first-line drugs, there are second-line drugs such as fluoroquinolones, amikacin and kanamycin. The regimens currently used for MDR- and XDR-TB are long, toxic, expensive and of limited efficacy (Zumla, Abubakar, *et al.* 2012b). Given the high prevalence of HIV-associated TB infection, it is also important to consider potential interactions of anti-TB drugs with drugs used in anti-retroviral therapy (ART).

Currently, the development pipeline has over a dozen new or repurposed anti-TB drugs in clinical trails or in preclinical development (Cole & Riccardi 2011). Phase 3 trials to investigate whether the current regimen can be shortened to four months by replacing ethambutol or isoniazid with gatifloxacin or moxifloxacin are also in progress (Figure 3). For the treatment of LTBI, three month regimens of rifapentine and isoniazid have yielded good results (Sterling *et al.* 2011). Keeping the drug pipeline' filled is essential for further progress in TB treatment.

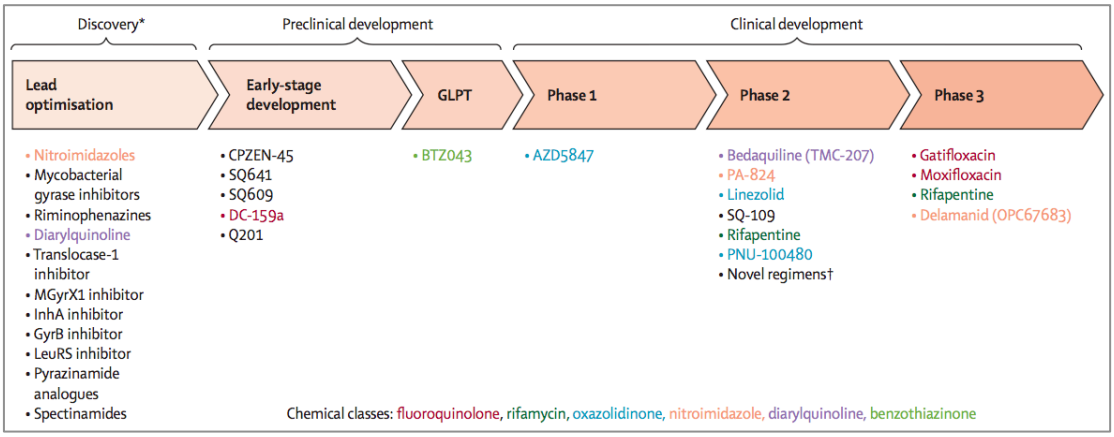


Figure 3. Development Pipeline for new tuberculosis drugs (Stop TB Partnership, 2012; Raviglione *et al.* 2012)
*Ongoing projects without a lead compound series. †Combination regimens

There is an urgent need to develop TB regimens that can cure all forms of TB, that are safe and compatible with ART and effective against LTBI. This calls for greater emphasis in understanding the fundamental aspects of *M. tb* biology, its interaction with the host and its life cycle. At present, the major research priorities in the field of TB drug development are:

(1) Identify the mechanisms of action of anti-TB drugs that are currently being used or being developed.

(2) Develop models of *M. tb* metabolism and physiology, and understand persistence mechanisms of the bacilli.

(3) Understand the relation between active and latently persisting bacilli and develop drugs that specifically target LTBI.

2.1.3.3 Prevention

Since 1921, *Mycobacterium bovis* bacillus Calmette-Guérin (BCG) has been the only vaccine used against tuberculosis. It is a live attenuated vaccine that was derived upon 230 repeated subcultures of a strain of *Mycobacterium bovis* on potato slices soaked in glycerol and ox bile, leading to the *in vitro* accumulation of mutations and ultimately attenuation (Calmette 1931). Over 4 billion people have been vaccinated with BCG so far making it the most widely used vaccine in the world (Kaufmann & Hussey 2010). BCG is effective against severe forms of TB in children (Trunz *et al.* 2006), but it offers very little protection against adult pulmonary TB, the most prevalent form of the disease (Fine 1995). In addition, BCG is not recommended for use in children diagnosed as HIV-positive owing to the risk of disseminated BCG disease. There is a dire need for new vaccines that are safe in HIV-infected children and adults, and effective against all age groups and populations. At present there are at least 12 TB vaccine candidates in clinical trials (Figure 3) with the aim of either replacing or enhancing the existing BCG vaccine (Kaufmann & Hussey 2010). The vaccine candidates can be classified into 4 main types (Brennan 2012; Kaufmann & Hussey 2010):

(1) prime vaccines, to be given as a replacement to BCG, (2) booster vaccines to be given weeks, or years after vaccination with BCG or its replacement, prior to TB exposure, (3) post-infection vaccines to target latent infection or prevent re-activation in those who have been pre-exposed to TB, (4) immunotherapy vaccines to be given as an adjunct to chemotherapy (Figure 3) (Brennan 2012).

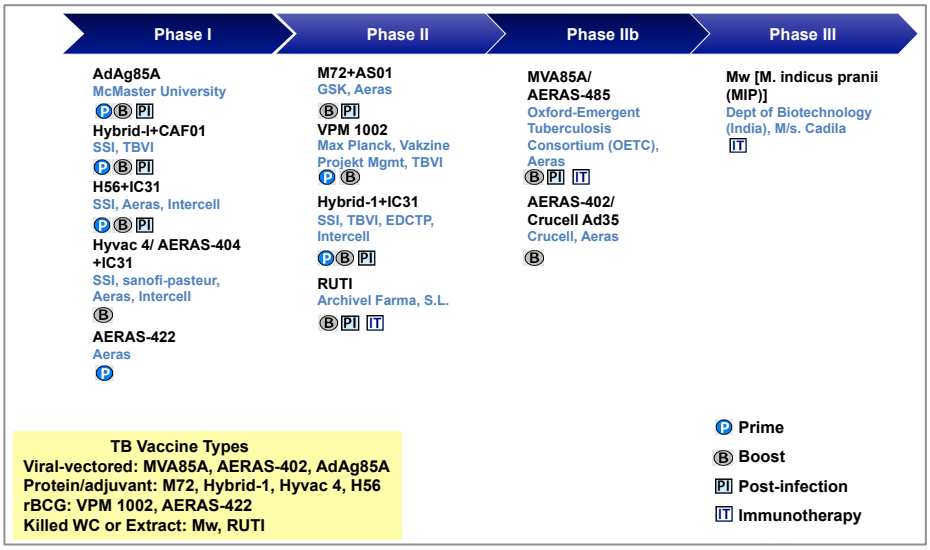


Figure 4. Global TB Vaccine Pipeline in 2011 (Stop TB Partnership, 2011a).

The objective of fundamental research in vaccine development is to understand how to manipulate the host immune response to control *M. tb* infection and disease. Therefore, it is critical to identify components of the host immune system that are required for elimination of the bacteria. At present, the major research priorities in the area of vaccine development are:

(1) Evaluate and compare immune responses to BCG and to new vaccines in different populations and age groups.

(2) Understand host-pathogen interactions during disease progression and to determine components of the host's immune system that lead to elimination of *M. tb* and prevent reactivation of LTBI.

(3) Identify immunodominant antigens that are associated with different metabolic states of *M. tb*.

2.2 *Mycobacterium tuberculosis (M. tb)*

Mycobacterium tuberculosis (*M. tb)*, the causative agent of TB in humans, is an aerobic, slow-growing, acid-fast, rod-shaped bacterium. The genus *Mycobacterium* is within the order Actinomycetales that comprises a large number of well-characterized species, several of which are associated with human and animal disease. Other human pathogens include, *Mycobacterium leprae,* the cause of leprosy, *Mycobacterium ulcerans*, the causative agent of Buruli ulcer, and *Mycobacterium avium,* which is an opportunistic pathogen often reported to be involved in HIV-infection (Horsburgh 1991).

M. tb has a slow doubling time of approximately 24 hours and takes three to four weeks to form colonies *in vitro*. In humans, it undergoes a complex life cycle, which can involve a dormant or latent phase where metabolism is thought to be greatly reduced as the bacteria are contained within granulomas, organised arrangements of immune cells. The dormant bacteria can reactivate to cause full-blown disease many years later, following immunosuppression or aging. The slow replication rate and ability to persist in a latent state contribute to the chronic nature of TB infection. A characteristic feature of mycobacteria is their extremely resilient cell envelope that contains a rich variety of lipids such as mycolic acids, glycolipids, and polysaccharides. This unique cell wall is also responsible for the acid fast staining property of mycobacteria.

2.2.1 The *Mycobacterium tuberculosis* complex

Mycobacterial species that cause TB in humans and other mammalian hosts are collectively known as tubercle bacilli and are grouped within the *Mycobacterium tuberculosis* complex (MTBC). In humans, TB is primarily caused by *M. tuberculosis* and *M. africanum* (sub-Saharan Africa). In addition, several animal-adapted members of MTBC have been identified. These include, *M. bovis* (cattle) and its attenuated derivative *M. bovis* BCG (vaccine strain), *M. microti* (voles and other small rodents), *M. pinnipedii* (seals and sea lions), and *M. caprae* (goats and sheep) (Schaaf & Zumla 2009).

The genomic revolution in the last decade has facilitated the application of a variety of powerful tools, such as large-scale transcriptomics, proteomics, comparative genomics, and structural genomics studies in the field of TB research that have led to great advances in our understanding of the biology and pathogenesis of the tubercle bacilli.

2.2.2 Genome Biology of *M. tb* H37Rv

In 1998, the paradigm reference strain H37Rv became the first fully sequenced *M. tb* strain (S. Cole *et al.* 1998). The circular *M. tb* H37Rv genome consists of 4,411,532 bp and displays a high G+C content of 65% (Figure 5). It contains over 4000 protein-coding and 50 stable RNA genes, using around 91% of its potential coding capacity. It displays a gene density at one gene per kilobase, which is comparable to other prokaryotes. Genes are evenly distributed on the forward and reverse strands, and 59% of the transcription is in the same polarity as the replication fork. More than half of the coding sequences in *M. tb* have arisen from gene duplication or domain-shuffling events (Tekaia *et al.* 1999).

Analysis of the genome sequence highlighted several characteristics of the biochemistry, physiology, genetics and immunology of *M. tb*. Being the most widely used strain in TB research, continued efforts have been carried out (S. T. Cole & Barrell 1998; J.-C. Camus *et al.* 2002b; J. Camus *et al.* 2002a; Lew *et al.* 2011) to maintain an accurate and up-to-date genome annotation of *M. tb* H37Rv using a combination of bioinformatics, data mining, comparative genomics, scientific literature and manual curation. For over 10 years, the TubercuList database (Lew *et al.* 2011) (http://tuberculist.epfl.ch) has been providing access to gene-based annotation for *M. tb* H37Rv. The current release (R25 – April 2012) contains 4095 annotated features, which includes 4022 protein genes and 73 RNAs (tRNAs, ribosomal RNAs, small RNAs, stable RNAs).

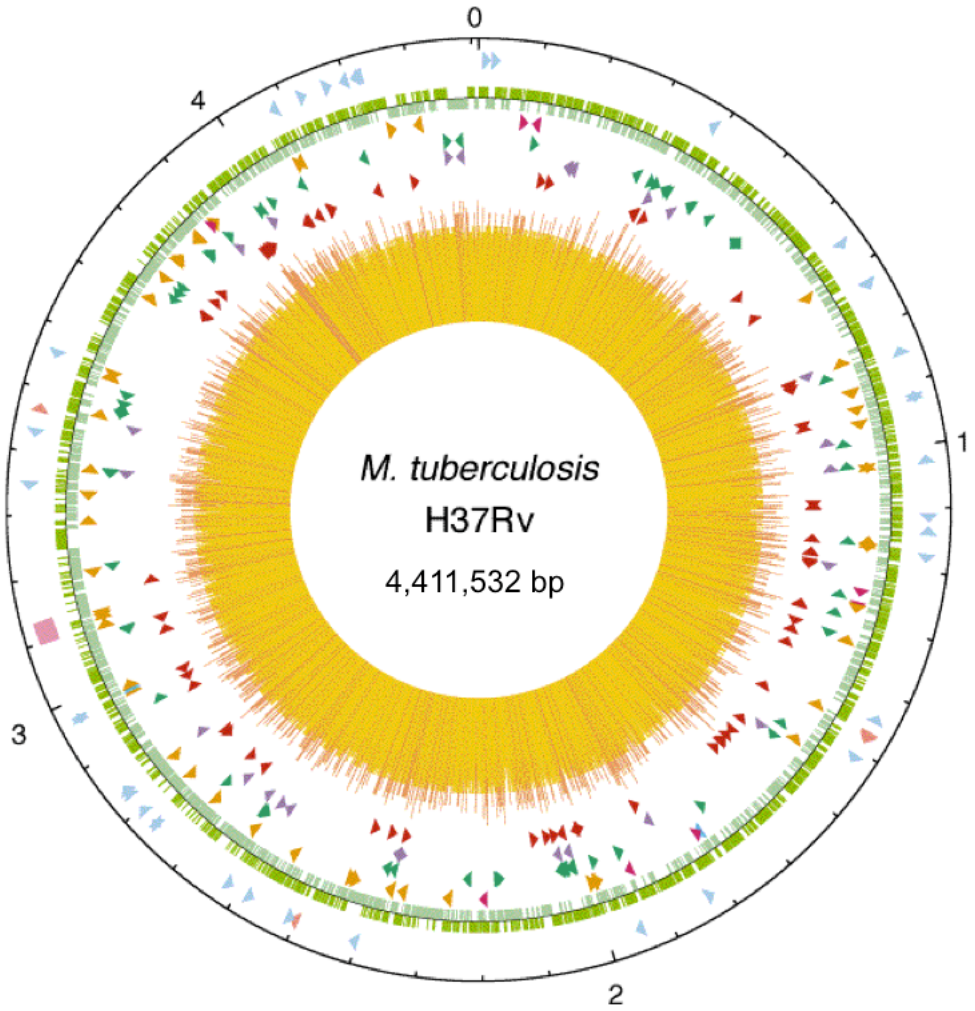


Figure 5. Circular map of the chromosome of *M. tuberculosis* H37Rv (S. Cole *et al.* 1998).

The outer circle shows the scale in megabases, with 0 representing the origin of replication. The first ring from the exterior denotes the positions of stable RNA genes (tRNAs are blue, and others are pink) and the direct-repeat region (pink cube); the second ring shows the coding sequence by strand (clockwise, dark green; anticlockwise, light green); the third ring depicts repetitive DNA (insertion sequences, orange; 13E12 REP family, dark pink; prophage, blue); the fourth ring shows the positions of the PPE family members (green); the fifth ring shows the positions of the PE family members (purple, excluding PGRS); and the sixth ring shows the positions of the PGRS sequences (dark red). The histogram (center) represents the G+C content, with <65% G+C in yellow and >65% G+C in red (S. Cole *et al.* 1998).

The *M. tb* genes have been broadly classified into eleven functional categories (Table 1). The genome sequence highlighted several characteristics of the biology of *M. tb,* which have been summarized below for each functional category.

Table 1. Broad classification of *M. tb* genes according to the TubercuList database (Lew *et al.* 2011)

Category	Function	Number of features	% of total
0	Virulence, detoxification, and adaptation	238	5.8
1	Lipid metabolism	272	6.6
2	Information pathways	242	5.9
3	Cell wall and cell processes	773	18.9
4	Stable RNAs	73	1.8
5	Insertion seqs and phages	147	3.6
6	PE/PPE proteins	168	4.1
7	Intermediary metabolism and respiration	936	22.9
8	Unknown	16	0.4
9	Regulatory proteins	198	4.8
10	Conserved hypotheticals	1032	25.2

Virulence, detoxification, and adaptation

This functional category includes a number of virulence factors, heat shock proteins, also known as molecular chaperones, oxidative stress response proteins, and toxin-antitoxin systems. An important group of *M. tb* virulence factors involved in host cell invasion is the mammalian cell entry (mce) protein family. The *M. tb* H37Rv genome contains four *mce* operons that are similar in sequence and organization. Inactivation of certain *mce* genes has been shown to reduce the ability of *M. tb* to invade and/or persist in host cells, resulting in attenuation of *M. tb* virulence in mice (Gioffré *et al.* 2005). Homologs of Mce proteins are found in pathogenic as well as non-pathogenic mycobacteria suggesting that their function might not be limited to cell entry (F. Zhang & Xie 2011).

The *M. tb* genome has evolved strategies to counteract the toxic effects of oxidative damage. These include production of compounds like mycothiol, and enzymes, superoxide dimutase (SodAB) and catalase (KatG). The *katG*

gene encodes catalase and peroxidase, which have been shown to be involved in growth and persistence of *M. tb* in mice and guinea pigs (Manca *et al.* 1999). KatG is also essential for activation of isoniazid (INH). Resistance to INH has often been associated with mutations in the *katG* gene (Y. Zhang *et al.* 1992).

It is interesting to note that more than half the CDSs (52%) annotated in this functional category belong to the toxin-antitoxin (TA) locus. TA systems typically consist of a two-gene operon encoding a toxin protein that targets an essential cellular function and an antitoxin that counteracts the effects of the toxin (van Embden *et al.* 1993; Gerdes *et al.* 2005). At present, 62 pairs of TA genes are annotated in the *M. tb* H37Rv genome, nearly all of which are conserved among the MTBC, but absent from species outside of this complex (Lew *et al.* 2011). The largest family of TA systems, VapBC encodes toxins that act by inhibiting translation via mRNA cleavage. A subset of TA systems has been shown to be induced in response to stress and implicated in *M. tb* pathogenesis and adaptation (Ramage *et al.* 2009).

Lipid metabolism

M. tb devotes 8% of its coding capacity to lipid biosynthesis and lipid degradation. There are representatives of almost every known lipid biosynthetic system encoded in the *M. tb* genome, including homologues of enzymes found in mammals and plants. The *M. tb* genome contains roughly 250 enzymes involved in fatty acid metabolism, compared to only 50 in *Escherichia coli.* A large proportion of these encode components of fatty acid oxidation systems, including the canonical FadA/FadB β-oxidation complex, revealing the importance of degradation of host-cell lipids in intracellular survival of *M. tb* (S. Cole *et al.* 1998). Mycobacteria are unique among

bacteria, as they possess both a mammalian type fatty acid synthase I (FASI) enzyme as well as a bacterial type fatty acid synthase II (FASII). These enzymes are involved in the synthesis of mycolic acids, which are key components of the mycobacterial cell wall (S. Cole *et al.* 1998). While the FASs are involved in biosynthesis of primary metabolites, there is another family of genes known as the polyketide synthases (PKSs), which often catalyze the formation of secondary metabolites. The *pps* gene cluster (*ppsABCDE*) and the *mas* gene cluster produce phthiocerol and mycocerosic acid that esterify to form phthiocerol dimycocerosate (PDIM), which is one of the most abundant cell wall-associated lipids (S. Cole *et al.* 1998). Over the years, *M. tb* lipids have also been shown to play key roles in virulence and pathogenesis of the bacillus.

Information pathways

This functional class comprises genes involved in the fundamental cellular processes, such as DNA replication, DNA repair, transcription and translation. About 40% of the genes in this category encode ribosomal proteins, which make up subunits of the ribosome in conjunction with ribosomal RNA. A number of commonly used TB drugs target functions belonging to this category. For example, rifampin targets the DNA-dependent RNA polymerase RpoB, fluoroquinolones target DNA gyrase, and streptomycin targets the ribosomal protein RpsL (Schaaf & Zumla 2009).

The prokaryotic core RNA polymerase (RNAP) is composed of four distinct subunits, β, β', ω, and an α dimer. The fifth subunit is the σ factor, which can associate reversibly with core RNAP to form the holoenzyme. It provides the promoter recognition function by directing the holoenzyme to specific genes. Bacterial genomes generally encode a principal σ factor that is involved in

transcription of housekeeping genes, and a variable number of alternative σ factors. The *M. tb* genome encodes 13 σ factors, each with its own specificity. The principal σ factor, SigA is essential for general transcription of housekeeping genes, while the other σ factors (SigB to SigM) are nonessential (S. Cole *et al.* 1998). One common mechanism by which σ factors are regulated is inhibition by proteins referred to as anti-σ factors. Anti-σ factors for SigE, SigH, SigL, SigK and SigF have been annotated in the *M. tb* genome (Lew *et al.* 2011). The *sigB* gene encodes an alternative-principal σ factor, which is induced under various stress conditions. SigF is a stress response sigma factor, which is required for full virulence of *M. tb* (Lee *et al.* 2008). The ten other σ factors are classified as extracytoplasmic function σ factors, which control a variety of functions in response to specific extracellular environmental signals, such as, nutrient starvation (SigD), heat stock (SigE), antibiotic exposure (SigJ), and oxidative stress (SigH) (S. Cole *et al.* 1998; Manganelli *et al.* 2004).

Cell wall and cell processes

Given the distinctive architecture and composition of the mycobacterial cell wall, it is not surprising that *M. tb* devotes a significant proportion (19%) of its genes to cell wall and cell processes (Lew *et al.* 2011). This functional class contains genes involved in cell wall biosynthesis, along with lipoproteins, transmembrane- and membrane-associated proteins, and secreted proteins. The *M. tb* cell wall is composed of peptidoglycan (PG) and arabinogalactan (AG), which form the cell wall core (mAGP) by attaching to mycolic acids (Brennan 2003). In addition there are a number of cell wall-associated lipids, such as phosphatidyl-myoinositol mannosides (PIMs) and lipoglycans, termed lipomannan (LM) and lipoarabinomannan (LAM) (Brennan 2003). In addition to their physiological role these cell wall components also play an

immunomodulatory role by interacting with receptors of the host immune system (Brennan 2003; Mishra *et al.* 2011). Many of the proteins involved in biosynthesis of cell wall components in *M. tb* are essential for its survival and therefore represent excellent drug targets (Sassetti *et al.* 2003). Ethambutol is a first-line TB drug which blocks the synthesis of arabinogalactan, whereas cycloserine, a second-line TB drug, inhibits peptidoglycan synthesis (Schaaf & Zumla 2009).

An important family of genes in this functional class encodes specialized secretion systems, known as type VII secretion systems (T7SS), which enable transport of extracellular proteins across the cell wall. *M. tb* possesses five T7SS, named ESX-1 to ESX-5, which display a conserved genomic organization characterized by the presence of tandem gene pairs encoding ESX family proteins located immediately downstream of PE/PPE genes (Abdallah *et al.* 2007). The ESX-1 system encodes and secretes the prototypic ESX proteins, namely the 6 kDa early secreted antigenic target, ESAT-6 (EsxA) and the 10 kDa culture filtrate protein, CFP-10 (EsxB), which are involved in host-pathogen interactions and play a crucial role in *M. tb* virulence (Alderson *et al.* 2000). Incidentally, ESX-1 is also conserved in other pathogenic mycobacteria such as *M. leprae* and *M. marinum*. Loss of the region of difference 1 (RD1) containing the ESX-1 locus contributes to the attenuation of the vaccine strains of *M. bovis* BCG (Pym *et al.* 2002). Of the other systems, ESX-5 is known to be necessary for the secretion of PE and PPE proteins in *M. marinum* (Abdallah *et al.* 2008) and plays a similar role in *M. tb,* where its disruption results in loss of PPE protein secretion, altered cell wall integrity and attenuation (Bottai *et al.* 2012). ESX-3 is essential for *in vitro* growth and may be involved in iron/zinc homeostasis (Siegrist *et al.* 2009).

Stable RNAs

This category is dedicated to the genes encoding functional RNA molecules in *M. tb.* It presently contains 73 genes encoding ribosomal RNAs, tRNAs, small RNAs and other stable RNAs (Figure 5). Most eubacteria usually contain several copies of the ribosomal operon located close to the origin of replication, in order to exploit the gene-dosage effect obtained during replication. In contrast, *M. tb* has a single ribosomal RNA operon, located 1500 kb from the origin of replication (Bercovier *et al.* 1986). It is likely that this contributes to the slow-growth of mycobacteria. Genes encoding tRNAs in *M. tb* recognize 43 of 61 possible sense codons. Over the last few years, a number of small non-coding RNAs have been identified in *M. tb*, including, anti-sense transcripts and intergenic small RNAs (sRNAs) (Arnvig & Young 2009). While functional characterization of sRNAs in *M. tb* has only been recently initiated, some sRNAs have been found at increased abundance during stress conditions, infection or in the stationary phase, implicating their involvement in host adaptation and virulence (Arnvig & Young 2012; DiChiara *et al.* 2010).

Insertion sequences and phages

About 3.5% of the *M. tb* genome is composed of insertion sequences (IS), prophages, and a family of repetitive sequences, which exhibit some characteristics of mobile genetic elements (Figure 5). There are 56 copies of IS elements belonging to at least nine different families, some of which are highly similar to those found in other actinomycetes. The most abundant IS elements in the *M. tb* genome are IS*6110* (16 copies) and IS*1081* (6 copies). While most IS elements are stable, IS*6110* frequently transposes, and can lead to inactivation of genes upon insertion or deletion of chromosomal fragments as a result of homologous recombination. The IS*6110* copy number can vary

from 0-25 between *M. tb* strains, therefore making them useful molecular markers for genotyping strains from the MTBC (van Embden *et al.* 1993; Prod'hom *et al.* 1997). There are two prophage-like elements, phiRv1 and phiRv2, in the genome of *M. tb*, which share a similar genetic organization and are both over 10 kb in length. The phiRv1 prophage is absent in the *M. bovis* BCG vaccine strain. The functional class also includes seven copies of a repetitive sequence referred to as REP13E12 or the '13E12' family (Figure 5).

PE/PPE proteins

The most prominent multi-gene families in *M. tb*, which are unique to mycobacteria, encode the glycine-rich PE and PPE proteins. These genes account for 10% of the coding capacity of the *M. tb* genome (Figure 5). PE and PPE members share a conserved N-terminal domain with characteristic Pro-Glu (PE) or Pro-Pro-Glu (PPE) motifs and can be divided into subfamilies according to the homology and presence of characteristic motifs in their C-terminal domains. The largest of these subfamilies contains multiple copies of polymorphic GC-rich repetitive sequences at the C-terminal end, and is called the PE_PGRS family of proteins. The second largest is the major polymorphic tandem repeat or MPTR PPE family. A subset of the PE and PPE genes is associated with the ESX loci, which encode components of type VII secretion systems. It has been suggested that the PE and PPE proteins may contribute to antigenic variation by interfering with the T-cell immune responses (Ramakrishnan *et al.* 2000).

Intermediary metabolism and respiration

Around 23% of genes belong to this class, making it the largest functional category (excluding conserved hypothetical proteins) in the *M. tb* genome. The abundance of genes in this category, and their functional diversity reflect

the high metabolic capacity of *M. tb.* Genes necessary for synthesis of all essential amino acids, vitamins and enzyme co-factors are present in *M. tb.* In addition it can synthesize enzymes necessary for the anabolic pentose phosphate pathway, the catabolic Kreb's cycle, glycolysis, and the glyoxylate cycle (S. Cole *et al.* 1998). It contains electron transport chains using oxidative phosphorylation, as well as several anaerobic electron transport chains. These include gene clusters that encode nitrate reductase (*narGHJI*), nitrite reductase *(nirBD*), and fumarate reductase (*frdABCD*) (S. Cole *et al.* 1998; Schaaf & Zumla 2009). It has been demonstrated that the capacity for anaerobic respiration might contribute to *M. tb* virulence (Stermann *et al.* 2004). Several studies have suggested that *M. tb* switches its intermediary metabolism from carbon sources such as glucose and glycerol to fatty acids and host lipids during the course of infection (Schnappinger *et al.* 2003; Muñoz-Elías *et al.* 2006). The ability of the bacillus to adapt its metabolism to the number of different environments, such as the lung, the macrophage, and the granuloma, plays a key role in its pathogenesis.

Regulatory proteins

Gene expression in *M. tb* is controlled at the level of transcription initiation by a diverse array of regulatory elements. The *M. tb* genome encodes about 200 regulatory proteins, which include eleven two-component systems, five unpaired response regulators, eleven protein kinases and over 100 putative transcriptional regulators (S. Cole *et al.* 1998). The eukaryotic-like serine/threonine protein kinases (STPKs), namely, *pknA*, *pknB* and *pknD-pknL* have been found to be involved in various processes including cell growth, morphology, molecular transport, as well as transcription regulation (Sharma *et al.* 2006; Kumar *et al.* 2009). *M. tb* encodes eleven complete two-component systems, including the DosR (DevR) and DosS (DevS) system,

which controls the expression of a number of genes that are induced during hypoxia and in response to nitric oxide stress. Another two-component system, encoding PhoP and PhoR has been demonstrated to be essential for the virulence of *M. tb.* PhoP positively regulates the biosynthesis of virulence-associated lipids, and also affects the function of ESX-1 through the *espA-espC-espD* locus, which is crucial for ESX-1 function and virulence (Frigui *et al.* 2008). EspR is an important transcriptional regulator that upregulates ESX-1 activity by binding upstream of the *espA-espC-espD* locus (Raghavan *et al.* 2008). In addition, a number of transcription regulator genes have been shown to be expressed in *M. tb* upon infection.

Unknown & Conserved hypotheticals

Due to the increasing availability of complete mycobacterial genome sequences and scientific literature mining, the proportion of unknown CDSs and conserved hypothetical proteins in the *M. tb* genome has been steadily decreasing. However, approximately one quarter of the CDSs still fall into these two categories. The 'Conserved hypothetical proteins' category refers to CDSs that are conserved in closely related species of the MTBC, or in other mycobacteria, such as, *M. smegmatis* and *M. marinum*, that are not members of the MTBC. The 'Unknown' category refers to genes with no known orthologues and includes only 16 CDSs, at present.

2.2.3 Comparative and Functional Genomics of the *M. tb* complex

The availability of complete genome sequences of *M. tb* H37Rv and related strains combined with advances in high-throughput experimental technologies have led to tremendous advances in our knowledge of the evolution, population biology, and virulence of *M. tb,* which has been applied to the development of new diagnostic tests, better drugs and vaccines against TB.

As of August 2012, the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/) and the European Bioinformatics Institute (http://www.ebi.ac.uk/) databases hold complete genome sequences for 48 mycobacterial strains and partial sequences (scaffolds/contigs) for 33 strains. Whole genome sequences for several clinical and laboratory strains of *M. tb* and other members of the MTBC are available (Table 2).

Table 2. List of completely sequenced genomes of MTBC members

Organism	Genome size (MB)	No. of Genes
Mycobacterium tuberculosis str. Erdman ATCC 35801	4.39	4,246
Mycobacterium tuberculosis CDC5079	4.40	3,695
Mycobacterium tuberculosis CDC5180	4.41	3,638
Mycobacterium tuberculosis CDC1551	4.40	4,293
Mycobacterium tuberculosis CTRI-2	4.40	4,001
Mycobacterium tuberculosis F11	4.42	3,998
Mycobacterium tuberculosis H37Ra	4.42	4,084
Mycobacterium tuberculosis H37Rv	4.41	4,062
Mycobacterium tuberculosis KZN 605	4.40	4,071
Mycobacterium tuberculosis KZN 1435	4.40	4,107
Mycobacterium tuberculosis KZN 4207	4.39	4,191
Mycobacterium tuberculosis RGTB327	4.38	3,739
Mycobacterium tuberculosis RGTB423	4.41	3,670
Mycobacterium tuberculosis UT205	4.42	3,814
Mycobacterium bovis AF2122/97	4.35	4,001
Mycobacterium bovis BCG str. Mexico	4.35	4,031
Mycobacterium bovis BCG str. Pasteur 1173P2	4.37	4,033
Mycobacterium bovis BCG str. Tokyo 172	4.37	4,027
Mycobacterium africanum GM041182	4.39	3,983

(Source: NCBI and EBI databases)

2.2.3.1 Genotyping of the *M. tb* complex

Discrimination between strains of pathogenic bacteria is crucial from an epidemiological perspective. In the 1990s, advances in the field of molecular biology led to the development of DNA-fingerprinting methods, which relied on strain-specific molecular markers, in order to identify and differentiate between strains of the MTBC. Molecular 'strain typing' or genotyping tools have been used to study disease outbreaks and transmission (Small *et al.* 1994), to assess the relative contributions of reinfection and reactivation to disease (van Rie *et al.* 1999), and to address a number of other issues relevant to TB control. Several countries routinely collect genotyping data as part of their TB surveillance programs. Three of the most commonly used genotyping methods are described below, and the genetic elements that are used as strain-specific markers in these methods are summarized in Figure 6.

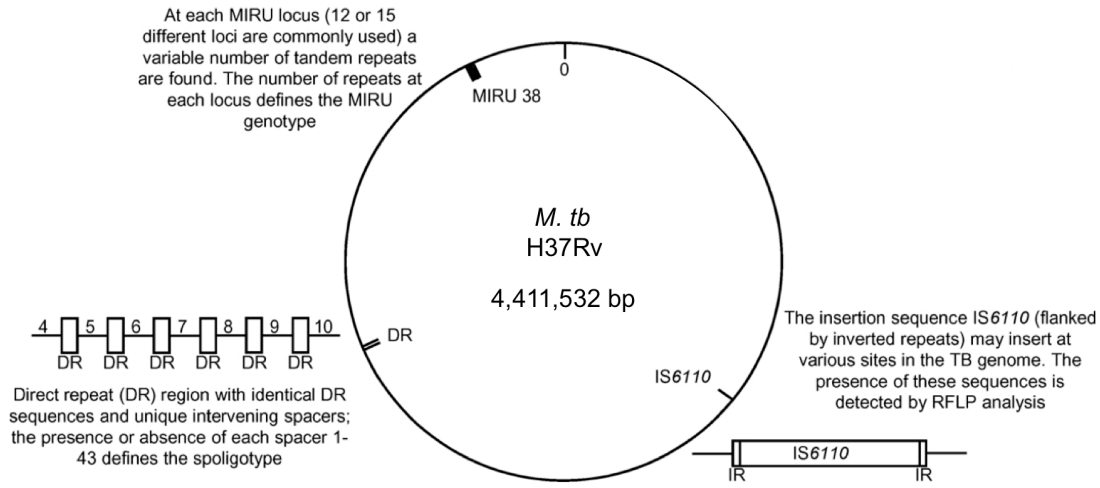


Figure 6. Schematic representation of the genetic basis of genotyping techniques.
Modified from (Nicol & Wilkinson 2008)) The circular chromosome of the reference strain *M. tb* H37Rv is shown together with examples of major genetic elements used for strain genotyping. For clarity, only one mycobacterial interspersed repetitive unit (MIRU) locus and one insertion sequence (IS*6110*) are shown.

IS*6110*-based RFLP and ligation-mediated PCR

The repetitive, mobile insertion sequence IS*6110* is present in the genomes of the MTBC in varying numbers of copies (Figure 6). Most members carry between 5-20 copies at different positions, although some strains with one or zero IS*6110* copies have also been identified. The IS*6110*-based restriction fragment length polymorphism (RFLP) technique makes use of the variation in the number and genomic position of IS*6110* elements to generate strain-specific hybridization patterns (van Embden *et al.* 1993). A major drawback of the RFLP procedure is that it is the time-consuming and expensive. The preferred alternative is ligation-mediated PCR which uses one primer specific for IS*6110* and a second specific for a linker ligated to SalI-restricted genomic DNA, making it faster and cheaper compared to RFLP (Prod'hom *et al.* 1997).

MIRU-VNTR typing

Multiple loci on the genome of MTBC strains contain variable numbers of tandem repeats (VNTRs) known as the mycobacterial interspersed repetitive units (MIRUs) (Supply *et al.* 1997). MIRU-typing is based on the polymorphisms observed in 41 VNTRs ranging from 40-100 bp dispersed on the genome of the MTBC (Figure 6). The number of repeats observed at 12, 15 or 24 selected MIRU loci can be used for genotyping using a PCR-based method (Supply *et al.* 2000). As the length of the repeats is known, the number of repeats at each locus is combined to generate a unique numerical code. The discriminatory power of MIRU-VNTR analysis is typically proportional to the number of loci evaluated. The 24-locus system of tandem repeats has been proposed as a high-resolution tool to capture the genetic diversity of strains (Supply *et al.* 2006).

Spoligotyping

MTBC strains contain a genomic region known as the Direct Repeat (DR) region that consists of 36 bp repeats interspersed by 43 unique spacer DNA sequences of 35-41 bp (Groenen *et al.* 1993) (Figure 6). The DR locus is a representative of the Clustered Regulatory Short Palindromic Repeats (CRISPR) family, which has been shown to confer resistance against viruses and bacteriophages in bacteria and archaea (Sorek *et al.* 2008). In the MTBC strains, loss of spacers can occur due to homologous recombination or transposition of IS*6110* insertion sequences (Legrand *et al.* 2001). Spacer oligonucleotide typing or spoligotyping is a PCR-based reverse hybridization technique that exploits the variation in the spacers within the conserved DRs in order to differentiate MTBC strains (Kamerbeek *et al.* 1997). The results are represented as a 43-digit binary string constructed on the basis of the presence or absence of spacers. Spoligotyping has low discriminatory power compared to the two other typing methods, but it is widely used because it is easy to perform, and highly reproducible.

Centralized databases have been constructed to store DNA fingerprinting data for TB using a standardized nomenclature in order to facilitate comparison of data from different epidemiological studies and enable researchers to analyze the genetic diversity and biogeographic distribution of MTBC strains across the world (Shabbeer *et al.* 2012). The two main repositories of *M. tb* genotyping data are, SpolDB4, which contains close to 40,000 spoligo patterns for MTBC isolates from more than 120 countries (Brudey *et al.* 2006), and MIRU-VNTR*plus*, which is a highly curated database containing detailed profiles (IS*6110* RFLP fingerprint, spoligotype, 24-locus MIRU profile, etc.) of 186 strains representing all MTBC lineages (Allix-Béguec *et al.* 2008).

It has been shown that different strains of *M. tb* have distinctive epidemiological and clinical characteristics in terms of pathogenicity, clinical representation and varying behaviour in animal models (Lopez *et al.* 2003). Given the importance of pathogen variation in the outcome of infection and disease, understanding the evolution of a pathogen can allow for better epidemiological predictions. Genotyping methods have a propensity for convergent evolution which limits their use for phylogenetics and strain classification (Comas *et al.* 2009). Availability of the whole genome sequence of *M. tb* H37Rv and the development of DNA microarray technology offered new opportunities to study strain diversity in the MTBC by comparative genomics (S. V. Gordon *et al.* 1999; Behr *et al.* 1999; S. T. Cole 2002b).

2.2.3.2 Genetic diversity in the *M. tb* complex

Whole genome comparison allowed us to infer the impact of processes such as mutation, deletion, duplication and selection on the evolution of the MTBC. Soon after the completion of the genome of *M. tb* H37Rv (H37Rv), sequences of the clinical strain *M. tb* CDC1551 (CDC1551) (Fleischmann *et al.* 2002), and the causative agent of bovine TB, *M. bovis* AF2122/97 (Garnier 2003) became available. DNA sequence data can also be exploited to study the nature and strength of selective forces shaping the genetic diversity within populations. Proteins are generally highly conserved and therefore amino acid substitutions are primarily driven by positive functional selection. Identification of proteins under such selective pressure may lead to unravelling of important virulence mechanisms as well as the genetic events that lead to variable forms of disease caused by the same organism.

Most of the comparative genomics studies have been performed on the canonical laboratory strains H37Rv, H37Ra, and *M. tb* Erdman, the reference clinical strains CDC1551 and *M. tb* HN878, and the vaccine strain *M. bovis* BCG. Genome comparison revealed that members of the MTBC share 99.9% identity at the DNA level as well as identical 16s rRNA (Brosch *et al.* 2002; Fleischmann *et al.* 2002). Single nucleotide polymorphisms (SNPs) and large sequence polymorphisms (LSPs), such as chromosomal deletions, are the main source of inter-strain variability.

The CDC1551 strain, is a highly transmissible clinical isolate that caused a TB outbreak in a rural area of the United States in the 1990s (Valway *et al.* 1998). The strain was capable of inducing greater immunoreactivity, and was less virulent compared to *M. tb* H37Rv in animal models (Manca *et al.* 2001). The complete genome of CDC1551 was sequenced in order to identify

polymorphisms with potential relevance to disease, pathogenesis, immunity and evolution. Comparison of H37Rv and CDC1551 revealed 86 InDels (longer than 10 bp) and 1075 SNPs, of which 579 were nonsynonymous, presenting important leads to associate genotypic changes with phenotypic differences (Fleischmann *et al.* 2002).

M. tb H37Ra is the avirulent counterpart of the virulent strain H37Rv and both strains were derived from their virulent parent H37, which was originally isolated from a patient with chronic pulmonary TB (Brosch *et al.* 1999). The H37Ra genome is highly similar to that of H37Rv with only 198 SNPs that occur between H37Ra and H37Rv (Zheng *et al.* 2008). In fact, 119 of them are identical between H37Ra and CDC1551 and 3 are due to H37Rv strain variation, leaving only 76 H37Ra-specific SNPs that affect 32 genes (Garnier 2003). It was later shown that the attenuation in virulence in H37Ra was primarily due to a nonsynonymous mutation in the *phoP* gene leading to the inactivation of the positive transcriptional regulator PhoP (Frigui *et al.* 2008).

The genome sequence of *M. bovis* was found to be 99.95% identical to the genomes of H37Rv and CDC1551, but the genome size was slightly smaller due to a number of deletions. Direct comparison of 2,504 coding sequences (CDS) across the three genomes revealed that approximately 1600 *M. bovis* CDS are identical to H37Rv and CDC1551 respectively. There were about 2400 SNPs between *M. bovis* and the two *M. tb* strains, of which roughly 800 were nsSNPs (Fleischmann *et al.* 2002). This analysis highlighted the conservation of gene sequence across the MTBC, and also the divergence of *M. bovis* and *M. tb.*

Comparison of virulent and avirulent or attenuated MTBC strains made it possible to identify genetic differences that might explain the molecular mechanisms of pathogenicity. Comparison of genome sequences of the vaccine strain *M. bovis* BCG Pasteur 1173P2 (BCG Pasteur) with H37Rv, CDC1551, and *M. bovis* AF2122/97 uncovered LSPs that led to the loss of 133 genes in BCG Pasteur (Behr *et al.* 1999; Brosch *et al.* 2007). There were only 736 SNPs between the two *M. bovis* isolates, and around 2400 SNPs between BCG and *M. tb* strains.

Further insights on the virulence determinants came from comparison of *M. tb* with *M. leprae,* the causative agent of leprosy. *M. leprae* is an obligate intracellular pathogen that essentially displays the same cellular tropism and host range as the tubercle bacillus. The *M. leprae* genome has undergone massive gene decay as a result of extreme reductive evolution, leaving only 1604 functional protein-coding genes in the leprosy bacillus (S. T. Cole *et al.* 2001). The minimal gene set retained by *M. leprae* despite the massive gene decay suggests that these gene functions are essential for *M. leprae* and possibly for other pathogenic mycobacteria. *In silico* comparisons among the 1439 genes common to *M. tb* and *M. leprae* revealed a set of 219 "core" genes unique to mycobacteria making them potential targets for developing specific anti-mycobacterial drugs (Marmiesse *et al.* 2004).

The MTBC is considered to be clonal and there is little evidence for on-going horizontal transfer in members of this complex (Supply *et al.* 2003; Smith *et al.* 2003; Hirsh *et al.* 2004), therefore variations occur almost exclusively by genomic mutations. SNPs occur at a very low rate of 1 per 2000-4000bp in the MTBC. LSPs represent irreversible genetic events in the evolution of the MTBC. Each strain has roughly 200 intergenic SNPs that can potentially

alter gene expression, 500 genes that are affected by at least one nsSNP, and 100 genes that are lost or inactivated due to LSPs (Fleischmann *et al.* 2002; Tsolaki *et al.* 2004; Ernst *et al.* 2007).

There is clear evidence that SNPs and LSPs both contribute to strain variability, for example, most of the drug-resistance mutations are due to SNPs (Y. Zhang *et al.* 1992; Ramaswamy *et al.* 2000; Cavusoglu *et al.* 2002), and LSPs have been shown to be responsible for variability in virulence of *M. tb* strains (Marmiesse *et al.* 2004). SNPs are distributed throughout the genome and have a low reverse mutation rate and homoplasy index, which makes them ideal to investigate genetic diversity (Comas *et al.* 2009). In a study of genomic deletions within clinical isolates of *M. tb*, mycobacterial clones shared the same genomic deletions (LSPs), suggesting that deletions could be used as evolutionary marker (Mostowy *et al.* 2002).

2.2.3.3 Evolutionary scenario for the *M. tb* complex

Studies have made use of LSPs, together with SNPs and repetitive element patterns (spoligotypes) to decipher the genealogy of the MTBC (Ernst *et al.* 2007; Brosch *et al.* 2002; Mostowy *et al.* 2002; Kamerbeek *et al.* 1997). In this approach, the successive and unidirectional loss of DNA in representative strains reveals the order in which members of the complex descended from their ancient common ancestor.

Large sequence polymorphisms (LSPs) between MTBC members were first identified using techniques such as subtractive hybridization (Mahairas *et al.* 1996), bacterial artificial chromosome (BAC) arrays (Brosch *et al.* 1998; S. V. Gordon *et al.* 1999), microarrays (Behr *et al.* 1999), and pulsed-field gel electrophoresis (S. V. Gordon *et al.* 1999; Brosch *et al.* 1999). These studies uncovered 14 regions of difference (RD1 to RD14), ranging in size from 2 to 12.7 kb, that were absent from *M. bovis* BCG strains relative to the *M. tb* H37Rv genome. The RD1 region, in particular, is absent from all BCG strains, but present in all members of the MTBC. It was later shown to be the primary attenuation event in the derivation of *M. bovis* BCG from *M. bovis* (Pym *et al.* 2002). Comparison of other members of the MTBC with *M. tb* H37Rv revealed six regions, including, five H37Rv-specific deletions (RvD1-RvD5) and one *M.tb* specific deletion (TbD1). The TbD1 region, which is characterized by the absence of a 2 kb fragment truncating genes *mmpS6* and *mmpL6*, is the only genomic region present in *M. bovis* and absent from the *M. tb* strains. Based on sequence comparison of 26 structural genes across over 800 isolates from the MTBC, Sreevatsan *et al.* identified two high frequency polymorphisms in the genes *katG* and *gyrA* that could be used to classify MTBC strains into three principal genetic groups (PGGs) (Sreevatsan *et al.* 1997). Brosch *et al.* examined the distribution of 20 variable regions (14

RD, 5 RvD, TbD1) across 100 diverse strains belonging the MTBC. In addition, sequences of five genes, namely, *oxyR*, *pncA*, *katG*, *gyrA* and *mmpl6*, were also analyzed for all the strains as these genes had been shown to vary between different tubercle bacilli (Sreevatsan *et al.* 1996; Scorpio *et al.* 1997; Sreevatsan *et al.* 1997; Brosch *et al.* 2002).

An evolutionary pathway of the tubercle bacilli was proposed based on the presence or absence of conserved deleted regions and sequence polymorphisms in the five genes (Figure 7). The *M. tb* strains can be divided into two groups, "ancestral" and "modern", based on the presence or absence of the TbD1 region. All strains of the PGG-2 and PGG-3 as defined by Sreevatsan *et al.* belong to "modern" *M. tb* strains. The loss of RD9 corresponds to a major divergence leading to a main lineage encompassing *M. africanum* and all animal-adapted members of the MTBC. Successive deletions led to *M. microti, M. pinnipedii, M. caprae, M. bovis* and eventually, *M. bovis* BCG. The evolutionary scheme was refined later on to include another molecular marker in the *pks15/1* gene which is responsible for the production of an immunomodulatory phenolic glycolipid (PGL) (Marmiesse *et al.* 2004). A 7 bp deletion, inactivating the *pks15/1* gene was observed in all "modern" *M. tb* strains, whereas a 6 bp deletion that had on the function of the *pks15/1* gene occurred in the *M. africanum* → *M. bovis* lineage. The 6 bp deletion occurred after the RD9 and before RD10 deletion suggesting that it arose independently from the 7 bp deletion. It was also shown that spoligotypes of the RD9 deleted lineage could also be used as phylogenetic markers, and the resulting phylogeny was congruent to the one based on LSPs, suggesting that recombination is rare or absent between strains of this lineage (Smith *et al.* 2006).

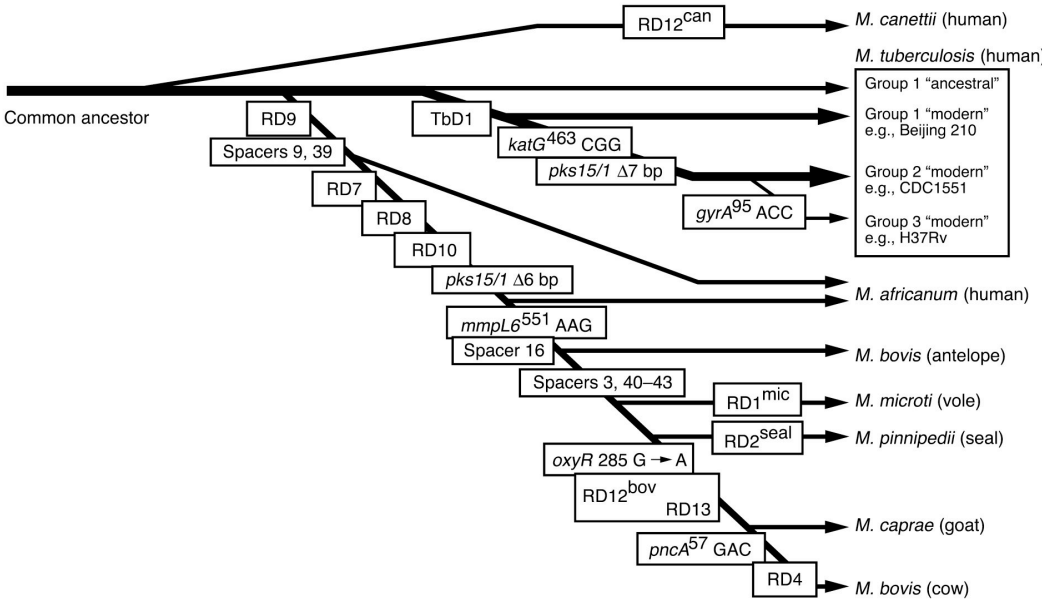


Figure 7. New evolutionary scheme of the MTBC from Ernst *et al* (Ernst *et al.* 2007).

This is a refined evolutionary scheme after Brosch *et al.*, based on informative markers, such as RD regions, SNPs, and spoligotype spacers (Brosch *et al.* 2002; Marmiesse *et al.* 2004; Smith *et al.* 2006).

The evolutionary scheme highlighted that *M. tb* and *M. bovis,* as well as other members of the MTBC have evolved from a common ancestor. Additionally, they confirmed that members of the MTBC show a clonal population structure, with very little horizontal gene transfer.

The use of regions of difference (RDs), SNPs, and strain-specific deletions and duplications, also allowed the construction of the BCG genealogy, which classified the BCG strains into four major groups (Figure 8). After the original BCG vaccine was derived in 1921, it was distributed to different laboratories across the world, which maintained their own daughter strains by passaging, until the introduction of archival seed lots in the 1960s. Comparative genomics uncovered regions of difference and SNPs between the daughter strains, dividing them into early strains represented by BCG Japan, Birkhaug, Sweden and Russia, and late strains, including BCGs Pasteur, Danish, Glaxo and Prague (Mahairas *et al.* 1996; S. V. Gordon *et al.* 1999;

Mostowy 2003). The most prominent genomic polymorphisms in the *M. bovis* BCG Pasteur strain were two large tandem duplications, DU1 (≈ 26 kb) and DU2 (≈ 36 kb), making it partially diploid for 58 genes (Brosch *et al.* 2007). Sequence comparison of *M. bovis* BCG Pasteur with other BCG strains, showed that, DU1 is restricted to BCG Pasteur, while four forms of DU2 existed, DU2-1 is confined to early BCG vaccines, whereas DU2-III and DU2-IV occur in late vaccines (Brosch *et al.* 2007).

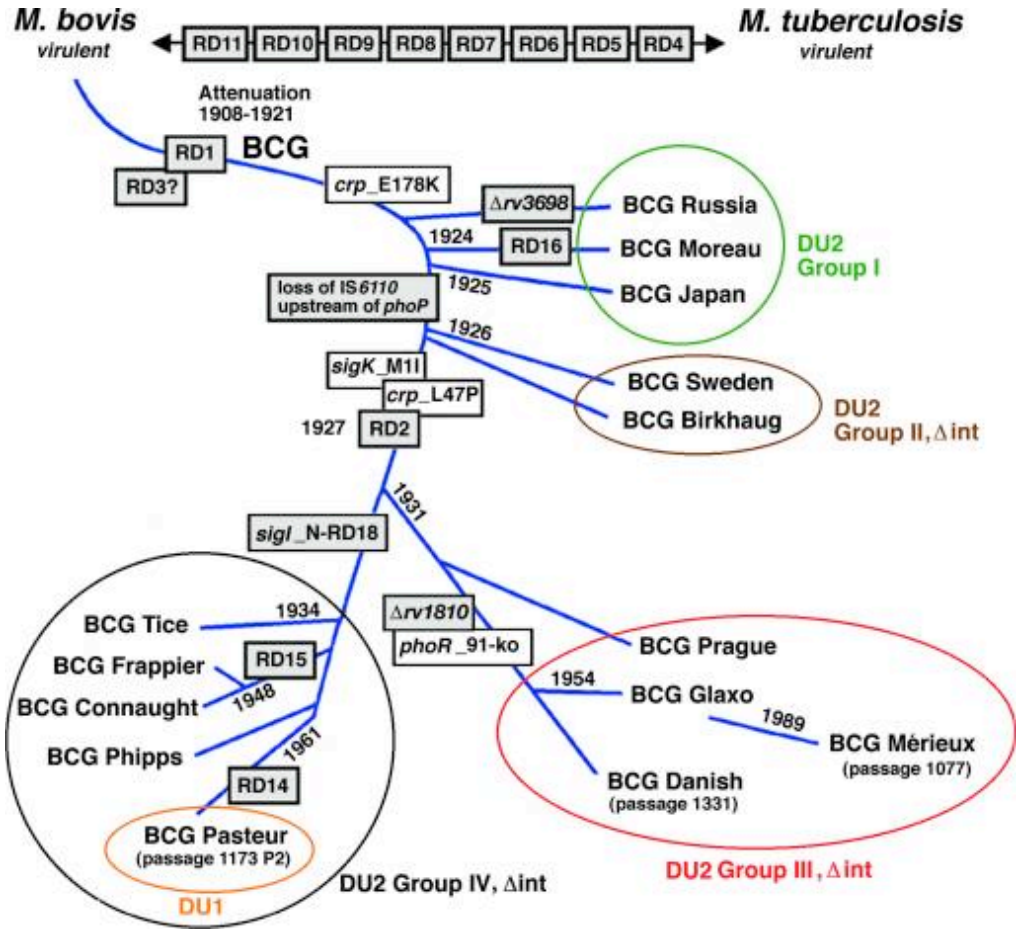


Figure 8. Refined genealogy of BCG vaccines (Brosch *et al.* 2002).
The scheme shows the position of genetic markers identified in this work, RD markers, some strain-specific deletions, SNPs, and the distribution of vaccines into the four groups.

A number of successive studies employed SNPs and LSPs as phylogenetically informative markers to infer a robust *M. tb* phylogeny from analysis of globally sampled *M. tb* isolates, and obtained comparable results. Knowledge

of the genetic population structure can yield an association between *M. tb* lineages and important traits such as clinical phenotypes, transmissibility, likelihood of drug resistance, and adaptation to distinct human populations.

Baker *et al.* investigated SNPs in seven housekeeping genes (*katG, gyrA, rpoB, oxyR, ahpC, pncA*, and *rpsL*), and identified 36 synonymous SNPs, which allowed them to construct a phylogenetic tree that divided the strains into four distinct groups (I – IV) (Baker *et al.* 2004). Gutacker *et al.* used *insilico* whole genome comparisons to identify 36 synonymous SNPs in 5069 clinical isolates, and defined nine main lineages (Gutacker *et al.* 2002). Filliol *et al.* used the same strategy to identify 159 synonymous SNPs in 219 *M. tb* and *M. bovis* isolates, which led to ten major groupings, including a lineage specific to *M. bovis* (Filliol *et al.* 2006). Gagneux *et al.* obtained similar findings by using phylogenetically informative LSPs to screen a global sample of 857 clinical isolates of *M. tb* from 80 countries (Gagneux *et al.* 2006).

In addition to defining a robust phylogeny, these studies also demonstrated a strong association between the phylogenetic lineages and particular geographical regions. Given the clonal population structure of the MTBC, the selectively neutral sequence variation (sSNPs) was chromosomally linked to the selected genetic variation (nsSNPs), resulting in perfectly congruent phylogenetic trees using sSNPs and nsSNPs (Gutacker *et al.* 2002).

Evaluation of the global phylogenies inferred from the aforementioned studies, together with the analysis of the large spoligotyping database (Brudey *et al.* 2006) led to the identification of six main strain lineages of *M. tb* and *M. africanum* that are associated with particular geographic regions (Table 2).

Table 2. Comparison of terminology and molecular markers for the six main lineages of *M. tb* and *M. africanum* strains.

Criteria	Lineage 1	Lineage 2	Lineage 3	Lineage 4	Lineage 5	Lineage 6
SNP Sreevatsan *et al.*	PGG 1	PGG 1	PGGs 2 & 3	PGG 1	PGG 1	PGG 1
SNP Baker *et al.*	Lineage IV	Lineage I	Lineage III	Lineage II	Not done	Not done
LSP Gagneux *et al.*	Indo-Oceanic lineage	East Asian lineage	East African-Indian lineage	Euro-American lineage	West African lineage I	West African lineage II
SNP Gutacker *et al.*	Cluster I	Cluster II	Cluster II.A	Cluster III-VII	Not done	Not done
SNP Filliol *et al.*	Cluster group 1	Cluster group 2	Cluster group 3a	Cluster group 3b-6b	Not done	Not done
Spoligotyping Brudey *et al.*	EAI	Beijing	CAS	Haarlem, LAM, T, X	AFRI2	AFRI1
Geographical Association	East Africa, Southeast Asia, South India	East Asia, Russia, South Africa	East Africa, North India, Pakistan	Americas, Europe, North Africa, Middle east	Ghana, Benin, Nigeria, Cameroon	Senegal, Guinea-Bissau, The Gambia

Hershberg *et al.* (2008) performed an in-depth survey of the genetic diversity in MTBC by *de novo* sequencing 89 genes in a global collection of 108 human-adapted MTBC strains. This analysis refined the previous MTBC phylogeny and also suggested an association between the genetic diversity in human-adapted MTBC strains and ancient human migrations out of Africa, as well as more recent population movements in the past few hundred years.

Deeper insight into the MTBC phylogeny has been obtained from next-generation sequencing (NGS) studies. Comas *et al.* (2010) carried out whole genome re-sequencing of 20 strains representative of the six *M. tb* lineages, and established a comprehensive MTBC phylogeny based on 9,037 variable common nucleotide positions. Schürch *et al.* (2010) successfully reconstructed the phylogeny of the *M. tb* Beijing genotype family based on NGS data from six Beijing strains. The impact of NGS on *M. tb* research, especially in the areas of molecular epidemiology and phylogenetic analysis has been elaborated in Section 2.2.4.

2.2.3.4 Overview of functional genomics

Comparative genomics has provided us with a catalogue of genomic diversity in the form of mutations, deletions, duplications and their impact on the evolution of the MTBC. In particular, the identification of genes or loci that differ between virulent and avirulent or attenuated strains of *M. tb* can allow us to identify the molecular mechanisms of pathogenicity, and also provide leads for the development of new therapies for the disease. The aim of functional genomics is to understand the functional consequences of the genetic diversity at the levels of genes, transcripts and proteins using a genome-wide approach. This section will highlight the functional genomic approaches used for studying gene expression and transcription regulation in the MTBC and their contribution to improving our understanding of *M. tb* biology and pathogenesis.

The development of DNA microarray technology was a major breakthrough in the study of gene expression and regulation. It enabled researchers to move beyond the study of individual genes and proteins to the functioning of the entire system. Microarrays have been a very important tool for identification of genes expressed under a variety of growth conditions as well as for elucidation of the function of regulatory proteins by mutation analysis.

Genome-wide expression profiling of *M. tb* has been used to address the basic question – how does *M. tb* survive and thrive in the host? Typically, expression analysis compares the level of RNA transcripts in two pools of bacteria grown in different conditions by hybridization to microarrays or takes advantage of isogenic mutant and wild-type strains. The rationale behind studying the changes in gene expression assumes that the changes reflect the importance of genes in responding to the environment. Also, genes

that are specifically expressed during infection or co-regulated with known virulence genes are likely to be candidate virulence genes.

Another application of microarrays is to characterize the role of transcription factors in gene regulation. Traditionally this was done by comparing expression profiles for wild-type and regulatory mutant strains. However, this approach is unable to distinguish between primary and secondary regulatory effects, in cases where one regulator controls the expression of the gene for another. A powerful alternative was introduced by combining chromatin immunoprecipitation with DNA microarrays, known as the ChIP-on-chip technology. In this approach, proteins in contact with DNA are chemically cross-linked and the DNA is fragmented by sonication or enzymatic digestion. The cross-linked proteins are immunoprecipitated with antibodies specific for the proteins of interest. The immunoprecipitated DNA is then reverse cross-linked, purified, labelled, and hybridized to microarrays in order to detect enrichment of the sequences bound by the proteins.

The pathogenesis of *M. tb* involves a complex interaction with the host. In order to establish a successful infection the pathogen must adapt and survive the plethora of stresses encountered in the host by fine-tuning gene expression. Indeed, *M. tb* has an impressive repertoire of transcription regulators, including 13 sigma factors that recognize different classes of promoter and nearly 200 regulatory proteins that respond to environmental cues or physiological changes. Many of these environmental cues are relayed via reversible protein phosphorylation cascades catalysed by serine-threonine protein kinases or two-component systems (S. Cole *et al.* 1998).

Initial transcriptomic studies focused on capturing the transcription in *in vitro* conditions that might mimic *in vivo* situations. Hypoxia is thought to be a key factor in triggering latency *in vivo.* Microarray studies identified over 100 genes whose expression is rapidly altered by defined hypoxic conditions (Sherman *et al.* 2001). A similar approach was used to identify genes affected by nitric oxide stress, which is known to inhibit bacterial respiration (Voskuil *et al.* 2003). Interestingly, the set of genes found to be induced by hypoxia as well as nitric oxide stress were later shown to be regulated by DosR (Dormancy survival regulatory), and are collectively known as the DosR or dormancy regulon. A combination of microarrays and proteomics was used to investigate the response of *M. tb* to nutrient starvation. This provided evidence for slowdown of the transcription apparatus, energy metabolism, lipid biosynthesis, and cell division in addition to the induction of genes that may play a role in maintaining long-term survival of *M. tb* (Betts *et al.* 2002). The first *in vivo* microarray-based analysis of the *M. tb* transcriptome identified genes that are differentially expressed in naive and activated macrophage phagosomes in mice. The findings indicate that the intraphagosomal environment is rich in lipids, poor in iron and carbohydrates, and a strong source of oxidative and nitrosative stress (Schnappinger *et al.* 2003).

Genome-wide expression profiling has also been used to study drug-induced alteration of the *M. tb* transcriptome. Some of the drugs that have been studied in this manner include isoniazid, thiolactomycin, triclosan, and tetrahydrolipstatin (M. Wilson *et al.* 1999; Betts *et al.* 2003; Waddell *et al.* 2004). Another extensive study used genome-wide expression profiling to map adaptability of *M. tb* to diverse environmental changes (different drug types) that interrupt metabolism in different ways (Boshoff *et al.* 2004). It was

observed that distinct transcriptional signatures were generated upon treatment with particular drug types, which can be used to identify the target of an inhibitor whose mode of action is unknown.

Sassetti and colleagues developed a novel use for microarrays to identify essential genes in *M. tb* at the whole genome level. The method involved construction of transposon mutant libraries, followed by the transposon site hybridization (TraSH) to map sites of transposon insertions and identify genes required for growth. They reported that 614 *M. tb* genes are required *in vitro,* and 194 are required for *in vivo* growth in the mouse model (Sassetti *et al.* 2003).

A number of transcriptional regulators have been studied using microarrays. Regulatory mutants for the iron-dependent repressor protein, IdeR, the DNA repair protein (Rodriguez *et al.* 2002) and DosR, the hypoxic response regulator (H.-D. Park *et al.* 2003), have been used to define regulons for these proteins. The so-called dormancy regulon comprises 48 genes controlled by DosR, which binds to DNA following phosphorylation by one of two cognate sensor kinases (Roberts *et al.* 2004; Saini *et al.* 2004). A number of sigma factors have also been implicated in virulence based on microarray analysis. For example, SigE has been shown to contribute to pathogen survival following severe distinct stresses. SigE regulates the expression of genes involved in maintenance of the cell wall, and was shown to be upregulated in *M. tb* after phagocytosis (Manganelli *et al.* 2001; Schnappinger *et al.* 2003).

Recently, the ChIP-on-chip technology has been used to define the regulons of several transcription regulators in *M. tb,* including, BlaI, which responds to β-lactam antibiotics (Sala *et al.* 2009), Lsr2, which is a nucleoid-associated

protein that regulates *M. tb* virulence (B. R. G. Gordon *et al.* 2010), and the stationary phase sigma factor, SigF (Hartkoorn *et al.* 2012).

2.2.4 Impact of high-throughput sequencing on *M. tb* research

Technological breakthroughs in the field of sequencing in the last few years have transformed biological research. These technologies are collectively referred to as 'next-generation' sequencing (NGS) technologies, which perform high-throughput sequencing to generate whole genome data at an unprecedented scale and speed. Sequencing based genomics has several advantages compared to microarray analysis:

1) Microarrays require *a priori* knowledge of the genome, whereas NGS data can also be assembled *de novo.*

2) Cross-hybridization between similar sequences restricts the use of microarray analysis to non-repetitive fractions of the genome. NGS offers single-nucleotide resolution and demonstrates a greater ability to distinguish between isoforms, and reveal sequence variants.

3) Competitive hybridization on microarrays is a relative measure, while quantification of signal from NGS approaches is a direct measure, based on counting sequence tags, making it better for detection of transcripts produced at high as well as low levels.

4) Microarrays require more starting material (micrograms) compared to high throughput sequencing (nanograms).

The massive sequence output, relatively low cost per base considering the small genome size of *M. tb*, and the ability to generate large quantities of data make NGS an attractive choice for re-sequencing microbial genomes. The major applications of NGS in *M. tb* research have been in genome re-sequencing. NGS has allowed the detection of genetic diversity in the MTBC at an unprecedented resolution. NGS has been applied in studies on evolution, transmission and treatment of *M. tb*. SNP typing from NGS has been used to estimate selective pressures on antigens (Comas *et al.* 2010), to

characterize of MDR- and XDR-TB strains (Ioerger *et al.* 2009), to identify mutations responsible for drug resistance (Hartkoorn 2012), and to establish a refined phylogeny of closely related *M. tb* strains (Schürch 2011). Based on NGS data, Niemann *et al.* (2009) uncovered a significant amount of genetic diversity in drug sensitive and MDR isolates of *M. tb* exhibiting identical *IS6110-RFLP* patterns, which may have important clinical implications. In addition, NGS provides a powerful alternative to gene expression microarrays in the form of RNA-seq, which can be used for mapping and quantifying transcripts in biological samples. For detecting small RNAs, preferential isolation via size selection or small RNA-enrichment can be performed. After sequencing, reads are aligned to a reference genome, compared with known transcript sequences, or assembled *de novo* to construct a transcription map of the genome. RNA-seq has been successfully applied to identify non-coding and small RNAs in the *M. tb* transcriptome (Arnvig & Young 2009; Pellin *et al.* 2012).

An alternative to ChIP-on-chip technology is the ChIP-seq technology, where chromatin immunoprecipitated DNA is subjected to NGS sequencing instead of being hybridized to microarrays. The obtained reads can be mapped to the reference genome of interest to generate a high-resolution, genome-wide protein-binding map. The higher resolution offered by ChIP-seq is evident by its capacity to identify novel binding sites and confirm previously identified binding sites and sequence motifs. ChIP-seq has been used in *M. tb* to define the regulons, and study the genome-wide distribution of EspR (Blasco *et al.* 2012), which is a regulator of the ESX-1 secretion system, and LexA, which is a key regulator of the DNA damage (SOS) response (Smollett *et al.* 2012).

3. Overview of Methods

The overall objectives of this thesis were to analyse genome-wide datasets generated from the application of microarray and next-generation sequencing technologies in comparative and functional genomics of the *M. tb* complex. The work carried out as part of this thesis can be classified into three main research areas:

1) Comparative genomics

- Assessment of the genetic diversity in the *esx* gene family among *M. tb* clinical isolates from SNP data generated using automated Sanger sequencing.
- Analysis of SNP distribution in *M. bovis* and *M. bovis* BCG strains by SNP-genotyping and phylogenetic clustering.

2) Global transcription and its regulation

- Integration of ChIP-seq and RNA-seq data to study global transcription and dynamics of the transcription machinery in *M. tb*.
- Identification of genome-wide transcription factor binding sites using ChIP-seq and ChIP-on-chip data.

3) Whole-genome re-sequencing (High-throughput SNP typing)

- SNP and InDel (insertions and deletions) detection from whole-genome re-sequencing data to identify genetic mutations responsible for drug resistance in *M. tb*, and contributing to phenotypic differences between *M. tb* strains.

This chapter will focus on the microarray and next generation sequencing approaches used in this thesis. It includes a brief description of the nature of the data generated by these technologies followed by an overview of the bioinformatics strategies for data analysis.

3.1 Microarrays

A 'microarray' refers to a collection of thousands of probes or oligonucleotide fragments bound to a solid-surface support that hybridize to target molecules, which are fluorescently labelled. The result of successful hybridization between the labelled target and the immobilized probe is an increase of fluorescence intensity over a background level, which can be measured using a fluorescent scanner. Oligonucleotide microarrays usually contain probes for annotated genes, whereas tiling arrays consist of partially overlapping or non-overlapping probes that span the entire genome. Many commercial oligonucleotide arrays and custom printed cDNA microarrays are now available (Harshman & MARTÍNEZ-A 2002; Harrington *et al.* 2000).

As described in the previous chapters, microarrays have been widely used for expression profiling of *M. tb* and more recently in studying protein-DNA interactions (ChIP-on-chip). In cDNA microarrays, mRNA from biological samples is reverse-transcribed into cDNA and labelled separately with two fluorescent dyes Cy3 (green) and Cy5 (red). Upon hybridization of the cDNA to the array, the relative fluorescent intensities are measured and captured to produce a composite image. The image is then used to decipher the hybridization efficiency of each spot on the microarray, which correlates to the relative abundance of the target gene in the sample of interest (Trachtenberg *et al.* 2012).

Microarray analysis starts with the assessment and quantification of the image acquired after the reaction on the microarray surface. The first step involves use of image analysis software to assess the quality of the image acquired upon scanning of the hybridized microarray. Poor quality spots are flagged and removed in order to improve the accuracy of the subsequent

analysis. The images are then transformed into numerical values representing the Cy3 and Cy5 intensities for each spot. Most commercial microarray scanners provide software to handle image processing; there are several additional image-processing packages available (Harshman & MARTÍNEZ-A 2002; Quackenbush 2001; Harrington *et al.* 2000).

The next step is data normalization, which attempts to compensate for the systematic variation resulting from the technology rather than from biological differences in the samples. These include differences in labelling, detection, hybridization or washing procedures, and in the quantity of the samples examined (Trachtenberg *et al.* 2012; Smyth & Speed 2003; López-Campos *et al.* 2012). Several normalization strategies have been proposed for microarray data; these methods can be broken into two groups, linear and non-linear methods (D. L. Wilson *et al.* 2003).

1) Linear methods rely on the concept that the bias introduced during the experiments can be summarized by means of a "global" constant or scaling factor; therefore, the intensities are corrected for global multiplicative effects. There are different ways to calculate the global normalization constant (mean, median, Z-score), which is used to re-scale the intensity for each probe in the array (D. L. Wilson *et al.* 2003; López-Campos *et al.* 2012).

2) Non-linear methods are more commonly used when normalizing two-colour cDNA microarrays. This Loess/LOWESS normalization methods stem from the Magnitude versus Amplitude plot (MA plot), where M is the log intensity ratio (difference in expression values) and A is the average log intensity for each spot (Figure 9A). MA plots are used to visualize the intensity-dependent ratio of raw microarray data. A normalization curve is fitted to this MA plot

by using the Loess method which performs local regression. The fits based on the normalization curve are subsequently subtracted from the M values (Figure 9B). Loess/LOWESS normalization algorithms correct for sub-array spatial variation and for intensity-dependent artifacts (Allison *et al.* 2006).

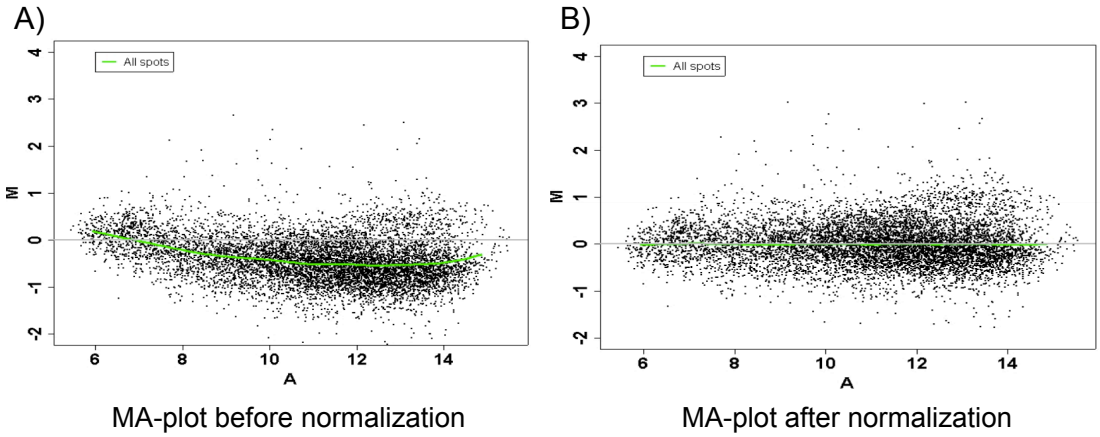


Figure 9. MA-plot of microarray data, A) before, and B) after LOESS normalization.

After normalization, the data are typically reported as an 'expression ratio' for microarrays or an 'enrichment ratio' for ChIP-on-chip. These ratios are usually log transformed (usually to the base 2), as it results in treating differential up-regulation and down-regulation equally, and also has a continuous mapping space ((Smyth & Speed 2003; M. Wilson *et al.* 1999).

The R/Bioconductor framework provides a lot of open source packages that implement most of the normalization methods (Gentleman *et al.* 2004). While preliminary data processing and standard normalization methods are common to most microarray datasets, the downstream analysis and data interpretation vary depending on the type of experiment, i.e. transcription profiling or ChIP-on-chip, as they answer different biological questions.

1) In the case of microarrays, the next step is to identify differentially expressed genes under two experimental conditions. Most published studies tend to use a post-normalization cut-off of two-fold increase or decrease, although there is no firm theoretical basis for selecting this level as significant (Harshman & MARTÍNEZ-A 2002; Quackenbush 2001; Harrington *et al.* 2000). There are several approaches for reporting the degree of reliability of the differential expression results. These include the t-test and its variants, which compare the difference in the mean expression levels between the two groups, taking into account the variability of the data (Trachtenberg *et al.* 2012; Pan 2002). Another widely used approach is to report the false discovery rate (FDR) (Quackenbush 2001; Benjamini & Hochberg 1995). It is an estimate of the fraction of false positives within the resulting differentially expressed elements. One of the most common software applications for FDR control is the SAM (statistical analysis of microarrays) tool (Smyth & Speed 2003; Tibshirani 2006; López-Campos *et al.* 2012).

2) In the case of ChIP-on-chip, the next step is peak detection and localization of binding sites. It is important to consider certain features of the ChIP-on-chip technique for effective data analysis. Prior to immunoprecipitation of enriched DNA, the ChIP procedure involves sonication, which generates DNA fragments ranging from 300 to 700 bp. Therefore, the sheared DNA fragments can bind to several adjacent probes on the microarray (average probe length is 60bp) introducing a positive correlation among probe hybridization densities. This feature is termed as the 'neighbour effect' and results in high enrichment ratios centered over the site of protein-DNA interaction and spanning several genomically adjacent probes (Figure 10).

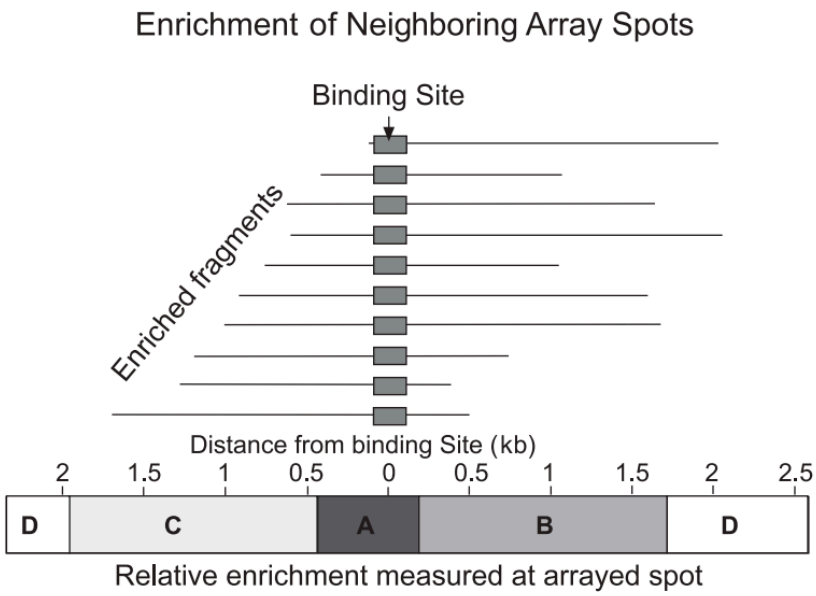


Figure 10. The neighbour effect in ChIP-chip data.
After IP enrichment, DNA fragments bound by the protein of interest will be of varying lengths. Spot 'A' contains the actual binding, as seen by the high enrichment ratio (black = high ratio, white = low ratio). Spots B and C, which are within 1 kb of the binding site will also be enriched. Spot B will have a higher Cy5/Cy3 ratio then spot C, since the binding site is closer to the B element. The two D spots are too far from the binding site to be enriched. (Buck & Lieb 2004).

Regions containing a cluster of peaks are more likely to be true binding sites than a single peak. Considering this characteristic of ChIP-chip data, a powerful method for peak detection is the sliding window approach. A window of selected size is slid across a region of the genome and the average $\log_2$ ratio of all the probes falling within the window is determined. The window is then moved downstream and the calculation is repeated iteratively for the entire length of the genome (Figure 11). A confidence value is assigned for each peak based on the number of independent probes used to construct the peak (D. L. Wilson *et al.* 2003; Buck & Lieb 2004; López-Campos *et al.* 2012). For functional interpretation of the ChIP-chip data, the predicted binding regions are linked to the genome annotation to identify the genes associated with the peaks (Figure 11).

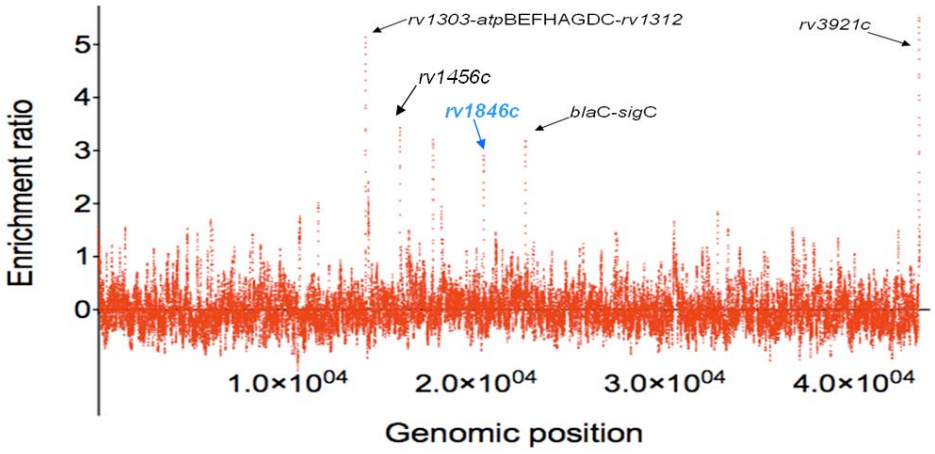


Figure 11. ChIP-on-chip data visualization.
ChIP-on-chip results for the BlaI (Rv1846c) protein of *M. tb* represented as a plot of the enrichment ratios for each probe calculated as sliding window averages against the genomic position of the probes (Sala *et al.* 2009). The genes showing high enrichment have been annotated.

Most transcription factors recognize specific DNA sequence patterns in their target sites, known as motifs. Comparison of DNA sequences retrieved from all the binding regions can help identify conserved motifs. Tools like MEME (Bailey *et al.* 2009), Gibbs Motif Sampler (Thompson *et al.* 2005) can be used for a *de novo* motif search. The genome-wide datasets be can visualized in the context of the genome annotation using stand-alone browsers such as Artemis (Carver *et al.* 2012) or online browsers like the UCSC genome browser (Dreszer *et al.* 2012).

3.2 Next-generation Sequencing

The wave of next generation sequencing technologies started with the development of the 454 pyrosequencing by Roche, followed by the Solexa sequencing-by-synthesis, which was commercialized by Illumina, and the ligation-based SOLiD sequencing by Applied Biosystems. The technologies from Roche and Illumina are most widely used, although new platforms are now emerging, such as single-molecule sequencing by Pacific Biosciences, and the Ion Torrent semiconductor-based pH sequencing by Life Technologies (Mardis 2008; Metzker 2009; Pareek *et al.* 2011) .

A common characteristic of all NGS platforms is that they do not rely on the Sanger sequencing chemistry. Instead they sequence clonally amplified DNA fragments or single DNA molecules in a flow cell in a massively parallel fashion. This is done using a step-wise iterative process or in a continuous real-time manner (Voelkerding *et al.* 2009). All sequencers produce megabases to gigabases of nucleotide sequence output in the form of reads: sequences of single-letter base calls plus a numeric quality score (Phred score) for each base call (Ewing & Green 1998). The major differences between these platforms lie in the specifics of the DNA sequencing reaction, the resulting read length, sequencing accuracy and throughput per run (Metzker 2009). A broad comparison of the NGS technologies is shown in Table 3.

Table 3. Comparison of NGS technologies by Dunne *et al* (Dunne *et al.* 2012).

Generation[a]	**Chemistry**	**Platform**	**Throughput per run (bases)**	**Error rate (Phred score)**	**Read length**
Second	Pyrosequencing	Life Sciences 454	450 Mb	Q30[b]	450
Second	Dye termination/ synthesis	Illumina Solexa	120,000 Mb	Q15[b]	2 x 150 paired
Second	Ligation	AB SOLiD	20,000	Q27[b]	50
Third	Semiconductor	Ion Torrent	10-100 Mb	Q15-Q20 est.	100
Third	Direct detection	Pacific Biosciences RS	75 Mb, estimate	Unknown	1,500 est.

[a]With "first generation" single-chain sequencing (Sanger dideoxy chain termination sequencing) as reference, "second generation" sequencing analyzes an ensemble of DNA molecules simultaneously by wash and scan techniques, and "third generation" sequencing interrogates molecules without the need to halt between read steps.

[b]Estimated Phred score, Q0−10 log10(P(≥n/s), where s is the observed signal and n is the length of the homopolymer that produced the signal, serves here as a common-denominator indicator of sequence quality that is familiar from low-throughput dideoxy or Sanger sequencing. A score of 20 is equivalent to an error rate of 1 in 100 bases.

All the NGS performed as part of this thesis made use of the Illumina sequencing technology (http://www.Illumina.com), which will be reviewed here in more detail. The Solexa/Illumina Genome Analyzer sequencing workflow (Figure 12) begins with fragmentation of genomic DNA and addition of adapter sequences to the ends of the fragments. The DNA molecules are attached to a slide and clonally amplified using the bridge amplification method. The sequencing reaction makes use of reversible terminator nucleotides, the fluorescently labelled nucleotides contain a termination moiety that prevents addition of other nucleotides to the synthesized strand such that only one nucleotide incorporation event occurs per fragment population per sequencing cycle. Post-incorporation fluorescence is recorded at the end of each cycle. This method generates short sequencing

reads of 36 – 100 bp, and an overall sequence output of around one billion base pairs per analytical run. The majority of published NGS papers to date have described methods using the short sequence data, produced with the Genome Analyzer.

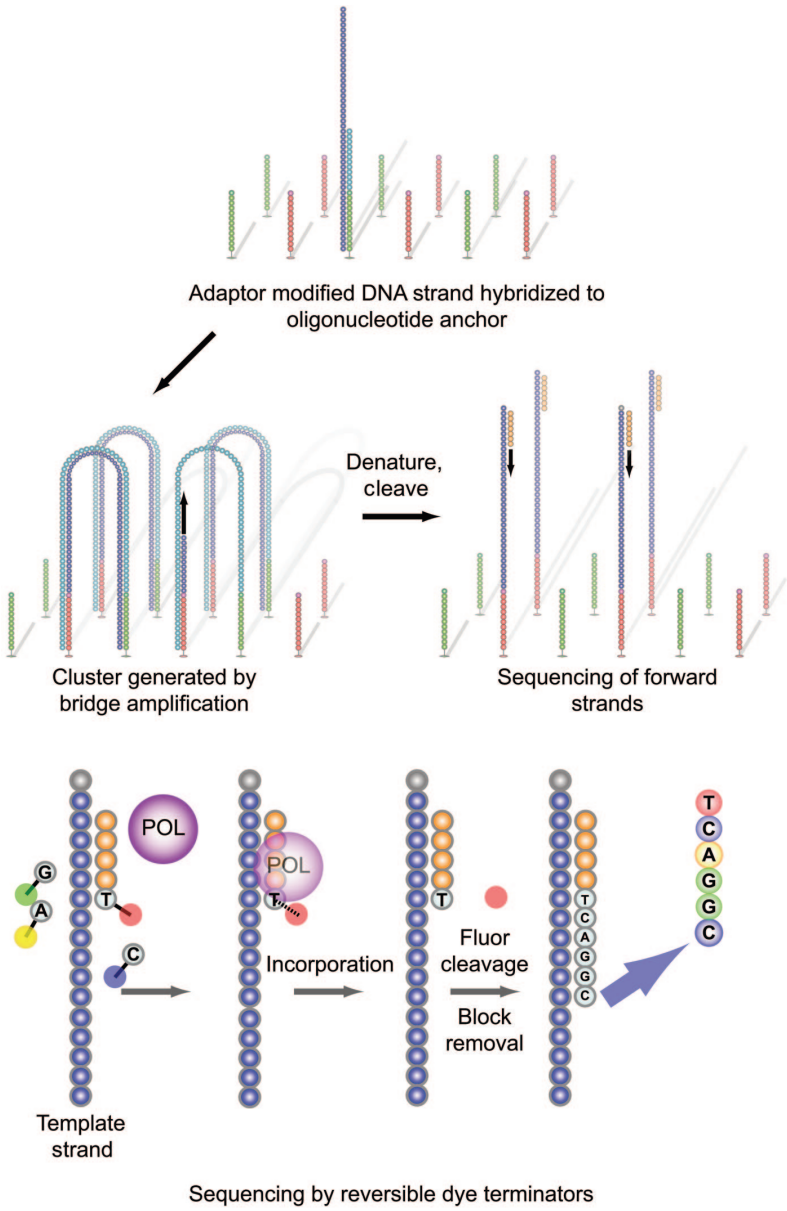


Figure 12. Illumina/Solexa Genome Analyzer sequencing (Voelkerding *et al.* 2009) Adapter-modified, single-stranded DNA is added to the flow cell and immobilized by hybridization. Bridge amplification generates clonally amplified clusters. Clusters are denatured and cleaved; sequencing is initiated with addition of primer, polymerase (POL) and 4 reversible dye terminators. Postincorporation fluorescence is recorded. The fluor and block are removed before the next synthesis cycle.

The applications of Illumina/Solexa sequencing in *M. tb* research can be classified into three main categories: 1) Whole genome re-sequencing to catalogue genetic variants such as SNPs and InDels, 2) RNA-seq for transcriptome profiling by sequencing cDNA, and 3) ChIP-seq to elucidate binding sites of individual transcription regulators, or study the genome-wide distribution of global transcription factors such as RNA polymerase. Analysis and interpretation of the vast amounts of data generated by NGS requires a multi-step strategy involving the use of bioinformatics and statistical tools, which needs to be tailored to address the different biological applications of this technology.

The first step common to all NGS data is alignment or *de novo* assembly of the sequence reads. Most alignment algorithms construct auxiliary data structures, called indices, either for the reference or the reads, which allow fast mapping of the reads to the genome sequence. Some of the commonly used alignment tools are, MAQ (H. Li *et al.* 2008a), Bowtie (Langmead *et al.* 2009), SOAP (R. Li *et al.* 2008b), etc. *De novo* assembly of short reads is more challenging given the minimal overlap between the reads. Some of the currently used *de novo* assembly programs are Velvet (Zerbino & Birney 2008), SSAKE (Warren *et al.* 2007), and ABySS (Simpson *et al.* 2009), all of them employ principles of graph theory. The choice of alignment parameters affects the results of the mapping procedure. These parameters include, the number of mismatches allowed for mapping, the maximum number of distinct matches allowed per sequence read, and the seed length (minimum portion of the read to be used to establish a candidate mapping site for a read).

Accurate alignment of sequence reads is of utmost importance in the case of variant-discovery. At least two factors can lead to incorrect read alignment.

The first one concerns the problem of mapping reads accurately to repetitive sequences. For example, given the high amount of repetitive DNA in the *M. tb* genome, a sequence read could map equally well to more than one genomic site. The second one involves the introduction of sequencing errors or mutations, which could lead to placement of the read at the wrong site. Usually, reads that align equally well at multiple sites can be randomly distributed to the sites or in some cases discarded, depending on the alignment software used. In *de novo* assembly these reads are typically discarded which leads to gaps in coverage and multiple aligned read groups (contigs) (Voelkerding *et al.* 2009). Based on the base-calling confidence scores, alignment programs estimate the likelihood of sequencing errors and calculate the probability of a read being mapped to the wrong place, known as the alignment score (H. Li *et al.* 2008a). Many alignment programs provide extended capabilities for SNP and InDel calling which take into consideration the alignment score and the sequencing depth for filtering low quality variants (H. Li *et al.* 2008a; R. Li *et al.* 2008b). Once the SNPs and/or InDel variants have been identified they need to be mapped to the genome sequence and analyzed in the context of the genome annotation. Polymorphisms occurring in protein-coding genes can be characterized as non-synonymous or synonymous based on whether they have an impact on the resulting amino acid or not.

Interpretation of the ChIP-seq and RNA-seq datasets requires further processing. Once the reads are mapped to the reference genome, the alignments are usually transformed to genomic densities, namely tags-to-positions maps (e.g. SAM files (H. Li *et al.* 2009)) which are in turn converted into count-per-position data (e.g. WIG files (Dreszer *et al.* 2012)), i.e. the read depth at every genomic position. WIG files can be used for

visualizing the NGS data aligned on the genome using tools such as the UCSC genome browser (Dreszer *et al.* 2012) or the Integrative Genome Viewer (Thorvaldsdóttir *et al.* 2012).

Most NGS experiments involve analysis of data from at least two or more biological/technical replicates. Different samples can yield significantly different numbers of unique mapped reads depending on the amount of starting material, the sequencing error rate, the severity of amplification biases, etc. Therefore, a normalization step is included to make sure results from different datasets are comparable. In most cases, each dataset is normalized to the total mapped reads.

In the case of RNA-seq data, the expression level for each transcript is determined by the total number of reads mapping to the transcript. A commonly used measure to determine the relative expression level of a transcript is RPKM (Reads per Kilobase of transcript per Million mapped reads) (Fang *et al.* 2012). As RPKM considers the length of the transcript, it enables comparison among different transcripts in the same dataset. As with most expression profiling experiments, RNA-seq is used to identify differentially expressed transcripts between different experimental conditions. A number of statistical methods have been proposed to detect differentially expressed genes from a counts table. These methods differ in their underlying data distribution and handling of biological replicates. Two of the commonly used methods include DEGSeq (Wang *et al.* 2009) and edgeR (Anders & Huber 2010), which are based on negative binomial distribution.

An additional processing step is required for ChIP-seq data, prior to estimation of binding-sites. As fragments are sequenced at the 5′ end, the

locations of mapped reads form two distributions, one on the forward strand and the other on the reverse strand, with a consistent distance between the peaks of the distributions (Rougemont & Naef 2012). In order to identify enriched regions, a combined profile of the two strands is calculated usually by shifting each distribution towards the centre (Figure 13). Programs like MACS (Lun *et al.* 2009), QuEST (Valouev *et al.* 2008) can perform peak detection in ChIP-seq data. Once the peaks are identified, motif searches can be performed using tools like MEME (Bailey *et al.* 2009) or Gibbs Motif Sampler (Thompson *et al.* 2005).

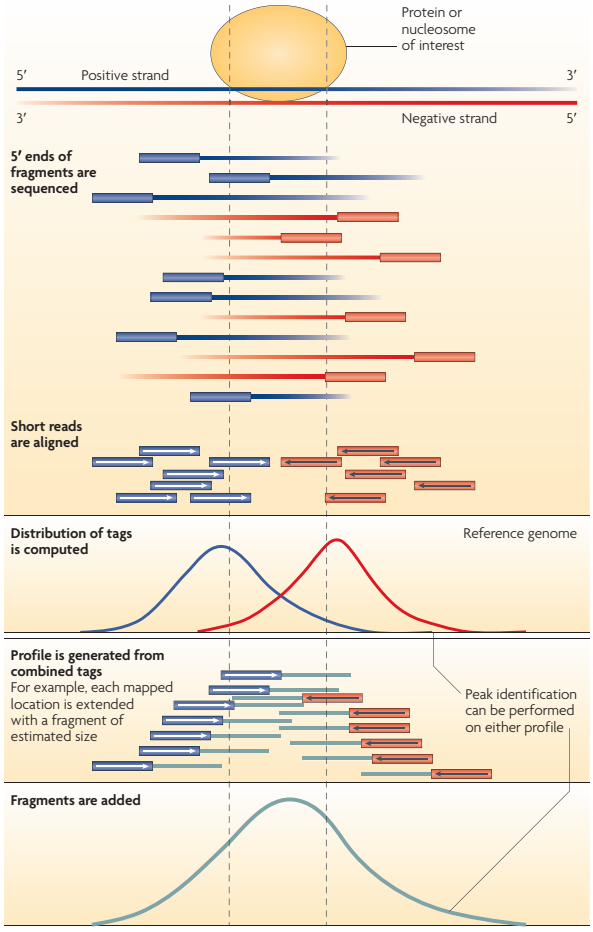


Figure 13. Strand-specific profiles at enriched sites in the ChIP-seq experiment (P. J. Park 2009).

The high spatial resolution of ChIP-seq makes it more efficient at identifying over-represented DNA sequence motifs compared to the ChIP-on-chip method.

The scheme below summarizes the technologies and experimental methods used, the nature of the biological datasets, and examples of the results obtained as part of this thesis.

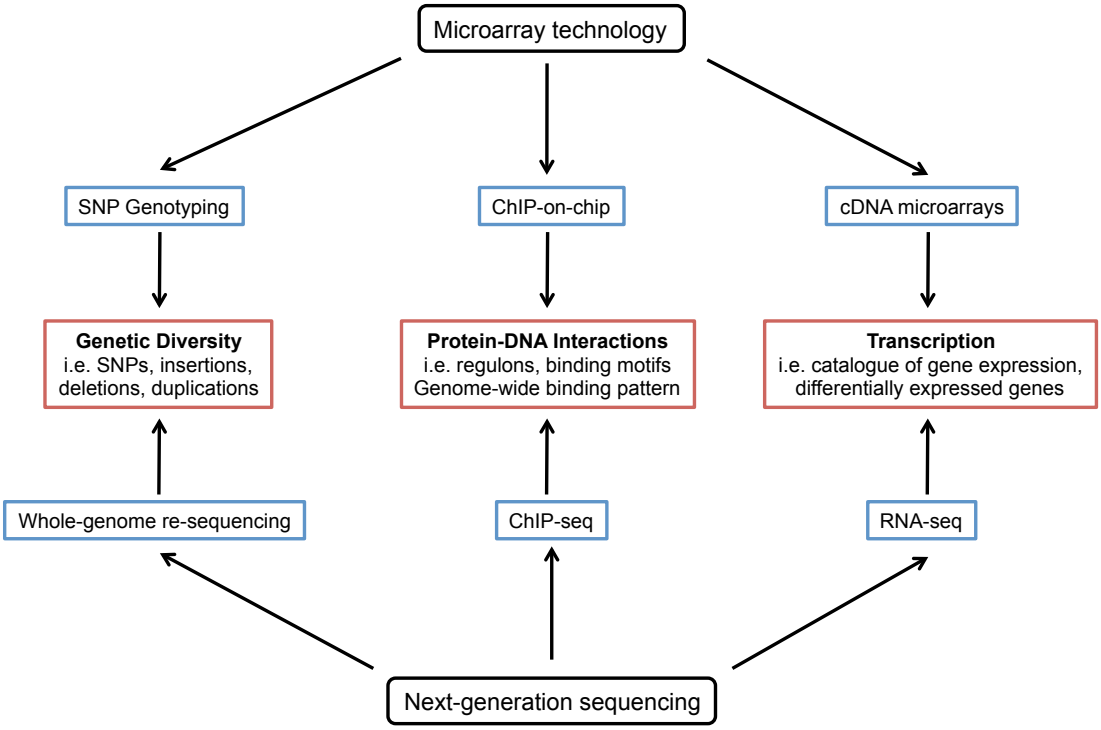


Figure 14. Overview of analysis strategies.

4. Results and Discussion

The publications resulting from my Ph.D work have been listed below. The following sections will include a brief summary of the major publications and my contribution to them.

List of Publications

1. Garcia Pelayo MC*, **Uplekar S***, Keniry A, Mendoza Lopez P, Garnier T, Nunez Garcia J, Boschiroli L, Zhou X, Parkhill J, Smith N, Hewinson RG, Cole ST & Gordon SV. A comprehensive survey of single nucleotide polymorphisms (SNPs) across *Mycobacterium bovis* strains and *M. bovis* BCG vaccine strains refines the genealogy and defines a minimal set of SNPs that separate virulent *M. bovis* strains and *M. bovis* BCG strains. *Infect. Immun.* 77, 2230–2238 (2009).
2. **Uplekar S**, Heym B, Friocourt V, Rougemont J & Cole ST. Comparative genomics of Esx genes from clinical isolates of *Mycobacterium tuberculosis* provides evidence for gene conversion and epitope variation. *Infect. Immun.* 79, 4042–4049 (2011).
3. Hartkoorn RC, Sala C, **Uplekar S**, Busso P, Rougemont J & Cole ST. Genome-wide definition of the SigF regulon in *Mycobacterium tuberculosis. J Bacteriol.* 194, 2001-9 (2012).
4. Blasco B, Chen JM, Hartkoorn R, Sala C, **Uplekar S**, Rougemont J, Pojer F & Cole ST. Virulence regulator EspR of *Mycobacterium tuberculosis* is a nucleoid-associated protein. *PLoS Pathog.* 8, e1002621 (2012).
5. Kirksey MA, Tischler AD, Siméone R, Hisert KB, **Uplekar S**, Guilhot C & McKinney JD. Spontaneous phthiocerol dimycocerosate-deficient variants of *Mycobacterium tuberculosis* are susceptible to gamma interferon-mediated immunity. *Infect. Immun.* 79, 2829–2838 (2011).
6. Massouras A, Hens K, Gubelmann C, **Uplekar S**, Decouttere F, Rougemont J, Cole ST & Deplancke B. Primer-initiated sequence synthesis to detect and assemble structural variants. *Nat. Methods* 7, 485–486 (2010).
7. Chen JM, **Uplekar S**, Gordon SV, Cole ST. A Point Mutation in cycA Partially Contributes to the D-cycloserine Resistance Trait of Mycobacterium bovis BCG Vaccine Strains. *PLoS ONE,* 2012.
8. Hartkoorn RC, Sala C, Neres J, Pojer F, Magnet SJ, Mujherjee R, **Uplekar S**, Boy-Rottger S, Altmann KH, & Cole ST. Towards a new tuberculosis drug: Pyridomycin – Nature's Isoniazid. *EMBO Molecular Medicine*, July 2012

Accepted

9. **Uplekar S**, Rougemont J, Cole ST & Sala C. High-resolution transcriptome and genome-wide dynamics of RNA polymerase and NusA in *Mycobacterium tuberculosis. Nucleic Acids Research.*

4.1. Comparative Genomics

4.1.1 A comprehensive survey of single nucleotide polymorphisms (SNPs) across *M. bovis* strains and *M. bovis* BCG vaccine strains

Contribution: Analysis of the SNP dataset, assessment of the SNP distribution across different M. bovis *BCG and* M. bovis *strains, construction of SNP based phylogenies, and phylogenetic analysis to quantify the selection pressure on the BCG strains.*

Mycobacterium bovis bacille Calmette-Guérin (BCG) is the only available vaccine against tuberculosis and the most widely used vaccine in the world. While BCG protects children effectively against early manifestations of TB, its success in adults is debatable. The protective efficacy against adult pulmonary TB ranges from 0-80%. One of the factors contributing to the variable efficacy of BCG could be the heterogeneity of the BCG daughter strains. After the original BCG vaccine was derived in 1921, it was distributed to different laboratories across the world, which maintained their own daughter strains by passaging, until the introduction of archival seed lots in the 1960s. The widespread distribution and subsequent subcultures led to the generation of a number of daughter strains named after their geographical origin (hence, BCG Russia, BCG Denmark, etc.) (Oettinger *et al.* 1999). The protective efficacy of the different BCG daughter strains has been shown to vary in both laboratory models (Lagranderie *et al.* 1996) and epidemiological studies (Davids *et al.* 2006). In order to promote the development of TB vaccines, it is necessary to improve our understanding of the genetic mechanisms of BCG attenuation and how they translate into variation in the efficacy of individual vaccine strains. Subtractive hybridization, and later comparative hybridization using DNA microarrays revealed a region of

difference (RD1) encoding the ESX-1 protein secretion system that is absent from all BCG strains and shown to play a major role in the attenuation of BCG (Mahairas *et al.* 1996; Lewis *et al.* 2003). However, the re-introduction of ESX-1 in BCG did not restore full virulence (Pym *et al.* 2002), suggesting the involvement of other attenuating mutations.

The evolutionary scheme based on strain-specific genetic markers divides BCG vaccine strains into four main sub-groups (Behr *et al.* 1999; Brosch *et al.* 2007). Comparison of the whole genome sequences of *M. bovis* BCG Pasteur (Brosch *et al.* 2007) with the virulent *M. bovis* 2122/97 (Garnier 2003) enabled high-resolution detection of sequence polymorphisms between the two strains. Of the 736 SNPs identified between the two strains, 56% were non-synonymous (nsSNPs) and could have functional consequences on the genes they affected (Brosch *et al.* 2007). All BCG strains contain a frameshift mutation in the *phoT* gene; inactivation of *phoT* in *M. bovis* attenuates the strain, and so this *phoT* mutation may also contribute to the loss of virulence of BCG (Collins 2003). SNPs unique to some BCG strains have also been shown to have functional consequences. For example, the *sigK* gene of BCG Pasteur has incurred a missense mutation in its start codon leading to a reduced expression of the major antigens MPB70 and MPB83 in this strain (Charlet *et al.* 2005). These data clearly indicate that SNPs have played a significant role in the attenuation of BCG. However, in order to discover SNP candidates that are likely to be responsible for the attenuation of BCG we need to consider SNPs that differentiate all attenuated BCG strains from virulent *M. bovis* strains.

We employed a high-throughput SNP screening methodology, the Sequenom genotyping platform, (Sequenom Inc., San Diego, CA) to map all SNPs identified between BCG Pasteur and *M. bovis* 2122/97 across a selection of *M. bovis* isolates and BCG daughter strains. Informative data sets were obtained for 658 SNPs from 21 virulent *M. bovis* strains from UK and France, and 13 BCG daughter strains. The distribution of SNPs across the British, French and BCG lineages allowed clustering of these strains to generate a linear phylogeny that was consistent with the geographical origin of the strains and previous evolutionary schemes for BCG. Our data has refined the previous genealogies of BCG by revealing a closer relationship between BCG Tice and BCG Pasteur than was previously appreciated, and also positioning BCG Beijing within the group of BCG Denmark-derived strains.

We identified a minimal set of 115 nsSNPs SNPs between virulent *M. bovis* strains and all BCG strains, affecting important functions such as transcriptional regulation and central metabolism, which might impact on virulence. In conclusion, the SNPs identified from our study provide a rich source of genetic variation that can be mapped to functional differences across BCG strains and provide further insight on the molecular basis for the attenuation of BCG.

Last year, Orduna *et al.* sequenced the whole genome of the BCG Mexico 1931 using 454 pyrosequencing. The aim of their study was to characterize the genomic and immune proteomic profile of the BCG vaccine strain used in Mexico. Comparison of the BCG Mexico sequence with the genomes of BCG Pasteur and BCG Tokyo revealed 33 SNPs, of which 23 were also reported in our BCG analysis. The results were consistent with our SNP-based phylogeny, which placed BCG Mexico 1931 in the same group as BCG Tice.

Two recent publications have characterized the functional effect of some of the nsSNPs identified in our study between the virulent *M. bovis* and the attenuated BCG strains,

1) Mendoza Lopez *et. al* (2010) analysed the nsSNP in BCG3145, which is a global gene regulator belonging to the SARP (*Streptomyces* antibiotic regulatory protein) family of proteins. The SNP replaces a highly conserved glutamic residue at position 159 with glycine (E159G), in the bacterial transcriptional activation (BTA) domain of BCG3145. Microarray analysis of transcriptome changes upon overexpression of BCG3145 and Rv3124 (orthologue of BCG3145 in *M. tb* H37Rv) revealed that the two proteins were positive regulators of molybdopterin biosynthesis. Based on this study, Rv3124 was renamed *moaR1* as it was found to regulate expression of the *moa1* locus. The E159G mutation in BCG3145 was shown to decrease, but not abolish, the ability of the regulator to induce expression of the *moa1* locus (Mendoza Lopez *et al.* 2010).
2) Chen *et. al* (2012) studied the nsSNP in the *cycA* gene of all BCG strains that results in a glycine to serine substitution at position 122 (G122S) in the CycA protein, which is an amino acid transporter. CycA is involved in the uptake of the antibiotic D-cycloserine. Compared to wild-type *M. tb* and *M. bovis*, BCG has been found to be more resistant to growth inhibition by D-cycloserine (DCS). Using genetic approaches, it was shown that a merodiploid strain of BCG expressing *M. tb* CycA had increased sensitivity to DCS. In addition, *M. smegmatis* strains heterologously expressing either *M. tb* or *M. bovis* CycA but not BCG CycA were rendered more sensitive to DCS. These findings demonstrate that the BCG *cycA* SNP partially contributes to DCS resistance in the vaccine strain (Chen *et al.* 2012).

INFECTION AND IMMUNITY, May 2009, p. 2230–2238 Vol. 77, No. 5
0019-9567/09/$08.00+0 doi:10.1128/IAI.01099-08

A Comprehensive Survey of Single Nucleotide Polymorphisms (SNPs) across *Mycobacterium bovis* Strains and *M. bovis* BCG Vaccine Strains Refines the Genealogy and Defines a Minimal Set of SNPs That Separate Virulent *M. bovis* Strains and *M. bovis* BCG Strains▽†

M. Carmen Garcia Pelayo,[1]‡ Swapna Uplekar,[2]‡ Andrew Keniry,[3]§ Pablo Mendoza Lopez,[1,4] Thierry Garnier,[5] Javier Nunez Garcia,[1] Laura Boschiroli,[6] Xiangmei Zhou,[7] Julian Parkhill,[3] Noel Smith,[1,8] R. Glyn Hewinson,[1] Stewart T. Cole,[2] and Stephen V. Gordon[1,9,10,11,12]*

TB Research Group, VLA Weybridge, New Haw, Surrey KT15 3NB, United Kingdom[1]; Global Health Institute, EPFL, CH-1015 Lausanne, Switzerland[2]; The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, United Kingdom[3]; Department of Biochemistry and Molecular Biology, Universidad de Granada, Granada, Spain[4]; Unité de Génétique et Génomique des Insectes Vecteurs, Institut Pasteur, 28 Rue du Dr Roux, Paris 75015, France[5]; Agence Française de Sécurité Sanitaire des Aliments, 23 avenue du Général-de-Gaulle, 94706 Maisons-Alfort Cedex, France[6]; College of Veterinary Medicine, China Agricultural University, Yuanmingyuan West Road No. 2, Haidian District, Beijing, People's Republic of China, 100193[7]; Centre for the Study of Evolution (CSE), University of Sussex, Brighton BN1 9QL, United Kingdom[8]; and UCD Schools of Agriculture, Food Science and Veterinary Medicine,[9] Medicine and Medical Science,[10] and Biomolecular and Biomedical Science,[11] College of Life Sciences, and UCD Conway Institute of Biomolecular and Biomedical Research,[12] University College Dublin, Dublin 4, Ireland

Received 3 September 2008/Returned for modification 26 October 2008/Accepted 4 March 2009

To further unravel the mechanisms responsible for attenuation of the tuberculosis vaccine *Mycobacterium bovis* BCG, comparative genomics was used to identify single nucleotide polymorphisms (SNPs) that differed between sequenced strains of *Mycobacterium bovis* and *M. bovis* BCG. SNPs were assayed in *M. bovis* isolates from France and the United Kingdom and from different BCG vaccines in order to identify those that arose during the attenuation process which gave rise to BCG. Informative data sets were obtained for 658 SNPs from 21 virulent *M. bovis* strains and 13 BCG strains; these SNPs showed phylogenetic clustering that was consistent with the geographical origin of the strains and previous schemes for BCG genealogies. The data revealed a closer relationship between BCG Tice and BCG Pasteur than was previously appreciated, while we were able to position BCG Beijing within a grouping of BCG Denmark-derived strains. Only 186 SNPs were identified between virulent *M. bovis* strains and all BCG strains, with 115 nonsynonymous SNPs affecting important functions such as global regulators, transcriptional factors, and central metabolism, which might impact on virulence. We therefore refine previous genealogies of BCG vaccines and define a minimal set of SNPs between virulent *M. bovis* strains and the attenuated BCG strain that will underpin future functional analyses.

Mycobacterium bovis bacillus Calmette-Guérin (BCG) is the only vaccine available against tuberculosis and is the most widely used vaccine in the world. It was derived by the repeated subculture of a strain of *Mycobacterium bovis* on potato slices soaked in glycerol and ox bile (10), leading to the in vitro accumulation of mutations and ultimately attenuation. Despite the widespread use of BCG, the precise genetic lesions that led to attenuation are not defined. Furthermore, the success of BCG led to its distribution from the Institut Pasteur to laboratories around the world, each of which continued the subculturing process, thereby leading to the generation of a number of daughter strains named after their geographical origin (hence BCG Tokyo, BCG Russia, etc.). The protective efficacy of these strains has been shown to vary in both laboratory models and epidemiological studies (6, 18, 36).

As BCG is the only vaccine currently available against tuberculosis, there is a clear need to understand the molecular basis of attenuation and variable efficacy afforded by BCG. The first study that attempted to identify mutations linked to attenuation was performed by Mahairas and colleagues, who identified three deletions, RD1 to RD3, from the genome of BCG strain Connaught (39). The RD1 locus was shown to be deleted from all BCG strains but present in all virulent strains of *M. bovis* and *Mycobacterium tuberculosis* studied. Subsequent work has shown that this deletion played a major role in the attenuation of BCG (38, 46). However, complementation of BCG with RD1 does not restore virulence to wild-type levels, suggesting that other attenuating mutations exist. Indeed, all BCG strains contain a frameshift mutation in the *phoT* gene; inactivation of *phoT* in *M. bovis* attenuates the strain, and so this *phoT* mutation may also contribute to the loss of virulence of BCG (15). This observation demonstrates that single nucleotide poly-

* Corresponding author. Mailing address: School of Agriculture, Food Science and Veterinary Medicine, College of Life Sciences, University College Dublin, Dublin 4, Ireland. Phone: 353 1 7166181. Fax: 353 1 7166185. E-mail: stephen.gordon@ucd.ie.

† Supplemental material for this article may be found at http://iai.asm.org/.

‡ These authors made equal contributions to this work.

§ Present address: Laboratory of Developmental Genetics and Imprinting, The Babraham Institute, Cambridge CB22 3AT, United Kingdom.

▽ Published ahead of print on 16 March 2009.

morphisms (SNPs) may play a significant part in the attenuation of BCG.

A major step toward defining the molecular basis of attenuation in BCG was the completion of the genome sequence of *M. bovis* BCG Pasteur (9). Genomic comparison of BCG Pasteur with *M. bovis* 2122/97 (22) identified a range of mutational differences, including deletions, duplications, and SNPs. In further work the configuration of two large duplications, DU1 and DU2, was shown to vary across BCG daughter strains, in a manner that was congruent with the previous BCG phylogeny defined by Behr and colleagues using deletions (7, 9, 41).

From the complete chromosome sequences, 736 SNPs were identified between the BCG Pasteur vaccine strain and the virulent *M. bovis* strain 2122/97 (9). However, only those SNPs that are unique to all BCG strains are good candidates for mutations involved in the attenuation of BCG; such SNPs presumably occurred during the attenuation of BCG after it was derived from wild-type *M. bovis*. The ideal experiment would be a comparison using the chromosome of the wild-type progenitor from which BCG strains were derived; unfortunately, this strain is unavailable. The *M. bovis* 2122/97 strain is derived from the clonal complex of *M. bovis* common in the British Isles (provisionally called Eu1 [53]), while the BCG strain is derived from the French lineage, which is phylogenetically distinct from the Eu1 clonal complex. Many of the 736 SNPs that differ between BCG Pasteur and *M. bovis* 2122/97 may be lineage specific to either BCG or the Eu1 clonal complex and therefore unlikely to be involved in the attenuation of BCG.

To identify those SNPs that are unique to BCG strains and therefore possible candidates for the attenuation of BCG, we mapped the phylogenetic position of all SNPs identified between *M. bovis* 2122/97 and BCG Pasteur across a population of United Kingdom and French *M. bovis* isolates, as well as BCG daughter strains. We used a high-throughput SNP screening methodology to screen hundreds of SNPs across the population and generated a distribution of SNPs across British, French, and BCG *M. bovis* strains.

It has previously been shown that some SNPs unique to BCG have functional effects: a nonsynonymous SNP (nsSNP) in the *mmaA3* gene, which encodes a mycolic acid methyltransferase, results in the loss of methoxymycolic acids from late strains of BCG (5); an SNP in *sigK* leads to reduced synthesis of MPB83 and MPB70 in late BCG strains (11); and an SNP in the *pykA* gene reverts a null mutation in *M. bovis* and allows BCG to grow on glycerol (33). Finally, SNPs in the cyclic AMP receptor protein (CRP) transcriptional regulator in some BCG strains affect the binding of the regulator to DNA (4, 30). Following these examples, we highlight SNPs that may impact on phenotypic differences between virulent *M. bovis* and BCG.

MATERIALS AND METHODS

SNP identification. The sequences of *M. bovis* 2122/97 and *M. bovis* BCG Pasteur were compared using the DIFFSEQ application from the EMBOSS package (http://emboss.sourceforge.net/). This tool is less complex than many genome comparison tools but can be used here, as the two genomes are entirely colinear. This allowed three classes of mutations to be identified: (i) transitions and transversions; (ii) insertions or deletions (InDels); and (iii) "block" substitutions, where a block of sequence of >1 bp replaces another. The InDels were further subdivided into two groups, the mIns and mDels, which are the insertions or deletions of a single base, respectively. We defined SNPs as the total number of transitions and transversions (736) plus the mIns and mDels (46), which gave a total of 782 positions to be investigated. Each SNP was initially verified by checking the original sequence trace files from the BCG Pasteur and *M. bovis* 2122/97 sequencing projects.

Bacterial strains. Strains are shown in Table 1, with a spoligotype phylogeny shown in Fig. 1 to illustrate the relationships across the strains. British *M. bovis* strains were selected from the VLA Weybridge strain collection to represent the British *M. bovis* population structure. French isolates were selected from the strain collection of the Agence Française de Sécurité Sanitaire des Aliments (AFSSA), such that the spoligotype of the strain was the same as that of BCG (SB0120, as defined by the international *M. bovis* spoligotype database [www.mbovis.org]). BCG daughter strains were obtained from the VLA Weybridge strain collection, Marcel Behr (McGill University, Montreal, Canada), or the Statens Serum Institut (Copenhagen, Denmark). A total of 34 strains were examined.

Molecular typing. Strains were typed by both spoligotyping and variable-number tandem repeat (VNTR) (ETR-A to -F) typing. Spoligotyping and VNTR were performed as described previously (21, 32).

Sequenom genotyping. Genotyping was performed with IPLEX chemistry, on the Sequenom genotyping platform (Sequenom Inc., San Diego, CA). During the IPLEX reaction, oligonucleotide primers anneal directly adjacent to the SNP of interest. Allele-specific extension products are then produced by single base extension of the oligonucleotide with terminator nucleotides, each of unique mass. Multiplexed IPLEX assays of between 1 and 28 assays per plex were designed to detect 701 single nucleotide base changes using the Sequenom Assay Design v3.0.2.0 package. Genomic DNA was extracted from BCG and other *M. bovis* strains using a standard cetyltrimethylammonium bromide method and diluted in Tris-EDTA to a final concentration of 4 ng/μl. SNP-containing loci were amplified from genomic DNA by PCR. Unincorporated nucleotides were removed by treatment with shrimp alkaline phosphatase, followed by the IPLEX extension reaction, per the manufacturer's instructions. The allele-specific products resulting from the IPLEX reaction were desalted through the addition of an anion-exchange resin and then analyzed by matrix-assisted laser desorption ionization–time of flight mass spectrometry. Genotypes were assigned in real time and then evaluated using the SpectroCALLER and SpectroACQUIRE software (Sequenom), respectively. Selected SNPs were confirmed using standard capillary sequencing.

Phylogenetic analysis. SNP calls were parsed to extract SNPs specific to each strain and concatenated into a single 701-bp sequence. Sequence alignment was carried out using ClustalW, and phylogenetic analyses were performed using the MEGA 4.0 software. Discrepancies in the data (i.e., call rates lower than 100% or ambiguous calls) were replaced by "?" for the purposes of alignment. Distance-based analysis was conducted by applying the neighbor-joining algorithm using the number of nucleotide differences. A consensus parsimony tree was generated using the maximum parsimony method with bootstrapping.

dN/dS estimations. The sequence of the common ancestor (anc1) of the British lineage and the French lineage was reconstructed at the polymorphic sites using the *M. tuberculosis* H37Rv strain as an outgroup. The reconstructed ancestral sequence was then compared to the sequence found in strain 2253 (a representative of the British lineage, SB0134 [Table 1]), and the ratio of the number of nonsynonymous to synonymous changes per site (dN/dS ratio) was calculated for 48 synonymous and nonsynonymous mutations that had happened during the descent of this strain from anc1 to its divergence from the lineage that led to the *M. bovis* sequenced strain 2122. In a similar way the dN/dS ratio was calculated for 87 synonymous and nonsynonymous mutations in strain F1.2 (representing the French lineage, SB0120 [Table 1]) that happened between anc1 and the divergence of F1.2 from the lineage that leads to BCG Pasteur. Finally, the dN/dS ratio was calculated for 161 synonymous and nonsynonymous mutations between the divergence of strain F1.2 and BCG Russia in the lineage leading to BCG Pasteur. dN/dS ratios were calculated with the S.T.A.R.T 2 package (31) using the Nei-Gojobori method and the Jukes-Cantor correction (43).

RESULTS

Sequenom analysis. SNPs were queried using oligonucleotides that anneal −1 from the base of interest; allele-specific extension products were then analyzed via matrix-assisted laser desorption ionization mass spectrometry to identify the base at each SNP position across the panel of strains (Fig. 2). From the total of 736 SNPs, 35 SNPs failed and 20 gave ambiguous calls (i.e., many strain-calls missing or both alleles called at the same

TABLE 1. Strains used in this study

M. bovis strain	Source	Spoligotype	VNTR	Comment
Non-BCG strains				
2122/97	VLA Weybridge, United Kingdom	SB0140	8555*33.1	Genome-sequenced strain
F1.2	AFSSA, Alfort, France	SB0120	5654*33.1	French clinical isolate
F3	AFSSA, Alfort, France	SB0120	5553*33.1	French clinical isolate
F4	AFSSA, Alfort, France	SB0120	5554*43.1	French clinical isolate
F5	AFSSA, Alfort, France	SB0120	5554*33.1	French clinical isolate
F6	AFSSA, Alfort, France	SB0120	5554*33.1	French clinical isolate
F7	AFSSA, Alfort, France	SB0120	ND[a]	French clinical isolate
F8	AFSSA, Alfort, France	SB0120	4554*33.1	French clinical isolate
F9	AFSSA, Alfort, France	SB0120	5554*33.1	French clinical isolate
F10	AFSSA, Alfort, France	SB0120	5554*33.1	French clinical isolate
F11	AFSSA, Alfort, France	SB0120	5554*33.1	French clinical isolate
F12	AFSSA, Alfort, France	SB0120	5554*33.1	French clinical isolate
F13	AFSSA, Alfort, France	SB0120	5554*33.1	French clinical isolate
F14	AFSSA, Alfort, France	SB0120	5554*33.1	French clinical isolate
2253	VLA Weybridge, United Kingdom	SB0134	3534*33.1	United Kingdom clinical isolate
1766	VLA Weybridge, United Kingdom	SB0134	3534*33.1	United Kingdom clinical isolate
681	VLA Weybridge, United Kingdom	SB0129	6554*23.1	United Kingdom clinical isolate
1198	VLA Weybridge, United Kingdom	SB0140	6554*33.1	United Kingdom clinical isolate
1094	VLA Weybridge, United Kingdom	SB0140	7554*33.1	United Kingdom clinical isolate
393	VLA Weybridge, United Kingdom	SB0140	7554*33.1	United Kingdom clinical isolate
AN5	VLA Weybridge, United Kingdom	SB1417	6554*33.2	Tuberculin production strain
BCG strains				
Pasteur	VLA Weybridge, United Kingdom	SB0120	556233.1	Genome-sequenced strain
Sweden	Marcel Behr, McGill University, Canada	SB0120	5553*33.1	Vaccine strain
Tokyo	VLA Weybridge, United Kingdom	SB0120	5553*33.1	Vaccine strain
Denmark	Statens Serum Institut, Denmark	SB0120	5552/3*33.1	Vaccine strain
Tice	Marcel Behr, McGill University, Canada	SB0120	555233.1	Vaccine strain
Frappier	Marcel Behr, McGill University, Canada	SB0120	555133.1	Vaccine strain
Glaxo	VLA Weybridge, United Kingdom	SB0120	5553*33.1	Vaccine strain
Russia	VLA Weybridge, United Kingdom	SB0120	ND	Vaccine strain
Beijing	China Agricultural University	SB0120	ND	Vaccine strain
Connaught	VLA Weybridge, United Kingdom	SB0120	ND	Vaccine strain
Birkhaug	Marcel Behr, McGill University, Canada	SB0120	ND	Vaccine strain
Prague	Marcel Behr, McGill University, Canada	SB0120	ND	Vaccine strain
Moreau	Marcel Behr, McGill University, Canada	SB0120	ND	Vaccine strain

[a] ND, not determined.

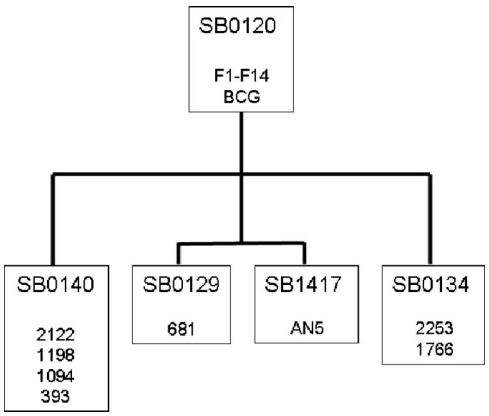


FIG. 1. Spoligotype phylogeny of *M. bovis* strains. The figure depicts relationships across the *M. bovis* strains used in this study based solely on spoligotype. Strain numbers are shown clustered into their spoligotype designations ("SB" numbers, based on the international *M. bovis* spoligotype database at www.mbovis.org). SB0120 is the most replete *M. bovis* spoligotype pattern, with all other patterns used in this study derivatives of this pattern. Branch lengths are not phylogenetically informative and are shown merely for the purposes of clustering related strains.

locus); this gave a total of 681 SNPs that gave usable data. Of these 681 SNPs, 23 were called as invariant across BCG Pasteur and other *M. bovis* strains, suggesting errors in the original genome sequences. However these 23 SNPs were contained within repeated sequences such as IS*1081*, REP13E12, and PE-PGRS genes, suggesting that the Sequenom method may also be at fault. To determine whether the sequence or Sequenom calls were correct, we examined the original sequence trace files for the BCG Pasteur and *M. bovis* 2122/97 genome sequences. This confirmed that 14 of the 23 invariant SNPs were indeed point mutations between BCG Pasteur and *M. bovis* 2122/97; four SNPs appeared to be errors in *M. bovis*, while the remaining five SNPs could not be called because of poor sequence reading coverage. As we could not determine the distribution of these 23 SNPs across the remaining strains, they were excluded from the analysis. Hence, we obtained informative data from 658 SNPs. The complete results for all of the 701 SNPs that gave data (including ambiguous calls) are shown in Table S1 in the supplemental material, while the distribution of SNPs across the functional classification of BCG genes is shown in Fig. S1 in the supplemental material.

Phylogenetic analysis. The phylogenetic distribution of SNPs allowed the strains to be clustered into related groupings,

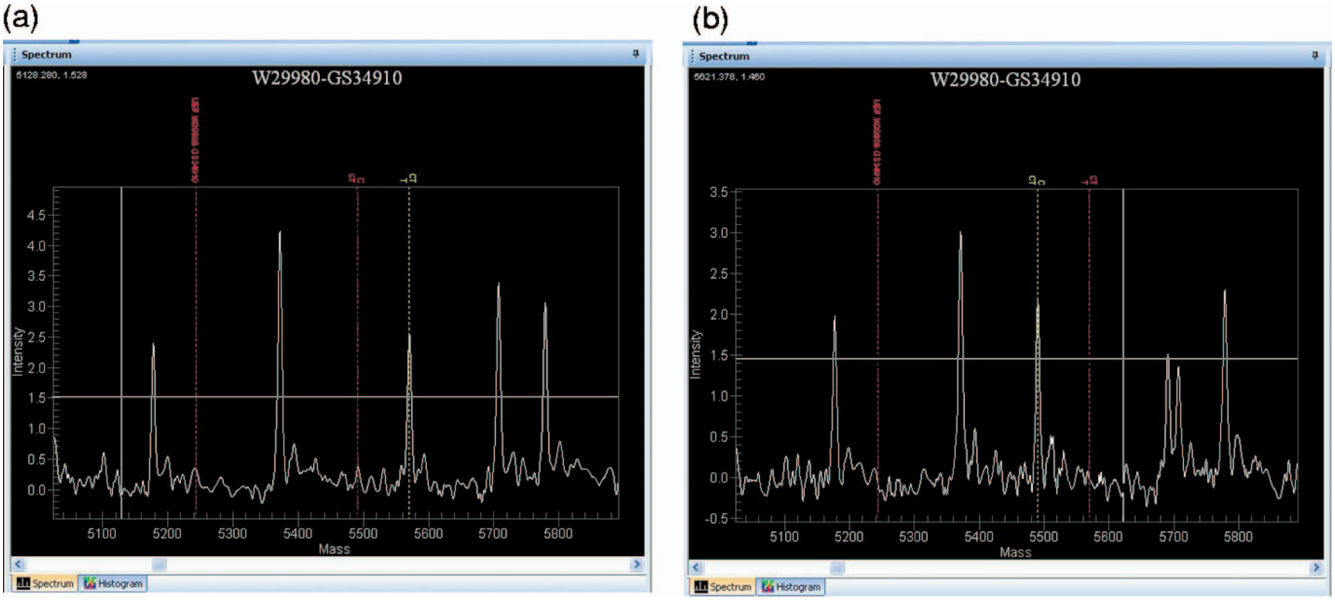


FIG. 2. Sequenom output. Images of the MassARRAY TyperAnalyzer v3.3 software (Sequenom) output for two samples. (a) Mass spectrometry output for BCG Pasteur, highlighting a T allele call for SNP GS34910, which is at BCG genome position 565209, located at the start of the *sigK* gene. (b) Mass spectrometry output for BCG Russia SNP GS34910, showing a C allele call for this SNP. This SNP gives rise to the differential expression of the SigK regulon between early and late BCG strains.

showing as expected that *M. bovis* strains isolated in Britain, those isolated in France, and BCG daughter strains formed three distinct clades (Fig. 3). As the SNPs were selected on comparison of the genomes of BCG Pasteur and *M. bovis* 2122/97, it should be noted that we generated a "linear" phylogeny, with these two strains being most distant from each other and all other strains falling between them.

The SNP phylogeny recapitulated BCG genealogies proposed using deletions or duplications (7, 9). Hence, "early" (pre-1927) or "late" (post-1927) BCG strains group together, although the SNPs did not allow the resolution of DU2 groups I and II as defined by Brosch et al. (9). BCG strains Russia, Tokyo, and Moreau (group I) and Sweden and Birkhaug (group II) clustered together. DU2 group III (BCG strains Denmark, Glaxo, Beijing, and Prague) formed a discrete group, with 10 SNPs distinguishing them from groups I and II. One SNP separated BCG Connaught/Frappier from group II, while seven SNPs separated these strains from BCG Tice. Finally, BCG Pasteur had eight strain-specific SNPs. One SNP, a 1-bp insertion at position 842687 (mIns-842687; see Table S1 in the supplemental material), was present in BCG Pasteur and BCG Beijing only. As this SNP is homoplasic, the Sequenom results were verified using standard sequencing and confirmed to be correct. Whether this indicates a selective pressure at this locus is unclear.

BCG Beijing was included in our SNP analysis as this strain had so far, to our knowledge, not been genetically studied. It was known, however, that BCG Beijing was derived from BCG Denmark and shows protective efficacy similar to that of BCG Denmark in animal models (56). The SNP analysis showed, as expected, that it belonged to the DU2-III group with BCG strains Denmark, Prague, and Glaxo. However, comparative genomic hybridization (data not shown) using an *M. tuberculosis* complex microarray (25) revealed that BCG Beijing had the characteristic RD-Denmark locus intact (41). Hence, it would appear that BCG Beijing was derived from a BCG Danish seed-lot before the RD-Denmark deletion had occurred. Using high-resolution NimbleGen arrays, Leung et al. also recently screened the genomes of BCG Beijing and 12 other BCG strains for InDels and also placed BCG Beijing in the BCG Denmark-derived clade (37).

The *M. bovis* progenitor attenuated by Calmette and Guérin is not available, having been lost from the Institut Pasteur archives. Therefore, we selected a panel of French *M. bovis* strains on the basis that they had the same spoligotype as BCG (SB0120) and similar VNTR profiles. Hence, the SNP screen showed minimal variation between the 13 French *M. bovis* strains. However, they were distinguished from the British strains by 158 SNPs and from BCG by 186 SNPs; we will deal below with the functional implications of these latter 186 SNPs.

The British strains were chosen on the basis of our detailed knowledge of the population structure of *M. bovis*, with molecular types SB0129, SB0134, and SB0140 representing the major groups of *M. bovis* in Britain (Fig. 1). In agreement with previous phylogenies of *M. bovis* based on spoligotypes and VNTR (53, 57), types SB0129 and SB0134 are more closely related to French strains than to the SB0140 clonal complex. The sequenced strain, 2122/97, showed 66 unique SNPs compared to other strains with the same SB0140 pattern.

One hundred fifty-eight SNPs separated all of the French strains from all of the British *M. bovis* strains. Using BLASTn (3), we determined the status of each of these SNPs in an outgroup, the genome sequence of *M. tuberculosis* H37Rv, to establish where the SNPs occurred in the evolution of the British and French lineages (Fig. 3). This analysis showed that 53 SNPs had occurred during the evolution of the British *M. bovis* branch (to SB0134), with 105 SNPs on the French branch. Hence, it appears that the British SB0134 strains are more

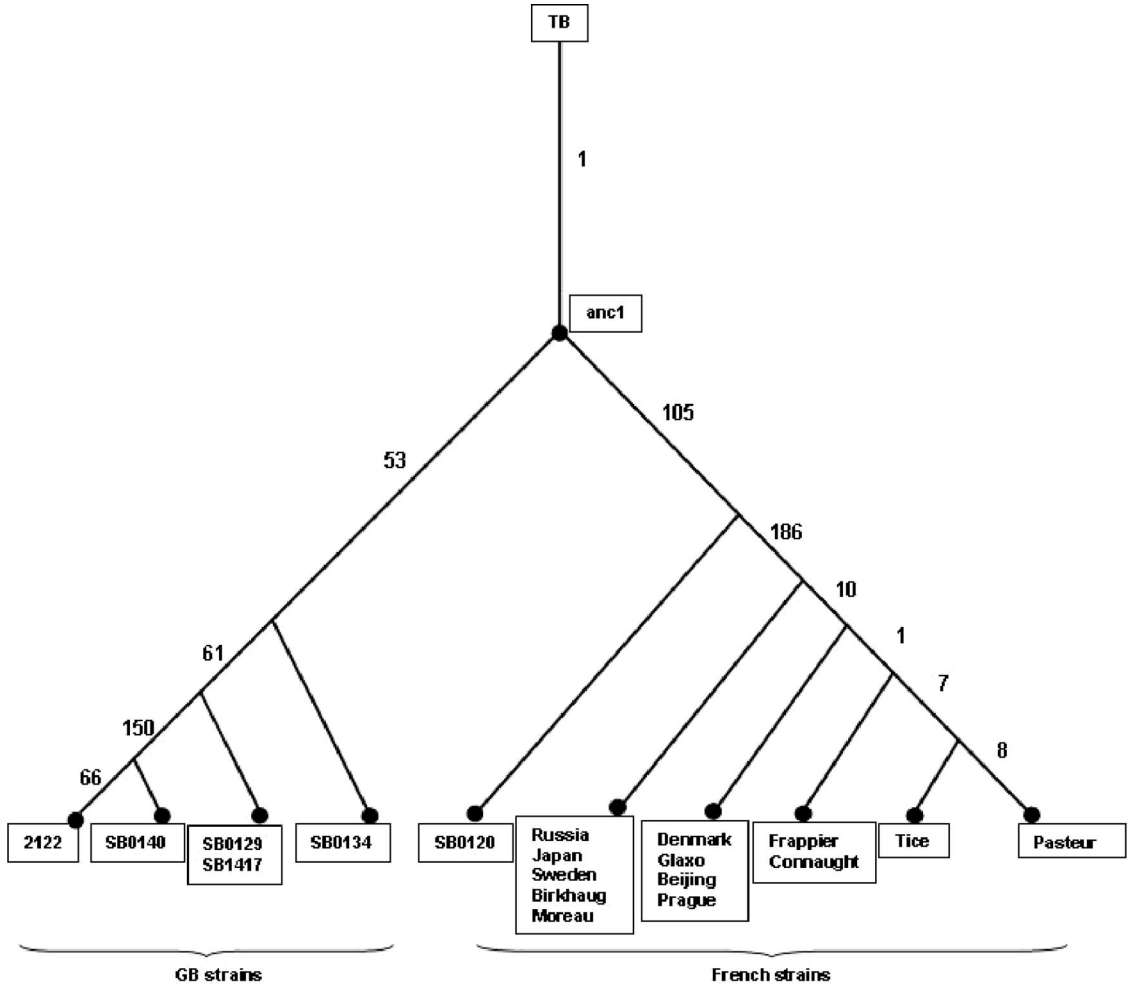


FIG. 3. Distribution of SNPs across British, French, and BCG lineages. A linear phylogeny is shown, where the sequenced strains that were used to derive the SNPs, *M. bovis* 2122/97 and BCG Pasteur, lie at the extremes. Numbers of SNPs that separate each grouping are shown. Branch lengths are not phylogenetically informative; the figure shows merely the clustering of strains generated by the SNP analysis.

similar to the common ancestor of the British-French lineage (anc1 in Fig. 3) than to the French strains selected in this study.

Determining the ratio of synonymous and nonsynonymous mutations per site between sequences (dN/dS ratio) can indicate the strength and direction of selection. Purifying selection is normally stronger on nonsynonymous changes, reducing the dN/dS ratio and giving values of <1; in contrast, positive or directional selection would give values for dN/dS of >1. Strains of BCG have been cultured in vitro for long periods, and therefore, dN/dS calculations may reveal differences in selection pressures between in vivo and in vitro growth. Calculations of dN/dS ratios gave values of 0.48 for anc1 (Fig. 3) to strain 2253 (SB0134, 48 SNPs), 0.68 from anc1 to strain F1.2 (SB0120, 87 SNPs), and 0.63 for the branch between the divergence of the F1.2 wild strain and the divergence of BCG Russia (161 SNPs). For such closely related strains, the interpretation of high dN/dS values is complex and the values are open to several interpretations (48). Nevertheless, two conclusions can be reached from our data. First, there appears to be no evidence for increased relaxed selection during the in vitro cultivation of BCG strains compared to in vivo-isolated *M. bovis* strains (0.63 in vitro versus 0.68 in vivo). Second, the relatively high dN/dS ratios are further evidence for relaxed purifying selection within the *M. tuberculosis* complex, presumably in response to a low effective population size (53). This latter point has been elegantly demonstrated by Hershberg et al. (29), who performed a multilocus sequencing analysis on a worldwide collection of 108 *M. tuberculosis* complex strains. The average pairwise dN/dS ratio across these strains for 370 SNPs was 0.57, comparable to the values that we obtained and, they concluded, likely a consequence of reduced selective constraint.

Interestingly, the *ugpAEBC* region contains SNPs that differentiate the British, French, and BCG lineages. For example, taking the *M. bovis* 2122/97 genome sequence as the reference, SNP mIns-3095955 is a 1-bp insertion in *ugpA* that is present in SB0134 and SB0120 *M. bovis* (i.e., including BCG); this polymorphism is in fact a 1-bp deletion from type SB0129, AN5, and *M. bovis* 2122/97 that frameshifts *ugpA*; hence, SNP mIns-3095955 delineates SB0129 and its descendants. Similarly, SNP 3093642 is an nsSNP in *ugpB* found only in the British SB0140 clonal complex. Finally, SNPs 3092465 and 3093840 are found only in BCG strains (see Table S1 in the supplemental mate-

TABLE 2. Genes defined as essential in vivo by transposon site hybridization that contain nsSNPs across all BCG strains[c]

Gene	Rv coding sequence	BCG coding sequence	SNP	Function
nrp	Rv101	BCG0134	L1365M	Nonribosomal peptide synthase
senX3	Rv0490	BCG0531	F109S	Two-component sensor
kdpD	Rv1028c	BCG1085c	P83S; N776D[a]	Two-component sensor
murI	Rv1338	BCG1400	R154L	Glutamate racemase
lysX	Rv1640c	BCG1679c	D769E[a]	Lysyl-tRNA synthase
pks12	Rv2048c	BCG2067c	S2964R	Mannosyl-β-1-phosphoisoprenoid synthase
fadE22	Rv3061c	BCG3086c	K488E[a,b]; S497C[b]	Acyl coenzyme A dehydrogenase
Rv3335c	Rv3335c	BCG3406c	A86V[a]	Integral membrane protein
Rv3616c	Rv3616c	BCG3680c	A4V[a]	ESX-1 secreted antigen

[a] Conservative substitution.
[b] Position with reference to H37Rv protein sequence as *M. bovis* 2122/97 allele is frameshifted.
[c] Transposon site hybridization was performed as described previously (50).

rial). Hence, focused sequencing of these SNPs provides a simple way to cluster British, French, and BCG strains. Possible functional implications of these mutations are discussed below.

Functional inferences. Of the 186 SNPs identified between virulent *M. bovis* strains and BCG, 115 are nonsynonymous and 55 are synonymous, 13 are intergenic, and three are located in pseudogenes. While synonymous SNPs (sSNPs) may have functional consequences in rare cases (34), nsSNPs, frameshifts, or intergenic SNPs that affect gene expression or protein structure are the most likely source of phenotypic variation.

Virulence. In a global analysis of genes required for in vivo survival, Sassetti and Rubin identified 194 major candidate genes (50); eight of these genes contain nsSNPs in BCG (Table 2). While many of these nsSNPs encode conservative amino acid substitutions, some may have functional consequences. For example, *kdpD* encodes the histidine kinase sensor of the two-component system KdpDE that regulates turgor pressure and potassium homeostasis. The BCG *kdpD* allele contains two nsSNPs, P83S and N776D, the latter of which is a conservative substitution. The N-terminal domain of KdpD contains two Walker nucleotide binding motifs which are important in ATP binding (27). While the P83S substitution does not disrupt these motifs, the P83 residue is conserved across KdpD homologues in many bacterial species (27). Functional analysis of the BCG KdpD protein is therefore warranted.

The SenX3 histidine kinase is also implicated in virulence (44, 50) and contains an F109S mutation; its linked response regulator, the RegX3 protein, also contains an nsSNP (A18T). The RegX3 mutation occurs at an alanine residue which is conserved across many similar two-component regulators. The function of SenX3-RegX3 in *M. bovis* BCG is unknown, but in *Mycobacterium smegmatis* SenX-RegX regulates the expression of the high-affinity *pstSCAB* phosphate uptake system (23, 24). If SenX3-RegX3 also functions in phosphate control in BCG, and as the *pst* high-affinity system appears to be nonfunctional in BCG, the accumulation of mutations in the *senX3-regX3* locus may be indicative of relaxed selection acting at this locus.

It is worth noting that the *pks12* gene, which encodes a mannosyl-β-1-phosphoisoprenoid synthase (40), contains 11 nsSNPs between BCG and other *M. bovis* strains; however, only one of these SNPs, S2964R, is present in all BCG strains but absent from all other *M. bovis* strains studied. Furthermore, there is no evidence of any structural difference between the mannosyl-β-1-phosphoisoprenoid synthases synthesized by BCG and those synthesized by *M. tuberculosis* (40), suggesting that this accumulation of nsSNPs in *pks12* has no functional effect.

Lesions in metabolism. The *M. bovis* genome encodes 17 cytochrome P450 oxidase (CYP) enzymes, with an 18th gene, the CYP142 gene, frameshifted (22). All BCG strains contain a frameshifted CYP123 gene, with an extended C-terminal portion that would be expected to disrupt correct protein folding and function. The role of mycobacterial CYP enzymes is still being elucidated, but they are expected to be integral to lipid metabolism. The function of CYP123 is unknown, but it was found to be upregulated in response to heat stress (55). The genes for three further CYP enzymes, the CYP126, CYP128, and CYP135B1 genes, contain nsSNPs, but their functional consequences are unknown. CYP128 was shown to be essential for in vitro growth of *M. tuberculosis* (49), so presumably the L203F substitution present in the BCG protein has no major functional effects.

M. bovis and *M. bovis* BCG strains cannot catabolize alanine due to a frameshift mutation in the *aldA* gene, which encodes alanine dehydrogenase. Chen and colleagues have also shown that BCG strains exhibit defects in serine metabolism, with BCG Pasteur and Frappier being unable to catabolize serine (12). The genetic basis for this defect in serine catabolism remains unidentified; indeed, *sdaA*, which encodes serine deaminase, is identical between *M. bovis* and *M. bovis* BCG strains, which suggests a regulatory defect (12). Analysis of SNPs that are shared between BCG Frappier and Pasteur did not reveal any obvious candidates that would explain the serine catabolism phenotype.

The *M. tuberculosis* complex contains three systems for the biosynthesis of trehalose, namely, the OtsAB, TreS, and TreXYZ systems (14, 42). It appears that the OtsAB system is the principal pathway for trehalose biosynthesis in *M. tuberculosis*; inactivation of the TreXYZ system had no in vitro or in vivo effect on growth (42). The *M. bovis* 2122/97 strain contains an internal deletion in the *treY* gene that leads to an inactive protein product (22, 26); however, this mutation is present only in strains that are closely related to 2122/97. It is surprising, therefore, that all BCG strains have an independent mutation in the TreXYZ system, with *treZ* frameshifted. Hence, the same biosynthetic pathway is mutated in both wild-type virulent *M. bovis* and in vitro-attenuated BCG. As the TreXYZ pathway is not required for virulence, its loss may simply reflect the removal of biosynthetic redundancy.

Growth on glycerol. A key metabolic selection placed on the *M. bovis* progenitor of BCG was the utilization of glycerol as a carbon source. We have previously shown that *M. bovis* strains contain an nsSNP in the gene encoding pyruvate kinase (*pykA*) that prevents the conversion of phosphoenolpyruvate to pyruvate and hence blocks glycolysis from feeding into the tricarboxylic acid cycle (22, 33). BCG strains all contain an Asp220Glu mutation that restores activity to pyruvate kinase, a mutation that was selected by the glycerol-based medium used by Calmette and Guérin.

In a chemostat analysis of *Escherichia coli* strains grown on glycerol, mutations in *glpK* (encoding glycerol kinase) that increased enzyme efficiency were selected for (28). The *glpK* gene of *M. bovis* 2122/97 is frameshifted, while that of BCG is in frame (22, 33). This latter mutation was not, however, selected for during BCG's in vitro growth, as French and related British *M. bovis* strains have an in-frame *glpK*; hence, it is merely a mutation in *M. bovis* 2122/97. The bovine tuberculin production strain, *M. bovis* AN5, was included in our SNP screen since, like BCG, it is a glycerol-adapted strain of *M. bovis* (45); SNPs shared between BCG and AN5 may favor growth of *M. bovis* on glycerol. However, apart from the previously described Asp220Glu mutation in *pykA* we could identify only one homoplasic SNP shared between BCG and AN5 (SNP 1369616), which was located in the gene for PPE18. However, it appears unlikely that this represents a glycerol-adaptive mutation.

As noted above, the *ugpAEBC* locus contains a number of SNPs with functional consequences. Hence, *ugpA* is frameshifted in British type SB0129 and the SB0140 clonal complex because of a 1-bp insertion; the resulting protein will not localize correctly in the membrane and is therefore nonfunctional. All BCG strains have an independent null mutation in the same operon, with *ugpB* containing an in-frame stop codon, and they also contain an nsSNP in *ugpC*. Hence, in British and BCG lineages the UgpAEBC transporter is nonfunctional. This is interesting because *ugpAEBC* encodes a glycerol-3-phosphate transporter that in related actinobacteria is responsive to phosphate starvation (35); hence, BCG contains null mutations in both the high-affinity Pst phosphate uptake system and the Ugp system. This may reflect BCG's in vitro growth conditions but will undoubtedly impair in vivo phosphate acquisition and hence may be implicated in attenuation.

Transcriptional regulators. Comparison of the transcriptomes of *M. bovis* 2122/97 and BCG Pasteur revealed that 133 genes showed a minimum twofold difference in expression across the strains (9). It is probable that some of these expression differences reflect differences between the United Kingdom and French clades of *M. bovis* rather than a difference between virulent and attenuated strains. However, the genes for three transcriptional regulators, BCG3734, BCG3145, and BCG2507c, show nsSNPs across all BCG strains compared to virulent *M. bovis* strains, mutations which may explain global expression differences between *M. bovis* BCG and *M. bovis*.

BCG3734 encodes the CRP, a global gene regulator. Previous work has shown that there are two nsSNPs in BCG3734 compared to the *M. bovis* and *M. tuberculosis* genes, which encode E178K and L47P substitutions (54). These mutations have been shown to enhance the binding of BCG CRP to its DNA binding sites (4, 30); however, this enhanced binding does not appear to play a role in attenuation of BCG (30).

The LuxR family regulator BCG2507c contains an N-terminal adenylate/guanylate cyclase catalytic domain, a putative ATPase domain (COG3903 superfamily), and a C-terminal helix-turn-helix domain. The D535E mutation in BCG2507c falls in the ATPase domain; as it is a conservative substitution, its functional consequences are expected to be minimal.

BCG3145 is a member of the AfsR/DnrI/SARP (*Streptomyces* antibiotic regulatory protein) class of transcriptional regulators. This class also contains EmbR, the regulator of three arabinosyltransferases that are the targets of the front-line tuberculosis drug ethambutol (8). The structure of EmbR has been elucidated, revealing DNA binding, bacterial transcriptional activation (BTA), and forkhead-associated domains (1). While BCG3145 lacks the forkhead-associated domain, the E159G mutation in BCG3145 mutates to glycine a conserved glutamic acid residue located in a tetratricopeptide repeat in the BTA domain (region T3). Tetratricopeptide repeat domains are associated with protein-protein interactions (16), and a conserved core (helices T1 to T7) of the BTA domain seems to be required for proper function of SARP family proteins (1, 51). Hence, the E159G mutation may affect the ability of BCG3145 to regulate transcription.

Cycloserine resistance. Growth on D-cycloserine can be used to differentiate *M. bovis* from *M. bovis* BCG strains, with *M. bovis* being sensitive to 0.02 mg/ml cycloserine while *M. bovis* BCG strains are resistant. The molecular basis for this phenotype is unknown. The emergence of cycloserine resistance is usually due to mutations in the *alrA* gene encoding alanine racemase or the *ddlA* gene encoding D-alanyl-D-alanine ligase (20); however, the sequences of *alrA* and *ddlA* in *M. bovis* BCG and other *M. bovis* strains are identical, as are their expression levels (9). The other possibility is that the cycloserine transporter CycA (which also transports D-serine, D-alanine, and glycine) is defective for cycloserine transport in BCG. CycA does in fact contain an nsSNP across all BCG strains, introducing a G122S mutation. It is noteworthy that in all sequenced mycobacterial CycA proteins this position is occupied by glycine or sometimes, in other bacteria, by alanine. David showed that resistance to cycloserine in *M. tuberculosis* can be due to defective transport of the antibiotic (17), and so it is tempting to speculate that the G122S mutation plays some role in the inherent cycloserine resistance of BCG. This would also suggest that BCG may be defective in D-alanine, glycine, and D-serine transport, but this has not been reported.

DISCUSSION

Phylogenetic relationships based on SNPs identified by comparing two sequenced strains will inevitably place these latter strains as most distant in any phylogeny (2, 58). Phylogenies with these characteristics can therefore be deemed "preselected" or "linear" phylogenies. This is evident in our phylogeny (Fig. 3). However, this does not invalidate the resulting tree; our prior knowledge of the population structure of *M. bovis* allowed us to select strains which encompass the diversity within the population and to address our primary question of which SNPs were acquired during the initial derivation of BCG. Indeed, the SNP phylogeny is congruent with trees con-

structed using spoligotyping and deletion typing (52, 53). The one divergence from previous trees is the positioning of BCG Tice and Frappier; BCG Tice and Pasteur shared six SNPs not seen in Frappier, hence placing Tice closer to BCG Pasteur. In a previous schema BCG Frappier had been positioned closer to BCG Pasteur than Tice (9). However, Rosenthal is known to have mixed BCG Pasteur with BCG Tice in the early 1950s to correct for overattenuation (19). From the SNP data presented here, it appears that this mixture of BCG Tice and BCG Pasteur resolved as a variant of BCG Pasteur. In a recent DNA array-based comparative study of BCG vaccine strains, Leung et al. note that BCG Tice has a distinctive duplication (DU-Tice), providing a further marker to differentiate BCG Tice (37).

We describe a range of SNPs with putative functional effects in this study. While it is straightforward to ascribe functional effects to SNPs that, for example, frameshift genes, predicting the impact of intergenic SNPs or nsSNPs is problematic. We have therefore been conservative in our interpretation of the SNP data and instead see them as the basis for focused experimental validation. Furthermore, we are mindful that 35 SNP detection reactions did not generate usable data, as is common in high-throughput projects. Hence, it is possible that informative SNPs are missing from the data set presented here, but future scans will revisit these SNPs.

What are the implications of this research for the BCG vaccine? The fact that BCG daughter strains are genetically distinct is evident from our study and previous work; however, how these genetic differences translate into variation in vaccine efficacy requires considerable functional analyses. Indeed, clarity is first needed on the question of the efficacy of BCG substrains. In a recent review of published data on the efficacy of different BCG strains in human and animal studies, Ritz and colleagues showed that BCG Pasteur, Denmark, and Glaxo were associated with better protection against challenge in the mouse model while BCG Denmark appeared the best in guinea pig models (47). However, standardization of animal protocols for such variables as the dose and route of vaccine administration, interval between immunization and challenge, and dose and route of challenge strain(s) is needed if true head-to-head comparisons of BCG daughter strains are to be made. Similarly, results from human studies suggest that there are distinct differences in immunogenicity across BCG strains, although the lack of standardization across trials again complicates interpretation. For example, in a study of neonatal vaccination in Mexico using BCG strains Brazil, Denmark, and Japan, it appeared that BCG Japan was the least immunogenic (59); however, in a South African study of neonatal vaccination using BCG strains Japan and Denmark, BCG Japan was more immunogenic than BCG Denmark (18).

Against this background it is difficult to connect genetic differences in BCG strains to vaccine efficacy. However, this is not to say that linkages cannot be made, once clear phenotypes are defined. An elegant example of the link between BCG genotype and phenotype was provided by Leung et al. (37), who disclosed a deletion in the BCG Moreau strain that explained the lack of phenolic glycolipid and phthiocerol dimycocerosates from this strain (13). As these latter lipids are potent immunomodulators, their absence from BCG Moreau may explain the lack of BCG-associated complications reported with this strain (13). Hence, the SNP differences defined herein will provide a rich source of genetic variation that can be mapped to functional differences across BCG strains.

Using high-throughput SNP screening, we have therefore identified SNPs that refine the phylogeny of British and French *M. bovis* strains, confirm previous genealogies of BCG vaccines, and define a minimal set of SNPs between virulent *M. bovis* strains and the attenuated BCG. These SNPs will therefore facilitate future studies of *M. bovis* phylogeography and the molecular basis for the attenuation of BCG.

ACKNOWLEDGMENTS

This work was funded by the Department of Environment, Food and Rural Affairs (project SE3224); the Fondation Raoul Follereau; the Wellcome Trust; and SystemsX.ch. Xiangmei Zhou was funded by BBSRC grant CPA1497, "Genomic and Post-genomics of Salmonella and Mycobacterium as paradigms of intracellular pathogens," awarded to Paul Barrow, University of Nottingham, United Kingdom.

We also thank one anonymous referee for helpful comments.

REFERENCES

1. **Alderwick, L. J., V. Molle, L. Kremer, A. J. Cozzone, T. R. Dafforn, G. S. Besra, and K. Futterer.** 2006. Molecular structure of EmbR, a response element of Ser/Thr kinase signaling in Mycobacterium tuberculosis. Proc. Natl. Acad. Sci. USA **103:**2558–2563.
2. **Alland, D., T. S. Whittam, M. B. Murray, M. D. Cave, M. H. Hazbon, K. Dix, M. Kokoris, A. Duesterhoeft, J. A. Eisen, C. M. Fraser, and R. D. Fleischmann.** 2003. Modeling bacterial evolution with comparative-genome-based marker systems: application to *Mycobacterium tuberculosis* evolution and pathogenesis. J. Bacteriol. **185:**3392–3399.
3. **Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman.** 1990. Basic local alignment search tool. J. Mol. Biol. **215:**403–410.
4. **Bai, G., M. A. Gazdik, D. D. Schaak, and K. A. McDonough.** 2007. The *Mycobacterium bovis* BCG cyclic AMP receptor-like protein is a functional DNA binding protein in vitro and in vivo, but its activity differs from that of its *M. tuberculosis* ortholog, Rv3676. Infect. Immun. **75:**5509–5517.
5. **Behr, M. A., B. G. Schroeder, J. N. Brinkman, R. A. Slayden, and C. E. Barry III.** 2000. A point mutation in the *mma3* gene is responsible for impaired methoxymycolic acid production in *Mycobacterium bovis* BCG strains obtained after 1927. J. Bacteriol. **182:**3394–3399.
6. **Behr, M. A., and P. M. Small.** 1997. Has BCG attenuated to impotence? Nature **389:**133–134.
7. **Behr, M. A., M. A. Wilson, W. P. Gill, H. Salamon, G. K. Schoolnik, S. Rane, and P. M. Small.** 1999. Comparative genomics of BCG vaccines by whole-genome DNA microarray. Science **284:**1520–1523.
8. **Belanger, A. E., G. S. Besra, M. E. Ford, K. Mikusova, J. T. Belisle, P. J. Brennan, and J. M. Inamine.** 1996. The embAB genes of Mycobacterium avium encode an arabinosyl transferase involved in cell wall arabinan biosynthesis that is the target for the antimycobacterial drug ethambutol. Proc. Natl. Acad. Sci. USA **93:**11919–11924.
9. **Brosch, R., S. V. Gordon, T. Garnier, K. Eiglmeier, W. Frigui, P. Valenti, S. Dos Santos, S. Duthoy, C. Lacroix, C. Garcia-Pelayo, J. K. Inwald, P. Golby, J. N. Garcia, R. G. Hewinson, M. A. Behr, M. A. Quail, C. Churcher, B. G. Barrell, J. Parkhill, and S. T. Cole.** 2007. Genome plasticity of BCG and impact on vaccine efficacy. Proc. Natl. Acad. Sci. USA **104:**5596–5601.
10. **Calmette, A., and C. Guérin.** 1909. Sur quelques propriétés du bacille tuberculeux d'origine, cultivé sur la bile de boeuf glycérinée. C. R. Acad. Sci. Paris **149:**716–718.
11. **Charlet, D., S. Mostowy, D. Alexander, L. Sit, H. G. Wiker, and M. A. Behr.** 2005. Reduced expression of antigenic proteins MPB70 and MPB83 in Mycobacterium bovis BCG strains due to a start codon mutation in sigK. Mol. Microbiol. **56:**1302–1313.
12. **Chen, J. M., D. C. Alexander, M. A. Behr, and J. Liu.** 2003. *Mycobacterium bovis* BCG vaccines exhibit defects in alanine and serine catabolism. Infect. Immun. **71:**708–716.
13. **Chen, J. M., S. T. Islam, H. Ren, and J. Liu.** 2007. Differential productions of lipid virulence factors among BCG vaccine strains and implications on BCG safety. Vaccine **25:**8114–8122.
14. **Cole, S. T., R. Brosch, J. Parkhill, T. Garnier, C. Churcher, D. Harris, S. V. Gordon, K. Eiglmeier, S. Gas, C. E. Barry III, F. Tekaia, K. Badcock, D. Basham, D. Brown, T. Chillingworth, R. Connor, R. Davies, K. Devlin, T. Feltwell, S. Gentles, N. Hamlin, S. Holroyd, T. Hornsby, K. Jagels, B. G. Barrell, et al.** 1998. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature **393:**537–544.
15. **Collins, D. M., R. P. Kawakami, B. M. Buddle, B. J. Wards, and G. W. de Lisle.** 2003. Different susceptibility of two animal species infected with isogenic mutants of Mycobacterium bovis identifies phoT as having roles in tuberculosis virulence and phosphate transport. Microbiology **149:**3203–3212.

16. **D'Andrea, L. D., and L. Regan.** 2003. TPR proteins: the versatile helix. Trends Biochem. Sci. **28:**655–662.
17. **David, H. L.** 1971. Resistance to D-cycloserine in the tubercle bacilli: mutation rate and transport of alanine in parental cells and drug-resistant mutants. Appl. Microbiol. **21:**888–892.
18. **Davids, V., W. A. Hanekom, N. Mansoor, H. Gamieldien, S. J. Gelderbloem, A. Hawkridge, G. D. Hussey, E. J. Hughes, J. Soler, R. A. Murray, S. R. Ress, and G. Kaplan.** 2006. The effect of bacille Calmette-Guerin vaccine strain and route of administration on induced immune responses in vaccinated infants. J. Infect. Dis. **193:**531–536.
19. **Dubos, R. J., and C. H. Pierce.** 1956. Differential characteristics in vitro and in vivo of several substrains of BCG. I. Multiplication and survival in vitro. Am. Rev. Tuberc. **74:**655–666.
20. **Feng, Z., and R. G. Barletta.** 2003. Roles of *Mycobacterium smegmatis* D-alanine:D-alanine ligase and D-alanine racemase in the mechanisms of action of and resistance to the peptidoglycan inhibitor D-cycloserine. Antimicrob. Agents Chemother. **47:**283–291.
21. **Frothingham, R., and W. A. Meeker-O'Connell.** 1998. Genetic diversity in the Mycobacterium tuberculosis complex based on variable numbers of tandem DNA repeats. Microbiology **144:**1189–1196.
22. **Garnier, T., K. Eiglmeier, J. C. Camus, N. Medina, H. Mansoor, M. Pryor, S. Duthoy, S. Grondin, C. Lacroix, C. Monsempe, S. Simon, B. Harris, R. Atkin, J. Doggett, R. Mayes, L. Keating, P. R. Wheeler, J. Parkhill, B. G. Barrell, S. T. Cole, S. V. Gordon, and R. G. Hewinson.** 2003. The complete genome sequence of *Mycobacterium bovis*. Proc. Natl. Acad. Sci. USA **100:**7877–7882.
23. **Gebhard, S., S. L. Tran, and G. M. Cook.** 2006. The Phn system of Mycobacterium smegmatis: a second high-affinity ABC-transporter for phosphate. Microbiology **152:**3453–3465.
24. **Glover, R. T., J. Kriakov, S. J. Garforth, A. D. Baughn, and W. R. Jacobs, Jr.** 2007. The two-component regulatory system *senX3-regX3* regulates phosphate-dependent gene expression in *Mycobacterium smegmatis*. J. Bacteriol. **189:**5495–5503.
25. **Golby, P., K. A. Hatch, J. Bacon, R. Cooney, P. Riley, J. Allnutt, J. Hinds, J. Nunez, P. D. Marsh, R. G. Hewinson, and S. V. Gordon.** 2007. Comparative transcriptomics reveals key gene expression differences between the human and bovine pathogens of the Mycobacterium tuberculosis complex. Microbiology **153:**3323–3336.
26. **Gordon, S. V., K. Eiglmeier, T. Garnier, R. Brosch, J. Parkhill, B. Barrell, S. T. Cole, and R. G. Hewinson.** 2001. Genomics of Mycobacterium bovis. Tuberculosis (Edinburgh) **81:**157–163.
27. **Heermann, R., K. Altendorf, and K. Jung.** 2003. The N-terminal input domain of the sensor kinase KdpD of Escherichia coli stabilizes the interaction between the cognate response regulator KdpE and the corresponding DNA-binding site. J. Biol. Chem. **278:**51277–51284.
28. **Herring, C. D., A. Raghunathan, C. Honisch, T. Patel, M. K. Applebee, A. R. Joyce, T. J. Albert, F. R. Blattner, D. van den Boom, C. R. Cantor, and B. O. Palsson.** 2006. Comparative genome sequencing of Escherichia coli allows observation of bacterial evolution on a laboratory timescale. Nat. Genet. **38:**1406–1412.
29. **Hershberg, R., M. Lipatov, P. M. Small, H. Sheffer, S. Niemann, S. Homolka, J. C. Roach, K. Kremer, D. A. Petrov, M. W. Feldman, and S. Gagneux.** 2008. High functional diversity in Mycobacterium tuberculosis driven by genetic drift and human demography. PLoS Biol. **6:**e311.
30. **Hunt, D. M., J. W. Saldanha, J. F. Brennan, P. Benjamin, M. Strom, J. A. Cole, C. L. Spreadbury, and R. S. Buxton.** 2008. Single nucleotide polymorphisms that cause structural changes in the cyclic AMP receptor protein transcriptional regulator of the tuberculosis vaccine strain *Mycobacterium bovis* BCG alter global gene expression without attenuating growth. Infect. Immun. **76:**2227–2234.
31. **Jolley, K. A., E. J. Feil, M. S. Chan, and M. C. Maiden.** 2001. Sequence type analysis and recombinational tests (START). Bioinformatics (Oxford) **17:** 1230–1231.
32. **Kamerbeek, J., L. Schouls, A. Kolk, M. van Agterveld, D. van Soolingen, S. Kuijper, A. Bunschoten, H. Molhuizen, R. Shaw, M. Goyal, and J. van Embden.** 1997. Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. J. Clin. Microbiol. **35:**907–914.
33. **Keating, L. A., P. R. Wheeler, H. Mansoor, J. K. Inwald, J. Dale, R. G. Hewinson, and S. V. Gordon.** 2005. 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. **56:**163–174.
34. **Kimchi-Sarfaty, C., J. M. Oh, I. W. Kim, Z. E. Sauna, A. M. Calcagno, S. V. Ambudkar, and M. M. Gottesman.** 2007. A "silent" polymorphism in the MDR1 gene changes substrate specificity. Science **315:**525–528.
35. **Kocan, M., S. Schaffer, T. Ishige, U. Sorger-Herrmann, V. F. Wendisch, and M. Bott.** 2006. Two-component systems of *Corynebacterium glutamicum*: deletion analysis and involvement of the PhoS-PhoR system in the phosphate starvation response. J. Bacteriol. **188:**724–732.
36. **Lagranderie, M. R., A. M. Balazuc, E. Deriaud, C. D. Leclerc, and M. Gheorghiu.** 1996. Comparison of immune responses of mice immunized with five different *Mycobacterium bovis* BCG vaccine strains. Infect. Immun. **64:**1–9.
37. **Leung, A. S., V. Tran, Z. Wu, X. Yu, D. C. Alexander, G. F. Gao, B. Zhu, and J. Liu.** 2008. Novel genome polymorphisms in BCG vaccine strains and impact on efficacy. BMC Genomics **9:**413.
38. **Lewis, K. N., R. Liao, K. M. Guinn, M. J. Hickey, S. Smith, M. A. Behr, and D. R. Sherman.** 2003. Deletion of RD1 from Mycobacterium tuberculosis mimics bacille Calmette-Guerin attenuation. J. Infect. Dis. **187:**117–123.
39. **Mahairas, G. G., P. J. Sabo, M. J. Hickey, D. C. Singh, and C. K. Stover.** 1996. Molecular analysis of genetic differences between *Mycobacterium bovis* BCG and virulent *M. bovis*. J. Bacteriol. **178:**1274–1282.
40. **Matsunaga, I., A. Bhatt, D. C. Young, T. Y. Cheng, S. J. Eyles, G. S. Besra, V. Briken, S. A. Porcelli, C. E. Costello, W. R. Jacobs, Jr., and D. B. Moody.** 2004. *Mycobacterium tuberculosis* pks12 produces a novel polyketide presented by CD1c to T cells. J. Exp. Med. **200:**1559–1569.
41. **Mostowy, S., A. G. Tsolaki, P. M. Small, and M. A. Behr.** 2003. The in vitro evolution of BCG vaccines. Vaccine **21:**4270–4274.
42. **Murphy, H. N., G. R. Stewart, V. V. Mischenko, A. S. Apt, R. Harris, M. S. McAlister, P. C. Driscoll, D. B. Young, and B. D. Robertson.** 2005. The OtsAB pathway is essential for trehalose biosynthesis in Mycobacterium tuberculosis. J. Biol. Chem. **280:**14524–14529.
43. **Nei, M., and T. Gojobori.** 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. **3:**418–426.
44. **Parish, T., D. A. Smith, G. Roberts, J. Betts, and N. G. Stoker.** 2003. The senX3-regX3 two-component regulatory system of Mycobacterium tuberculosis is required for virulence. Microbiology **149:**1423–1435.
45. **Paterson, A. B.** 1948. The production of bovine tuberculoprotein. J. Comp. Pathol. **58:**302–313.
46. **Pym, A. S., P. Brodin, R. Brosch, M. Huerre, and S. T. Cole.** 2002. Loss of RD1 contributed to the attenuation of the live tuberculosis vaccines *Mycobacterium bovis* BCG and *Mycobacterium microti*. Mol. Microbiol. **46:**709–717.
47. **Ritz, N., W. A. Hanekom, R. Robins-Browne, W. J. Britton, and N. Curtis.** 2008. Influence of BCG vaccine strain on the immune response and protection against tuberculosis. FEMS Microbiol. Rev. **32:**821–841.
48. **Rocha, E. P., J. M. Smith, L. D. Hurst, M. T. Holden, J. E. Cooper, N. H. Smith, and E. J. Feil.** 2006. Comparisons of dN/dS are time dependent for closely related bacterial genomes. J. Theor. Biol. **239:**226–235.
49. **Sassetti, C. M., D. H. Boyd, and E. J. Rubin.** 2003. Genes required for mycobacterial growth defined by high density mutagenesis. Mol. Microbiol. **48:**77–84.
50. **Sassetti, C. M., and E. J. Rubin.** 2003. Genetic requirements for mycobacterial survival during infection. Proc. Natl. Acad. Sci. USA **100:**12989–12994.
51. **Sheldon, P. J., S. B. Busarow, and C. R. Hutchinson.** 2002. Mapping the DNA-binding domain and target sequences of the Streptomyces peucetius daunorubicin biosynthesis regulatory protein, DnrI. Mol. Microbiol. **44:**449–460.
52. **Smith, N. H., J. Dale, J. Inwald, S. Palmer, S. V. Gordon, R. G. Hewinson, and J. M. Smith.** 2003. The population structure of Mycobacterium bovis in Great Britain: clonal expansion. Proc. Natl. Acad. Sci. USA **100:**15271–15275.
53. **Smith, N. H., S. V. Gordon, R. de la Rua-Domenech, R. Clifton-Hadley, and R. G. Hewinson.** 2006. Bottlenecks and broomsticks: the molecular evolution of Mycobacterium bovis. Nat. Rev. Microbiol. **4:**670–681.
54. **Spreadbury, C. L., M. J. Pallen, T. Overton, M. A. Behr, S. Mostowy, S. Spiro, S. J. Busby, and J. A. Cole.** 2005. Point mutations in the DNA- and cNMP-binding domains of the homologue of the cAMP receptor protein (CRP) in *Mycobacterium bovis* BCG: implications for the inactivation of a global regulator and strain attenuation. Microbiology **151:**547–556.
55. **Stewart, G. R., L. Wernisch, R. Stabler, J. A. Mangan, J. Hinds, K. G. Laing, D. B. Young, and P. D. Butcher.** 2002. Dissection of the heat-shock response in *Mycobacterium tuberculosis* using mutants and microarrays. Microbiology **148:**3129–3138.
56. **Wang, G. Z., V. Balasubramanian, and D. W. Smith.** 1988. The protective and allergenic potency of four BCG substrains in use in China determined in two animal models. Tubercle **69:**283–291.
57. **Winder, C., S. V. Gordon, J. Dale, R. G. Hewinson, and R. Goodacre.** 2006. Metabolic fingerprints of Mycobacterium bovis cluster with molecular type: implications for genotype-phenotype links. Microbiology **152:** 2757–2765.
58. **Worobey, M.** 2005. Anthrax and the art of war (against ascertainment bias). Heredity **94:**459–460.
59. **Wu, B., C. Huang, L. Garcia, A. Ponce de Leon, J. S. Osornio, M. Bobadilla-del-Valle, L. Ferreira, S. Canizales, P. Small, M. Kato-Maeda, A. M. Krensky, and C. Clayberger.** 2007. Unique gene expression profiles in infants vaccinated with different strains of *Mycobacterium bovis* bacille Calmette-Guerin. Infect. Immun. **75:**3658–3664.

Editor: J. L. Flynn

4.1.2 Comparative genomics of *esx* genes from clinical isolates of *M. tb*

Contribution: Analysis of the SNP data, assessment of the SNP distribution across the esx *family and across different lineages of M. tb, mapping the SNP data to known Esx epitopes, phylogenetic analyses to quantify the selection pressure acting on the* esx *genes, and testing for gene conversion.*

In humans, $CD4^+$ and $CD8^+$ T-cells confer protective immunity against *M. tb* by mediating antigen-specific immune responses. Several secreted proteins from *M. tb* contain short peptide fragments bearing epitopes that bind to major histocompatibility complex (MHC) molecules, which are recognized by T-cells (Andersen *et al.* 1991). In order to evade the host immune system, variants of pathogens emerge, with alterations in epitope regions recognized by critical host immune cells.

The Esx family of *M. tb* comprises 23 proteins (EsxA to EsxW) involved in the host-pathogen interactions of *M. tb.* Two of its members, EsxA or ESAT-6 (early secretory antigenic target of 6 kDa) and EsxB or CFP-10 (culture filtrate protein of 10 kDa), are among the most frequently recognized T-cell antigens from *M. tb* (Andersen *et al.* 1995). There are 11 pairs of tandem genes encoding paralogous Esx proteins on the *M. tb* H37Rv genome. Five of the tandem *esx* gene pairs are contained within the conserved ESX-1 to ESX-5 genetic loci, encoding components of type VII secretion systems (Abdallah *et al.* 2007). The 13 *esx* genes (six gene pairs and a singleton) that are not part of the conserved ESX loci have arisen from singular duplication events (Gey Van Pittius *et al.* 2001). Consequently, the Esx family can be classified into the Mtb9.9, QILSS, and TB10.4 subfamilies based on high sequence identity. The TB10.4 subfamily comprises EsxH, which is a component of the

ESX-3 locus, and two of its paralogs (Skjøt *et al.* 2002). The Mtb9.9 and QILSS subfamilies have five members each that include gene duplicates of EsxN and EsxM, respectively, both of which belong to the ESX-5 locus. These gene duplicates show a striking level of amino acid similarity (93 to 98%), indicating recent duplication from part of the ESX-5 locus (Gey Van Pittius *et al.* 2001).

The strong antigenic properties of some of the Esx proteins have been utilized in the development of diagnostic tests (EsxA and EsxB) and protein-based vaccines (EsxH). Antigenic epitopes in several Esx proteins have been characterized experimentally. Despite the high level of sequence conservation in the Esx family, there is heterogeneity in T-cell responses to different Esx antigens. On the one hand, there is evidence of cross reactivity between highly similar epitopes in duplicated proteins from the QILSS and Mtb9.9 subfamilies, but on the other hand, even single-residue differences in the epitope sequences have been shown to alter the responder frequencies to these antigens. For example, it has been shown that human $CD4^+$ T-cell lines specific for EsxL (G22 and L23) fail to recognize peptides from EsxN (A22 and S23) and EsxV (G22 and S23), all of which are from the Mtb9.9 subfamily (Alderson *et al.* 2000).

The aim of this study was to characterize sequence diversity in *esx* genes isolated from clinical *M. tb* samples in order to identify substitutions that may impact immunogenicity. A total of 108 clinical strains isolated from TB patients at Hôpital Ambroise Paré in Boulogne-Billancourt, France were included in this study. MIRU-VNTR typing revealed that the clinical strains represented the main geographical lineages of *M. tb.* Sequencing of the *esx* genes from the clinical samples was carried out using the automated Sanger

sequencing method and the sequences obtained were compared to the *M. tb* reference strain H37Rv in order to detect nucleotide variation. A total of 109 unique SNPs were observed in the clinical dataset, 59 of which were nonsynonymous.

Assessment of the SNP distribution across the Esx proteins indicated high genetic variability in the Mtb9.9 and QILSS subfamilies, and more conservation in the ESX-1 to ESX-4 loci. Comparison of the DNA sequences of variable *esx* genes provided clear evidence for gene conversion events between duplicated paralogous genes in the QILSS and Mtb9.9 subfamilies. Mapping the SNPs to known Esx epitopes revealed that the majority of the nonsynonymous SNPs occurred in regions of the gene coding for known epitopes. We identified a nonsynonymous SNP (E68K) in EsxB present in 18 clinical isolates belonging to different lineages. This substitution occurs within a known human T-cell epitope and may influence responses to EsxB peptides that are used in commercially available diagnostic tests for TB. Estimation of the dN/dS ratio (ratio of synonymous and nonsynonymous mutations per site) suggested that some of the Esx epitopes might be under positive selection. In conclusion, our work revealed a number of previously unknown sequence polymorphisms in the Esx family, some of which may have an impact on the immunogenicity of the corresponding proteins.

In a previous investigation, Comas *et al.* uncovered more than 9,000 SNPs by whole-genome sequencing of 21 strains using the Illumina genome analyzer. Combining the SNP data with the epitope information from the IEDB database, they estimated low dN/dS values indicating that human T-cell epitopes are under purifying selection and are highly conserved (Comas *et al.* 2010). On the contrary, our study has reported that the some of the

immunodominant Esx proteins are highly polymorphic, in part because of high level of recombination. For technical reasons, we believe that Comas *et al.* have underestimated the amount of variation in the Esx family. One of the limitations of the Illumina technology is sequence assembly and identification of SNPs in repetitive regions due to the short lengths of the sequence reads. This could mask any polymorphisms generated due to gene conversion events in duplicated paralogous genes, due to erroneous mapping of short reads. The use of Sanger sequencing with primers specific for the highly similar *esx* genes helped us overcome the pitfalls of short read sequencing technology. As we report lineage-specific sequence polymorphisms within the Esx family, comparative gene expression studies of clinical isolates by Homolka *et. al* (2010) have shown lineage-specific transcription patterns in response to host-derived stimuli.

Mathematical modelling analyses by Cohen *et al.* (2008) suggest that a failure to consider mycobacterial strain diversity could have a significant negative impact on vaccine efficacy, due to strain replacement by *M. tb* variants not targeted by the vaccine. Therefore, characterization of the molecular and phenotypic differences and similarities between clinical isolates is a crucial component in the design of new control measures.

INFECTION AND IMMUNITY, Oct. 2011, p. 4042–4049 Vol. 79, No. 10
0019-9567/11/$12.00 doi:10.1128/IAI.05344-11

Comparative Genomics of *esx* Genes from Clinical Isolates of *Mycobacterium tuberculosis* Provides Evidence for Gene Conversion and Epitope Variation▽†

Swapna Uplekar,[1,4] Beate Heym,[2] Véronique Friocourt,[2] Jacques Rougemont,[3,4] and Stewart T. Cole[1]*

Global Health Institute, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland[1]; Service de Microbiologie-Hygiène, Hôpital Ambroise Paré, 9 Avenue Charles de Gaulle, 92100 Boulogne-Billancourt, France[2]; School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland[3]; and Swiss Institute of Bioinformatics, Bâtiment Génopode, Université de Lausanne, Lausanne, Switzerland[4]

Received 17 May 2011/Returned for modification 6 June 2011/Accepted 20 July 2011

The 23-membered Esx protein family is involved in the host-pathogen interactions of *Mycobacterium tuberculosis*. These secreted proteins are among the most immunodominant antigens recognized by the human immune system and have thus been used to develop vaccines and immunodiagnostic tests for tuberculosis (TB). Gene pairs for 10 Esx proteins are contained in the ESX-1 to ESX-5 loci, encoding type VII secretion systems. A subset of Esx proteins can be further classified into the Mtb9.9, QILSS, and TB10.4 subfamilies. To survey genetic diversity in the Esx family and its potential for antigenic variation, we sequenced all *esx* genes from 108 clinical isolates of *M. tuberculosis* from different clades by using a targeted approach. A total of 109 unique single nucleotide polymorphisms (SNPs) were observed, and 59 of these were nonsynonymous. Some of the resultant amino acid substitutions affect known Esx epitopes, including two in the EsxB (CFP-10) and EsxH (TB10.4) antigens. Assessment of the SNP distribution across the Esx proteins revealed high genetic variability, especially in the Mtb9.9 and QILSS subfamilies, and more conservation in the ESX-1 to ESX-4 loci. Comparison of the DNA sequences of variable *esx* genes provided clear evidence for recombination events between different genes in the same strain, some of which are predicted to truncate the corresponding protein. Many of these polymorphisms escape detection by ultrahigh-throughput sequencing using short sequence reads, as such approaches cannot distinguish between closely related genes. The *esx* gene family is dynamic, and sequence changes likely lead to immune variation.

Development of an effective tuberculosis (TB) control strategy requires detailed understanding of the biology of *Mycobacterium tuberculosis* and its interaction with human hosts. In humans, $CD4^+$ and $CD8^+$ T cells are known to mediate antigen-specific immune responses that are essential for conferring protective immunity against *M. tuberculosis* (7, 36). Several secreted proteins from mycobacteria have been shown to induce strong cellular immune responses due to the presence of short peptide fragments bearing epitopes that bind to major histocompatibility complex (MHC) molecules, which are recognized by T lymphocytes (7).

Two of the most frequently recognized T cell antigens from *M. tuberculosis* are the small secreted proteins EsxA or ESAT-6 (early secretory antigenic target of 6 kDa) and EsxB or CFP-10 (culture filtrate protein of 10 kDa), the prototypes of the Esx family (6). Genes encoding ESAT-6 (*esxA*) and CFP-10 (*esxB*) are located directly adjacent to each other and known to be cotranscribed (10). Analysis of the genome sequence of *M. tuberculosis* H37Rv revealed 11 pairs of tandem genes encoding paralogous Esx proteins located immediately downstream of the PE/PPE genes (15, 48). The *esx* family has 23 members (11 gene pairs and a singleton, *esxQ*) named *esxA* to *esxW*. Although the level of sequence identity varies between the Esx proteins (35% to 98%), all of them belong to the WXG100 family, which is characterized by a size of ~100 amino acids and the presence of a Trp-Xaa-Gly (W-X-G) motif (37).

EsxA and EsxB interact to form a 1:1 heterodimer, which appears to be essential for their secretion (10, 43). Proteins encoded by two other paralogous gene pairs, EsxR-EsxS and EsxH-EsxG, also form 1:1 complexes, suggesting that this may be typical of all Esx protein couplets (8, 32). Five of the 11 tandem gene pairs are contained within conserved genetic loci ESX-1 to ESX-5, encoding components of a type VII secretory apparatus (Table 1) (2, 26, 48). The ESX-1 system, which is responsible for the secretion of EsxA and EsxB, has been extensively studied due to its important role in *M. tuberculosis* pathogenesis (11, 12, 27). Loss of the region of difference 1 (RD1) containing the ESX-1 locus contributes to the attenuation of the vaccine strains *Mycobacterium bovis* BCG and *Mycobacterium microti* (12, 40, 41). Of the other systems, ESX-5 is known to be necessary for the secretion of PE and PPE proteins in *Mycobacterium marinum* and for macrophage subversion (3, 4). ESX-3 is essential for *in vitro* growth and may be involved in iron/zinc homeostasis (44), while the functions of ESX-2 and ESX-4 remain unknown. Comparative genomic analysis suggested that the ESX loci in mycobacteria resulted from a series of duplication events, where ESX-4 was the progenitor (26).

* Corresponding author. Mailing address: Global Health Institute, Ecole Polytechnique Fédérale de Lausanne, Station 19, CH-1015 Lausanne, Switzerland. Phone: 41 21 693 1851. Fax: 41 21 693 1790. E-mail: stewart.cole@epfl.ch.

† Supplemental material for this article may be found at http://iai.asm.org/.

▽ Published ahead of print on 1 August 2011.

TABLE 1. Overview of the *esx* family of *M. tuberculosis*[a]

Conserved ESX locus	*esxA* (ESAT-6) paralogs		*esxB* (CFP-10) paralogs	
	Inside the ESX locus	Outside the ESX locus	Inside the ESX locus	Outside the ESX locus
ESX-1	*esxA* (*rv3875*)		*esxB* (*rv3874*)	
ESX-2	*esxC* (*rv3890c*)		*esxD* (*rv3891c*)	
ESX-3	***esxH*** (***rv0288***)	***esxR*** (***rv3019c***), ***esxQ***[b] (***rv3017c***)	*esxG* (*rv0287*)	*esxS* (*rv3020c*)
ESX-4	*esxT* (*rv3444c*)		*esxU* (*rv3445c*)	
ESX-5	*esxN* (*rv1793*)	*esxI* (*rv1037c*), *esxL* (*rv1198*), *esxO* (*rv2346c*), *esxV* (*rv3619c*)	***esxM*** (***rv1792***)	***esxJ*** (***rv1038c***), ***esxK*** (***rv1197***), ***esxP*** (***rv2347c***), ***esxW*** (***rv3620c***)
None		*esxE*[c] (*rv3904c*)		*esxF*[c] (*rv3905c*)

[a] Bold, TB10.4 subfamily; underlined, Mtb9.9 subfamily; bold and underlined, QILSS subfamily.
[b] *esxQ* does not occur in tandem with a *cfp-10* homologue.
[c] *esxE* and *esxF* show no significant homology to any conserved ESX loci.

The 13 *esx* genes that are not part of the ESX-1 to ESX-5 loci seem to have arisen from singular duplication events (26). Consequently, the Esx family can be classified into distinct subfamilies based on high sequence identity between certain members (Table 1). The TB10.4 subfamily comprises *esxH*, which is a component of the ESX-3 locus, and two of its paralogs (45). The Mtb9.9 (31) and QILSS (48) subfamilies have 5 members each that include gene duplicates of *esxN* and *esxM*, respectively, both of which belong to the ESX-5 locus. These gene duplicates show a striking level of amino acid similarity (93 to 98%), indicating recent duplication from part of the ESX-5 locus (26). Comparative proteomics between *M. tuberculosis* H37Rv and the attenuated strain *M. tuberculosis* H37Ra revealed the absence of some EsxM and EsxN paralogs in *M. tuberculosis* H37Ra (28). Along with the ESX-1 system, two other loci, encoding paralogs of EsxM and EsxN, have been deleted from the vaccine strain *M. bovis* BCG: the RD5 and RD8 loci comprising *esxP-esxO* and *esxW-esxV*, respectively (9, 28).

The identification of EsxA and EsxB as potent T cell antigens prompted several immunological studies in different animal models and humans (1, 6, 21, 42, 49), and it has been shown that other Esx proteins also display antigenic qualities (12). ESAT-6 and CFP-10 have been utilized in the development of diagnostic tests that can distinguish between infected and vaccinated individuals (33, 49), while EsxH (TB10.4) has been used to obtain a fusion protein-based subunit vaccine (20). Furthermore, antigenic epitopes in several Esx proteins have been characterized experimentally. Investigation of TB10.4 paralogs revealed unique antigenic epitopes in the three proteins despite the high sequence similarity (~75%) (45). Interestingly, the Mtb9.9 subfamily proteins, which display an even higher degree of amino acid similarity (≤93%), have also been shown to induce heterogeneous human T cell responses (5).

The success of *M. tuberculosis* as a pathogen has been due largely to its ability to survive in spite of a host immune response. In order to evade the host immune system, variants of pathogens emerge, with alterations in epitope regions recognized by critical host immune cells. Despite the high level of sequence conservation in the Esx family, there is heterogeneity in T cell responses to different Esx antigens. The aim of this study was to characterize sequence diversity in *esx* genes isolated from clinical *M. tuberculosis* samples in order to identify substitutions that may impact immunogenicity.

MATERIALS AND METHODS

TB patients. The clinical strains used in this study were isolated from tuberculosis patients at Hôpital Ambroise Paré in Boulogne-Billancourt, France, over a period of 3 years from 2001 to 2004. A total of 103 patients were included, among whom 58 were male and 45 were female. The patient ages ranged from 12 to 96 years, with a median age of 38 years. Localization of tuberculosis was pulmonary in 84 cases and extrapulmonary in the remaining 19 cases (pleural TB, 3 cases; lymphatic TB, 8 cases; bone and joint TB, 3 cases; meningitis, 2 cases; miliary TB, 2 cases; digestive TB, 1 case). The HIV status was unknown for 31 patients, 67 patients tested HIV negative, and 5 patients tested HIV positive.

Strains. With the exception of two patients, one single isolate per patient was included in this study. For one of the patients, 2 isolates were included, which had been isolated at a 3-week interval, while for another patient, 5 isolates were included, which had been isolated over a period of 3 years. The respiratory samples were decontaminated by the *N*-acetyl-cysteine–NaOH method. Routine laboratory procedures included microscopic examination after fluorescence staining and culture on solid Lowenstein-Jensen (Bio-Rad, Marnes-la Coquette, France) and liquid (MGIT; Becton Dickinson, Le Pont-de-Claix, France) media. Solid media were maintained at 37°C for 3 months and examined for growth once a week. The MGIT media were incubated in the Bactec MGIT 960 system and examined as soon as a positive signal was emitted. In the case of negativity, the liquid cultures were maintained in the MGIT system for 45 days. Species identification was done using the commercially available Accuprobe system (bioMérieux, Marcy l'Etoile, France), and antibiotic susceptibility testing was carried out with the Bactec MGIT 960 system.

Molecular typing. For molecular typing, a loopful of Lowenstein-Jensen culture was suspended in 150 μl of Tris-EDTA (TE) buffer and heated at 95°C for 15 min. This crude lysate was used for PCRs. The *M. tuberculosis* H37Rv reference strain was included for control purposes. Amplification of the direct repeat (DR) regions and hybridization for spoligotyping were carried out as previously described (30), and hybridization profiles were expressed using the octal code and compared with the database, SpolDB4, to determine the corresponding spoligo international types (14). Amplification of the mycobacterial interspersed repetitive-unit (MIRU) loci for MIRU–variable-number tandem-repeat (VNTR) typing was carried out as described in the past (47). The fragment sizes of the amplification products were estimated by agarose gel electrophoresis in relation to molecular weight markers, and the number of MIRU copies was determined with reference to the MIRU-VNTR allele table (see Table S1 in the supplemental material).

PCR and sequencing. Fragments bearing all *esx* genes were amplified using the sets of primers listed in Table S2A in the supplemental material. Amplification was performed using 25 μl of ReddyMix PCR Master Mix (Thermo Fisher Scientific, Inc.) and 1 μl of each primer (10 pmol). Dideoxy sequencing of the amplified gene fragments was carried out on both strands with the BigDye Terminator cycle sequencing kit (Applied Biosystems, Foster City, CA), using the primers listed in Table S2B in the supplemental material, and with an ABI 3700 DNA analyzer.

SNP detection. Sequences of *esx* genes from the clinical strains were compared with the corresponding sequences from the *M. tuberculosis* reference strain H37Rv. The positions of variant nucleotides were recorded as single nucleotide polymorphisms (SNPs) using the BLAST function on the TubercuList website (http://tuberculist.epfl.ch). By comparison of the amino acid resulting from the substitution with the reference amino acid sequence, these SNPs were further characterized as being synonymous (sSNPs; no change in amino acid) or nonsynonymous (nsSNPs; resulting in an amino acid change).

Epitope identification. A total of 93 peptides belonging to the Esx proteins comprising human T cell epitopes and major histocompatibility complex (MHC)-binding peptides were obtained from the Immune Epitope Database and Analysis Resource (50) (see Table S3 in the supplemental material). The database was accessed on 20 July 2010 (http://www.immuneepitope.org/).

Phylogenetic analysis. Comparison of the synonymous substitution rate per site (*dS*) to the nonsynonymous substitution rate per site (*dN*) via the ratio $\omega = dN/dS$ allows quantification of selection pressures on codon alignments. The number of SNPs observed in individual *esx* genes was not sufficient to determine a gene-specific *dN*/*dS* ratio from our data set. Instead, we concatenated all the codons containing SNPs to generate a single sequence for each isolate, which was used for subsequent analysis. Phylogenetic trees were obtained using the PHYLIP package (22) by implementing the neighbor-joining algorithm based on distances calculated under the K80 model. The neighbor-joining tree topologies and the sequence alignment were subjected to codon-based likelihood analysis using the CODEML (51) program in the PAML package (52). The M0 model was used to estimate a single ω ratio for all branches of the tree.

In order to screen for recombination, the genetic algorithm recombination detection (GARD) tool (38) was used, which is available on the web server of the HyPhy package (39) (http://www.datamonkey.org/). GARD searches for putative recombination breakpoints using a multiple-sequence alignment and generates the phylogenies for each nonrecombinant segment in order to assess a goodness of fit based on the Akaike information criterion (AIC) and AIC_c (AIC derived from a maximum likelihood model fit to each segment) (46). Information from all the fitted models is combined to assign a level of support to the placement of breakpoints and for different phylogenies inferred among nonrecombinant segments.

RESULTS

Diversity of bacterial strains. A total of 108 clinical samples were included in this study to investigate the genetic diversity in the 23 genes constituting the *esx* family. Recent clinical isolates were used in order to assess the current situation and avoid possible sequence biases that could have been introduced in heavily passaged laboratory strains. The clinical isolates originated from different geographical locations in Europe, North Africa, America, and Asia and represented 75 different spoligo international types. MIRU-VNTR typing revealed the distribution of the isolates across different genotypic families, representing the main geographical lineages of *M. tuberculosis* (24). A subset of the isolates could be identified as members of the Beijing (6 isolates), Haarlem and Haarlem-like (23 isolates), East African Indian (EAI; 6 isolates), Latin American (LAM; 4 isolates), West African (7 isolates), and East African (4 isolates) families, representing the breadth of genomic diversity in the species (see Table S1 in the supplemental material). One of the clinical isolates was an *M. bovis* strain. The *esx* gene sequences obtained from the clinical samples and the *M. tuberculosis* reference strain H37Rv were compared in order to detect any variation at the nucleotide level. Gene sequences of *esx* from the attenuated strain H37Ra are identical to those of H37Rv (53).

Comparative *esx* genomics of clinical isolates. In the entire clinical data set, a total of 797 substitutions were identified, corresponding to 109 unique SNPs across the 23 *esx* genes, each of which occurred in one or several isolates (see Table S4 in the supplemental material). Analysis of the 109 SNPs revealed 50 sSNPs and 59 nsSNPs. Based on the spoligotype and MIRU-VNTR information available for the clinical isolates, we investigated whether the highly prevalent SNPs were specific to any of the geographical lineages represented in our data set (see Table S1 and Fig. S1 in the supplemental material). Interestingly, three SNPs in *esxV*, including two nsSNPs, Q20L and S23L, and an sSNP in codon 57, occurred in 74 isolates which belonged to different lineages. The high prevalence of these SNPs in the clinical data set indicates that these positions in *esxV* are lineage markers. An E68K substitution in EsxB (CFP-10) was observed in 19 isolates, mostly belonging to the Haarlem, Haarlem-like, and LAM families, all of which represent the Euro-American lineage (24). On the other hand, there were also several SNPs with a very low prevalence that appeared to be specific to particular strains or lineages. The only *M. bovis* isolate in the data set showed two nsSNPs, M82V in EsxE and W58stop in EsxF, which were not observed in the other isolates. These were confirmed to be *M. bovis* specific upon comparison with the genome sequences of *M. bovis* AF2122/97 and *M. bovis* BCG Pasteur 1173P2 (13, 25).

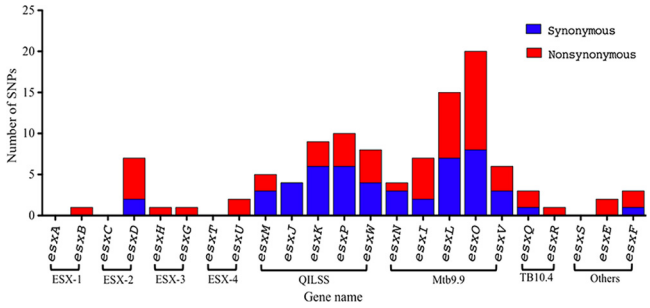


FIG. 1. Distribution of nonsynonymous (red) and synonymous (blue) SNPs across 23 *esx* genes. Brackets indicate genes encoded in the ESX-1 to ESX-4 loci and members of the Mtb9.9, QILSS, and TB10.4 subgroups.

SNPs involving stop codons can have a significant impact on the structure and function of a protein. Apart from EsxF, two other *esx* genes had substitutions that introduced stop codons. These include EsxL, containing a Q76stop, which was observed in three isolates, and EsxW, containing a Q59stop, which was observed in one isolate. In contrast to EsxW, a stop59Q substitution in EsxM was present in nine clinical isolates. Another nsSNP, A71S in EsxH, was seen in only two isolates, both of which represented West African lineages (I and II). Three out of six Beijing strains harbored an nsSNP, P63S in EsxU, that was not present in any of the other isolates.

SNP distribution across *esx* genes. Analysis of 108 clinical isolates revealed SNPs in 19 of the 23 *esx* genes. Sequences for *esxA*, *esxC*, *esxT*, and *esxS* for all clinical isolates were invariant. The distribution of the SNPs was not uniform across the *esx* family members (Fig. 1). Grouping of the *esx* genes into their subfamilies revealed that the four secretion systems ESX-1 to ESX-4 displayed a low level of variation in general and a strikingly low level of synonymous substitutions. On the other hand, genes belonging to the Mtb9.9 and the QILSS subfamilies, including the ESX-5 system, accounted for the majority of the variation, displaying a large number of sSNPs as well as nsSNPs. On average, from whole-genome comparisons of two strains of *M. tuberculosis* (such as H37Rv and CDC1551 [23]),

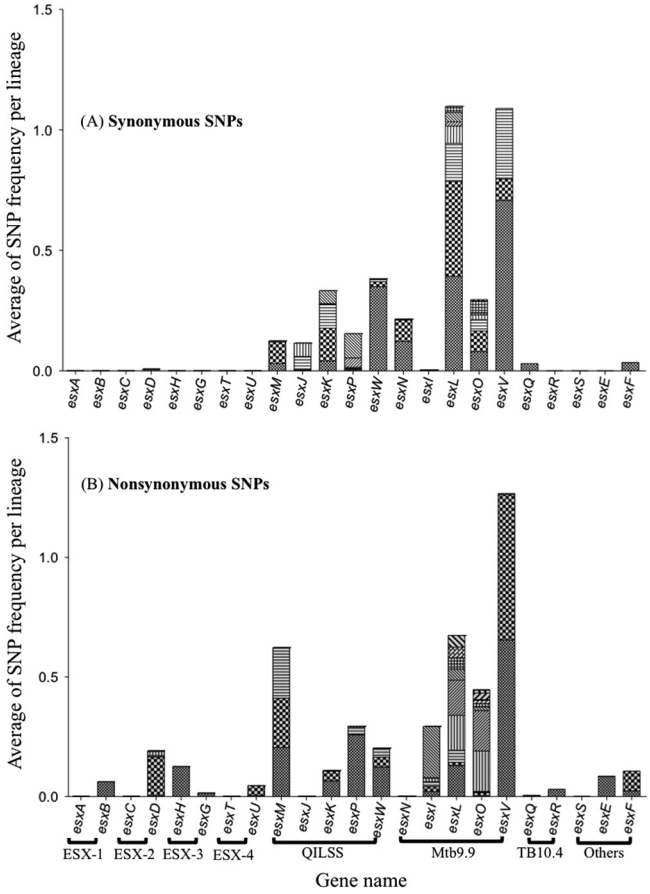


FIG. 2. Incidence of synonymous (A) and nonsynonymous (B) SNPs in 23 *esx* genes seen across 108 clinical isolates. Each pattern in the bars indicates a unique SNP, and the height of each bar correlates with the average frequencies per lineage.

the SNP frequency is 1 per 4.1 kb. In contrast, in our *esx* data set, the frequency compared to H37Rv ranges from a median of 1 per kb to a maximum of 3 per kb depending on the clinical isolate examined.

To estimate the incidence of SNPs across the clinical data set, while accounting for lineage-specific polymorphisms, we divided the clinical isolates into groups based on the lineage information. The isolates that could not be successfully genotyped were considered one separate group. The frequency of occurrence of an SNP was calculated separately for each lineage and represented as the average of SNP frequency per lineage. SNPs were grouped together based on the gene in which they occurred (Fig. 2). As seen with the number of polymorphisms, SNPs in the Mtb9.9 and QILSS genes were present in a large number of clinical isolates. In contrast, SNPs in genes encoded in the ESX-1 to ESX-4 loci and the TB10.4 paralogs were observed in very few isolates. Due to differences in the proportion of SNPs observed in different *esx* subfamilies, we carried out phylogenetic analysis separately on the two main classes of *esx* genes: the ESX-5 paralogs (Mtb9.9 and QILSS members) and components of the ESX-1 to ESX-4 loci.

Gene conversion in ESX-5 paralogs. A particular feature of SNPs observed in *esx* genes belonging to the Mtb9.9 and QILSS subfamilies was the cooccurrence of two or more SNPs in neighboring codons. All the isolates containing an sSNP in codon 40 of *esxJ* also had an sSNP substitution in codon 41. The same was true for sSNPs in codons 6 and 8 of *esxL*. The highly prevalent Q20L and S23L SNPs in *esxV* occurred in over 70% of the clinical isolates, all of which harbored an sSNP in codon 57 of *esxV*. Reciprocal substitutions, Q59stop and stop59Q, were also seen in the *esxM* and *esxW* genes.

Sequences for *esxP* from 17 clinical isolates contained an sSNP in the second codon followed by an nsSNP (T3S) in the third codon (Fig. 3A). The resulting codons in the *esxP* clinical sequences were identical to the *M. tuberculosis* H37Rv sequences of the paralogous genes *esxK*, *esxJ*, and *esxM* from the QILSS subfamily. The nucleotide sequences in *esxP*, *esxK*, and *esxJ* are also identical for 108 bp downstream of the SNPs. Sequences up to 80 bp upstream of the SNPs, including the intergenic regions, are identical only between *esxP* and *esxK* (Fig. 3A). The *esxK* sequences of the 17 clinical isolates and *M. tuberculosis* H37Rv were identical, but note that in one other clinical isolate, codons 40, 41, and 42 all contain the same sSNPs as those present in *esxP* of *M. tuberculosis* H37Rv. We subjected the 17 *esxK* and *esxP* sequences to a recombination screen using the GARD tool. The results revealed the presence of one significant recombination breakpoint at the seventh position (codon 3) and two suggestive potential breakpoints between codons 40 and 42 (Fig. 3B).

Phylogenetic analysis. In order to calculate the ω (*dN*/*dS*) ratio to quantify the selective pressure acting on the *esx* genes, we chose the random-site model, which accounts for variable selective pressures across codons and can detect amino acid residues under either positive or negative selection. Analysis of codon alignments, including SNPs observed in all *esx* genes, gave a low ω of 0.35, because the *dS* value of the ESX-5 paralogs is very high. Mtb9.9 and QILSS genes harbor 46 of the 50 sSNPs which are observed in a large proportion of the clinical isolates. However, the codon-based analysis assumes a single tree topology for the whole alignment and does not account for recombination. Therefore, we generated a codon alignment for SNPs in *esx* genes belonging to the ESX-1 to ESX-4 loci and reanalyzed them using CODEML. The estimated ω value of 1.66 ($dN/dS > 1$) indicated diversifying selection in this subset of *esx* genes.

SNPs occurring in Esx epitopes. In order to identify whether any of the SNPs observed in our clinical data set could impact the immunogenicity of the Esx proteins, we searched the IEDB database for experimentally confirmed human T cell and MHC-binding peptides specific to the Esx family. A total of 93 peptides were obtained, 80 of which comprised overlapping peptides that belonged to EsxA and EsxB and spanned the entirety of these proteins. Sequence alignments of the EsxA (excluding EsxQ due to a low level of sequence homology to other EsxA paralogs) and EsxB paralogs (Fig. 4) highlight the antigenic epitopes and positions of all the SNPs observed in the clinical data set. There were 3 sSNP and 15 nsSNPs occurring in known epitope regions. Excluding the 80 epitopes from EsxA and EsxB, 9 of the 13 known epitopes in the other Esx proteins had at least one or more SNPs (Table 2). Mtb9.9 genes, which harbored 12 of the 18 SNPs, were overrepresented. The corresponding proteins share an MHC-binding peptide, MIRAQA [GA][SL]LEA, at positions 16 to 26 that is subject to se-

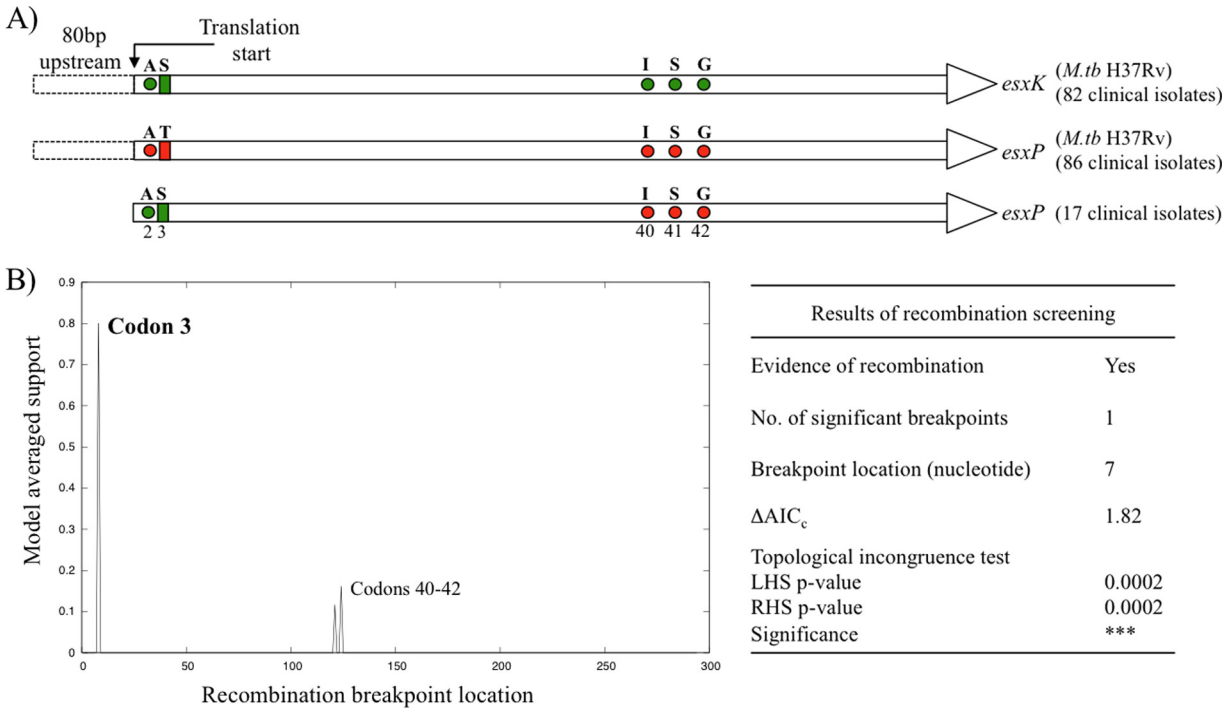


FIG. 3. Putative recombination between *esxP* and *esxK* genes. (A) Schematic representation of sequence variation in *esxK* and *esxP* from *M. tuberculosis* H37Rv and *esxP* sequences from clinical isolates. The *M. tuberculosis* H37Rv sequences for *esxK* and *esxP* are identical except at the positions indicated with circles (silent polymorphisms) and rectangles (amino acid differences). Positions of the SNPs observed in *esxP* from 17 clinical isolates are color coded depending on their similarity to the corresponding gene in *M. tuberculosis* H37Rv. Nucleotide sequences up to 80 bp upstream of the gene are also identical in *esxP* and *esxK*. (B) Plot and summary of GARD results showing the location of the putative recombination breakpoint. ΔAIC$_c$ indicates improvement of the AIC$_c$ score compared to that of the model with zero breakpoints. Significance of the difference in topologies between the partitions to the left and right of the breakpoint is indicated with *P* values for both sides.

quence variation. Interestingly, three of the Mtb9.9 proteins, EsxI, EsxV, and EsxO, showed a Q20L polar-to-nonpolar substitution. One of the two variable positions in the MHC-binding peptide reflected an amino acid change, S23L in EsxI and EsxV, and a reverse L23S change in EsxO. In fact, we were able to identify several recurring positions both inside and outside known epitopes (Fig. 4) at which SNPs were observed in two or more Esx proteins.

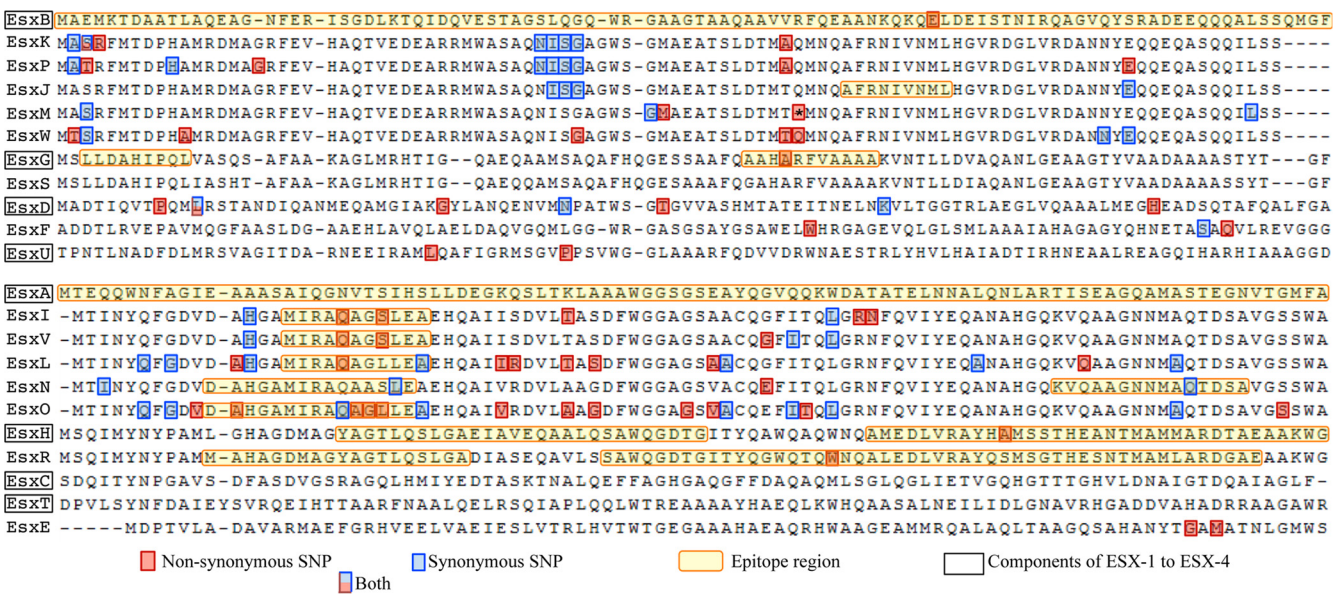


FIG. 4. Multiple sequence alignment of EsxB (CFP-10) and EsxA (ESAT-6) paralogs from *M. tuberculosis* H37Rv. Positions containing nonsynonymous (red boxes) and synonymous (blue boxes) changes have been annotated for each protein. Known immune epitope regions are highlighted in yellow.

TABLE 2. SNPs affecting known epitope regions

Protein	SNP	Type of change	Amino acid sequence[a]		No. of isolates
			T cell epitope	MHC-II binding	
EsxB	E68K	Acidic to basic	KQELDEISTNIRQAG		19
EsxG	A56T	Nonpolar to polar	AAHARFVAA		1
EsxH	A71S	Nonpolar to polar	AMEDLVRAYHAMSSTHEA		2
EsxQ	S90A	Polar to nonpolar	RCRRALRQIGVLERPVGDSS		1
EsxQ	L76P	Nonpolar to nonpolar	RCRRALRQIGVLERPVGDSS		1
EsxR	W58C	Nonpolar to polar	YQGWQTQWNQALEDLVRAYQ		3
EsxI	S23L	Polar to nonpolar		MIRAQAGSL	1
EsxI	Q20L	Polar to nonpolar		MIRAQAGSL	13
EsxL	Q20L	Polar to nonpolar		MIRAQAGLL	1
EsxN	24	Synonymous		DAHGAMIRAQAAS*L*E	4
EsxN	83	Synonymous		KVQAAGNNMA*Q*TDSA	1
EsxO	L23S	Nonpolar		MIRAQAGLL	6
EsxO	G22A	Polar to nonpolar		MIRAQAGLL	6
EsxO	A21P	Nonpolar to nonpolar		MIRAQAGLL	1
EsxO	20	Synonymous		MIRA*Q*AGLL	1
EsxO	A12D	Nonpolar to acidic		DAHGAMIRAQAGLLE	1
EsxV	S23L	Polar to nonpolar		MIRAQAGSL	77
EsxV	Q20L	Polar to nonpolar		MIRAQAGSL	74

[a] Obtained from the IEDB resource (49). Amino acids from codons containing nsSNPs are underlined, and sSNPs are italicized.

DISCUSSION

In this study, all 23 *esx* genes from 108 clinical samples were sequenced in order to identify substitutions that may have an impact on the immunogenicity or function of Esx proteins. A total of 797 substitutions corresponding to 109 distinct SNPs were observed in the entire clinical data set. Our analysis indicated that *esx* genes encoded within the ESX-1 to ESX-4 loci displayed less variation overall than the *esx* genes located outside these loci. EsxA and EsxB have been most extensively studied owing to their crucial role in *M. tuberculosis* pathogenesis and applications in the development of vaccines and diagnostics. A past study by Musser et al. involving sequencing of 24 important *M. tuberculosis* antigens from 16 clinical strains revealed no variation in EsxA and EsxB sequences (35). The commercially available gamma interferon (IFN-γ) release assays (IGRA) used for diagnosing TB, QuantiFERON gold test (33), and T-SPOT.TB (34) employ peptides from ESAT-6 (EsxA) and CFP-10 (EsxB). While no sequence variation has yet been observed in EsxA, there is an amino acid substitution (E68K) in EsxB present in 18 of the 108 strains representing diverse lineages. This substitution occurs within a known human T cell epitope, so it may influence responses to EsxB peptides in the IGRA. Although the strains containing the EsxB SNP represent different lineages, they also share three sSNPs in codons 6 and 8 of EsxL and codon 57 of EsxV.

A more recent study by Davila et al. to assess the genetic diversity of EsxA and EsxH genes among 88 clinical isolates revealed no variation in either of the two genes (17), and this is essentially in agreement with our findings (Table 2). Our results showed that three out of five EsxA paralogs encoded by the ESX-1 to ESX-5 loci were invariant across the entire clinical data set. There was also an absence of silent substitutions in the *esx* components of the ESX-1 to ESX-4 loci with the exception of *esxD*. Interestingly, the majority of the nsSNPs occur in regions of the gene coding for known epitopes. Our estimation of dN/dS ($\omega = 1.66$) indicated that some of these epitopes might be under positive selection. These findings are in contrast with those in a recent publication by Comas et al. demonstrating the hyperconservation of human T cell epitopes among 21 strains representing the six main geographical lineages of *M. tuberculosis* (16). Their investigation uncovered more than 9,000 SNPs by whole-genome sequencing of 21 strains using the Illumina genome analyzer. Using the epitope information from the IEDB database, they estimated low ω values ($dN/dS < 1$) in antigens compared to those for essential and nonessential genes.

For technical reasons, we believe that Comas et al. (16) have underestimated the amount of variation in the Esx family. One of the limitations of the Illumina technology is sequence assembly and identification of SNPs in repetitive regions due to the short lengths of the sequence reads. This could mask any polymorphisms generated due to gene conversion events in duplicated paralogous genes, due to erroneous mapping of short reads. This includes genes in the Mtb9.9 and QILSS subfamilies. Our study involved traditional sequencing with primers specific for the different *esx* genes and significantly longer sequence reads than those generated by the Illumina technology and thus avoids this pitfall. Interestingly, one of the three outliers in the Comas et al. study was *esxH*, which displayed a high level of variation in epitope regions ($dN/dS > 1$).

Members of the Mtb9.9 and QILSS subfamilies show the highest levels (93 to 98%) of amino acid identity among the Esx family. Although several immunogenic epitopes were located in regions of 100% sequence identity, some of the major immunodominant epitopes were also identified in regions of sequence diversity in proteins from the Mtb9.9 and QILSS subfamilies. A recent study by Jones et al. confirmed that the Esx proteins, whose genes were not part of the ESX-1 to ESX-5 loci, showed similar levels of immunogenicity to the Esx proteins from the ESX-1 to ESX-5 loci (29). Strikingly, even single-residue differences in the epitope sequences altered the responder frequencies to these antigens. The amino acid residues critical for antigenicity include T58 for the QILSS subfamily and G22 and S23 for the Mtb9.9 subfamily. In the

clinical data set, we observed an A58T mutation in EsxP and EsxK and a converse T58A substitution in EsxW. S23L substitutions in EsxI and EsxV and L23S and G22A substitutions in EsxO were also found in several isolates. A study using human $CD4^+$ T cells performed by Alderson et al. demonstrated that T cell lines specific for EsxL (G22 and L23) failed to recognize peptides from EsxN (A22 and S23) and EsxV (G22 and S23) (5). On the other hand, peptide fragments from EsxV, which is absent in *M. bovis*, have been shown to induce IFN-γ responses in cattle infected with TB (29), suggesting cross-reactivity between highly similar epitopes in duplicated proteins. The conservation of silent as well as nonsynonymous SNPs between paralogs and orthologs of the Mtb9.9 family, as seen in *esxP* and *esxM*, respectively, suggests that even minor variation within the Mtb9.9 and QILSS families could significantly alter the length and expression of this protein subfamily. Based on the clinical SNPs, we were able to provide evidence for putative recombination between *esxK* and *esxP*. It is possible that gene conversion events may occur between other duplicated paralogous genes in the QILSS and Mtb9.9 subfamilies. Gene conversion has been described in members of the PE-PGRS gene family in *M. bovis*, where homogenotization has been reported for genes *Mb1485* and *Mb1487* (19). Frequent homologous recombination events in the highly repetitive PE/PPE multigene families with a potential role in antigenic variability have been described in *M. tuberculosis* (18), and the present work indicates that the *esx* gene family is also dynamic.

Amino acid substitutions encoded by duplicated genes may allow for genetic drift, by regulating expression of the functionally similar protein paralogs that differ in their immunodominant epitopes. Although *in silico* prediction of T cell binding epitopes is possible, we included only the experimentally confirmed epitope data in our analysis. With the exception of EsxA and EsxB, there is very little or no epitope data in IEDB for other members of the Esx family. As new Esx epitopes are identified in the future, we should be able to assess the role of the other nsSNPs observed in our clinical data set. In conclusion, our analysis of *esx* genes has revealed a number of previously unknown sequence polymorphisms in the highly immunogenic Esx family, including some in known epitope regions, which may affect the immunogenicity of the corresponding protein. This is the case with EsxB (CFP-10) and EsxH (TB10.4) and has implications for the fields of vaccines and diagnostics.

ACKNOWLEDGMENTS

We thank Megan Murray for helpful discussions and comments on the manuscript.

This study was supported by funding from the European Community's Seventh Framework Program (FP7/2007-2013) under grant agreement no. 201762 and by SystemsX.ch and the Swiss National Science Foundation (31003A-125061).

S.U. and J.R. are affiliated with the Swiss Institute of Bioinformatics.

REFERENCES

1. **Aagaard, C., M. Govaerts, L. Meng Okkels, P. Andersen, and J. M. Pollock.** 2003. Genomic approach to identification of *Mycobacterium bovis* diagnostic antigens in cattle. J. Clin. Microbiol. **41:**3719–3728.
2. **Abdallah, A. M., et al.** 2007. Type VII secretion—mycobacteria show the way. Nat. Rev. Microbiol. **5:**883–891.
3. **Abdallah, A. M., et al.** 2008. The ESX-5 secretion system of *Mycobacterium marinum* modulates the macrophage response. J. Immunol. **181:**7166–7175.
4. **Abdallah, A. M., et al.** 2006. A specific secretion system mediates PPE41 transport in pathogenic mycobacteria. Mol. Microbiol. **62:**667–679.
5. **Alderson, M. R., et al.** 2000. Expression cloning of an immunodominant family of *Mycobacterium tuberculosis* antigens using human CD4(+) T cells. J. Exp. Med. **191:**551–560.
6. **Andersen, P., A. B. Andersen, A. L. Sørensen, and S. Nagai.** 1995. Recall of long-lived immunity to *Mycobacterium tuberculosis* infection in mice. J. Immunol. **154:**3359–3372.
7. **Andersen, P., D. Askgaard, L. Ljungqvist, M. W. Bentzon, and I. Heron.** 1991. T-cell proliferative response to antigens secreted by *Mycobacterium tuberculosis*. Infect. Immun. **59:**1558–1563.
8. **Arbing, M. A., et al.** 2010. The crystal structure of the *Mycobacterium tuberculosis* Rv3019c-Rv3020c ESX complex reveals a domain-swapped heterotetramer. Protein Sci. **19:**1692–1703.
9. **Behr, M. A., et al.** 1999. Comparative genomics of BCG vaccines by whole-genome DNA microarray. Science **284:**1520–1523.
10. **Berthet, F. X., P. B. Rasmussen, I. Rosenkrands, P. Andersen, and B. Gicquel.** 1998. A *Mycobacterium tuberculosis* operon encoding ESAT-6 and a novel low-molecular-mass culture filtrate protein (CFP-10). Microbiology **144**(Pt. 11)**:**3195–3203.
11. **Brodin, P., et al.** 2005. Functional analysis of early secreted antigenic target-6, the dominant T-cell antigen of *Mycobacterium tuberculosis*, reveals key residues involved in secretion, complex formation, virulence, and immunogenicity. J. Biol. Chem. **280:**33953–33959.
12. **Brodin, P., et al.** 2004. Enhanced protection against tuberculosis by vaccination with recombinant *Mycobacterium microti* vaccine that induces T cell immunity against region of difference 1 antigens. J. Infect. Dis. **190:**115–122.
13. **Brosch, R., et al.** 2007. Genome plasticity of BCG and impact on vaccine efficacy. Proc. Natl. Acad. Sci. U. S. A. **104:**5596–5601.
14. **Brudey, K., et al.** 2006. *Mycobacterium tuberculosis* complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology. BMC Microbiol. **6:**23.
15. **Cole, S. T., et al.** 1998. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature **393:**537–544.
16. **Comas, I., et al.** 2010. Human T cell epitopes of *Mycobacterium tuberculosis* are evolutionarily hyperconserved. Nat. Genet. **42:**498–503.
17. **Davila, J., L. Zhang, C. F. Marrs, R. Durmaz, and Z. Yang.** 2010. Assessment of the genetic diversity of *Mycobacterium tuberculosis esxA*, *esxH*, and *fbpB* genes among clinical isolates and its implication for the future immunization by new tuberculosis subunit vaccines Ag85B-ESAT-6 and Ag85B-TB10.4. J. Biomed. Biotechnol. **2010:**208371.
18. **Delogu, G., and M. J. Brennan.** 2001. Comparative immune response to PE and PE_PGRS antigens of *Mycobacterium tuberculosis*. Infect. Immun. **69:**5606–5611.
19. **Delogu, G., S. T. Cole, and R. Brosch.** 2008. The PE and PPE gene protein families of *M. tuberculosis*, p. 131–150. *In* S. H. E. Kaufmann, P. van Helden, E. Rubin, and W. J. Britton (ed.), Handbook of tuberculosis. Wiley-VHC, Weinheim, Germany.
20. **Dietrich, J., et al.** 2005. Exchanging ESAT6 with TB10.4 in an Ag85B fusion molecule-based tuberculosis subunit vaccine: efficient protection and ESAT6-based sensitive monitoring of vaccine efficacy. J. Immunol. **174:**6332–6339.
21. **Elhay, M. J., T. Oettinger, and P. Andersen.** 1998. Delayed-type hypersensitivity responses to ESAT-6 and MPT64 from *Mycobacterium tuberculosis* in the guinea pig. Infect. Immun. **66:**3454–3456.
22. **Felsenstein, J.** 1989. PHYLIP—phylogeny inference package (version 3.2). Cladistics **5:**164–166.
23. **Fleischmann, R. D., et al.** 2002. Whole-genome comparison of *Mycobacterium tuberculosis* clinical and laboratory strains. J. Bacteriol. **184:**5479–5490.
24. **Gagneux, S., and P. M. Small.** 2007. Global phylogeography of *Mycobacterium tuberculosis* and implications for tuberculosis product development. Lancet Infect. Dis. **7:**328–337.
25. **Garnier, T., et al.** 2003. The complete genome sequence of *Mycobacterium bovis*. Proc. Natl. Acad. Sci. U. S. A. **100:**7877–7882.
26. **Gey Van Pittius, N. C., et al.** 2001. The ESAT-6 gene cluster of *Mycobacterium tuberculosis* and other high G+C Gram-positive bacteria. Genome Biol. **2**(10)**:**44.1–44.18.
27. **Harboe, M., T. Oettinger, H. G. Wiker, I. Rosenkrands, and P. Andersen.** 1996. Evidence for occurrence of the ESAT-6 protein in *Mycobacterium tuberculosis* and virulent *Mycobacterium bovis* and for its absence in *Mycobacterium bovis* BCG. Infect. Immun. **64:**16–22.
28. **He, X. Y., Y. H. Zhuang, X.-G. Zhang, and G. L. Li.** 2003. Comparative proteome analysis of culture supernatant proteins of *Mycobacterium tuberculosis* H37Rv and H37Ra. Microbes Infect. **5:**851–856.
29. **Jones, G. J., S. V. Gordon, R. G. Hewinson, and H. M. Vordermeier.** 2010. Screening of predicted secreted antigens from *Mycobacterium bovis* reveals the immunodominance of the ESAT-6 protein family. Infect. Immun. **78:**1326–1332.
30. **Kamerbeek, J., et al.** 1997. Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. J. Clin. Microbiol. **35:**907–914.
31. **Louise, R., V. Skjøt, E. M. Agger, and P. Andersen.** 2001. Antigen discovery and tuberculosis vaccine development in the postgenomic era. Scand. J. Infect. Dis. **33:**643–647.

32. **Mahmood, A., et al.** 2011. Molecular characterization of secretory proteins Rv3619c and Rv3620c from *Mycobacterium tuberculosis* H37Rv. FEBS J. **278**(2)**:**341–353.
33. **Mazurek, G. H., et al.** 2001. Comparison of a whole-blood interferon gamma assay with tuberculin skin testing for detecting latent *Mycobacterium tuberculosis* infection. JAMA **286:**1740–1747.
34. **Meier, T., H. P. Eulenbruch, P. Wrighton-Smith, G. Enders, and T. Regnath.** 2005. Sensitivity of a new commercial enzyme-linked immunospot assay (T SPOT-TB) for diagnosis of tuberculosis in clinical practice. Eur. J. Clin. Microbiol. Infect. Dis. **24:**529–536.
35. **Musser, J. M., A. Amin, and S. Ramaswamy.** 2000. Negligible genetic diversity of *Mycobacterium tuberculosis* host immune system protein targets: evidence of limited selective pressure. Genetics **155:**7–16.
36. **North, R. J., and Y. J. Jung.** 2004. Immunity to tuberculosis. Annu. Rev. Immunol. **22:**599–623.
37. **Pallen, M. J.** 2002. The ESAT-6/WXG100 superfamily—and a new Gram-positive secretion system? Trends Microbiol. **10:**209–212.
38. **Pond, S. L. K.** 2005. A genetic algorithm approach to detecting lineage-specific variation in selection pressure. Mol. Biol. Evol. **22:**478–485.
39. **Pond, S. L. K., S. D. W. Frost, and S. V. Muse.** 2005. HyPhy: hypothesis testing using phylogenies. Stat. Methods Mol. Evol. **21**(5)**:**676–679.
40. **Pym, A. S., P. Brodin, R. Brosch, M. Huerre, and S. T. Cole.** 2002. Loss of RD1 contributed to the attenuation of the live tuberculosis vaccines *Mycobacterium bovis* BCG and *Mycobacterium microti*. Mol. Microbiol. **46:**709–717.
41. **Pym, A. S., et al.** 2003. Recombinant BCG exporting ESAT-6 confers enhanced protection against tuberculosis. Nat. Med. **9:**533–539.
42. **Ravn, P., et al.** 1999. Human T cell responses to the ESAT-6 antigen from *Mycobacterium tuberculosis.* J. Infect. Dis. **179:**637–645.
43. **Renshaw, P. S., et al.** 2005. Structure and function of the complex formed by the tuberculosis virulence factors CFP-10 and ESAT-6. EMBO J. **24:**2491–2498.
44. **Siegrist, M. S., et al.** 2009. Mycobacterial Esx-3 is required for mycobactin-mediated iron acquisition. Proc. Natl. Acad. Sci. U. S. A. **106:**18792–18797.
45. **Skjøt, R. L. V., et al.** 2002. Epitope mapping of the immunodominant antigen TB10.4 and the two homologous proteins TB10.3 and TB12.9, which constitute a subfamily of the ESAT-6 gene family. Infect. Immun. **70:**5446–5453.
46. **Sugiura, N.** 1978. Further analysis of the data by Akaike' s information criterion and the finite corrections. Commun. Stat. Theory Methods **7:**13–26.
47. **Supply, P., et al.** 2000. Variable human minisatellite-like regions in the *Mycobacterium tuberculosis* genome. Mol. Microbiol. **36:**762–771.
48. **Tekaia, F., et al.** 1999. Analysis of the proteome of *Mycobacterium tuberculosis* in silico. Tuber. Lung Dis. **79:**329–342.
49. **van Pinxteren, L. A., P. Ravn, E. M. Agger, J. Pollock, and P. Andersen.** 2000. Diagnosis of tuberculosis based on the two specific antigens ESAT-6 and CFP10. Clin. Diagn. Lab. Immunol. **7:**155–160.
50. **Vita, R., et al.** 2010. The immune epitope database 2.0. Nucleic Acids Res. **38:**D854–D862.
51. **Yang, Z.** 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol. Biol. Evol. **15:**568–573.
52. **Yang, Z.** 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. **24:**1586–1591.
53. **Zheng, H., et al.** 2008. Genetic basis of virulence attenuation revealed by comparative genomic analysis of *Mycobacterium tuberculosis* strain H37Ra versus H37Rv. PLoS One **3:**e237.

Editor: J. L. Flynn

4.2 Global Transcription and its Regulation

4.2.1 High-resolution transcriptome and genome-wide dynamics of RNA polymerase and NusA in *M. tb*

Contribution: Processing, quantification, visualization, and comparison of datasets, design and implementation of a workflow for identification of transcriptional units upon integration of the ChIP-seq and RNA-seq data, qualitative analyses to study global patterns of RNAP and NusA distribution across the TUs.

The ability of *M. tb* to adapt to the numerous environments it encounters in the host is reflected by the complexity of its regulatory machinery. Gene expression in *M. tb* is primarily controlled at the level of transcription initiation by a diverse array of regulatory elements, which interact with RNA polymerase (RNAP). In order to fully understand *M. tb* pathogenesis, it is important to decipher how regulatory proteins control gene expression in different developmental and environmental conditions, and to locate where they bind to the genome. NusA is an essential modulator of gene expression that binds directly to RNAP and has been shown to play a role in elongation, pausing, termination and anti-termination of transcription.

Previous studies in *Escherichia coli* (Mooney *et al.* 2009) and *Bacillus subtilis* (Ishikawa *et al.* 2010) have made use of ChIP-on-chip (chromatin immunoprecipitation followed by hybridization to microarrays) technology to study the genome-wide distribution of RNAP and NusA and elucidate the dynamics of the transcriptional complex. We used a novel high-resolution approach involving a combination of two complementary techniques to investigate the assembly, distribution and activity of the transcriptional

complex throughout the *M. tb* genome.

The genome-wide dynamics of the transcriptional complex in *M. tb* was investigated by carrying out chromatin immunoprecipitation experiments followed by high-throughput sequencing (ChIP-seq) in *M. tb* H37Rv, using antibodies specific for the β-subunit, RpoB of polymerase (RNAP), and for the transcriptional regulator NusA. To correlate the occupancy of the transcriptional complex with its activity, the global transcriptome was determined by using a strand-specific RNA-seq approach. ChIP-seq and RNA-seq data were obtained for exponential (Exp) as well as stationary (Stat) phase cultures. The physiological differences between the two growth phases are a result of major changes in transcriptional regulation. The availability of the ChIP-seq and RNA-seq datasets from two different growth conditions allowed us to compare transcription profiles in the two conditions and correlate the differences in activity of the transcriptional complex with RNA production.

All datasets were mapped to the H37Rv genome sequence and quantitative analysis was based on the genomic features annotated in the TubercuList database (http://tuberculist.epfl.ch) which comprises 4019 protein coding sequences (CDS), 73 genes encoding ribosomal RNAs, small RNAs (sRNAs), tRNAs and other stable RNAs, 3080 intergenic regions (defined as regions flanked by two non-overlapping features). The level of ChIP-seq enrichment for each feature was determined as the ratio of the number of ChIP reads to the number of Input DNA (control sample) reads mapping to the feature. In the case of RNA-seq, there was a striking abundance of reads mapping to the ribosomal RNA operon ($> 95\%$) in both exponential and stationary phases. Therefore, each dataset was normalized to the number of reads remaining

after subtracting the number of ribosomal RNA operon reads from the total number of mapped reads. The gene expression values were quantified in terms of RPM (reads per million) which can be defined as the total number of reads mapping to the feature divided by feature length (in bp) normalized to the number of remaining reads (in millions).

ChIP-seq results showed that RNAP and NusA occur ubiquitously across the genome, with NusA mirroring RNAP distribution in both the exponential and stationary phases of growth. Comparison of the transcriptome profiles in exponential and stationary phases across different functional categories (based on TubercuList) revealed an abundance of stable RNAs and information pathways in both phases of growth. The stationary phase profile was similar to the expression profile observed in the nutrient starvation condition, that has been previously characterized (Betts *et al.* 2002). There was a clear under-representation of genes involved in intermediary metabolism and respiration in the stationary phase. The high resolution and accuracy offered by RNA-seq helped us to uncover anti-sense and intergenic transcripts corresponding to previously un-annotated features. These independent features were confirmed by the presence of RNAP and NusA at their promoters and could also be validated by alternative methods.

A systematic workflow was designed to identify transcription units (TUs) across the genome based on the RNA-seq data and map putative promoters based on enrichment of RNAP and NusA at the start of these units. This led to the identification of 606 high-quality TUs that were widely distributed along the *M. tb* genome. Analysis of RNAP and NusA binding across the promoter and body of TUs and their correlation with transcription uncovered interesting biological aspects of the transcriptional complex in *M. tb.*

Generally, promoter-proximal peaks for RNAP and NusA were observed, followed by a decrease in signal strength reflecting transcriptional polarity. A significant association between expression levels and the presence of NusA throughout the gene body was demonstrated, confirming the peculiar transcription-promoting role of NusA. We also observed that increased NusA occupancy in the body on the unit contributes to higher levels of transcription. The whole-genome distribution of the TUs revealed a strong bias in the expression levels with respect to the direction of the replication forks. The majority of the highly transcribed TUs were localized on the leading strand showing that DNA polymerase and RNAP proceed in the same direction through most of the highly expressed features in the *in vitro* conditions tested.

Using a combination of the two high-throughput approaches, we have obtained an unprecedented depth of information on the transcription profile of *M. tb* genes during exponential and stationary phases of growth, and also offered new insights on the biological roles of RNAP and NusA in promoting transcription.

High-resolution transcriptome and genome-wide dynamics of RNA polymerase and NusA in *Mycobacterium tuberculosis*

Swapna Uplekar[1,3], Jacques Rougemont[2,3], Stewart T. Cole[1*] and Claudia Sala[1*]

[1] *Global Health Institute, Ecole Polytechnique Fédérale de Lausanne, Station 19, CH-1015 Lausanne, Switzerland*

[2] *Bioinformatics & Biostatistics Core Facility, Ecole Polytechnique Fédérale de Lausanne, Station 15, CH-1015, Lausanne, Switzerland*

[3] *Swiss Institute of Bioinformatics, Bâtiment Génopode, Université de Lausanne, CH-1015, Lausanne, Switzerland*

[*] These authors contributed equally to the work

Running title: RNAP and NusA ChIP-seq in *M. tuberculosis*

Keywords: ChIP-seq/*Mycobacterium tuberculosis*/NusA/RNA-seq/ tuberculosis

To whom correspondence should be addressed: Dr. C. Sala or Prof. S.T. Cole, Global Health Institute, EPFL, Station 19, CH-1015 Lausanne, Switzerland (claudia.sala@epfl.ch or stewart.cole@epfl.ch)

Abstract

To construct a regulatory map of the genome of the human pathogen, *Mycobacterium tuberculosis*, we applied two complementary high-resolution approaches: strand-specific RNA-seq, to survey the global transcriptome, and ChIP-seq, to monitor the genome-wide dynamics of RNA polymerase (RNAP) and the antiterminator NusA. Although NusA does not bind directly to DNA, but rather to RNAP and/or to the nascent transcript, we demonstrate that NusA interacts with RNAP ubiquitously throughout the chromosome and that its profile mirrors RNAP distribution in both the exponential and stationary phases of growth. Generally, promoter-proximal peaks for RNAP and NusA were observed, followed by a decrease in signal strength reflecting transcriptional polarity. Differential binding of RNAP and NusA in the two growth conditions correlated with transcriptional activity as reflected by RNA abundance. Indeed, a significant association between expression levels and the presence of NusA throughout the gene body was detected, confirming the peculiar transcription-promoting role of NusA. Integration of the datasets pinpointed transcriptional units, mapped promoters and uncovered new anti-sense and non-coding transcripts. Highly expressed transcriptional units were situated mainly on the leading strand, despite the relatively unbiased distribution of genes throughout the genome, thus helping the replicative and transcriptional complexes to align.

Introduction

In order to adapt and survive in a range of different environments, *Mycobacterium tuberculosis,* the etiologic agent of human tuberculosis (TB), has to fine-tune gene expression to support active growth, to survive periods of non-replicating persistence and to cope with the plethora of stresses encountered within the host (1-3).

In prokaryotes, regulation of gene expression mainly takes place at the transcriptional level, and this is particularly evident in *M. tuberculosis* that has 13 sigma factors, which direct RNA polymerase (RNAP) core enzyme to defined subsets of promoters, and nearly 200 potential transcriptional regulators (4). One particular class of regulators is represented by proteins affecting pausing, termination and anti-termination of transcription, the best characterized of them being NusA.

NusA is an essential RNAP- and RNA-binding modulator of gene expression, present in all bacteria whose genomes have been sequenced so far, and was named N-utilizing substance after the phage l N-protein mediated anti-termination process (5). Its role in preventing termination at ribosomal RNA operons was first demonstrated in the paradigm organism *Escherichia coli*, where the ability of NusA to bind to *nut*-like sites (Box A, B and C) and to protect the nascent naked transcript from Rho-dependent termination has been proved (6). Likewise, more recent studies in *M. tuberculosis* confirmed that this mechanism is conserved among different bacterial species (7). Conversely, NusA has also been implicated in pausing and termination at intrinsic or Rho-independent terminator sequences (8,9). Cardinale et al. later demonstrated that the protein plays a pivotal role in Rho-dependent silencing of foreign DNA in *E. coli* (10). This multi-functionality is reflected in the crystal structures of the *Thermotoga maritima* and *M. tuberculosis*

NusA proteins, which have a multidomain organization. The protein is in fact composed of an N-terminal RNAP-binding domain, connected through a flexible linker to the S1, KH1 and KH2 RNA-binding motifs (11,12).

In a comprehensive study aimed at understanding the dynamics of the transcriptional complex in *E. coli,* Mooney and colleagues demonstrated the genome-wide association of RNAP with NusA as RNA synthesis occurs and the enzyme moves away from the promoter (13). In addition, RNAP promoter-proximal peaks coincide with the distribution of NusA, therefore indicating that RNAP peaks reflect elongating rather than stalled transcriptional complexes. Recent work performed in *Bacillus subtilis* showed that, contrary to *E. coli* RNAP, the enzyme is distributed evenly from the promoter throughout the coding sequence in this species and does not generate promoter-proximal signals (14).

Whole-genome research conducted thus far in *M. tuberculosis* has focused on the final product of transcriptional regulation, the RNA, and has analyzed the differential expression of genes in specific growth conditions by microarrays (15-17). More recently, deep sequencing was applied to transcriptomic profiling and revealed the existence of previously unknown small RNAs whose expression changes during the transition from exponential to stationary phase, and may undergo deregulation during infection (18). However, little is known of the molecular mechanisms taking place at the DNA level and there is no genome-wide description of the RNAP and NusA interaction with the chromosome in different growth phases. In this work, we used two complementary, single-nucleotide resolution approaches, namely ChIP-seq and RNA-seq, to investigate the assembly, distribution and activity of the transcriptional complex throughout the *M. tuberculosis* genome.

Cluster Generation Kit v3 and TruSeq SBS Kit v3 (for NusA ChIP-seq). Data were processed with the Illumina Pipeline Software v1.60 (for RNAP ChIP-seq), v1.70 (for input samples), v1.80 (for NusA ChIP-seq).

Genome annotation

All analyses in this study were carried out using the *M. tuberculosis* H37Rv annotation from the TubercuList database (http://tuberculist.epfl.ch/). There are 4019 protein coding sequences (CDS) currently annotated in the genome, 73 genes encoding for stable RNAs, small RNAs and tRNAs. In order to quantify protein occupancy and transcription across the entire genome, 3080 intergenic regions (regions flanked by two non-overlapping CDS) were included, resulting in a total of 7172 features.

ChIP-seq data analysis

The single-ended sequence reads generated from ChIP-seq experiments were aligned to the *M. tuberculosis* H37Rv genome (NCBI accession NC_000962.2) using Bowtie (48) allowing up to 3 mismatches and up to 10 hits per read. Since the samples were sequenced using different protocols resulting in varied read lengths (38 – 50 nt) all the raw datasets were trimmed to 38 bases to enable unbiased comparison of experiments. Bowtie results were converted into SAM/BAM format using samtools (49). A custom perl script was then used to obtain the per-base coverage normalized to the total number of mapped reads for each dataset. The script also shifted (by 80 bp) and merged the read counts for forward and reverse strands to generate wig files containing single ChIP-seq profiles that were visualized on the UCSC Genome Browser Mycobacterium tuberculosis H37Rv 06/20/1998 Assembly (50). To compute the read count (RC) for each feature, the number of reads mapping to all positions in the feature were summed up and normalized to feature

length. The final read count for each feature was determined as the mean of the read counts of both replicates. To determine the level of ChIP-seq enrichment for each feature, an enrichment ratio (ER) was calculated by dividing the read count for the ChIP-seq sample by the read count for the Input (control) sample.

RNA extraction

Forty milliliters of either exponential or stationary phase *M. tuberculosis* cultures were pelleted and cells were flash frozen in liquid nitrogen and stored at -80°C until use. Bacteria were re-suspended in 1 ml Trizol (Invitrogen) and added to a 2-ml screw-cap tube containing 0.5 ml zirconia beads (BioSpec Products). Cells were disrupted by bead-beating twice for 1 minute with a 2-minute interval on ice. The cell suspension was then transferred to a new tube, where chloroform-isoamylalcohol (24:1) extraction was performed. RNA was precipitated by adding 1/10 volume of sodium acetate (2 M, pH 5.2) and 0.7 volume of isopropanol, washed with 70% ethanol, air-dried and resuspended in DEPC-treated water. DNase treatment was carried out twice using RQ1 RNase-free DNase (Promega), following the manufacturer's recommendations, and the reactions were subsequently cleaned up by phenol-chloroform extraction and ethanol precipitation. RNA was stored at -80°C in DEPC-treated water. Amount and purity of RNA were determined spectrophotometrically, integrity of RNA was assessed on 1% agarose gel.

Library preparation for RNA-seq analysis and Illumina high-throughput sequencing

100 ng of total RNA from exponential and stationary phase were mixed with 5x Fragmentation buffer (Applied Biosystems), incubated for 4 minutes

at 70°C and then transferred immediately on ice. RNA was purified using RNAClean XP beads (Beckman Coulter), according to the manufacturer's recommendations, and subsequently treated with Antarctic phosphatase (New England Biolabs). RNA was then re-phosphorylated at the 5'-end with polynucleotide kinase (New England Biolabs) and purified with Qiagen RNeasy MinElute columns. In order to ensure strand-specificity, v1.5 sRNA adapters (Illumina) were ligated at the 5'- and 3'-ends using RNA ligase. Reverse transcription was carried out using SuperScript III Reverse Transcriptase (Invitrogen) and SRA RT primer (Illumina). Twelve cycles of PCR amplification using Phusion DNA polymerase were then performed and the library was finally purified with AMPure beads (Beckman Coulter) as per the manufacturer's instructions. A small aliquot (2.5 ml) was analyzed on Invitrogen Qubit and Agilent Bioanalyzer prior to sequencing on Illumina Genome Analyzer IIx using the TruSeq SR Cluster Generation Kit v3 and TruSeq SBS Kit v3. Data were processed with the Illumina Pipeline Software v1.82.

RNA-seq data analysis

The single-ended sequence reads generated from RNA-seq experiments were aligned to the *M. tuberculosis* H37Rv genome (NCBI accession NC_000962.2) using Bowtie (48) allowing 1 mismatch and no more than 5 hits per read. Since the samples were sequenced using different protocols resulting in varied read lengths (40 – 50 nt) the raw datasets were trimmed to 38 bases to enable unbiased comparison of the experiments. Bowtie results were converted into SAM/BAM format using samtools (49). Preliminary analysis revealed a striking abundance of reads mapping to the ribosomal RNA operon (over 95%) in all samples. Consequently, the datasets were normalized to the number of remainder reads i.e. subtracting the number of

ribosomal RNA operon reads from total number of mapped reads for each dataset. The gene expression values were quantified in terms of RPM (reads per million) which can be defined as the total number of reads mapping to the feature divided by feature length (in bp) normalized to the number of remaining reads (in millions). The normalization factor was also used to generate wig files that were visualized on the UCSC Genome Browser (50). The mean RPM value was determined for the experimental replicates. The MA-plot was generated using a python script where the average expression in the exponential and stationary phase samples was plotted against the $\log_2$ fold-change between the two conditions. Analysis of differential expression was carried out using the DESeq package (51).

Transcriptional unit (TU) quantification

Enrichment of RNAP and NusA in TUs was quantified separately for the promoter and the body of the TU. The first feature was considered as the promoter, while the remaining features represented the body. The TU body read count and enrichment ratio were computed as the mean of the RCs and ERs for all features in the TU. The level of transcription of the TU was calculated as the average RPM of all features (including promoter) in the TU.

Reverse transcription

Two micrograms of *M. tuberculosis* RNA were incubated with 50 ng random primers and 1 mM dNTPs at 65°C for five minutes. After cooling on ice, 40 U RNase inhibitor, 10 mM DTT, 1x reaction buffer, 5 mM $MgCl_2$ and 200 U SuperScript III reverse transcriptase (Invitrogen) were added in a final volume of 20 ml. A control reaction without reverse transcriptase was included as control. Reactions were incubated at room temperature for ten minutes, at 50°C for one hour and at 55°C for one hour. Reverse transcriptase

was inactivated by incubation at 80°C for two minutes. RNase H treatment was carried out for twenty minutes at 37°C with 1 ml RNase H (Invitrogen). cDNA was stored at -20°C.

Quantitative PCR (qPCR) for ChIP-seq and RNA-seq data validation

All PCR primers were designed using Primer3 software (http://frodo.wi.mit.edu/primer3/). The 20 ml PCR reaction consisted of 1X Sybr Green PCR Master Mix (Applied Biosystems), 0.1 mM of each primer and 1 ml of cDNA or IP DNA from immunoprecipitation reactions. Reactions were carried out in duplicate in an Applied Biosystems 7900HT Sequence Detection System with the following protocol: denaturation at 95°C for 10 min, 40 cycles of denaturation at 95°C for 15 sec, annealing and extension at 60°C for 40 sec with data collection. Melting curves were constructed to ensure that only one amplification product was obtained.

Parallel reactions using different amounts of H37Rv chromosomal DNA were performed for each primer set in order to obtain the standard curve correlating the threshold cycle with the number of template molecules. The resulting equation was used to quantify the number of target molecules in the unknown samples.

In the case of qPCR for RNA-seq data confirmation, normalization was obtained to the total amount of RNA used in the reverse transcription reaction, as previously described (23). Results were expressed as the log2 ratio of the number of molecules determined in the exponential phase vs. the number of molecules in the stationary phase and correlated with the log2 ratio obtained with RPM values.

Regarding the qPCR for ChIP-seq data validation, the number of target molecules was normalized to the Input sample, after subtraction of the background represented by the mock-IP (no antibody control). Results were

expressed as the log2 enrichment ratio of the IP DNA vs. Input and correlated with the log2 enrichment ratio calculated as described earlier.

Statistical analysis

Statistical analyses were performed with the statistical language R and various Bioconductor packages (52) (http://www.bioconductor.org). Several plots were created using GraphPad Prism 5 software (www.graphpad.com). Custom Perl and Python scripts were developed by members of the EPFL Bioinformatics and Biostatistics core Facility. Correlations between replicates were based on Pearson's product moment correlation coefficient. The Wilcoxon signed-rank test (Mann-Whitney U test) was used to assess the differences between exponential and stationary phase values for the same set of features.

Data access

The ChIP-seq and RNA-seq datasets have been deposited in NCBI's Gene Expression Omnibus (53) under accession number GSE40862.

Results

General strategy

The genome-wide dynamics of the transcriptional complex in *M. tuberculosis* was investigated by carrying out chromatin immunoprecipitation experiments followed by deep sequencing (ChIP-seq) using antibodies specific for a core component of RNA polymerase (RNAP), the beta subunit RpoB, and for the transcriptional regulator NusA. To correlate the occupancy of the transcriptional complex with its activity, the global transcriptome was determined by using a strand-specific RNA-seq approach. ChIP-seq and RNA-seq data were obtained from exponential (Exp) and stationary (Stat) phase cultures. All datasets were mapped to the H37Rv genome sequence and further analyses were based on the annotation from the TubercuList database (http://tuberculist.epfl.ch) which comprises 4019 protein coding sequences (CDS), 73 genes encoding ribosomal RNAs, small RNAs (sRNAs), tRNAs and other stable RNAs. There are 3080 intergenic regions (IG), defined as regions flanked by two non-overlapping features, of which 2283 are ≥30 bases in length.

The genome-wide distribution of RNAP and NusA in *M. tuberculosis*

ChIP-seq experiments for RNAP and NusA, and sequencing of the Input DNA, were performed in duplicate and the results are summarized in Supplementary Table S1. Supplementary Table S2 shows the correlation coefficients for the biological replicates, confirming the extremely high reproducibility of the Input and NusA datasets and, to a lesser extent, of the RNAP results.

Inspection of RNAP data, displayed on the UCSC genome browser, revealed that signals were present along the entire genome although prominent accumulation of the enzyme at the putative promoter regions was detected in both of the conditions tested (promoter-proximal peaks). This is illustrated in Figure 1, where a global view of a 1 MB portion of the genome is shown (Figure 1A) together with three representative transcriptional units (Figure 1B-D). The NusA binding profile appeared to qualitatively match the RNAP distribution. Indeed, upon closer inspection of the RNAP and NusA tracks, extensive overlap was observed in both growth phases, as evidenced from the correlation reported in Supplementary Figure S1, thus proving that, although NusA does not bind to DNA directly, it is part of the transcriptional complex and associates with RNAP.

Head-to-head comparison of the Exp and Stat datasets was possible by calculation of the enrichment ratios relative to the Input sample (ER) for RNAP and NusA for each genomic feature. Results are reported in Supplementary Table S4. Approximately the same number of features was found to be enriched (ER ≥ 2) in Exp and Stat (1117 and 1062, respectively) and, in both phases, most of the signals were identified in IG regions ($> 60\%$ vs. $< 40\%$ inside CDSs and RNA-encoding genes), where promoter sequences are most likely present. The strongest peaks were detected at the promoters of the ribosomal RNA operon *rrn* (Figure 1B), the ribosomal protein operons, the ATP synthase (Figure 1C), PDIM biosynthesis and *nadABC* gene clusters in Exp phase, and preceding genes encoding sRNA and tRNAs. In both growth conditions, most of the regions showed enrichment for either RNAP and NusA together or RNAP only, whilst NusA alone was detected for a small fraction of signals (193 and 107 features in Exp and Stat, respectively). Independent support for the ChIP-seq findings was obtained by quantitative

PCR (qPCR) for a subset of CDSs and IGs, selected in order to cover a broad range of ER values (Supplementary Figure S2).

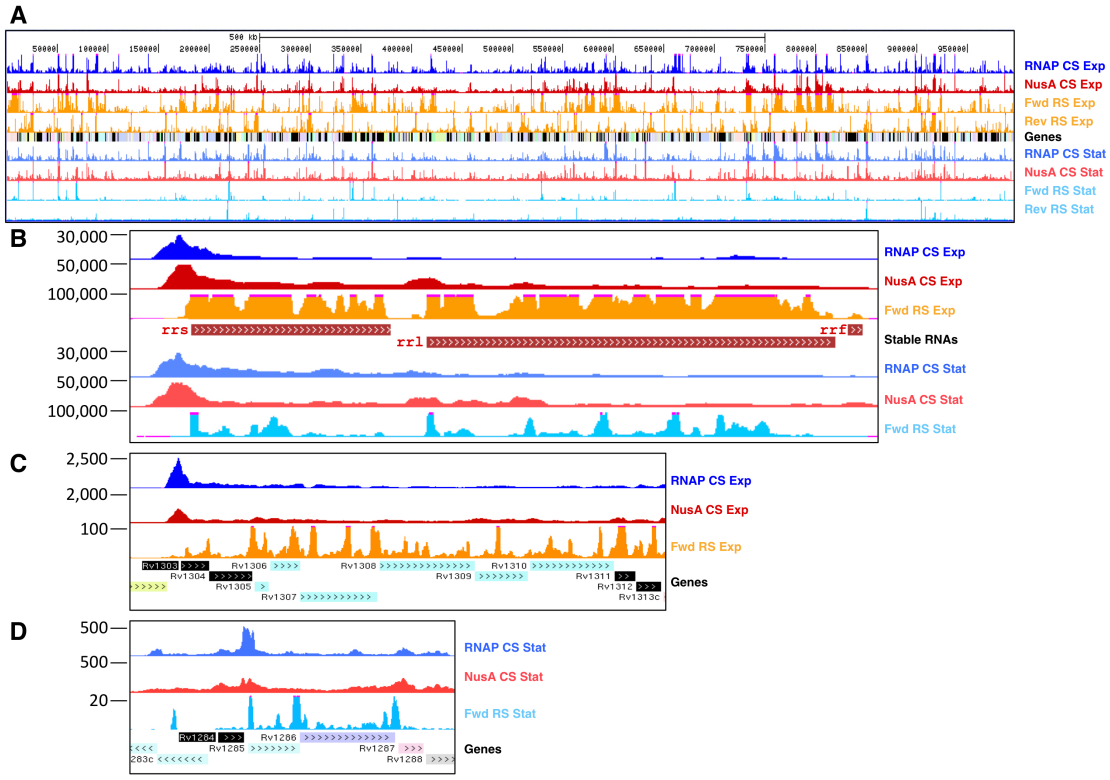


Figure 1. ChIP-seq profiles of RNAP and NusA and transcriptional levels measured by RNA-seq. (A) UCSC genome browser view of ChIP-seq (CS) profiles for RNAP and NusA and RNA-seq (RS) profiles on forward (Fwd) and reverse (Rev) strands across a 1 MB region of the *M. tuberculosis* H37Rv genome in Exponential (Exp) and Stationary (Stat) phases. The scale and the genome coordinates are reported at the top. (B) Profile of the ribosomal RNA operon in Exp and Stat phases. (C) Profile of the ATP synthase operon in Exp phase. (D) Profile of the *rv1285* to *rv1286* region in Stat phase. The scale for the number of reads is on the y-axis.

Detailed examination of the whole-genome profiles revealed unexpectedly high densities for RNAP and NusA at the end of convergent genes or inside CDSs, suggesting the existence of new features. For instance, these are the cases of the peak between *rv3661* and *rv3662c* or the peak inside *ino1* where, indeed, new sRNAs have been recently identified (18). In addition, the genomic regions encoding the putative excisionases of prophages phiRv1 and phiRv2 exhibited exceptionally high ER values for both RNAP

and NusA, mainly in the Exp growth phase (21 and 7 for RNAP and NusA respectively for *rv1584c*, 17 and 6.7 for RNAP and NusA respectively for *rv2657c*). Deeper understanding of these peaks and of the entire ChIP-seq dataset was possible upon transcriptomic profiling as described later.

Deep RNA-seq analysis

To correlate the presence of RNAP and NusA, inferred from ChIP-seq, with transcriptional activity, global transcriptomic analysis was performed by deep sequencing, RNA-seq. More than 98% of the sequence tags could be mapped to the annotated CDSs in the sense orientation, whilst less than 2% was attributed to the anti-sense orientation, thereby confirming the strand-specific nature of the protocol. The correlation coefficient for biological replicates revealed high reproducibility ($r^2 > 0.95$) and further mapping statistics are reported in Supplementary Table S3. As expected, most of the reads aligned to *rrn* in both Exp and Stat phases and there was a clear predominance of stable RNAs such as sRNAs, tRNAs, the RNA component of RNAse P and the tmRNA *ssr* under both growth conditions. The remaining reads could be assigned to CDSs on both strands in almost equal proportion and these provided sufficient coverage for quantification purposes. RPM (Reads Per Million reads aligned) values represent the unit chosen to allow comparison of the expression levels and these are presented in Supplementary Table S5 for all of the features in the sense and anti-sense orientation, in the two growth conditions.

Codon usage and tRNAs

Measuring tRNA expression levels was informative in terms of understanding genome biology and generating confidence in the quantitation procedure. We counted the occurrence of each codon in every CDS then

multiplied this count by the value of gene expression obtained by RNA-seq (RPM) before summing these expression values across the whole genome to yield a per-codon "expression level". This codon usage was then compared with the expression of the corresponding tRNA. Figure 2A shows a clear correlation between these two metrics in the Exp condition (with highly similar results in the Stat phase). Notably, AT-rich codons generally tend to be under-used as opposed to their GC-rich alternatives with the cognate tRNAs being considerably more abundant.

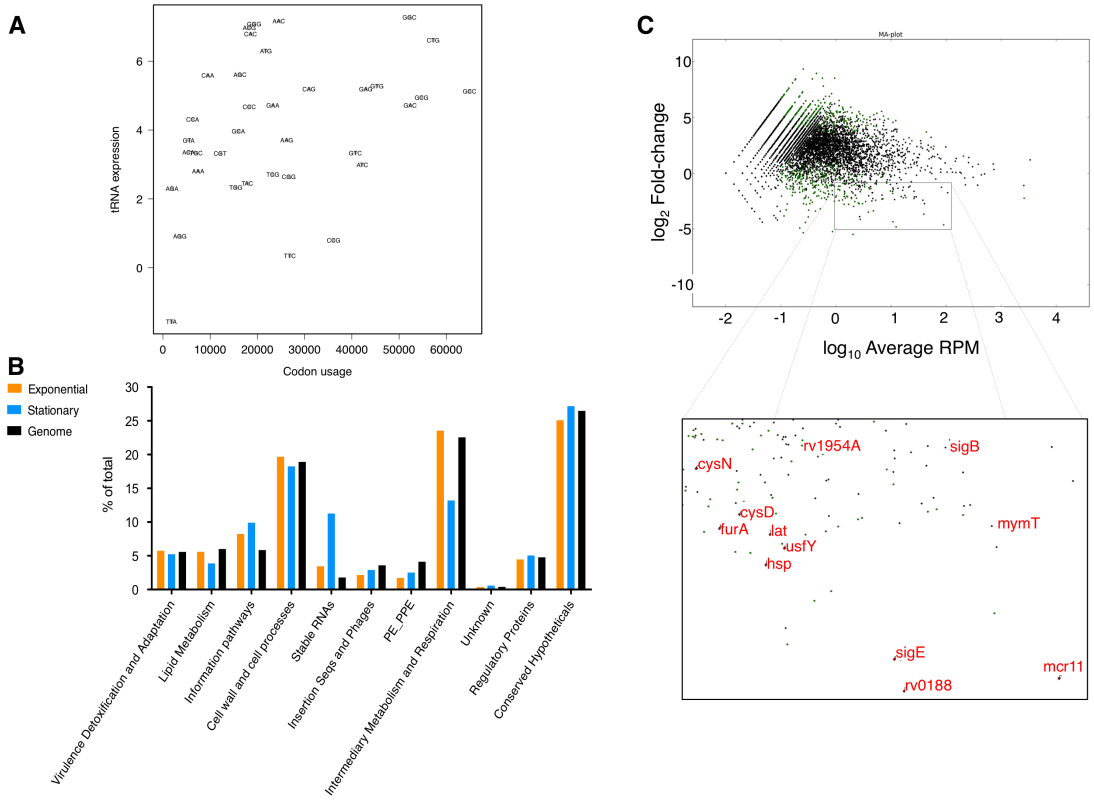


Figure 2. RNA-seq results. (A) Correlation between tRNA abundance and codon usage. (B) Distribution of genes transcribed in exponential and stationary phases across functional categories as defined in the TubercuList database (http://tuberculist.epfl.ch/) relative to the proportion observed in the *M. tuberculosis* H37Rv genome. (C) Magnitude-Amplitude-plot in which the mean ($\log_{10}$) expression value for each gene in exponential and stationary phase samples is plotted against the expression ratio ($\log_2$) between the two samples. Differentially expressed genes are colored in green. The inset box indicates some of the genes

induced in stationary phase. *rv1954A* indicates the anti-sense transcribed feature *rv1954*.

Comparison of the Exp vs. Stat phase transcriptomes

Grouping CDSs into functional categories according to TubercuList facilitated comparison of the total transcriptome in Exp and Stat phases. The box-and-whisker plots in Supplementary Figure S3 show that all categories are represented in both conditions, although the RPM dynamic range is higher in Exp phase, in particular for the lipid metabolism, information pathways and PE-PPE groups. On the contrary, features belonging to the virulence, detoxification and unknown subgroups are more enriched in Stat phase. Figure 2B illustrates the relative abundance of each category in terms of the proportion of genes with expression level (RPM) ≥1, the value chosen as cutoff. There is enrichment of genes involved in information pathways ($p<10^{-4}$ in Exp and Stat, Fisher's exact test) and an abundance of stable RNAs ($p<10^{-4}$ in Exp and $p<10^{-21}$ Stat, Fisher's exact test) in both Exp and Stat phases. There is also a clear underrepresentation of genes belonging to intermediary metabolism and respiration ($p<10^{-7}$) in the Stat phase.

Closer inspection of the differentially expressed genes revealed a Stat profile similar to the nutrient starvation condition previously described by Betts and colleagues (15), with down-regulation of the ATP synthase gene cluster, the *nad* genes, the housekeeping sigma factor *sigA*, the ribosomal protein operons, cell wall and cell membrane functions. By contrast, specific features induced in the Betts model were also found to be up-regulated in Stat phase, such as *lat*, *rv0188*, *usfY*, *hsp*, and the *cys* operon, while a steep increase in the expression level of the sigma factors *sigB* and *sigE* was observed (Figure 2C). Genes *rv1954c* and *rv2660c* had been reported as induced in starvation conditions (15) and our data confirmed this but in the

anti-sense orientation. In addition, the high resolution provided by deep sequencing allowed the quantification of small transcripts, namely *mcr11,* reported as inducible in Stat phase (18), the recently discovered gene *mymT* (19) and the sRNA encoded in the intergenic region between *rv3661* and *rv3662c* whose expression increases significantly in Stat phase (18).

To validate the quantification method, twelve features covering a broad range of RPM values were selected for quantitative reverse transcription PCR (qRT-PCR) analysis. Results are reported in Supplementary Figure S4 and confirm the good correlation ($r^2 > 0.91$) between the Exp/Stat ratio calculated from RPM values and the corresponding figures obtained from qRT-PCR, thus justifying the use of RPM values for absolute comparisons between the different growth phases.

Correlation between the RNAP and NusA profiles and transcription

The availability of the ChIP-seq and RNA-seq datasets from two different growth conditions allowed correlation studies to understand the relationship between the presence of the transcriptional complex and RNA production. For this purpose, features with RNAP or NusA ER ≥ 2 and with RPM ≥ 1 were selected. From the histogram reported in Figure 3 it is evident that most of the features enriched in RNAP and NusA in Exp phase were also associated with transcription (89%). A smaller percentage of features with ER ≥ 2 for RNAP but no detectable NusA were transcribed (71%), suggesting that, indeed, the presence of NusA in the transcriptional complex favors transcription. The values changed markedly in the Stat phase, where a reduced number of features enriched in both RNAP and NusA was found to be transcribed (76%). Even more relevant, less than half of the RNAP-only containing signals were associated with RPM values greater than 1. Overall,

the vast majority of the RNAP- and/or NusA-enriched features in Exp phase were also expressed, whereas approximately half of those enriched in Stat were associated with detectable transcripts, indicating that RNAP and NusA interact with the genome but are not involved in active transcriptional activity during the Stat condition. Regarding the NusA-only containing peaks (often intragenic), most of them correlated with a transcript at least in Exp phase, thereby leaving a small number (42) with unusual NusA binding.

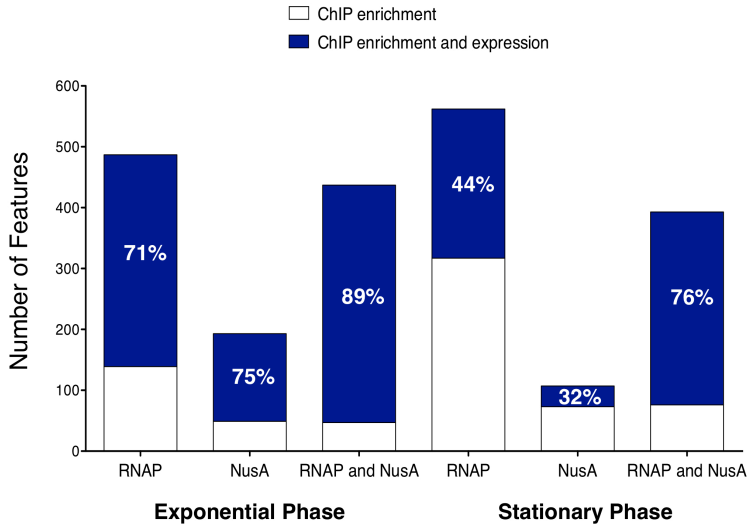


Figure 3. Overlap between enriched (ER ≥ 1) and transcribed (RPM ≥ 1) features in exponential and stationary phases. Stacked bar charts show the total number of features enriched with RNAP, NusA and both (RNAP as well as NusA) in the two phases. Blue stacks represent the proportion of enriched features that are also transcribed.

Most of the surprising signals previously reported for ChIP-seq could be explained upon direct comparison with the RNA-seq results. In the next section we will describe two subsets of these peculiar features: those corresponding to anti-sense and to intergenic transcription. Importantly, this information allowed refinement of the existing *M. tuberculosis* genome annotation (4).

Unusual ChIP-seq peaks reflect anti-sense transcription in *M. tuberculosis*

As mentioned earlier, less than 2% of the RNA-seq reads mapped to the anti-sense orientation as compared to the existing annotation. Quantification uncovered 134 anti-sense transcribed features in Exp phase and 41 in Stat phase (Supplementary Table S6), the majority of which could be explained as RNAs that are much longer than the corresponding annotated CDS, thereby generating anti-sense transcripts to the following, reverse-oriented, feature. In this section, we will not provide details of these but rather focus on the anti-sense RNAs that are transcribed independently. Upon clustering them into functional categories, two groups appeared prominent, the "Insertion sequences and phages" and the "Unknown" categories, where a >2- or 3-fold increase was noticed compared to their abundance in the genome, respectively.

The top-scoring feature in Exp phase is the anti-sense transcript inside the gene *ino1*. Interestingly, this region contains one of the features that was unexpectedly highly enriched in RNAP and NusA ChIP-seq studies (Supplementary Figure S5A).

A closer look into the fraction of RNAP- and NusA-enriched features with no corresponding sense transcript revealed that some of these were associated with an anti-sense RNA. This is exemplified by *rv0842*, where RNAP and NusA densities at the beginning of the gene do not correlate with expression of the CDS but rather with transcription of an anti-sense RNA that extends in the upstream region (Supplementary Figure S5B). Furthermore, an anti-sense transcript was mapped inside the phiRv1 excisionase-encoding gene *rv1584c*. Its expression level increased in Exp and was reduced to the background level in Stat phase. On the contrary, no obvious explanation could be obtained for the RNAP peak inside the phiRv2

gene *rv2657c*. RNAP and NusA may just be "sitting" there without transcribing (at least in the conditions tested) or the transcript could be undetectable because of stability issues. Finally, curious RNAP and NusA peaks were detected at the end of *rv0061*. After looking at the transcriptomic profile, these signals could be associated with anti-sense transcription of the whole CDS (Supplementary Figure S5C), thereby allowing re-annotation of this open reading frame as *rv0061c,* encoding a protein similar to *M. marinum* MMAR_3839, with 76% identity in 112 amino acid overlap.

Concerning the Stat condition, the top-scoring anti-sense RNA was noted inside *rv3684*. Accordingly, RNAP and NusA peaks are high in this position, thus cross-validating the finding (Supplementary Figure S6A). The genomic region encompassing the end of *rv1734c* was found to be highly transcribed in the anti-sense orientation in Stat phase (Supplementary Figure S6B). Correspondingly, the NusA density profile increases in this condition.

Intergenic ChIP-seq densities as hallmarks of intergenic transcripts

Within the 2283 IG regions of at least 30 bases in length, 1153 showed an RPM ≥ 1. We could identify most of these as part of transcriptional units as described later, IGs constituting the 5'- or 3'-UTR of flanking genes (12%) and "anti-sense IGs" (2%), i.e. IG regions transcribed in the opposite sense to both of the flanking CDSs (Supplementary Table S6).

Examples of long 5'-UTRs are represented by *rv3219* (Supplementary Figure S7A) and by the *rv3648c* mRNA (Supplementary Figure S7B). The latter feature, expressed at high levels in the Exp growth phase (20), showed high enrichment ratios for RNAP and NusA and suggested the existence of attenuation mechanisms in controlling gene expression.

Other examples include IG regions transcribed as independent features. This is the case of the IG at the end of the convergent genes *rv3661* and

rv3662c: an sRNA (18), highly induced in Stat phase, is encoded here and this is corroborated by enrichment of both RNAP and NusA in the ChIP-seq data.

The last type of IG directs anti-sense transcripts with respect to both of the flanking genes and therefore representing independent features such as sRNAs. The top-scoring instance is the IG between *rv1144* and *rv1145*, where an sRNA was identified and high ER values for both RNAP and NusA were calculated (Supplementary Figure S7C).

Integration of ChIP-seq and RNA-seq datasets

In order to provide more quantitative correlation patterns between the dynamics of RNAP and NusA and transcription levels, a subset of well-known highly expressed transcriptional units (TUs) was selected. Specifically, 24 ribosomal protein operons, comprising a total of 75 CDSs and 57 IGs, represented the test set in the systems biology approach described hereafter. The level of transcription and enrichment of RNAP and NusA were quantified for each operon, with the "operon promoter" being defined as the first feature of the TU (usually an IG), and the "body" representing the remaining features. RNA-seq data showed that the majority of these TUs were expressed in both Exp (21 out of 24) and Stat (18 out of 24) growth conditions as long, continuous transcripts but the RPM values were significantly higher in the Exp compared to the Stat phase ($p<10^{-6}$, Wilcoxon signed-rank test, Figure 4A). Promoter regions of all transcribed TUs showed a median ER greater than two for RNAP as well as NusA (the latter in Exp phase only). These values were considerably different in the body of the TUs: the median ER of RNAP and NusA was reduced, confirming the presence of the promoter-proximal peaks mentioned earlier rather than homogeneous distribution of the transcriptional complex throughout the operon. Interestingly, a significant higher level of NusA in Exp phase compared to Stat was detected in the body

($p<10^{-4}$), whereas the RNAP ER in the operon body did not vary between the two conditions.

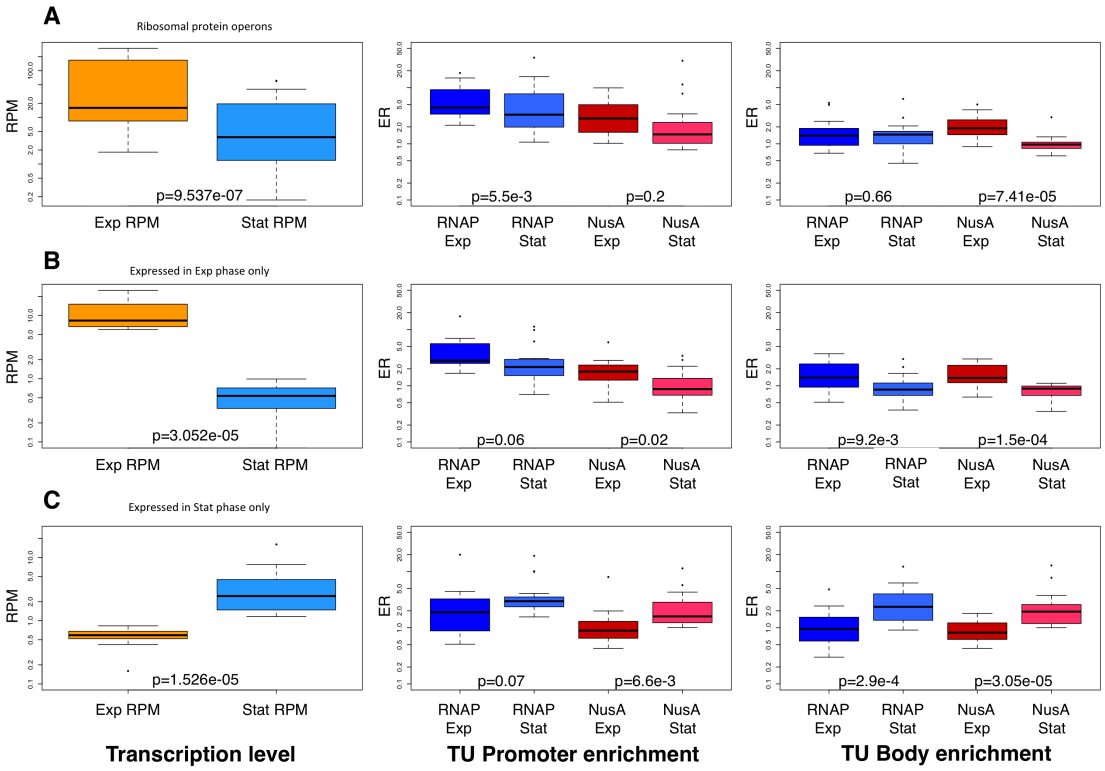


Figure 4. Box plots showing the distribution of RNAP and NusA in the promoter and body of transcriptional units (TUs) along with the transcription levels (RPM). (A) Ribosomal protein operons. (B) TUs expressed in exponential phase only and (C) TUs expressed in stationary phase only. P-values for comparison of different groups are based on the Wilcoxon signed-rank test.

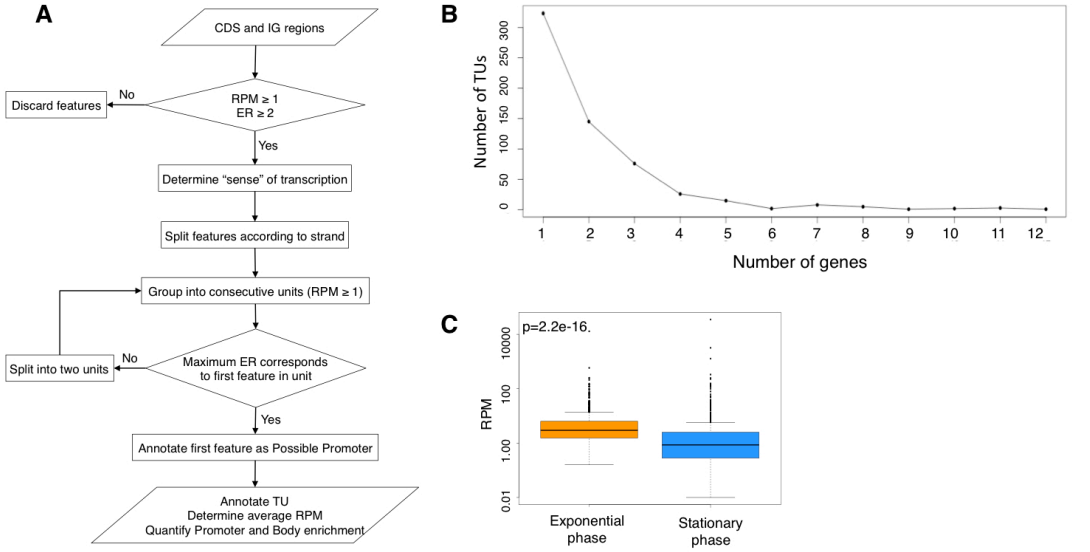
Altogether, these results suggest that increased NusA occupancy in the operon body contributes to higher levels of transcription in the Exp compared to the Stat growth phase. In addition, the criteria that define a TU (i.e. promoter at the first feature and uninterrupted RNA) and the patterns identified with this subgroup of operons were applied to and examined in a larger dataset, as described below.

Genome-wide TU identification

An algorithm (Figure 5A) was designed to discover TUs using the genome-wide ChIP-seq and RNA-seq profiles. Genomic features, including CDS and IGs, that displayed NusA or RNAP ER ≥ 2 and RPM ≥ 1 in at least one of the two growth phases were used as input. Features with RPM ≥ 1 were split into two groups based on the strand that was associated with higher RPM values. The forward and reverse sets of features were then sorted according to the genomic order and grouped into TUs. A TU was defined as a set of two or more consecutive features that are transcribed in the same direction and controlled by a promoter located at the first feature. Following the identification of transcription boundaries for all operons by means of RNA-seq data, promoters were mapped based on the RNAP and NusA ER. If the highest ER value was observed at the first feature, it was considered as a possible promoter. If the feature showing maximum ER did not correspond to the first one, the TU was split into two separate units at that feature, and promoter identification was repeated. This ensured that all TUs displayed a uniform organization characterized by enrichment of RNAP/NusA at the promoter and an uninterrupted transcript from the start to the end of the unit. Of the 817 continuous transcripts identified from RNA-seq, 606 were associated with significant promoter enrichment (ER ≥2) and therefore defined as "high-quality TUs". Of these, 301 were mapped on the forward strand and 305 on the reverse strand (Supplementary Table S7). Basic features of these TUs are reported in Figure 5. 323 of these were single-gene units, while the remainder comprised between 2 and 12 genes. The largest TUs were represented by the ATP synthase operon, the ESX-3 locus (*eccA3* to *eccE3*), and two ribosomal protein operons (from *rpsJ* to *rpsQ* and from *rplN* to *rplO*). 268 operons were transcribed in both Exp and Stat phases, while 272 were transcribed only in the Exp phase and 17 only in the Stat phase, for instance the latter included *lipX-PPE17*, *rv2557-rv2558*, *usfY*, *lat*,

cysD, rv2660c. Importantly, all of the previously described ribosomal protein operons were correctly identified with the developed workflow, underscoring the validity of the algorithm.

To appraise differences in RNAP and NusA distribution between Exp and Stat phase, the same parameters used for the characterization of the ribosomal protein operons were employed. Specifically, the RNAP and NusA ER values in the promoter and body of two subsets, namely the top 17 TUs transcribed in Exp phase only (Figure 4B) and the 17 TUs that were transcribed in Stat phase only (Figure 4C), were compared. The RPM and ER figures for those transcribed in Exp phase mirrored the observations made in the case of ribosomal protein operons (Figure 4A and 4B). Once again, the level of NusA in the body was significantly higher in Exp vs. Stat phase ($p<10^{-4}$, Wilcoxon signed-rank test). However, in contrast to what was observed for ribosomal protein operons, the RNAP ER in the TU body was also significantly increased in Exp phase ($p< 10^{-2}$), consistent with the lack of expression in Stat phase. Regarding the TUs transcribed in Stat phase only, complementary patterns to those in Figure 4B were observed (Figure 4C). Indeed, the levels of RNAP and NusA in the operon body were higher in the Stat phase compared to the Exp phase ($p< 10^{-3}$ and 10^{-4}, respectively) reflecting the difference in transcription between the two conditions. In conclusion, it is noteworthy that Exp phase-only and Stat phase-only transcribed TUs displayed RNAP ER at the promoters in both growth conditions, irrespective of the transcriptional levels, supporting the notion of a stationary enzyme in the non-transcribed phase.



D

Growth Phase	RNAP TR	NusA TR	RNAP:NusA promoter	RNAP:NusA body
Exponential	0.51	0.75	1.78	1.11
Stationary	0.50	0.76	1.73	1.16

Figure 5. Integration of ChIP-seq and RNA-seq data to map transcriptional units (TUs). (A) Workflow for TU identification. (B) Distribution of TU composition. (C) TU expression levels in Exp and Stat conditions. (D) Median traveling ratios (TR) for RNAP and NusA in Exp and Stat phases and RNAP:NusA ratio at the TU promoters and bodies.

Transition from transcription initiation to elongation is rate limiting

The rate of transition from transcription initiation to elongation was measured as the ratio of the RNAP and NusA ER in the body relative to the corresponding promoter ER. This value defines the traveling ratio (TR) and the medians for all of the TUs are reported in Figure 5D. The TR for RNAP was around 0.5, reflecting the presence of promoter-proximal peaks and implying that the transition from initiation to elongation is a rate-limiting step at most transcribed regions. The TR for NusA was slightly higher (0.75), conveying the idea that NusA-containing transcriptional complexes move more efficiently from the promoter throughout the TU. Further support for this concept came from the calculated RNAP:NusA ratio at promoters and in

gene bodies. Indeed, this was in favor of RNAP at promoter regions and close to 1 in bodies, suggesting that a proportion of RNAP molecules, probably not associated with NusA, contacts the promoter but either does not progress or leaves abortively.

Global RNAP and NusA profiles throughout the TUs

Global profiles for RNAP and NusA were obtained by averaging the read counts for all of the identified TUs. Two highly expressed and enriched loci, namely *rrn* and the sRNA *mcr11,* were excluded from the analysis, so as to avoid a biased output. The TUs were split into two groups based on their genomic orientation and each one was divided into 100 equal sized bins. The average number of reads for RNAP and NusA was calculated within each bin and the mean for all operons was used to generate a single operon profile. The average enrichment in 50 bp regions flanking each TU was also included. Figure 6A presents the averaged forward and reverse profiles for RNAP and NusA in both phases of growth. Prominent enrichment was evident close to the start of the TUs on both strands, as pointed out earlier, and the average level of RNAP was higher than that of NusA in agreement with the previously calculated ratios. The decrease in the level of RNAP and NusA along the operon body reflects the polarity of transcription. Since half of the identified operons are transcribed in both Exp and Stat phases, the profiles for RNAP and NusA in the two phases almost coincide.

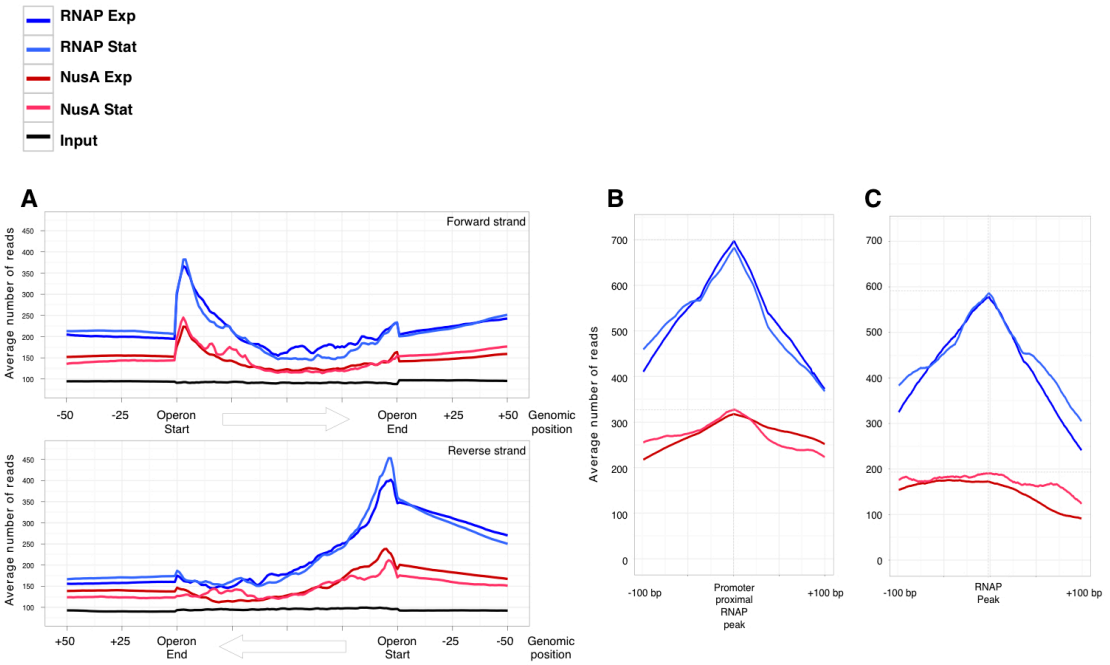


Figure 6. Global RNAP and NusA profiles of the transcriptional units. (A) Averaged profiles for RNAP and NusA across forward and reverse TUs. The y-axis indicates the average number of reads, the x-axis the genomic position with an arbitrary length for the TU. (B) Averaged RNAP and NusA profiles at promoter regions for RNAP-enriched transcribed features. (C) Averaged RNAP and NusA profiles at promoter regions for RNAP-enriched non-transcribed features. The legend is provided on top.

Finally, detailed promoter profiles were generated for transcribed TUs as follows: the average number of reads per nucleotide was calculated for the RNAP and NusA datasets in an interval spanning 200 bp centered at the maximum of the RNAP peak for each promoter (i.e. the first feature of the TU). The forward and reverse strand results were merged to generate a single Exp and Stat phase profile for RNAP and NusA (Figure 6B). The averaged plots not only confirmed enrichment of NusA at promoters, but also its co-localization with RNAP in both Exp and Stat phase. As a control, a subset of features enriched with RNAP (in both Exp and Stat phases) but not associated with transcription was chosen and the analysis repeated. Unlike the profile observed for transcribed TUs, the NusA enrichment around the RNAP

peak was much lower for non-transcribed features and did not display the same shape nor co-localize with the RNAP peak (Figure 6C), thereby confirming NusA as a transcription promoting factor in *M. tuberculosis* in both of the experimental conditions tested.

Discussion

Our aim was to map the genome of the major human pathogen, *M. tuberculosis*, by systematically studying the activity of the transcriptional complex and its interaction with DNA. Using high-resolution ChIP-seq analyses we demonstrated the ubiquitous association of RNAP and NusA with the genome and the significant overlap of the two binding profiles, thus confirming that NusA is part of the RNAP complex. ChIP-seq was an extremely powerful tool to monitor NusA, a protein which does not interact directly with DNA but rather with RNAP and/or with the nascent transcript.

Quantitative reproducibility between biological replicates was extremely high in the case of NusA in both of the conditions tested, whereas a reduced correlation was noted for RNAP. This could be the consequence of a lower affinity of the monoclonal antibody against RpoB as compared to the polyclonal NusA anti-serum or could result from intrinsic higher variability of the RNAP dynamics on the *M. tuberculosis* genome caused by non-specific or random interactions of the enzyme with DNA, as previously reported in *E. coli* (21,22). Interactions between the transcriptional complex and DNA might be stronger when NusA is present, thus explaining the increased reproducibility of the NusA profiles. Indeed, the density of RNAP only-containing peaks was found to be less reproducible than that of peaks associated with NusA, supporting our hypothesis.

In addition to ChIP-seq, RNA-seq was exploited to obtain a global view of transcriptional regulation in *M. tuberculosis* and to explain unusual

peaks observed in the ChIP-seq profiles. Quantification of RNA-seq results by RPM values with normalization to the total number of reads allowed absolute comparison of the conditions tested, in a way that resembles microarray-based transcriptomic approaches, where total fluorescence was used as normalizing factor (23). The Stat profile was similar to that of a starved culture (15) and hallmarks of this condition were readily identified. Discrepancies with previously published studies can be explained by the different culture conditions utilized. In our model, a four-week old *M. tuberculosis* culture may mimic the starvation condition since all of the nutrients have been consumed.

Importantly, the deep sequencing transcriptomic experiments reported here exhibited striking strand specificity, with a minor portion of the sequence tags ($< 2\%$) mapping to the CDSs in anti-sense orientation. We therefore considered this anti-sense transcription as genuine and exploited the data to improve the *M. tuberculosis* genome annotation (4). As an illustration, the CDS *rv0061* was renamed *rv0061c* following the results of this work. Several anti-sense transcripts mapped by RNA-seq found additional confirmation by ChIP-seq profiling. BLAST analysis performed on the top-scoring anti-sense-transcribed features revealed that some of them are specific to the MTB complex (e.g. *rv1374c*, *rv1734c*), whereas others, like *ino1*, showed more than 85% conservation at the nucleotide level in other mycobacterial species such as *M. marinum, M. smegmatis* and *M. leprae,* suggesting that the anti-sense RNA encoded there might be conserved. Curiously, two features (*rv2660c* and *rv1954c*) that were described as induced upon starvation (15) were also found to be up-regulated in Stat phase but in the anti-sense orientation. The strand-specificity of RNA-seq uncovered this discrepancy, whilst previously used PCR-based microarrays did not allow correct strand identification (15). Ironically, Rv2660c has been proposed as part of a multi-subunit TB vaccine (24) even though its gene is transcribed in the opposite direction.

Intergenic transcription and corresponding RNAP/NusA ER added further value. A number of long 5'-UTRs has been identified, such as those of *whiB1* (*rv3219*) and of *cspA* (*rv3648c*), suggesting the existence of *cis*-regulatory elements affecting transcription elongation or attenuation mechanisms, including those mediated by riboswitches. To date, some of these have been predicted or validated in *M. tuberculosis* (25-27). The presence of NusA at these genomic positions confirms the role of this protein in the transcription attenuation process, in agreement with earlier studies (28,29). The single-nucleotide resolution conferred by RNA-seq allowed the identification of intergenic sRNAs and independent confirmation by ChIP-seq was obtained. Most of the previously described small transcripts (18,30-32) were easily recognized in our study and some of the discrepancies encountered could be related to different culture conditions or sequencing methodologies. Prediction of the conservation and of the functional role carried out by these transcripts remains elusive, given the reduced sequence identity occurring in IGs among the various mycobacterial species.

A major strength of our work lies in the combination of ChIP-seq with RNA-seq technologies, and therefore the cross-validation of the respective datasets. Approximately the same number of features was found to be enriched in RNAP and/or NusA in the two conditions tested but there was good correlation with transcription in Exp phase only. On the contrary, roughly half of the RNAP/NusA signals were associated with RNA in Stat phase. Overall, while on the one hand RNA stability effects cannot be ruled out, on the other these data suggest a model where RNAP and NusA bind throughout the *M. tuberculosis* genome but their activity might be impeded by the lack of additional transcription factors and/or by the lack of nutrients, especially in Stat phase. It has been reported that approximately 23% of the RNAP signals in growing *E. coli* were not associated with RNA, thus leading

to the definition of "poised" RNAP (33). This is similar to the percentage (21%) of RNAP- and/or NusA-enriched features that do not correspond to a transcript in Exp phase in our study. Indeed, RNAP sitting on promoter sequences has to undergo several steps of abortive initiation and requires the presence of activators, such as GreA (34), and/or small molecules (35) or a particular DNA topology (36) in order to proceed fruitfully. A recent study identified transcription start site associated RNAs in *Mycoplasma pneumoniae* that overlap RNAP pausing sites (37). These small RNA molecules may exist in *M. tuberculosis* as well, thus explaining some of the unusual RNAP and/or NusA peaks that were not associated with a detectable transcript. Since the results of RNA-seq and RNAP ChIP-seq are not necessarily the same in *M. tuberculosis*, care should be taken with replacing transcriptomic profiling by detection of RNAP binding sites. Indeed, RNA studies (and nowadays powerful deep transcriptomic profiling) represent the gold standard for gene expression analysis.

The NusA-only containing signals deserve some discussion. The majority of these peaks were associated with RNA either in the same or in the neighboring feature whereas only 42 signals could not be correlated to transcription. Technical reasons could explain why RNAP was not detected above the background, for instance immunoprecipitation might be difficult because the epitope is masked by other transcription factors or by intricate chromatin structures in those regions.

Integration of ChIP-seq and RNA-seq data in a systems biology context provided the most relevant output to this work. Based on the criteria defined with the ribosomal protein operons (i.e. presence of an RNAP/NusA peak at the first feature and a continuous transcript throughout the TU) we developed an algorithm that identified 606 high-quality TUs widely distributed along the *M. tuberculosis* genome. More TUs may be present on

the chromosome but will have escaped identification since our algorithm was based on data generated *in vitro* in specific growth conditions. This is the first study of its kind to have performed genome-wide TU identification relying on empirical data rather than on predictions as done previously (38). We could confirm those TUs for which experimental evidence in the literature exists, for example the *furA-katG* operon (39), the *yrbE1A-mce1F* locus (40), the *pstB-pstA2* unit (41) and the *rv3134c-devS* operon (42). Our approach is complementary to the transcriptional start sites (TSS) mapping developed in *Helicobacter pylori* by means of a differential RNA-seq method (43). While the latter procedure allows precise identification of the 5'-end of primary transcripts, the combination of RNAP/NusA ChIP-seq and RNA-seq provides a more dynamic view of the interactions of the transcriptional complex with the genome and circumvents potential technical issues related to RNA stability.

By studying RNAP and NusA binding across TUs and their correlation with transcription, we established some common patterns in both growth phases that shed light on the biology of the transcriptional complex in *M. tuberculosis*. Firstly, promoter-proximal peaks for RNAP (and, to a lesser extent, for NusA) were recognized in the global TU profile. These signals may correspond to RNAP trapped at promoters before promoter clearance (i.e. abortive initiation), represent pause sites or result from premature transcription termination or attenuation mechanisms. Similar findings were previously described in *E. coli* (13), where the simultaneous occurrence of RNAP and NusA peaks allowed the classification of those promoter-proximal signals as elongating (EC) rather than stalled complexes. Given the strong analogy with our results, we infer that this is the case for *M. tuberculosis* as well, thus revealing Actinobacteria to be more similar to Proteobacteria than to the Firmicute *B. subtilis* (14) in the mechanistic basis of gene expression.

Secondly, significant abundance of NusA in the body of the transcribed TUs was observed, underlining its role in promoting transcription by anti-termination processes. In addition, the traveling ratio (TR) for RNAP was found to be <1, suggesting that the transition from initiation to elongation is a rate-limiting step in *M. tuberculosis* as well, similarly to *E. coli* (33,44). Slightly different from the RNAP TR, the TR for NusA was 0.75, suggesting that NusA-containing complexes move through the TUs faster. Indeed, the ratio between RNAP and NusA ER at promoters was in favor of RNAP, as witnessed by the global promoter profile, indicating that a proportion of the RNAP molecules does not associate with NusA and probably does not move along the genes. Overall, the importance of NusA in promoter clearance was also highlighted, since features lacking the NusA peak in the corresponding RNAP enrichment were not expressed. Finally, a striking polarity in the distribution of RNAP and NusA was noticed in the TU profiles.

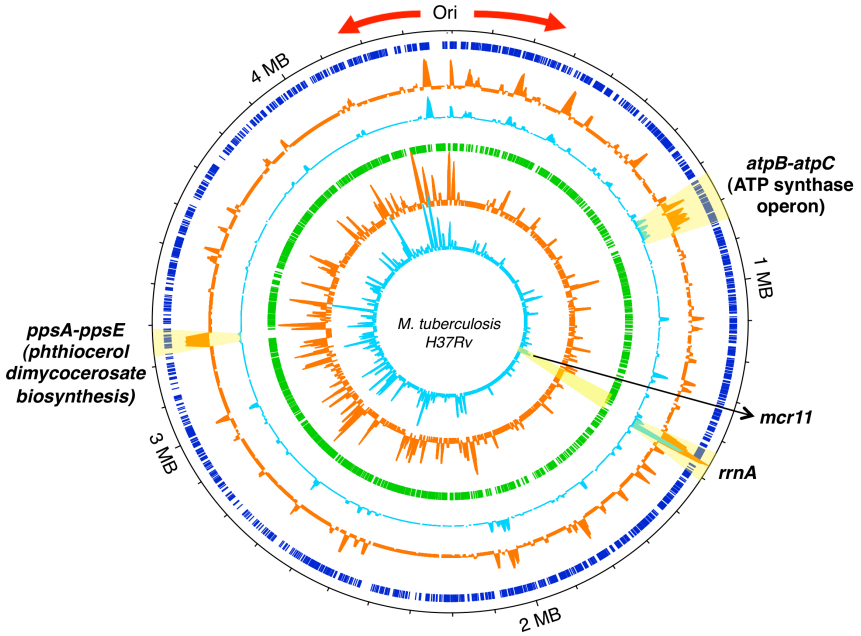


Figure 7. Circular genome view of the transcription profiles for transcriptional units in exponential and stationary phases. *Ori* indicates the origin of replication. The outermost blue circle represents CDSs on the forward strand, followed by the exponential phase expression profile for TUs on the forward strand in orange, the stationary phase expression profile for TUs on the forward strand in blue. The green

circle represents CDSs on the reverse strand, followed by the exponential phase expression profile for TUs on the reverse strand in orange and the stationary phase expression profile for TUs on the reverse strand in blue. Genomic coordinates are marked at every 1 MB. Red arrows indicate the direction of the replication forks. Some of the highly transcribed loci are highlighted in yellow and annotated.

The whole-genome distribution of the TUs (Figure 7) revealed a strong bias in the expression levels with respect to the direction of the replication forks. The majority of the highly transcribed TUs were localized on the leading strand thus resulting in an image where the *M. tuberculosis* chromosome is split into two non-symmetrical halves with the replication terminus located tentatively at ~2.1 MB. Interestingly, while the orientation of the CDSs in the H37Rv genome is only slightly biased (59% with the same polarity as DNA replication, (4)), their expression levels are more affected, as confirmed by applying a Kolmogorov-Smirnov test to the cumulative RPM along each strand resulting in a $p < 10^{-10}$ (Supplementary Figure S8). This is different from what was observed in other bacteria, such as *E. coli* (45) and *B. subtilis* (46), where the orientation bias is more pronounced. Our genome-wide transcriptional study shows that in a slow-growing pathogenic bacterium the DNA and RNA polymerases proceed in the same direction through most of the highly expressed features thus avoiding potential collisions.

Supplementary data

Supplementary Data are available at NAR online: Supplementary Tables 1-7, Supplementary Figures 1-8.

Acknowledgements

The authors would like to thank Dr. Kristine Arnvig and Dr. Ian Taylor (NIMR, London, United Kingdom) for providing purified NusA, Dr. Ida Rosenkrands (Statens Serum Institut, Copenhagen, Denmark) for generating anti-NusA antibodies, Dr. Keith Harshman (Lausanne Genomic Technologies Facility, University of Lausanne, Switzerland) for advice on RNA-seq experiments.

Funding

The research leading to these results has received funding from the European Community's Seventh Framework Programme ([FP7/2007-2013]) under grant agreement n°260872, SystemsX.ch and the Swiss National Science Foundation under grant n°31003A-125061.

Author contributions

STC and CS designed the study, CS performed the experiments, SU, JR, STC, CS analyzed data, SU, STC and CS wrote the paper.

Conflict of interest statement

The authors declare that no conflict of interest exists.

References

1. Russell, D.G. (2011) Mycobacterium tuberculosis and the intimate discourse of a chronic infection. *Immunol Rev*, **240**, 252-268.
2. Russell, D.G., VanderVen, B.C., Lee, W., Abramovitch, R.B., Kim, M.J., Homolka, S., Niemann, S. and Rohde, K.H. (2010) Mycobacterium tuberculosis wears what it eats. *Cell Host Microbe*, **8**, 68-76.
3. Stokes, R.W. and Waddell, S.J. (2009) Adjusting to a new home: Mycobacterium tuberculosis gene expression in response to an intracellular lifestyle. *Future Microbiol*, **4**, 1317-1335.
4. Cole, S.T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S.V., Eiglmeier, K., Gas, S., Barry, C.E., 3rd *et al.* (1998) Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature*, **393**, 537-544.
5. Greenblatt, J. and Li, J. (1981) The nusA gene protein of Escherichia coli. Its identification and a demonstration that it interacts with the gene N transcription anti-termination protein of bacteriophage lambda. *J Mol Biol*, **147**, 11-23.
6. Vogel, U. and Jensen, K.F. (1997) NusA is required for ribosomal antitermination and for modulation of the transcription elongation rate of both antiterminated RNA and mRNA. *J Biol Chem*, **272**, 12265-12271.
7. Arnvig, K.B., Pennell, S., Gopal, B. and Colston, M.J. (2004) A high-affinity interaction between NusA and the rrn nut site in Mycobacterium tuberculosis. *Proc Natl Acad Sci U S A*, **101**, 8325-8330.
8. Gusarov, I. and Nudler, E. (2001) Control of intrinsic transcription termination by N and NusA: the basic mechanisms. *Cell*, **107**, 437-449.
9. Schmidt, M.C. and Chamberlin, M.J. (1987) nusA protein of Escherichia coli is an efficient transcription termination factor for certain terminator sites. *J Mol Biol*, **195**, 809-818.
10. Cardinale, C.J., Washburn, R.S., Tadigotla, V.R., Brown, L.M., Gottesman, M.E. and Nudler, E. (2008) Termination factor Rho and its cofactors NusA and NusG silence foreign DNA in E. coli. *Science*, **320**, 935-938.
11. Gopal, B., Haire, L.F., Gamblin, S.J., Dodson, E.J., Lane, A.N., Papavinasasundaram, K.G., Colston, M.J. and Dodson, G. (2001) Crystal structure of the transcription elongation/anti-termination

factor NusA from Mycobacterium tuberculosis at 1.7 A resolution. *J Mol Biol*, **314**, 1087-1095.

12. Worbs, M., Bourenkov, G.P., Bartunik, H.D., Huber, R. and Wahl, M.C. (2001) An extended RNA binding surface through arrayed S1 and KH domains in transcription factor NusA. *Mol Cell*, **7**, 1177-1189.
13. Mooney, R.A., Davis, S.E., Peters, J.M., Rowland, J.L., Ansari, A.Z. and Landick, R. (2009) Regulator trafficking on bacterial transcription units in vivo. *Mol Cell*, **33**, 97-108.
14. Ishikawa, S., Oshima, T., Kurokawa, K., Kusuya, Y. and Ogasawara, N. (2010) RNA polymerase trafficking in Bacillus subtilis cells. *J Bacteriol*, **192**, 5778-5787.
15. Betts, J.C., Lukey, P.T., Robb, L.C., McAdam, R.A. and Duncan, K. (2002) Evaluation of a nutrient starvation model of Mycobacterium tuberculosis persistence by gene and protein expression profiling. *Mol Microbiol*, **43**, 717-731.
16. Dahl, J.L., Kraus, C.N., Boshoff, H.I., Doan, B., Foley, K., Avarbock, D., Kaplan, G., Mizrahi, V., Rubin, H. and Barry, C.E., 3rd. (2003) The role of RelMtb-mediated adaptation to stationary phase in long-term persistence of Mycobacterium tuberculosis in mice. *Proc Natl Acad Sci U S A*, **100**, 10026-10031.
17. Voskuil, M.I., Visconti, K.C. and Schoolnik, G.K. (2004) Mycobacterium tuberculosis gene expression during adaptation to stationary phase and low-oxygen dormancy. *Tuberculosis (Edinb)*, **84**, 218-227.
18. Arnvig, K.B., Comas, I., Thomson, N.R., Houghton, J., Boshoff, H.I., Croucher, N.J., Rose, G., Perkins, T.T., Parkhill, J., Dougan, G. *et al.* (2011) Sequence-based analysis uncovers an abundance of non-coding RNA in the total transcriptome of Mycobacterium tuberculosis. *PLoS Pathog*, **7**, e1002342.
19. Gold, B., Deng, H., Bryk, R., Vargas, D., Eliezer, D., Roberts, J., Jiang, X. and Nathan, C. (2008) Identification of a copper-binding metallothionein in pathogenic mycobacteria. *Nat Chem Biol*, **4**, 609-616.
20. Hu, Y., Butcher, P.D., Mangan, J.A., Rajandream, M.A. and Coates, A.R. (1999) Regulation of hmp gene transcription in Mycobacterium tuberculosis: effects of oxygen limitation and nitrosative and oxidative stress. *J Bacteriol*, **181**, 3486-3493.
21. deHaseth, P.L., Lohman, T.M., Burgess, R.R. and Record, M.T., Jr. (1978) Nonspecific interactions of Escherichia coli RNA polymerase with native and denatured DNA: differences in the binding behavior of core and holoenzyme. *Biochemistry*, **17**, 1612-1622.

22. Grigorova, I.L., Phleger, N.J., Mutalik, V.K. and Gross, C.A. (2006) Insights into transcriptional regulation and sigma competition from an equilibrium model of RNA polymerase binding to DNA. *Proc Natl Acad Sci U S A*, **103**, 5332-5337.
23. Manganelli, R., Dubnau, E., Tyagi, S., Kramer, F.R. and Smith, I. (1999) Differential expression of 10 sigma factor genes in Mycobacterium tuberculosis. *Mol Microbiol*, **31**, 715-724.
24. Aagaard, C., Hoang, T., Dietrich, J., Cardona, P.J., Izzo, A., Dolganov, G., Schoolnik, G.K., Cassidy, J.P., Billeskov, R. and Andersen, P. (2011) A multistage tuberculosis vaccine that confers efficient protection before and after exposure. *Nat Med*, **17**, 189-194.
25. Gardner, P.P., Daub, J., Tate, J.G., Nawrocki, E.P., Kolbe, D.L., Lindgreen, S., Wilkinson, A.C., Finn, R.D., Griffiths-Jones, S., Eddy, S.R. *et al.* (2009) Rfam: updates to the RNA families database. *Nucleic Acids Res*, **37**, D136-140.
26. Vitreschak, A.G., Mironov, A.A., Lyubetsky, V.A. and Gelfand, M.S. (2008) Comparative genomic analysis of T-box regulatory systems in bacteria. *RNA*, **14**, 717-735.
27. Warner, D.F., Savvi, S., Mizrahi, V. and Dawes, S.S. (2007) A riboswitch regulates expression of the coenzyme B12-independent methionine synthase in Mycobacterium tuberculosis: implications for differential methionine synthase function in strains H37Rv and CDC1551. *J Bacteriol*, **189**, 3655-3659.
28. Sha, Y., Lindahl, L. and Zengel, J.M. (1995) Role of NusA in L4-mediated attenuation control of the S10 r-protein operon of Escherichia coli. *J Mol Biol*, **245**, 474-485.
29. Yakhnin, A.V. and Babitzke, P. (2002) NusA-stimulated RNA polymerase pausing and termination participates in the Bacillus subtilis trp operon attenuation mechanism invitro. *Proc Natl Acad Sci U S A*, **99**, 11067-11072.
30. Arnvig, K.B. and Young, D.B. (2009) Identification of small RNAs in Mycobacterium tuberculosis. *Mol Microbiol*, **73**, 397-408.
31. DiChiara, J.M., Contreras-Martinez, L.M., Livny, J., Smith, D., McDonough, K.A. and Belfort, M. (2010) Multiple small RNAs identified in Mycobacterium bovis BCG are also expressed in Mycobacterium tuberculosis and Mycobacterium smegmatis. *Nucleic Acids Res*, **38**, 4067-4078.
32. Pellin, D., Miotto, P., Ambrosi, A., Cirillo, D.M. and Di Serio, C. (2012) A genome-wide identification analysis of small regulatory RNAs in Mycobacterium tuberculosis by RNA-Seq and conservation analysis. *PLoS One*, **7**, e32723.

33. Reppas, N.B., Wade, J.T., Church, G.M. and Struhl, K. (2006) The transition between transcriptional initiation and elongation in E. coli is highly variable and often rate limiting. *Mol Cell*, **24**, 747-757.
34. Stepanova, E., Lee, J., Ozerova, M., Semenova, E., Datsenko, K., Wanner, B.L., Severinov, K. and Borukhov, S. (2007) Analysis of promoter targets for Escherichia coli transcription elongation factor GreA in vivo and in vitro. *J Bacteriol*, **189**, 8772-8785.
35. Lee, S.J. and Gralla, J.D. (2004) Osmo-regulation of bacterial transcription via poised RNA polymerase. *Mol Cell*, **14**, 153-162.
36. Shin, M., Song, M., Rhee, J.H., Hong, Y., Kim, Y.J., Seok, Y.J., Ha, K.S., Jung, S.H. and Choy, H.E. (2005) DNA looping-mediated repression by histone-like protein H-NS: specific requirement of Esigma70 as a cofactor for looping. *Genes Dev*, **19**, 2388-2398.
37. Yus, E., Guell, M., Vivancos, A.P., Chen, W.H., Lluch-Senar, M., Delgado, J., Gavin, A.C., Bork, P. and Serrano, L. (2012) Transcription start site associated RNAs in bacteria. *Mol Syst Biol*, **8**, 585.
38. Roback, P., Beard, J., Baumann, D., Gille, C., Henry, K., Krohn, S., Wiste, H., Voskuil, M.I., Rainville, C. and Rutherford, R. (2007) A predicted operon map for Mycobacterium tuberculosis. *Nucleic Acids Res*, **35**, 5085-5095.
39. Pym, A.S., Domenech, P., Honore, N., Song, J., Deretic, V. and Cole, S.T. (2001) Regulation of catalase-peroxidase (KatG) expression, isoniazid sensitivity and virulence by furA of Mycobacterium tuberculosis. *Mol Microbiol*, **40**, 879-889.
40. Casali, N., White, A.M. and Riley, L.W. (2006) Regulation of the Mycobacterium tuberculosis mce1 operon. *J Bacteriol*, **188**, 441-449.
41. Torres, A., Juarez, M.D., Cervantes, R. and Espitia, C. (2001) Molecular analysis of Mycobacterium tuberculosis phosphate specific transport system in Mycobacterium smegmatis. Characterization of recombinant 38 kDa (PstS-1). *Microb Pathog*, **30**, 289-297.
42. Bagchi, G., Chauhan, S., Sharma, D. and Tyagi, J.S. (2005) Transcription and autoregulation of the Rv3134c-devR-devS operon of Mycobacterium tuberculosis. *Microbiology*, **151**, 4045-4053.
43. Sharma, C.M., Hoffmann, S., Darfeuille, F., Reignier, J., Findeiss, S., Sittka, A., Chabas, S., Reiche, K., Hackermuller, J., Reinhardt, R. *et al.* (2010) The primary transcriptome of the major human pathogen Helicobacter pylori. *Nature*, **464**, 250-255.
44. Wade, J.T. and Struhl, K. (2008) The transition from transcriptional initiation to elongation. *Curr Opin Genet Dev*, **18**, 130-136.

45. Blattner, F.R., Plunkett, G., 3rd, Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F. *et al.* (1997) The complete genome sequence of Escherichia coli K-12. *Science*, **277**, 1453-1462.
46. Kunst, F., Ogasawara, N., Moszer, I., Albertini, A.M., Alloni, G., Azevedo, V., Bertero, M.G., Bessieres, P., Bolotin, A., Borchert, S. *et al.* (1997) The complete genome sequence of the gram-positive bacterium Bacillus subtilis. *Nature*, **390**, 249-256.
47. Sala, C., Haouz, A., Saul, F.A., Miras, I., Rosenkrands, I., Alzari, P.M. and Cole, S.T. (2009) Genome-wide regulon and crystal structure of BlaI (Rv1846c) from Mycobacterium tuberculosis. *Mol Microbiol*, **71**, 1102-1116.
48. Langmead, B., Trapnell, C., Pop, M. and Salzberg, S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, **10**, R25.
49. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078-2079.
50. Dreszer, T.R., Karolchik, D., Zweig, A.S., Hinrichs, A.S., Raney, B.J., Kuhn, R.M., Meyer, L.R., Wong, M., Sloan, C.A., Rosenbloom, K.R. *et al.* (2012) The UCSC Genome Browser database: extensions and updates 2011. *Nucleic Acids Res*, **40**, D918-923.
51. Anders, S. and Huber, W. (2010) Differential expression analysis for sequence count data. *Genome Biol*, **11**, R106.
52. Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*, **5**, R80.
53. Edgar, R., Domrachev, M. and Lash, A.E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*, **30**, 207-210.

4.2.2 Genome-wide definition of the SigF Regulon in *M. tb*

Contribution: Processing of raw data, normalization, filtering, quantification and integration of the ChIP-on-chip and microarray results.

In order to adapt and survive the multitude of stresses imposed by the host immune system, *M. tb* employs a large number of transcriptional regulators to effectively control its gene expression. The *M. tb* genome encodes 13 putative sigma factors that modulate gene expression by directly altering the promoter preference of RNA polymerase (S. Cole *et al.* 1998). The primary sigma factor, SigA is essential for general transcription of housekeeping genes, while the other sigma factors (SigB to SigM) are nonessential and generally respond to various environmental and physiological stresses to help fine-tune expression of specific genes (Manganelli *et al.* 2004).

SigF is a stationary-phase stress response sigma factor required for full virulence of *M. tb.* Studies in *M. bovis* BCG showed induction of the *sigF* gene in response to cold shock, hypoxia and nitrogen depletion, but these findings could not be replicated in *M. tb*, suggesting differences in regulation of *sigF* between the two species. Although *sigF* is nonessential for bacterial growth *in vitro* and occurs as a pseudogene in *M. leprae* it is conserved in most mycobacterial species. Several approaches have been used to identify genes requiring SigF for their expression, but there is little agreement between the findings from these studies. However, all of them indicate that SigF controls its own transcription.

In this study, a combination of two genome-wide approaches was used to define the SigF regulon: ChIP-on-chip (chromatin immunoprecipitation followed by hybridization to microarrays) and expression profiling analysis

(using microarrays) of SigF-mediated transcripts. Since SigF is not an abundant protein in the logarithmic phase of growth, a pristinamyin IA-inducible system was used to control its expression. Experiments were performed using custom-made microarrays. Each array contained 43,450 probes (21,725 probes on each strand), located, on average, every 203 bases and covering the entire *M. tb* H37Rv genome.

Analysis of the ChIP-on-chip data revealed a total of 201 significantly enriched probes, which clustered in 67 distinct genomic loci. To understand SigF mediated gene regulation, transcriptome data were compared for *sigF*-induced and un-induced samples. We found 138 probes, corresponding to 51 genomic loci containing 33 protein-coding genes that were significantly upregulated upon *sigF* induction. No genes were found to be downregulated. On integration of the two datasets, we identified 16 genomic loci that were upregulated upon SigF induction and had a corresponding 5' SigF-binding site. Thus, we could distinguish between transcripts that are directly regulated by SigF from those that are controlled indirectly, through a regulatory cascade, possibly involving other transcriptional regulators. Quantitative PCR confirmed the microarray findings and SigF-specific transcription start points could be mapped using 5'-RACE (rapid amplification of cDNA ends). A canonical SigF consensus promoter sequence GGTTT-$N_{(15\text{-}17)}$-GGGTA was identified prior to eleven genes.

The SigF regulon defined using these data confirmed that SigF controls its own expression, along with the expression of genes involved in lipid and intermediary metabolism, virulence (*hbhA*), and at least one transcriptional regulator (*Rv2884*), possibly acting downstream of SigF. In addition, SigF was also found to direct the transcription of the gene for small RNA F6.

JB
Journals.ASM.org

Genome-Wide Definition of the SigF Regulon in *Mycobacterium tuberculosis*

Ruben C. Hartkoorn,[a] Claudia Sala,[a] Swapna Uplekar,[a] Philippe Busso,[a] Jacques Rougemont,[b] and Stewart T. Cole[a]

Global Health Institute, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland,[a] and Bioinformatics and Biostatistics Core Facility, Swiss Institute of Bioinformatics, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland[b]

In *Mycobacterium tuberculosis* the alternative sigma factor SigF controls the expression of a particular subset of genes by altering RNA polymerase specificity. Here, we utilize two genome-wide approaches to identify SigF-binding sites: chromatin immunoprecipitation (ChIP-on-chip) and microarray analysis of SigF-mediated transcripts. Since SigF is not an abundant protein in the logarithmic phase of growth, a pristinamyin IA-inducible system was used to control its expression. We identified 67 high-affinity SigF-binding sites and 16 loci where a SigF promoter directs the expression of a transcript. These loci include *sigF* itself, genes involved in lipid and intermediary metabolism and virulence, and at least one transcriptional regulator (*Rv2884*), possibly acting downstream of SigF. In addition, SigF was also found to direct the transcription of the gene for small RNA F6. Many loci were also found where SigF may be involved in antisense transcription, and in two cases (*Rv1358* and *Rv1870c*) the SigF-dependent promoter was located within the predicted coding sequence. Quantitative PCR confirmed the microarray findings and 5′-rapid amplification of cDNA ends was used to map the SigF-specific transcriptional start points. A canonical SigF consensus promoter sequence GGTTT-$N_{(15-17)}$-GGGTA was found prior to 11 genes. Together, these data help to define the SigF regulon and show that SigF not only governs expression of proteins such as the virulence factor, HbhA, but also impacts novel functions, such as noncoding RNAs and antisense transcripts.

Mycobacterium tuberculosis, the etiologic agent of human tuberculosis, is a slow-growing pathogen that survives and multiplies despite a huge barrage of stresses imposed by the immune system. Resisting these stresses requires the bacterium to be able to sense its environment and react accordingly. The complete genome sequence of *M. tuberculosis* (7) revealed at least 140 potential transcriptional regulators (36), and a total of 13 sigma factors that can directly alter the promoter preference of RNA polymerase. Among these sigma factors, SigA is the housekeeping sigma factor that is responsible for general transcription, while the other sigma factors (SigB to SigM) are nonessential but help fine-tune gene expression in response to various physiological stresses.

SigF was first described in *Mycobacterium bovis* BCG (BCG) by DeMaio et al. (10) as a stationary-phase stress response sigma factor that is also induced upon cold shock and nitrogen depletion. Although some of these findings were partly confirmed in BCG by others (34), they were not replicated in *M. tuberculosis* (23), with only nutrient starvation showing some induction of *sigF* (4). Although *sigF* is nonessential for bacterial growth *in vitro* and found as a pseudogene in *Mycobacterium leprae* (24, 31), *sigF* is conserved in most mycobacterial species. Infection studies in mice and guinea pigs show that *M. tuberculosis* lacking *sigF* is partially attenuated (5, 14, 19).

To date, a number of genome-wide approaches have been used to uncover the SigF transcriptome, but with contrasting results. SigF-regulated *M. tuberculosis* genes have been studied by comparing the transcriptomes of wild-type and *sigF* knockout mutant bacteria (14) and of wild-type *M. tuberculosis* with a strain carrying an acetamide-inducible *sigF* (20, 37). Alternative strategies used to discover SigF-controlled genes were ChIP-on-chip (chromatin immunoprecipitation and hybridization to microarrays) analysis of SigF-binding sites in BCG (30) and application of a two-plasmid system in *Escherichia coli* to identify genes controlled by SigF (16). Overall, there is little agreement between the findings of these studies except for confirmation of the initial finding that SigF controls its own transcription (18).

Here, a combination of both genome-wide transcriptome and ChIP-on-chip analyses was used to detect SigF-induced transcripts and identify SigF DNA-binding sites, respectively. The findings were subsequently confirmed by quantitative PCR and mapping the SigF-specific (or SigF-controlled) transcriptional start sites with the combined results leading to definition of the SigF regulon.

MATERIALS AND METHODS

Bacterial strains, reagents, and chemicals. All experiments were performed in either *M. tuberculosis* strain H37Rv or its isogenic mutant, H37Rv(Δ*rsbW/sigF*), lacking the *rsbW* and *sigF* operon, as described previously (15). These strains were transformed with either pMY769, an empty pristinamycin IA (P-IA)-inducible integrating vector, or pMYsigF, a pMY769 derivative with a P-IA-inducible *sigF* gene (15) to create H37Rv::pMY769, H37Rv::pMYsigF, H37Rv(Δ*rsbW/sigF*)::pMY769, and H37Rv(Δ*rsbW/sigF*)::pMYsigF. The strains were routinely grown in Middlebrook 7H9 medium supplemented with 10% albumin-dextrose-catalase (ADC), 0.2% glycerol, and 0.05% Tween 80. Bacterial growth was monitored spectrophotometrically at an optical density of 600 nm (OD_{600}).

P-IA was purified from pyostatin tablets as described previously (15). Rabbit anti-SigF polyclonal antibodies were kindly provided by Ida Rosenkrands (Statens Serum Institut, Copenhagen, Denmark), and a

Received 9 December 2011 Accepted 27 January 2012
Published ahead of print 3 February 2012
Address correspondence to Stewart T. Cole, stewart.cole@epfl.ch.
Supplemental material for this article may be found at http://jb.asm.org/.

doi:10.1128/JB.06692-11

mouse anti-HbhA polyclonal antibody was provided by Giovanni Delogu (Catholic University of the Sacred Heart, Rome, Italy). All other chemicals (except when mentioned) were obtained from Sigma-Aldrich.

Protein gels and Western blot analysis. To analyze the total protein content, bacteria were harvested, and the total cell extract was subjected to SDS-PAGE for Simply Blue staining or Western blot analysis as described previously (15). Protein bands were excised from the gel and identified by liquid chromatography-tandem mass spectrometry (LC-MS/MS) analysis following trypsin or chymotrypsin digestion.

SigF chromatin immunoprecipitation experiments. Log-phase cultures of H37Rv(Δ*rsbW/sigF*)::pMYsigF and H37Rv(Δ*rsbW/sigF*):: pMY769 were diluted to an OD_{600} of 0.05. After overnight incubation, P-IA (final concentration 2 μg/ml) was added to both cultures, and these were incubated for 3 days at 37°C with shaking. Chromatin immunoprecipitation was performed as previously described (33), with some modifications. Briefly, protein-DNA complexes were cross-linked using formaldehyde (final concentration, 1%; 10 min, 37°C) and quenched using glycine (final concentration, 125 mM). Bacteria were pelleted and washed twice with Tris-buffered saline (20 mM Tris-HCl [pH 7.5], 150 mM NaCl), and the pellets were stored at −80°C. The pellets were resuspended in 600 μl of immunoprecipitation buffer (50 mM HEPES-KOH [pH 7.5], 150 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.1% sodium deoxycholate, 0.1% SDS, Roche Antiprotease Mini) and disrupted by sonication at 4°C for 30 min (30-s pulses with a Bioruptor, Diagenode) to shear DNA to an average size of 200 to 400 bp. After centrifugation (2,000 × *g*, 15 min), the supernatant was diluted to 1.1 ml in immunoprecipitation buffer and mixed overnight (4°C) with 20 μl of polyclonal rabbit anti-SigF antibody. In parallel, per sample, 100 μl of protein A/G UltraLink resin (Thermo Scientific) was washed in immunoprecipitation buffer and incubated overnight (4°C) in 1 ml of immunoprecipitation buffer with 1 mg of bovine serum albumin/ml and 0.1 mg of salmon sperm DNA/ml. After overnight blocking, protein A/G UltraLink resin was pelleted (2,000 × *g*, 2 min) and mixed with the antibody and protein-DNA complex mixture (rotary mixer, 90 min, 4°C). The protein A/G UltraLink resin was pelleted (2,000 × *g*, 2 min) and washed twice with immunoprecipitation buffer, once with immunoprecipitation buffer plus 500 mM NaCl, once with wash buffer III (10 mM Tris-HCl [pH 8], 250 mM LiCl, 1 mM EDTA, 0.5% Nonidet P-40, 0.5% sodium deoxycholate), and once with Tris-EDTA buffer (pH 7.5). Immunoprecipitated complexes were eluted from the beads by treatment with 100 μl of elution buffer (50 mM Tris-HCl [pH 7.5], 10 mM EDTA, 1% SDS) at 65°C for 20 min. Samples were then treated with RNase A, and cross-links were reversed by incubation for 2 h at 56°C and 6 h at 65°C in 0.5× elution buffer plus 50 μg of proteinase K. DNA was extracted twice with phenol-chloroform, precipitated, and resuspended in 20 μl of water. DNA obtained from immunoprecipitation and total genomic DNA were labeled by Klenow random priming, with Cy3-dCTP or Cy5-dCTP (GE Healthcare), respectively, and purified with Qiagen MinElute kit, according to the manufacturer's protocol.

Microarray analysis of the SigF transcriptome. Log-phase H37Rv:: pMY769 and H37Rv::pMYsigF cultures were diluted to OD_{600} of 0.05 and split into two 50-ml cultures. After overnight incubation, P-IA (final concentration, 2 μg/ml) was added to one culture, and an equivalent volume of dimethyl sulfoxide was added to the other. After 72 h of incubation (rotary mixer, 4°C), the RNA was extracted using 6 M GTC (6 M guanidine thiocyanate, 0.75% sodium *N*-laurylsarcosine, 37.5 mM sodium citrate, 0.75% Tween 80, 0.15 M β-mercaptoethanol) and TRIzol (Invitrogen) as described previously (28). Extracted RNA was treated twice by RQ1 DNase (Promega) according to the manufacturer's instructions and cleaned up using an RNeasy minikit (Qiagen). RNA quantity and quality were assessed by using NanoDrop, agarose gel electrophoresis, and PCR to confirm the absence of DNA. mRNA was reverse transcribed and cDNA was labeled using RevertAID-H-minus reverse transcriptase (Fermentas) according to the manufacturer's instructions, using 1 to 4 μg of purified mRNA, random hexamers, and either Cy3-dCTP or Cy5-dCTP (GE Healthcare). Reverse transcription was performed using the following conditions: 10 min at 25°C, 90 min at 45°C, and 5 min at 70°C. Labeled cDNA was treated with RNase H and RNase A (30 min, 37°C), followed by cDNA cleanup using a GFX purification kit (GE Healthcare).

Microarray design, hybridization, and analysis. The same microarrays were used for SigF ChIP-on-chip and transcriptome analyses, and these were purchased from Oxford Gene Technology as described previously (33). Each array contained 43,450 probes (21,725 probes on each strand) covering the entire genome, located, on average, every 203 bases. Labeled DNA or cDNA was hybridized to microarrays in an Agilent Technologies Microarray chamber at 55°C for 72 h in buffer composed of 1 M NaCl, 50 mM MES (morpholineethanesulfonic acid), 20% formamide, 20 mM EDTA, and 1% Triton X-100. The arrays were washed once in 6× SSPE (1× SSPE is 0.18 M NaCl, 10 mM NaH_2PO_4, and 1 mM EDTA [pH 7.4])–0.005% *N*-laurylsarcosine and once in 0.06% SSPE–0.18% polyethylene glycol 200, both times for 5 min at room temperature. Finally, the microarrays were dried in isopropanol and immediately scanned using an Agilent Technologies microarray scanner, after which results were extracted using GenePix Pro 5.0 software.

For each ChIP-on-chip data set, the raw Cy3 and Cy5 intensity data obtained from the GenePix software were median normalized. Since no strand specificity is expected for ChIP-ed DNA (since the proteins bind double-stranded gDNA), the average of the sense and antisense intensities was calculated for each probe, and the enrichment was measured as the ratio of the averaged Cy3 and Cy5 intensities. The enrichment ratios were $\log_2$ transformed and further corrected using Lowess normalization in order to eliminate intensity-dependent dye bias. Since the number of enriched probes was low (<1%), the means and standard deviations of the $\log_2$-normalized enrichment ratios of the whole data set were used as a measure of variance of the data set. Using this, a Z-score and corresponding probability (*P* value) was calculated for each probe to be significantly different from the mean.

In addition, since DNA fragments obtained from ChIP-on-chip experiments were between 200 and 400 bp, they are expected to hybridize to probes adjacent to each other on the genome. Since the physical arrangement of probes on the array does not reflect their location on the *M. tuberculosis* chromosome, the data were reordered genomically. Putative SigF-binding sites were defined as a stretch of two or more consecutive probes with a *P* value of <0.01 in at least three out of four data sets. Probe enrichment was calculated as $2^{(\text{mean } \log_2 \text{ normalized enrichment ratios})}$. Positive probes were mapped to the current *M. tuberculosis* H37Rv genome annotation, as described in TubercuList (http://tuberculist.epfl.ch).

For each cDNA transcriptome data set, a similar data analysis was performed, but the sense and antisense probe intensities were kept separate in order to determine the direction of transcription. As described above, for each probe the raw Cy3 and Cy5 intensity data were median normalized, the enrichment ratios were calculated, $\log_2$ transformed, and Lowess normalized, and the statistical significance from the mean was determined by calculating the Z-score and the *P* value. In experiments to determine SigF mediated gene induction, (H37Rv::pMYSigF with or without P-IA), the probes were considered significantly enriched if they showed a *P* value below 0.005 in at least three of four data sets. In the experiment to determine P-IA mediated gene induction (H37Rv:: pMY769 with or without P-IA), the probes were considered significantly enriched if they showed a *P* value below 0.01 in at least two of the three data sets and if they were also enriched in the H37Rv::pMYSigF experiment, based on the previous criteria. To allow for the detection of smaller RNA species, in contrast to the ChIP-on-chip data set analysis, singletons (enriched probes with no genomically adjacent positive spot) were taken into account for the analysis. Positive probes were mapped, in the correct orientation, to the *M. tuberculosis* H37Rv genome annotation from TubercuList (http://tuberculist.epfl.ch).

All ChIP-on-chip and cDNA transcriptome data sets have been deposited in NCBI's Gene Expression Omnibus (12) and can be accessed through GEO Series accession number GSE35080, with subseries

GSE34919 (ChIP-on-chip microarray data) and GSE34922 (transcriptome microarray data).

RNA extraction for Q-PCR. To confirm the transcriptional regulation of genes by SigF, microarray data was confirmed by quantitative PCR (Q-PCR). Briefly, 40-ml cultures of H37Rv(Δ*rsbW/sigF*)::pMY769 and H37Rv(Δ*rsbW/sigF*)::pMYsigF were grown in the presence or absence of P-IA (2 μg/ml). After 0, 24, and 72 h induction, 10 ml of each culture was pelleted, flash frozen in liquid nitrogen, and subsequently extracted with TRIzol (according to the manufacturer's instructions) using bead beating to disrupt the bacteria, two rounds of DNase treatment to remove DNA (confirmed by PCR), and ethanol precipitation to clean and concentrate RNA (so as to retain small RNAs [sRNAs]). RNA was reverse transcribed using random primers and RevertAID-H-minus reverse transcriptase (Fermentas) according to the manufacturer's instructions. Biological triplicates were performed for each condition to allow for statistical analysis.

Q-PCR of purified cDNA was performed using SYBRgreen PCR mastermix (Applied Biosystems) and primer sets from Table S1 in the supplemental material. Q-PCR for each sample was performed in duplicate on a 7900HT sequence detection system (Applied Biosystems) using the following protocol: denaturation at 95°C for 3 min, followed by 40 cycles of denaturation at 95°C for 30 s and annealing/elongation with data collection at 60°C for 40 s. Melting curves were constructed to ensure that only one amplification product was obtained. To calculate enrichment ratios, mean threshold cycle (C_T) values were normalized against *sigA* C_T (ΔC_T) that was normalized against day 0 no P-IA ΔC_T ($\Delta\Delta C_T$). The statistical significance was determined by comparing $\Delta\Delta C_T$ values using a nonparametric paired *t* test (Wilcoxon test). The total fold enrichment was calculated as $2^{(-\Delta\Delta CT)}$.

Multiplex 5′-RACE. To determine the start point of transcription of SigF-regulated transcripts, multiplex 5′ rapid amplification of cDNA ends (5′-RACE) was performed on RNA extracted from H37Rv(Δ*rsbW/sigF*):: pMYsigF and H37Rv(Δ*rsbW/sigF*)::pMY769 induced with P-IA for 3 days (see above). For the first step of multiplex 5′-RACE, reverse transcription of purified mRNA was performed with sets of four specific primers for four different putative SigF-regulated genes. Briefly, RNA (500 ng), gene-specific primers (12.5 μM each), and deoxynucleoside triphosphates (10 mM) were mixed in 10 μl and heated (65°C, 5 min). After flash cooling on ice, RT buffer (2 μl), $MgCl_2$ (4 μl), dithiothreitol (2 μl) RNase OUT (1 μl), and Superscript III (1 μl) were added from the SuperScript III first-strand synthesis system (Invitrogen). First-stand synthesis was performed as follows (50°C for 1 h, 55°C for 1 h, and 70°C for 15 min). cDNA was purified with a High-Pure PCR product purification kit (Roche) and then used in the subsequent poly(A) tailing reaction (30 min at 37°C in the presence of 0.2 mM dATP and 80 U of terminal transferase [Roche]). For the amplification of specific genes, seminested PCR amplification was performed with an oligo(dT)-anchor primer and a second gene-specific primer. The products of the PCR were purified and sequenced by Sanger sequencing on an ABI 3130XL genetic analyzer (Applied Biosystems) with the second gene-specific primer.

For sequence motif identification, the BioProspector tool was used (21; http://ai.stanford.edu/~xsliu/cgi-bin/BPsearch.cgi), and the results were entered into Weblogo to generate a sequence logo (9; http://weblogo.berkeley.edu/logo.cgi).

RESULTS

We sought here to use two different yet complementary genome-wide approaches to identify genes regulated by RNA polymerase containing SigF: chromatin-immunoprecipitation with SigF-specific antibodies and gene expression profiling of *M. tuberculosis* cells that conditionally express this sigma factor.

Conditional overexpression of *sigF*. Quantitative reverse transcription-PCR (RT-PCR) showed that the level of *sigF* mRNA expression was the same in strains H37Rv and H37Rv::pMYsigF in the absence of the inducer P-IA and that no *sigF* transcript could be detected in H37Rv(Δ*rsbW/sigF*) or H37Rv(Δ*rsbW/sigF*):: pMYsigF without inducer. The addition of P-IA to cultures resulted in an at least 256-fold induction of *sigF* in H37Rv:: pMYsigF (after 3 days of P-IA exposure), whereas *sigF* transcripts were also readily detectable in induced H37Rv(Δ*rsbW/sigF*)::pMYsigF cells (data not shown). SigF protein induction could not be seen by Simply Blue staining of protein lysates run on an SDS-PAGE gel (Fig. 1A); however, Western blot analysis confirmed the induction of SigF protein after P-IA treatment, as well as tight regulation of *sigF* in its absence (Fig. 1B). As observed previously (15), the induction of *sigF* in both H37Rv::pMYsigF and H37Rv(Δ*rsbW/sigF*)::pMYsigF brought about a decrease in growth rate after 24 h, whereas control strains were not affected.

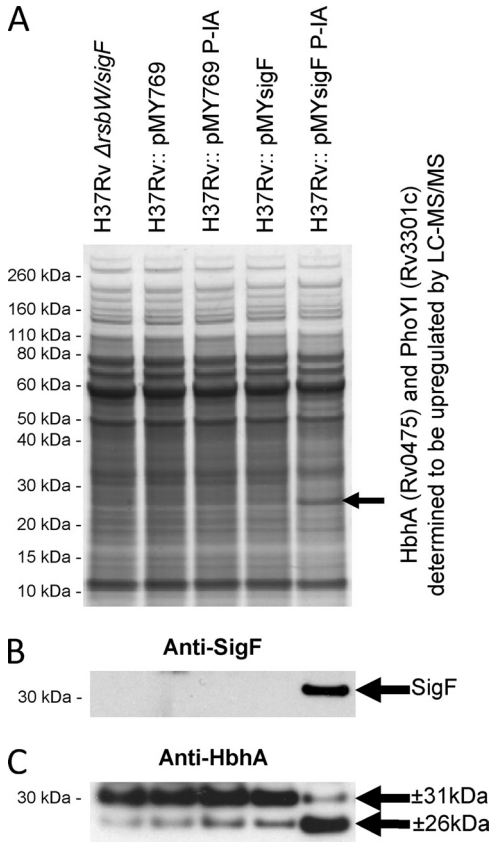


FIG 1 Protein expression profile of H37Rv(Δ*rsbW/sigF*), H37Rv::pMY769 (with or without P-IA), and H37Rv::pMYsigF (with or without P-IA). Analysis of total protein expression by SDS-PAGE gel and gel staining by Simply Blue (A) reveals the appearance of a protein band at ~26 kDa, later to be identified as HbhA by LC-MS/MS analysis. By LC-MS/MS analysis PhoYI was also found to be upregulated in this band. Confirmation of SigF expression and induction in the same set of lysates was performed by Western blot analysis with anti-SigF antibodies (B), while the induction of HbhA was confirmed with anti-HbhA antibodies (C). Note that in the case of HbhA, induction of SigF seems to lead not only to the induction of an HbhA band at 26 kDa but also the disappearance of an HbhA band at 31 kDa.

Analysis of the total protein content of *sigF*-induced and uninduced strains using staining SDS-PAGE gels revealed induction of a single protein band with an apparent molecular mass of ~26 kDa that was too small to be SigF (Fig. 1A). LC-MS/MS analysis of this band identified two proteins that were clearly induced in the presence of P-IA, namely, the heparin-binding hemagglutinin HbhA (Rv0475) and the phosphate-transport system regulatory

protein PhoYI (Rv3301c). Western blot analysis with a polyclonal anti-HbhA primary antibody confirmed the induction of HbhA at an apparent mass of 26 kDa but also revealed a decrease in the intensity of a higher-molecular-mass form at ~31 kDa (Fig. 1C) that may be methylated, since this has already been described (25, 29). Western blot analysis was not performed to confirm the identity of PhoYI due to the lack of specific antibodies.

SigF-dependent gene expression. To understand SigF-mediated gene regulation, we first looked at the changes in the bacterial transcriptome after *sigF* induction by hybridizing cDNAs to microarrays. These experiments were performed using H37Rv::pMY769 to control for P-IA-induced transcripts and H37Rv::pMYsigF to analyze SigF-induced transcripts. P-IA alone was found to induce six genes, i.e., *Rv3197A* (*whiB7*), *Rv0341* (*iniB*), *Rv1258c*, *Rv1988*, *Rv2416c* (*eis*), and *Rv2725c* (*hflX*), all of which were not found to be preceded by a SigF ChIP-on-chip binding site. Five of these genes are known to be induced by antibiotics through WhiB7 (27) and were not considered when analyzing SigF-induced transcripts.

When we compared RNA transcripts from uninduced and induced H37Rv::pMYsigF, we found no genes to be significantly downregulated, whereas 51 loci (represented by 138 probes) were upregulated after *sigF* induction. Of the 51 loci, 25 represent operons containing 33 annotated genes, one of which was only annotated in CDC1551 (*MT1083.2*) (see Table S2 in the supplemental material). The remaining probes mapped to 20 intergenic regions and to 6 antisense transcripts (compared to the annotated H37Rv genome) (see Table S3 in the supplemental material). The highest induction of an RNA transcript detected by microarrays was 19-fold for *Rv2884* (coding for a probable transcriptional regulatory protein), followed by *Rv1358* (14-fold) and *Rv2268c* (10-fold). The induction of consecutive genes allowed some operons to be identified, namely, *Rv0941c-Rv0940c* (Fig. 2A), *Rv1823-Rv1825* (Fig. 2B), *Rv1843c-Rv1841c*, and *Rv2268c-Rv2267c*. Interestingly, there were a number of transcripts, in particular *Rv1358*, where the transcript seems to start within the annotated gene (Fig. 2C). In addition, two examples of antisense RNA transcripts, found to be highly expressed after *sigF* induction, are found in *Rv1459c* (19-fold) and *Rv0450* (8-fold) (Fig. 2D and E). Analysis of the functional categories to which the enriched genes belong (6) revealed no particular bias.

ChIP-on-chip analysis. To understand which of the genes found to be upregulated after SigF induction were indeed controlled directly by SigF, rather than by a SigF-induced transcriptional regulator, ChIP-on-chip analysis was performed by comparing immunoprecipitated DNA from induced strains H37Rv (Δ*rsbW/sigF*)::pMY769 and H37Rv(Δ*rsbW/sigF*)::pMYsigF. Analysis of data from four independent biological replicates revealed 317 probes that were enriched in at least three cases. Of these probes, 116 were "singletons" and therefore discarded, leaving 201 significant probes that clustered in 67 loci. Analysis of ChIP-on-chip data showed that SigF-binding sites were evenly distributed throughout the genome. Of these 67 loci, 40 were intergenic, while 27 were intragenic (see Table S4 in the supplemental material). The biggest peak found by ChIP-on-chip analysis, with an enrichment of 81-fold, was between the two converging genes *Rv0243* (*fadA2*) and *Rv0244c* (*fadE5*). Although no transcript was detected in this location by RNA microarrays, it is known to encode an sRNA named F6 (2).

Integrating SigF binding and transcription. On integration of the two data sets, 16 loci (representing 24 genes) were clearly identified where a SigF-binding site was followed by an RNA transcript in the sense orientation, while nine regions were detected where a SigF-binding site was associated with an antisense transcript (Table 1). A Venn diagram showing the overlap between the two data sets is presented in Fig. 3.

Of the 51 identified loci whose expression was upregulated upon SigF induction, 26 were found not to have a corresponding 5′ SigF-binding peak. Possible reasons for this include their indirect regulation by other transcriptional regulators, such as Rv2884, or, alternatively, the stabilization of RNA by sRNA. More than 60% of the SigF-binding sites are not linked to genes displaying increased transcription. Reasons for this include additional levels of gene regulation or the existence of short genes or sRNAs, such as F6, which escape detection by the methods used.

Confirmation by Q-PCR. To obtain independent confirmation data of the SigF transcriptome and binding analysis, a number of genes were selected and subjected to Q-PCR analysis. In addition, a few loci were chosen that were previously reported to be controlled by SigF (e.g., *sigC*) but were not found here, as well as the F6 sRNA, for which no probe exists on the microarray. Expression profiles of H37Rv(Δ*rsbW/sigF*)::pMY769 and H37Rv (Δ*rsbW/sigF*)::pMYsigF were tested 0, 1, and 3 days after P-IA-mediated SigF induction. The data revealed that at day 0 there was no significant difference in the expression of any of the genes between the two strains (see Table S5 in the supplemental material). Likewise, after one or 3 days induction of H37Rv(Δ*rsbW/sigF*)::pMY769 no difference was seen for any of the genes tested (see Table S5 in the supplemental material). One day after the induction of H37Rv(Δ*rsbW/sigF*)::pMYsigF, significantly induced expression of all of the selected targets was seen except for *Rv1870c* and *sigC* (see Table S5 in the supplemental material), whereas after 3 days of induction all of the genes except *sigC* were upregulated (Fig. 4). At 3 days after SigF induction, the five most induced genes were *Rv1284* (194-fold), *Rv0941c* (142-fold), *Rv1358* (50-fold), F6 sRNA (35-fold), and *Rv2884* (34-fold).

Localizing promoters by 5′-RACE. Since genes can be expressed from multiple promoters, care needs to be taken not to confuse the SigF-specific transcriptional start point (TSP) with the TSP of another sigma factor, such as SigA. Consequently, comparisons were made of RNA extracted from cultures of H37Rv (Δ*rsbW/sigF*)::pMY769 and H37Rv(Δ*rsbW/sigF*)::pMYsigF at 3 days after induction. Multiplex 5′-RACE was performed for all 14 loci found to be regulated by SigF by Q-PCR, as well as for *rsbW*. The 5′-RACE analysis was unsuccessful in four cases (*Rv3254*, *Rv2268*, *Rv2140c*, and F6 sRNA) since the sequences obtained did not map to the corresponding genes. For 7 of the remaining 11 genes, a SigF-independent TSP was identified in H37Rv(Δ*rsbW/sigF*)::pMY769, thus implying the existence of a second promoter (see Table S6 in the supplemental material). The remaining four genes may have SigF-independent TSPs far from the priming site (as is true for *Rv1358*) or only be expressed from SigF-dependent promoters. 5′-RACE sequences obtained from H37Rv(Δ*rsbW/sigF*)::pMYsigF RNA revealed both a SigF-independent TSP (where there was one) and a SigF-specific TSP for all 11 genes. In the case of *Rv1358* and *Rv1870c*, the SigF-specific TSP was located internally.

To identify the SigF promoters for the SigF targets validated above, 40-bp sequences upstream of the identified TSPs were in-

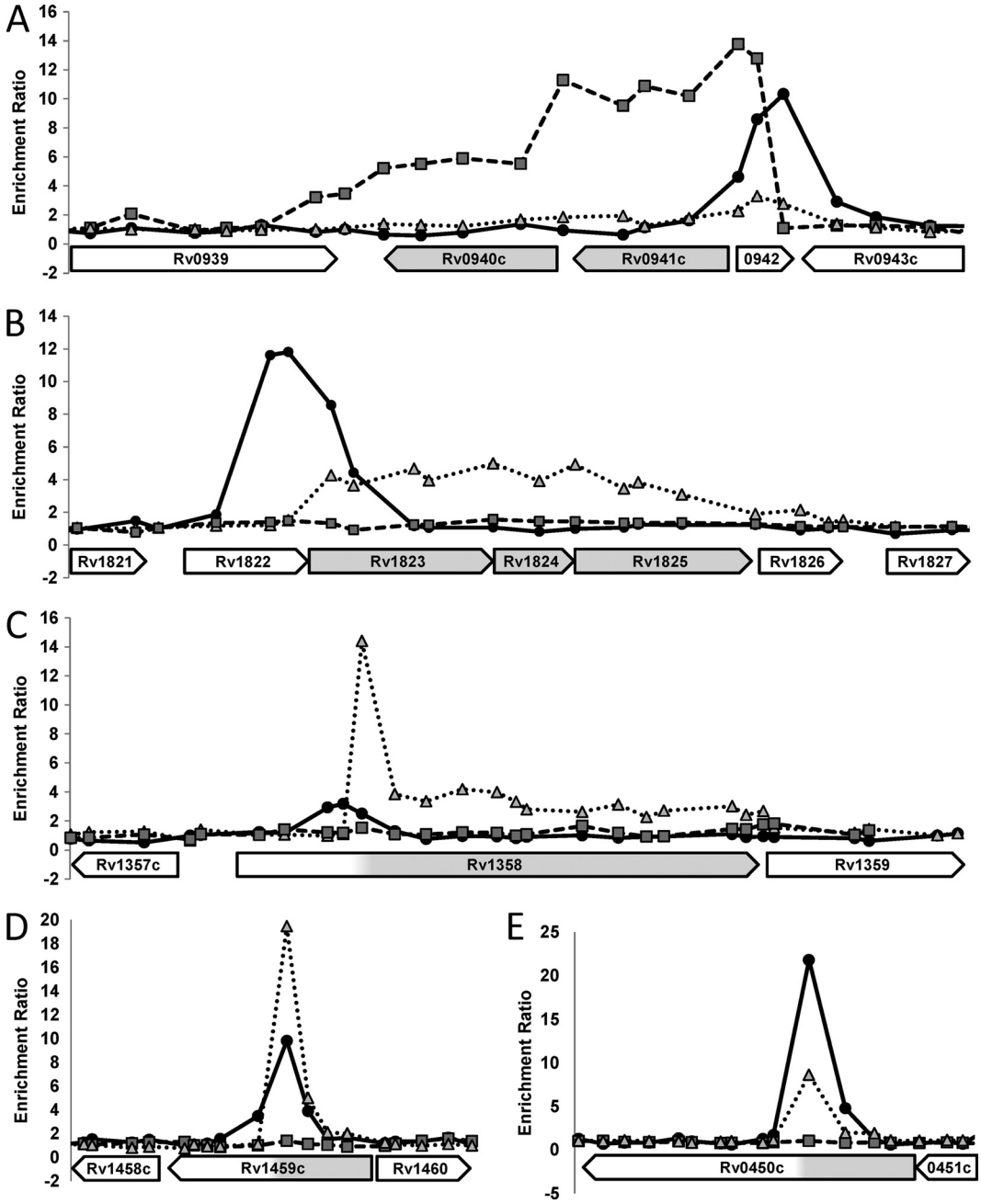


FIG 2 Representation of genomic regions identified by microarrays wherein a SigF-binding peak is detected, followed by an mRNA transcript. Examples include the binding of SigF prior to *Rv0941c-40c* and the transcription on the reverse strand (A) and the binding of SigF prior to *Rv1823-24-25* and the transcription of the forward strand (B). An example of a case where there is an alternative start of transcription is illustrated for *Rv1358*, where the SigF-binding peak is located in the middle of the gene, with subsequent transcription of a "truncated" mRNA on the sense strand (C). Also, two cases are shown (D and E) where there is SigF binding and subsequent RNA transcription in the antisense direction compared to the annotated genome. The data represent the mean of four biological replicates, and these values are plotted against the location of the probes on the chips. The SigF ChIP-on-chip profile is indicated by solid line with black circles, mRNA transcription in the reverse strand is indicated by dashed line with gray squares, and mRNA transcription in the forward strand is indicated by dotted line with gray triangles. Genes are annotated along the bottom, with predicted regulated regions colored in gray.

spected and, in all cases, the consensus sequence, GGTTT-$N_{(15\text{-}17)}$-GGGTA, was identified (Fig. 5).

DISCUSSION

We sought here to define the regulons of genes controlled by the alternative sigma factor SigF. Expression of SigF is low during the exponential growth phase, and the protein is likely to be largely inactive due to its binding to the anti-sigma factor RsbW, which is encoded by the same operon. Furthermore, there are no known stimuli that specifically increase the level of SigF. After considering these factors, we decided that in order to best study its role, *sigF* expression should be artificially induced by using a tightly regu-

TABLE 1 Summary of all SigF-induced RNA loci that are preceded by a SigF binding site, as determined by microarrays[a]

Locus	Strand annotation in *M. tuberculosis* genome	RNA			ChIP-on-chip	
		Orientation according to microarray	No. of probes	Fold enrichment	No. of probes	Fold enrichment
Rv0450c (*mmpl4*)	Forward	Antisense	1	8.6	2	21.8
Rv0475 (*hbhA*)	Forward	Sense	3	4.6	5	6.2
Rv0507 (*mmpl2*)	Reverse	Antisense	3	3.1	2	7.5
Rv0759c	Reverse	Sense	4	2.6	3	17.4
Rv0941c-40c	Reverse	Sense	12	3.2	4	10.3
MT1083.2	Forward	Sense	2	4.5	2	15.7
Rv1206 (*fadD6*)	Forward	Sense	1	3.4	2	3.5
Rv1284 (*canA*)	Forward	Sense	3	5.7	4	22.3
Rv1358[b]	Forward	Sense	12	14.4	3	3.2
Rv1397c (*vapC10*)	Forward	Antisense	1	4.0	6	11.9
Rv1455	Reverse	Antisense	1	7.9	2	4.3
Rv1459c	Forward	Antisense	2	19.4	3	9.8
Rv1823-25	Forward	Sense	10	4.3	4	11.8
Rv1843c-41c (*guaB1*)	Reverse	Sense	2	2.4	4	21.6
Rv1870c-69c	Reverse	Sense	3	2.3	2	23.0
Rv2140c (TB18.6)	Reverse	Sense	5	2.8	2	4.8
Rv2268c-67c (*cyp128*)	Reverse	Sense	2	10.6	2	4.6
Rv2722	Reverse	Antisense	1	3.8	3	16.2
Rv2884	Forward	Sense	4	19.3	4	61.9
Rv3049c	Forward	Antisense	1	3.7	2	4.4
Region between *Rv3210c* and *Rv3211*	Reverse	Antisense	1	8.1	3	5.0
Rv3254	Forward	Sense	4	2.5	2	3.2
Rv3287c-86c (*rsbW/sigF*)	Reverse	Sense	–[d]	–	2	32.0
Rv3301c (*phoYI*)	Reverse	Sense	3	3.0	1[c]	3.7
Rv3347c (*ppe55*)	Forward	Antisense	1	3.0	3	5.9

[a] The data for SigF-regulated RNA transcripts come from a comparison of the H37Rv::pMYsigF transcriptome with or without P-IA. The data for the SigF binding sites come from a comparison of immunoprecipitated DNA from H37Rv(Δ*rsbW/sigF*)::pMYsigF to H37Rv(Δ*rsbW/sigF*)::pMY769 (both cultures with P-IA). The data represent an analysis of four biological replicates.

[b] Transcription of this gene appears to start within the annotated gene (5′ truncated).

[c] For this ChIP-on-chip experiment, there was only one enriched probe.

[d] –, Induction of *rsbW/sigF* operon was observed, but the origin of the transcript (original genome or integrated pMYsigF vector) could not be deduced.

lated gene expression system. Consequently, we used the PipON system (13) because the Pptr promoter shows no leakiness in the absence of P-IA, as evidenced by our results.

At the protein level, which is relatively insensitive, only the heparin-binding hemagglutinin, HbhA (Rv0475), and PhoY1 were found to be upregulated following SigF overexpression. HbhA is a virulence factor that mediates extrapulmonary dissemination through binding to epithelial cells (26, 29). As expected, *hbhA* transcription was also upregulated upon SigF induction, and a SigF-binding site precedes its gene, as was also found in BCG (30). 5′-RACE analysis of the *hbhA* regulatory region revealed two promoters: a proximal promoter that was SigF-independent and a distal promoter that required SigF for expression. Curiously, two forms of the protein were found by Western blot analysis, with apparent molecular masses of 26 and 31 kDa. After SigF induction, there was an increase in abundance of the 26-kDa form but a decrease in the amount of 31-kDa HbhA. Extensive methylation of the heparin binding C-terminal domain has been reported to be important for its activity (17, 22), so SigF induction may lead to

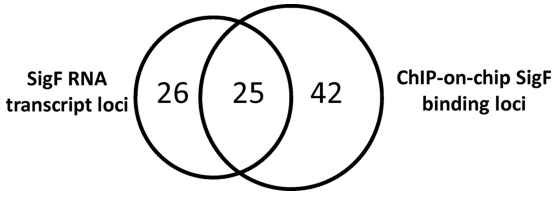


FIG 3 Venn diagrams illustrating the distribution of SigF-controlled loci, as identified by RNA microarrays and by ChIP-on-chip analysis.

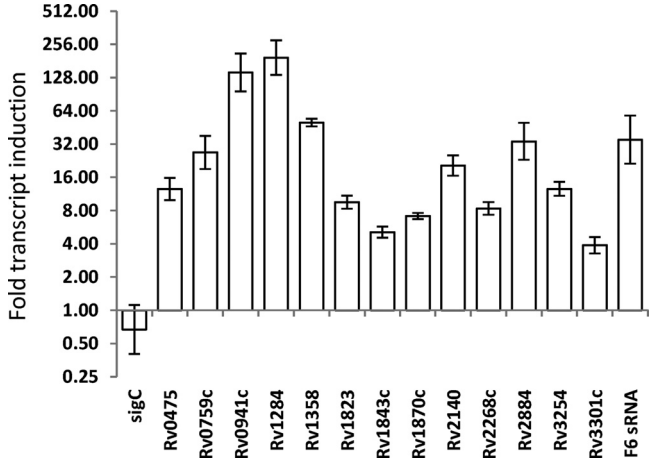


FIG 4 Q-PCR measurements for the fold induction of a number of genes in strain H37Rv(Δ*rsbw/sigF*)::pMYsigF after 3 days of induction with P-IA compared to no P-IA. The data represent the means and standard deviations of three biological replicates.

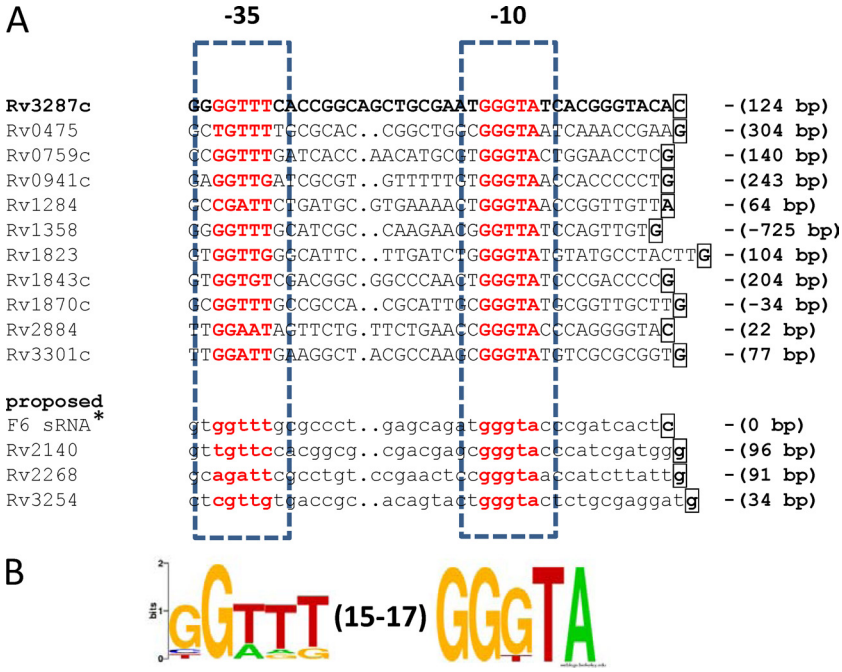


FIG 5 SigF-binding consensus sequence. (A) After 5′-RACE for the identification of the SigF-specific TSP (indicated by boxed base), a −10 and a −35 consensus (both shown in red) was identified using Bioprospector. The distance between the TSP and the annotated translation initiation codon is shown in brackets after the TSP. In addition, a proposed SigF binding site is given for the genes where 5′-RACE was not successful, including the TSP for F6 sRNA (*) described previously by Arnvig et al. (2). A graphical representation of the −10 and −35 SigF consensus sequence was generated using Weblogo (B).

accumulation of an unmethylated form of the protein and possibly also indirectly impact on the HbhA methylation status. It remains to be elucidated whether SigF, through the modulation of HbhA activity, is important *in vivo*, by playing a role in *M. tuberculosis* dissemination from the lungs.

Transcriptome analyses of SigF have previously been conducted in strains CDC1551 and H37Rv using various approaches but achieved very different results (14, 16, 20, 37). Our transcriptome study revealed 51 loci containing 33 protein-coding genes that were significantly upregulated on SigF induction, and yet the data set shows little overlap with the genes reported by Geiman et al. (1 of 99 genes) (14) or Lee et al. (1 of 5 genes) (20), some overlap (7 of 70 genes) with the genes identified by Williams et al. (*RsbW-SigF*, *Rv1824-Rv1825*, *Rv2884*, and *Rv3301c-Rv3300c*) (37), and good agreement with the genes found by Homerova et al. (using a two-plasmid system in *Escherichia coli*) (16). For the 60 genes that were suggested to be transcribed from SigF-dependent promoters by Williams et al. (37) but not found in the present study, only the regulatory region of *Rv0542c* is found to contain a SigF-binding site from our ChIP-on-chip data. Some of these discrepancies may be attributed to the use of different *M. tuberculosis* strains (CDC1551 versus H37Rv), to the RNA sample time points (12 h versus 3 days), and to the levels of *sigF* induction (acetamide versus P-IA). A particularly noteworthy difference is the negligible effect of SigF on the expression of *sigC*, in contrast to findings of a previous report (20). A full comparison of published SigF controlled transcripts in *M. tuberculosis* is shown in Table S7 in the supplemental material.

A strength of our study is the integration of two independent data sets: the SigF transcriptome with the SigF-binding sites on the genome defined by ChIP. Knowledge of the SigF-binding sites allows one to distinguish transcripts that are directly regulated by this sigma factor from those that are indirectly controlled through a regulatory cascade that may involve other transcriptional regulators such as Rv2884, whose gene is SigF dependent. Using the integrated approach we identified 16 loci where a SigF-binding site is followed by an induced transcript. These loci include previously published genes, such as *rsbW-sigF*, *MT1083.2*, *Rv1284*, *Rv1823-25*, *Rv2884*, and *Rv3301c* (16, 37). In addition, this work has identified 16 genes that have not previously been described to be under the control of SigF (Table 1). Furthermore, the regulation of the majority of these genes has also been confirmed by Q-PCR, as well as the identification of the SigF-specific TSP. 5′-RACE, followed by identification of a −10 sequence and −35 SigF-binding motif, resulted in the following consensus sequence: GGTTT-$N_{(15\text{-}17)}$-GGGTA located before all of the SigF-controlled genes, a motif very similar to the one described previously (3, 16, 20, 30, 32, 37). The majority of these genes controlled by SigF have no validated function, making it difficult to understand how they mediate a response to extracytoplasmic stress.

Since *sigF* has undergone pseudogenization in *M. leprae* and lost its function, it was of interest to examine whether any of the genes orthologous to those in the SigF regulon of *M. tuberculosis* remained active. *M. leprae* has undergone reductive evolution, and its genome has >1,300 pseudogenes (8, 35), with many others lost through deletions. Of the 25 genes that could be compared, 5 were still intact, 14 were present as pseudogenes, and 6 others may have been deleted. Interestingly, the genes that appear to have retained their function in *M. leprae*, such as *hbhA*, are those expressed from promoters recognized by SigF and another sigma factor.

A notable finding was the location of the biggest ChIP-on-chip

peak between two converging genes, *fadA2* and *fadE5* (*Rv0243* and *Rv0244c*), and this corresponds to the location of the gene for sRNA F6 (2), which is also known as *mcr14* (11). Q-PCR confirmed that this sRNA was indeed upregulated after SigF induction. Interestingly, overexpression of F6 has been reported to slow the growth of *M. tuberculosis* (2), thereby mirroring what we observed here when SigF was induced.

Our study revealed some regions where SigF-binding sites are not associated with RNA transcripts, and these were either intergenic or in the antisense orientation based on the current genome annotation. These regions may contain small open reading frames that have escaped previous detection or, more likely, code for sRNA. To test the latter possibility, we inspected the list of known sRNAs in *M. tuberculosis* (1, 2, 11), but none was identified. In two of the nine cases (intergenic peaks between *Rv1396c* and *Rv1397c* and between *Rv3210c* and *Rv3211*), a SigF-binding site is also present in BCG (30), and in seven cases a −10 GGGTA motif was detected near the ChIP peak. It should also be noted that some of the identified SigF binding sites may be artifacts of *sigF* overexpression, following its induction, due to higher levels than those observed physiologically. More work is required to determine whether these regions do indeed code for conditionally expressed sRNA.

The presence of SigF-dependent promoters within the predicted coding sequences of two protein-coding genes, *Rv1358* and *Rv1870c*, is intriguing. For the latter, this is most likely due to misattribution of the initiation codon since the mRNA identified in our work begins with a GTG codon that probably encodes the formyl-methionine. This example illustrates the utility of transcriptome studies for the refinement of genome annotation. Gene *Rv1358* is predicted to encode a transcriptional regulator of 1,159 amino acids belonging to the LuxR family that is well conserved among mycobacteria, including *M. leprae*. Translation of the SigF-dependent transcript of *Rv1358* would give rise to a protein lacking some 250 residues from its N-terminal domain. Whether two forms of the Rv1358 protein exist is currently unknown. In conclusion, our study has helped in defining the SigF regulon and characterizing its components.

ACKNOWLEDGMENTS

We thank Ida Rosenkrands (Statens Serum Institut, Copenhagen, Denmark) and Giovanni Delogu (Catholic University of the Sacred Heart, Rome, Italy) for kindly providing antibodies.

R.C.H. was the recipient of a postdoctoral fellowship from the Heiser Program for Research in Leprosy and Tuberculosis of the New York Community Trust. This study was supported in part by SystemsX.ch.

REFERENCES

1. **Arnvig KB, et al.** 2011. Sequence-based analysis uncovers an abundance of non-coding RNA in the total transcriptome of *Mycobacterium tuberculosis*. PLoS Pathog. **7**:e1002342.
2. **Arnvig KB, Young DB.** 2009. Identification of small RNAs in *Mycobacterium tuberculosis*. Mol. Microbiol. **73**:397–408.
3. **Beaucher J, et al.** 2002. Novel *Mycobacterium tuberculosis* anti-sigma factor antagonists control σF activity by distinct mechanisms. Mol. Microbiol. **45**:1527–1540.
4. **Betts JC, Lukey PT, Robb LC, McAdam RA, Duncan K.** 2002. Evaluation of a nutrient starvation model of *Mycobacterium tuberculosis* persistence by gene and protein expression profiling. Mol. Microbiol. **43**:717–731.
5. **Chen P, Ruiz RE, Li Q, Silver RF, Bishai WR.** 2000. Construction and characterization of a *Mycobacterium tuberculosis* mutant lacking the alternate sigma factor gene, *sigF*. Infect. Immun. **68**:5575–5580.
6. **Cole ST, BG Barrell.** 1998. Analysis of the genome of *Mycobacterium tuberculosis* H37Rv. Novartis Found. Symp. **217**:160–177.
7. **Cole ST, et al.** 1998. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature **393**:537–544.
8. **Cole ST, Supply P, Honore N.** 2001. Repetitive sequences in *Mycobacterium leprae* and their impact on genome plasticity. Lepr. Rev. **72**:449–461.
9. **Crooks GE, Hon G, Chandonia JM, Brenner SE.** 2004. WebLogo: a sequence logo generator. Genome Res. **14**:1188–1190.
10. **DeMaio J, Zhang Y, Ko C, Young DB, Bishai WR.** 1996. A stationary-phase stress-response sigma factor from *Mycobacterium tuberculosis*. Proc. Natl. Acad. Sci. U. S. A. **93**:2790–2794.
11. **DiChiara JM, et al.** 2010. Multiple small RNAs identified in *Mycobacterium bovis* BCG are also expressed in *Mycobacterium tuberculosis* and *Mycobacterium smegmatis*. Nucleic Acids Res. **38**:4067–4078.
12. **Edgar R, Domrachev M, Lash AE.** 2002. Gene Expr. Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. **30**:207–210.
13. **Forti F, Crosta A, Ghisotti D.** 2009. Pristinamycin-inducible gene regulation in mycobacteria. J. Biotechnol. **140**:270–277.
14. **Geiman DE, et al.** 2004. Attenuation of late-stage disease in mice infected by the *Mycobacterium tuberculosis* mutant lacking the SigF alternate sigma factor and identification of SigF-dependent genes by microarray analysis. Infect. Immun. **72**:1733–1745.
15. **Hartkoorn RC, et al.** 2010. Sigma factor F does not prevent rifampin inhibition of RNA polymerase or cause rifampin tolerance in *Mycobacterium tuberculosis*. J. Bacteriol. **192**:5472–5479.
16. **Homerova D, Surdova K, Mikusova K, Kormanec J.** 2007. Identification of promoters recognized by RNA polymerase containing *Mycobacterium tuberculosis* stress-response sigma factor sigma(F). Arch. Microbiol. **187**:185–197.
17. **Host H, Drobecq H, Locht C, Menozzi FD.** 2007. Enzymatic methylation of the *Mycobacterium tuberculosis* heparin-binding haemagglutinin. FEMS Microbiol. Lett. **268**:144–150.
18. **Jacques JF, Rodrigue S, Brzezinski R, Gaudreau L.** 2006. A recombinant *Mycobacterium tuberculosis* in vitro transcription system. FEMS Microbiol. Lett. **255**:140–147.
19. **Karls RK, Guarner J, McMurray DN, Birkness KA, Quinn FD.** 2006. Examination of *Mycobacterium tuberculosis* sigma factor mutants using low-dose aerosol infection of guinea pigs suggests a role for SigC in pathogenesis. Microbiology **152**:1591–1600.
20. **Lee JH, Karakousis PC, Bishai WR.** 2008. Roles of SigB and SigF in the *Mycobacterium tuberculosis* sigma factor network. J. Bacteriol. **190**:699–707.
21. **Liu X, Brutlag DL, Liu JS.** 2001. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of coexpressed genes. Pac Symp. Biocomput. **2001**:127–138.
22. **Locht C, Hougardy JM, Rouanet C, Place S, Mascart F.** 2006. Heparin-binding hemagglutinin, from an extrapulmonary dissemination factor to a powerful diagnostic and protective antigen against tuberculosis. Tuberculosis (Edinb.) **86**:303–309.
23. **Manganelli R, Dubnau E, Tyagi S, Kramer FR, Smith I.** 1999. Differential expression of 10 sigma factor genes in *Mycobacterium tuberculosis*. Mol. Microbiol. **31**:715–724.
24. **Manganelli R, et al.** 2004. Sigma factors and global gene regulation in *Mycobacterium tuberculosis*. J. Bacteriol. **186**:895–902.
25. **Menozzi FD, Bischoff R, Fort E, Brennan MJ, Locht C.** 1998. Molecular characterization of the mycobacterial heparin-binding hemagglutinin, a mycobacterial adhesin. Proc. Natl. Acad. Sci. U. S. A. **95**:12625–12630.
26. **Menozzi FD, et al.** 1996. Identification of a heparin-binding hemagglutinin present in mycobacteria. J. Exp. Med. **184**:993–1001.
27. **Morris RP, et al.** 2005. Ancestral antibiotic resistance in *Mycobacterium tuberculosis*. Proc. Natl. Acad. Sci. U. S. A. **102**:12200–12205.
28. **Parish T, Stoker NG, and SpringerLink.** 2001. *Mycobacterium tuberculosis* protocols and methods in molecular medicine. Humana Press, Inc., Totowa, NJ.
29. **Pethe K, et al.** 2001. The heparin-binding haemagglutinin of *Mycobacterium tuberculosis* is required for extrapulmonary dissemination. Nature **412**:190–194.
30. **Rodrigue S, et al.** 2007. Identification of mycobacterial sigma factor binding sites by chromatin immunoprecipitation assays. J. Bacteriol. **189**:1505–1513.
31. **Rodrigue S, Provvedi R, Jacques PE, Gaudreau L, Manganelli R.** 2006.

The sigma factors of *Mycobacterium tuberculosis*. FEMS Microbiol. Rev. **30**:926–941.
32. **Sachdeva P, Misra R, Tyagi AK, Singh Y.** 2010. The sigma factors of *Mycobacterium tuberculosis*: regulation of the regulators. FEBS J. **277**:605–626.
33. **Sala C, et al.** 2009. Genome-wide regulon and crystal structure of BlaI (Rv1846c) from *Mycobacterium tuberculosis*. Mol. Microbiol. **71**:1102–1116.
34. **Singh AK, Singh BN.** 2008. Conservation of sigma F in mycobacteria and its expression in *Mycobacterium smegmatis*. Curr. Microbiol. **56**:574–580.
35. **Singh P, Cole ST.** 2011. *Mycobacterium leprae*: genes, pseudogenes and genetic diversity. Future Microbiol. **6**:57–71.
36. **Tekaia F, et al.** 1999. Analysis of the proteome of *Mycobacterium tuberculosis* in silico. Tuberc. Lung Dis. **79**:329–342.
37. **Williams EP, Lee JH, Bishai WR, Colantuoni C, Karakousis PC.** 2007. *Mycobacterium tuberculosis* SigF regulates genes encoding cell wall-associated proteins and directly regulates the transcriptional regulatory gene *phoY1*. J. Bacteriol. **189**:4234–4242.

4.2.3 Virulence regulator EspR of *M. tb* is a nucleoid-associated protein

Contribution: Preliminary processing of the ChIP-seq data, qualitative analysis of the ChIP-seq peaks, and visualization and mapping in the context of the M. tb *H37Rv genome annotation.*

The ESX-1 system is the most studied type VII secretion system in *M. tb*. Located in the region of difference 1 (RD1), loss of ESX-1 genes is a major cause of attenuation of the vaccine strain *M. bovis* BCG (Mahairas *et al.* 1996; Lewis *et al.* 2003), and thus the principal virulence determinant in *M. tb*. ESX-1 secretion is required for early bacterial replication and virulence in mice (Siméone *et al.* 2009). Considering the role of ESX-1 in *M. tb* pathogenesis, much effort has been devoted to studying the composition, regulation and function of the ESX-1 secretion system. In addition to the genes encoded in the ESX-1 locus, an unlinked gene cluster, *espA-espC-espD* is essential for ESX-1 dependent protein secretion (Fortune *et al.* 2005; MacGurn *et al.* 2005).

EspR is a DNA-binding protein that positively regulates the *espA-espC-espD* cluster, by controlling expression of EspA protein (Raghavan *et al.* 2008). Recent structural studies performed in our laboratory suggest that EspR is a nucleoid-associated protein (NAP) (Blasco *et al.* 2011). NAPs are the bacterial equivalent of histones that organize the chromosome and act by stabilizing long-range structures in the genome through cooperative binding to multiple sites (Dillon & Dorman 2010). EspR employs an atypical DNA-recognition mechanism involving a dimer of dimers that can multimerize and recognize distal DNA binding sites in a cooperative manner (Blasco *et al.* 2011).

In order to confirm the role of EspR as a NAP and gain a better understanding into the extent of EspR-mediated regulation in *M. tb* we performed ChIP-seq (chromatin immunoprecipitation followed by high-throughput sequencing) analysis of EspR. Analysis and integration of data from two independent ChIP-seq experiments revealed 165 loci harboring 582 EspR binding peaks that were significantly enriched ($>$ 1.5 fold) compared to the control experiments. Mapping the peak locations to the genome annotation revealed that EspR binding was not restricted to promoter regions, but distributed almost evenly between genes (55%) and intergenic regions (45%). Along with EspA, the EspR regulon included EspR itself, the ESX-2 and ESX-5 systems, a host of diverse cell wall functions, such as production of the complex lipid PDIM (phenolthiocerol dimycocerosate) and the PE/PPE cell-surface proteins. A subset of the EspR-binding sites was validated experimentally and a binding motif, TTTGC[TC][GA] was deduced for EspR.

The ChIP-seq findings confirmed that EspR acts globally as a NAP and plays architectural and regulatory roles that impact *M. tb* pathogenesis through multiple genes.

OPEN ACCESS Freely available online

PLoS PATHOGENS

Virulence Regulator EspR of *Mycobacterium tuberculosis* Is a Nucleoid-Associated Protein

Benjamin Blasco[1], Jeffrey M. Chen[1], Ruben Hartkoorn[1], Claudia Sala[1], Swapna Uplekar[1,2], Jacques Rougemont[1,2], Florence Pojer[1], Stewart T. Cole[1]*

1 Global Health Institute, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland, **2** Swiss Institute of Bioinformatics, Lausanne, Switzerland

Abstract

The principal virulence determinant of *Mycobacterium tuberculosis* (*Mtb*), the ESX-1 protein secretion system, is positively controlled at the transcriptional level by EspR. Depletion of EspR reportedly affects a small number of genes, both positively or negatively, including a key ESX-1 component, the *espACD* operon. EspR is also thought to be an ESX-1 substrate. Using EspR-specific antibodies in ChIP-Seq experiments (chromatin immunoprecipitation followed by ultra-high throughput DNA sequencing) we show that EspR binds to at least 165 loci on the *Mtb* genome. Included in the EspR regulon are genes encoding not only EspA, but also EspR itself, the ESX-2 and ESX-5 systems, a host of diverse cell wall functions, such as production of the complex lipid PDIM (phenolthiocerol dimycocerosate) and the PE/PPE cell-surface proteins. EspR binding sites are not restricted to promoter regions and can be clustered. This suggests that rather than functioning as a classical regulatory protein EspR acts globally as a nucleoid-associated protein capable of long-range interactions consistent with a recently established structural model. EspR expression was shown to be growth phase-dependent, peaking in the stationary phase. Overexpression in *Mtb* strain H37Rv revealed that EspR influences target gene expression both positively or negatively leading to growth arrest. At no stage was EspR secreted into the culture filtrate. Thus, rather than serving as a specific activator of a virulence locus, EspR is a novel nucleoid-associated protein, with both architectural and regulatory roles, that impacts cell wall functions and pathogenesis through multiple genes.

Citation: Blasco B, Chen JM, Hartkoorn R, Sala C, Uplekar S, et al. (2012) Virulence Regulator EspR of *Mycobacterium tuberculosis* Is a Nucleoid-Associated Protein. PLoS Pathog 8(3): e1002621. doi:10.1371/journal.ppat.1002621

Editor: Eric J. Rubin, Harvard School of Public Health, United States of America

Received November 8, 2011; **Accepted** February 21, 2012; **Published** March 29, 2012

Funding: The research leading to these results has received funding from the European Community's Seventh Framework Programme ([FP7/2007-2013]) under grant agreement n°201762, SystemsX.ch and the Swiss National Science Foundation under grant n°31003A-125061. F.P. is a Swiss National Science Foundation MHV Post-doctoral Fellow. J.M.C. is a recipient of Canadian Thoracic Society and Canadian Institutes of Health Research Post-doctoral Fellowships. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: stewart.cole@epfl.ch

Introduction

Details of the genetic control mechanisms governing the pathogenicity of the etiological agent of human tuberculosis are starting to emerge [1]. It has been postulated that the DNA-binding protein EspR [2] controls the virulence of *Mycobacterium tuberculosis* (*Mtb*) by specifically regulating expression of EspA, an exported protein [3,4], which is required for the ESX-1 system to function normally. There are five genetically related ESX-systems in *Mtb* but functional information is scarce for all of them, although ESX-1 is by far the most studied [5,6]. ESX-1 is widely considered to be the principal virulence determinant of *Mtb* since it secretes the EsxAB (ESAT-6 and CFP10) proteins and ESX-1 secretion-associated proteins (Esps) [7]. Although mechanistic details are limited, some of these secreted proteins act as effector proteins that perturb host cell activities, permeabilize the phagosomal membrane and allow the tubercle bacillus to escape into the cytoplasm [6,8,9].

Structural studies revealed that EspR is a homodimer with two domains: an N-terminal DNA-binding domain with a helix-turn-helix (hth) motif and C-terminal domain that mediates dimerization [10,11]. Removal of 10 amino-acid residues from the C-terminus, as in the EspRΔ10 protein, does not affect DNA-binding activity but prevents dimerization and ablates activation of the *espACD* locus [2,10]. A model has been proposed, based on the results of co-crystallization with DNA and molecular dynamic simulations, wherein EspR employs an atypical DNA-recognition mechanism involving a dimer of dimers. Since, for sterical reasons, only one hth from each dimer is capable of inserting into the major groove of DNA at a given binding site, the second hth of each dimer remains free to act at other binding sites [10]. Consequently, dimers can dimerize then multimerize and recognize distal DNA binding sites in a cooperative manner as has been observed by atomic force microscopy (AFM) of EspR-nucleoprotein complexes at the *espA* locus, where DNA bending and bridging resulted in loop formation [10]. This behavior is characteristic of nucleoid-associated proteins (NAPs) rather than that of a classical gene activator protein [12,13]. NAPs are the bacterial equivalent of histones that organize the chromosome and act by stabilizing long-range structures in the genome through cooperative binding to multiple sites. This results in modulation of the accessibility of free DNA to the transcription machinery via the state of DNA compaction. NAPs thus function in a quite different manner to transcriptional activators, which typically recognize a limited number of sites in the genome and promote transcription through direct interaction with RNA polymerase.

To test the hypothesis that EspR might be a NAP and to gain more insight into the extent of EspR-mediated regulation in living

Author Summary

A major infection mechanism employed by the causative agent of tuberculosis, *Mycobacterium tuberculosis (Mtb)*, is the ESX-1 secretion system. It has been postulated that the DNA-binding protein EspR controls the virulence of *Mtb* by specifically regulating expression of the exported EspA protein, which is required for ESX-1 to function. Previous structural studies indicated that EspR forms dimers capable of multimerizing on DNA and forming loop structures, thus bringing together otherwise distant chromosomal regions. Such characteristics are reminiscent of nucleoid-associated proteins (NAPs), the histone equivalent in bacteria. Here we use ChIP-Seq technology to map EspR binding sites on the *Mtb* chromosome in living bacterial cells. Genome-wide analysis of EspR identified hundreds of binding-sites, with almost equal inter- and intra-genic distribution, and mostly found in proximity to genes associated with cell wall function. We validated a subset of EspR-binding sites experimentally and identified a consensus motif required for optimal binding affinity. Moreover, our study reveals that EspR expression varies with bacterial growth and that intracellular levels are not linked to EspR secretion. These findings corroborate the NAP nature of EspR and its dual roles, architectural and regulatory, that impact the *Mtb* chromosome and pathogenesis globally rather than the ESX-1 loci specifically.

mycobacteria, ChIP-Seq analysis of the *Mtb* H37Rv genome was performed. Unlike cDNA hybridization to micro-arrays, ChIP-Seq generates quantitative information with single nucleotide resolution. The complete repertoire of EspR-binding sites across the chromosome was established with the majority of the gene targets appearing to impact cell wall function generally rather than ESX-1 expression in particular. The ChIP-Seq findings were confirmed independently by ChIP followed by quantitative real-time PCR (ChIP-qPCR), by gel shift and DNase I footprinting assays, as well as by transcript analysis, and a binding motif was deduced using bioinformatics. Finally, the expression levels of EspR and its subcellular localization were monitored and quantified during the growth cycle. Altogether, these data provide a genome-wide view of EspR regulatory functions and place EspR firmly in the NAP family.

Results

Genome-wide analysis of the EspR regulon

We investigated EspR-binding to the chromosome of *Mtb* strain H37Rv during exponential growth by ChIP-Seq, chromatin immunoprecipitation followed by ultra-high throughput DNA sequencing [14]. Sequence reads obtained from two independent ChIP-Seq experiments using EspR-specific antibodies were mapped to the *Mtb* H37Rv genome. Based on the peak detection criteria, we identified 165 enriched loci harboring 582 EspR-binding peaks (Fig. 1, Table S2), that were enriched by >1.5-fold and these were not present in ChIP-Seq datasets from control experiments conducted without antibody or with unrelated antibodies (data not shown). These 165 loci occurred across the genome (Fig. 1 and Fig. 2A) and were sited both in intergenic regions (45%) and within genes (55%) implying that EspR is not a classical transcriptional regulator.

Diverse functions are encoded by genes where EspR bound upstream and classification by functional category reveals over-representation of cell wall/cell processes and the surface-exposed PE/PPE proteins (http://tuberculist.epfl.ch/; Fig. 2B). Internal sites were found within AT-rich genes encoding proteins belonging to the PPE family (Fig. 1), like *ppe24* (*rv1753c*), and some of these, such as *ppe58* (*rv3426*), also bind EspR at their 5′-ends. Binding sites were present within genes that are thought to have been acquired by horizontal transfer [15] like the *rv0986-rv0989c* region.

A survey of the ten top scoring peaks (Table 1) highlighted the major EspR-binding gene targets. Two of the top three sites (Fig. 1) occurred at a locus encoding an enzyme system that produces the complex lipids phthiocerol dimycocerosate (PDIM) and phenolic glycolipid (PGL) [16,17]. The second highest scoring site overlaps the translational start of *rv1490*, which encodes a membrane protein of unknown function, and this was followed by three other peaks of lower intensity, separated by ~300–400 bp, spread across *rv1490* (Fig. 1). The fourth and eighth highest scoring sites affect two genes, *pe-pgrs19* and *pe-pgrs20* [18], encoding mycobacteria-restricted PE_PGRS proteins, while the *espACD* locus, which is preceded by three EspR-binding sites (Fig. 1), occurred in the fifth position of the top ten ChIP-Seq hits (Table 1). The ninth peak is sited in the intergenic region between *lipF* (*rv3487c*), encoding a lipid esterase, and *rv3488*, whereas the last peak of the top ten ChIP-Seq list was found at the 3′-end of *fadB2* (*rv0468*) encoding a beta-hydroxybutyryl-CoA dehydrogenase and upstream of *umaA* (*rv0469*) coding for a mycolic acid synthase. EspR binds to multiple sites in the ESX-1, ESX-2 and ESX-5 loci (Fig. S2), as well as to two sites upstream of its own gene (Fig. 3), thus implying autogenous control. Taken together, these data suggest that EspR may be involved in regulating cell wall function.

Consensus sequence

An EspR consensus sequence was identified in the ChIP-Seq peaks using the MEME suite [19] and the binding motif deduced (Fig. 2C). FIMO (Find Individual Motif Occurrences) detected 736 occurrences of this motif (p-value≤0.001) distributed among 80% of the EspR peak sequences analyzed. Of these motifs, 59% were localized within open reading frames (ORF). The entire *Mtb* genome sequence was searched for potential EspR-binding sites using the TTTGC[TC][GA] consensus sequence and 199 putative intergenic and 827 intragenic sites were identified, of which 163 (43 intergenic and 120 intragenic) correspond to known ChIP-Seq peak sites. Further experimental support for the consensus sequence is available from footprinting studies of the *espA* [10] and *espR* promoter regions (see below), which revealed EspR protection from DNase I digestion at sites comprising at least one TTTGC-like motif.

Confirmation of *in vivo* EspR binding

To obtain independent confirmation for selected parts of the *in vivo* dataset, we initially focused on the EspR-dependent *espACD* operon [2]. Our previous *in vitro* work revealed two EspR binding sites separated by 19 bp and located between 506 and 444 bp upstream of *espACD* [10], consistent with the presence of a ChIP-Seq peak in this region (Fig. 1). On closer inspection, two additional major peaks of EspR-enrichment were found further upstream of *espA* (centered between −857 bp and −695 bp and between −1214 bp and −1113 bp, respectively). While this work was in progress, another report of the presence of two additional sites upstream of *espACD* appeared [11]. The existence of these sites also corroborates results we obtained previously using AFM to visualize nucleoprotein complexes of EspR and a 1360 bp *espACD* promoter fragment [10]. AFM revealed loop structures stabilized by multiple EspR dimer of dimers suggesting the presence of several distant EspR binding sites in the *espACD* upstream region. The 5′-end of the *espA* mRNA was located 66 bp upstream of the translation start codon using 5′ RACE (Fig. S3). Consequently, the

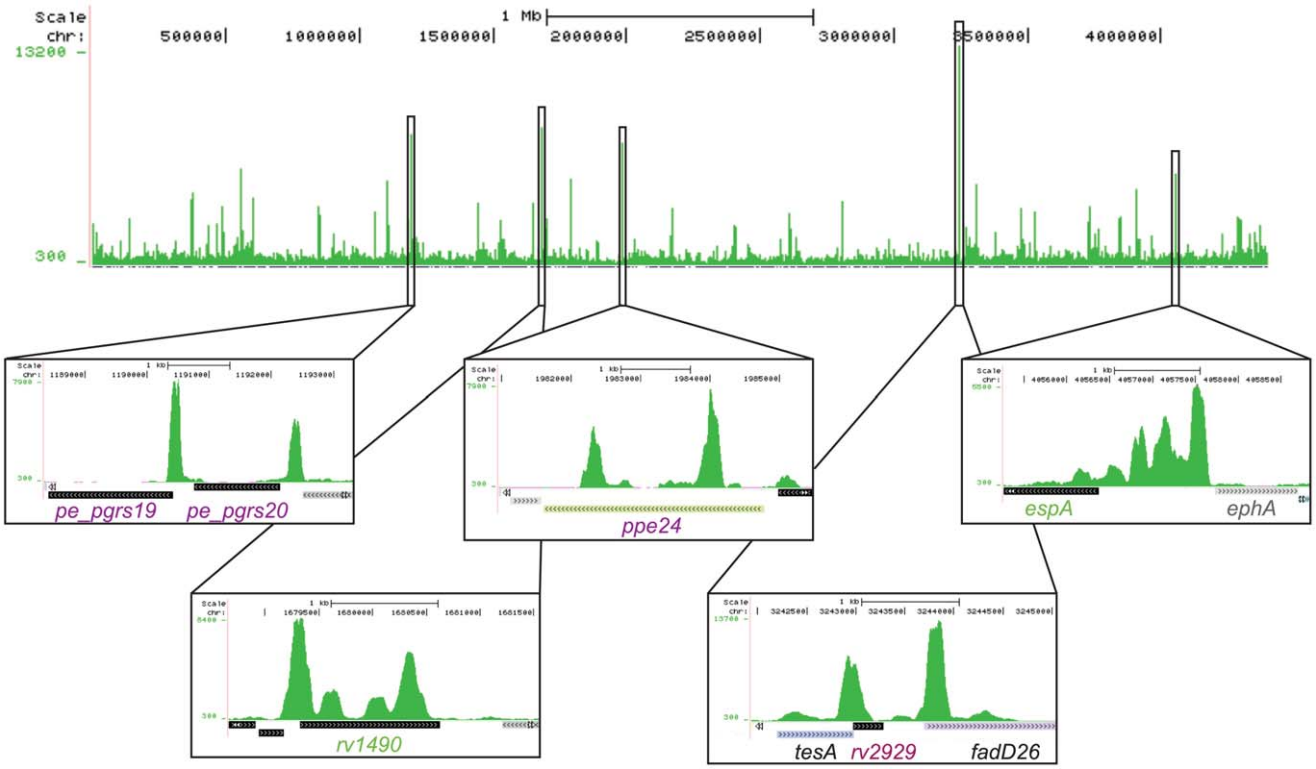


Figure 1. Genome-wide mapping of EspR binding sites. UCSC Genome Browser (http://genome.ucsc.edu) view of EspR binding across the *Mtb* genome as determined by ChIP-Seq. Peak height (y-axis) indicates the sequencing read depth at each genomic position (x-axis). Inset boxes show major EspR binding sites identified over (from left to right) the *pe_pgrs19* and *pe_pgrs20* genes, the *rv1490* gene, the *ppe24* gene, the *rv2929* and *fadD26* genes from the PDIM/PGL locus and the *espACD* operon promoter region.
doi:10.1371/journal.ppat.1002621.g001

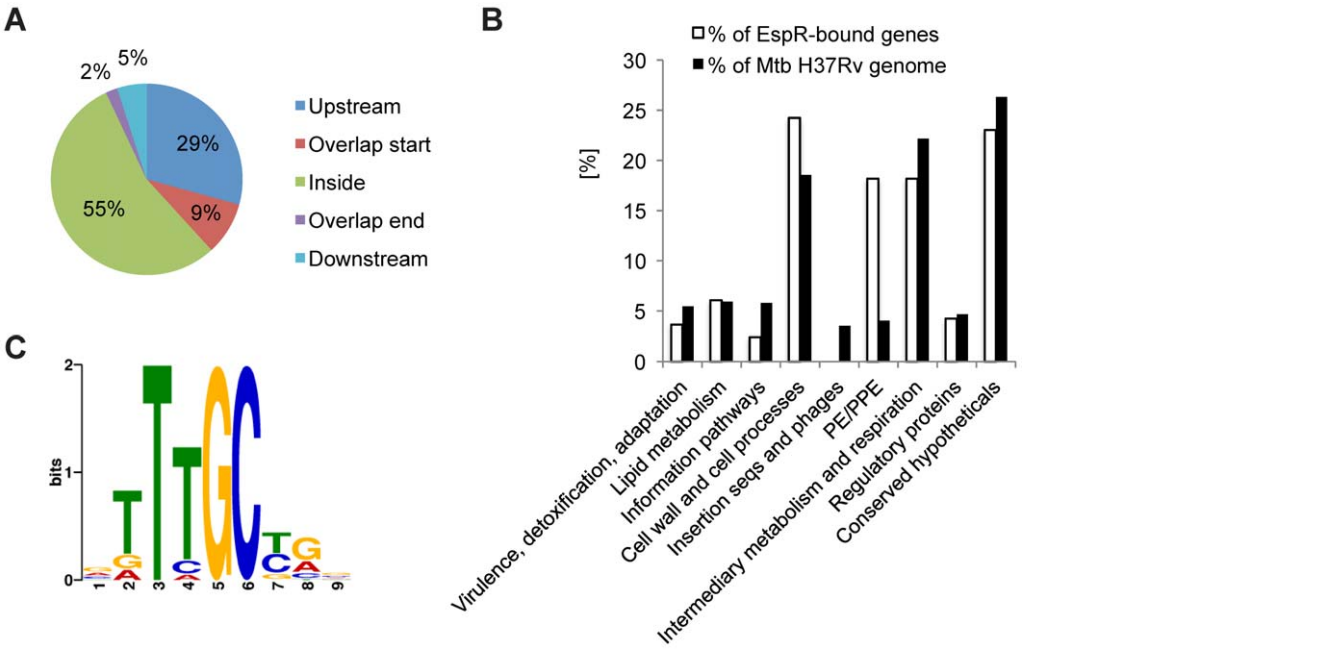


Figure 2. Binding of EspR to the *Mtb* chromosome. (**A**) Pie chart of EspR-binding peaks annotated relatively to the nearest gene translation start using the *ChIPpeakAnno* package [38]. (**B**) Bar chart displaying the percentage of genes of the EspR regulon compared to the *Mtb* H37Rv genome belonging to functional categories defined in the TubercuList database (http://tuberculist.epfl.ch/). (**C**) Most significant motif derived from ChIP-Seq binding sequences returned by the MEME tool [19]. The height of each letter represents the relative frequency of each base at different positions in the consensus.
doi:10.1371/journal.ppat.1002621.g002

Table 1. Top 10 EspR binding loci from ChIP-Seq.

Gene	Position	Gene description	Functional category
rv2930	inside	FadD26, fatty-acid-CoA synthetase, PDIM production	lipid metabolism
rv1490	overlaps start	Rv1490, probable membrane protein	cell wall and cell processes
rv2929	overlaps start	Rv2929, PDIM production	conserved hypotheticals
rv1067c	overlaps start	PE_PGRS family protein, PE_PGRS19	PE/PPE
rv3616c	upstream	EspA, ESX-1 secretion-associated protein A	cell wall and cell processes
rv1753c	inside	PPE family protein, PPE24	PE/PPE
rv0986	inside	ATP-binding protein of ABC transporter	cell wall and cell processes
rv1068c	upstream	PE-PGRS family protein, PE_PGRS20	PE/PPE
rv3487c	upstream	LipF, probable esterase/lipase	lipid metabolism
rv0469	upstream	UmaA, Mycolic methyltransferase A 1	lipid metabolism

doi:10.1371/journal.ppat.1002621.t001

nearest EspR binding site is positioned over 300 bp upstream of the promoter.

To further validate EspR-binding peaks, with varying degrees of enrichment, we performed ChIP followed by quantitative PCR on 11 selected sites (four located within intergenic regions, three within ORFs and three overlapping a translational start) and two non-peak regions (within *rv0888* and *sigA* ORF) as controls. All of the selected EspR-binding regions exhibited enrichment comparable to that observed from ChIP-Seq analysis (Fig. S4), thus confirming that all peaks were genuine EspR-targets.

To obtain further confirmation of the *in vivo* EspR binding sites, we performed electrophoretic mobility shift assays (EMSAs) using ~100 bp DNA sequences covering the top five binding sites (Fig. S5 and Table 1) and a DNA fragment of the same size from within the *espA* ORF as a negative control. EspR was shown to bind to all five sites in a concentration dependent manner, while the negative control fragment remained unbound at an equal protein concentration. However, clear differences in affinity between the fragments were visible. For example, the top-scoring *fadD26* peak bound EspR less strongly compared to the four others suggesting that other determinants, like long-range protein-protein or protein-DNA interactions, could contribute to the high-affinity binding observed *in vivo*.

Autogenous regulation at the *espR* promoter

The presence of twin peaks upstream of *espR* (**a**, **b** in Fig. 3A) is suggestive of autoregulation. To test this possibility, we performed EMSA, DNase I footprinting and 5′ RACE analysis of the *espR* promoter region. Peaks **a** and **b** were both shown to bind EspR using EMSA (Fig. 3B). Two regions within peak **a** were protected from DNase I by EspR; region I, covering 17 bp, and region II, 55 bp-long, are situated 101 and 76 bp upstream of the *espR* translational start codon, respectively (Fig. 3C and 3D). This binding pattern is reminiscent of that described previously at the *espA* promoter [10]. When incubated with the dimerization deficient EspRΔ10 protein, only part of region II and none of region I was protected ($IIa_{\Delta 10}$, 14 bp and $IIb_{\Delta 10}$, 12 bp, see Fig. S6). This implies that oligomerization enables cooperative binding between multiple EspR dimers, leading to the formation of higher-order oligomers. The zone protected by EspRΔ10 contains an inverted repeat of two consensus motifs: CAGCAAA<16> TTTGCTC.

5′ RACE analysis was employed to localize the *espR* promoter using RNA extracted from *Mtb* H37Rv grown to mid-log phase. The *espR* transcript starts with a poly-G (7) sequence 144 bp upstream of the translational start codon. The promoter is therefore situated in a region between peaks **a** and **b** so simultaneous occupation by EspR of both the **a** and **b** sites might form a repression loop. Expression data presented below indicate a negative effect of EspR on its own transcription.

Target gene regulation on EspR binding

To confirm the prediction that binding of EspR directly affects target gene expression, we exploited a pristinamycin-inducible system [20,21] to overexpress *espR* conditionally in *Mtb* (strain H37Rv::pMYespR; Fig. 4). Compared to the controls, it is noteworthy that *espR* over-expression significantly decreased growth after 24 h (Fig. 4A) while *espR* transcript and EspR protein levels were found to be ~8-fold and ~3-fold higher than in the control after 72 h, respectively (Figs. 4C and 4B). When the relative amounts of target transcripts in untreated and pristinamycin IA-treated H37Rv::pMYespR cells were measured by quantitative RT-PCR, significantly increased transcript levels were detected for *rv1490*, *espA*, and the ABC-transporter *rv0986* (Fig. 4D). Conversely, repression of *lipF* transcription was also observed upon EspR overexpression, whereas transcription of some target genes appeared unchanged (Fig. 4D). Using a discriminatory RT-PCR assay it was possible to measure the impact of EspR overproduction on expression of the chromosomal copy of *espR* and, again, this appeared to act negatively (Fig. 4C).

The combined findings suggest that EspR is capable of both positive and negative transcriptional regulation. Moreover, the inability to observe direct EspR-dependent regulation at some major EspR binding sites suggests that EspR has no or little effect on these genes in the conditions tested or that other regulators counter-balance the effect of increased EspR levels.

Growth phase dependent expression of EspR

It has been proposed that intracellular levels of EspR are regulated via its secretion by the ESX-1 system and that blocking EspR secretion results in enhanced EspR-mediated transcriptional effects [2]. This suggested that the intracellular requirements for EspR could change during the growth phase of *Mtb* since the secretion of other ESX-1 substrates, such as EsxA (ESAT-6), is known to occur early in the growth cycle. To determine whether levels were constant or variable during growth and to estimate the number of EspR molecules per cell, kinetic experiments were performed. We monitored EspR protein levels by quantitative Western blotting at different time points corresponding to the

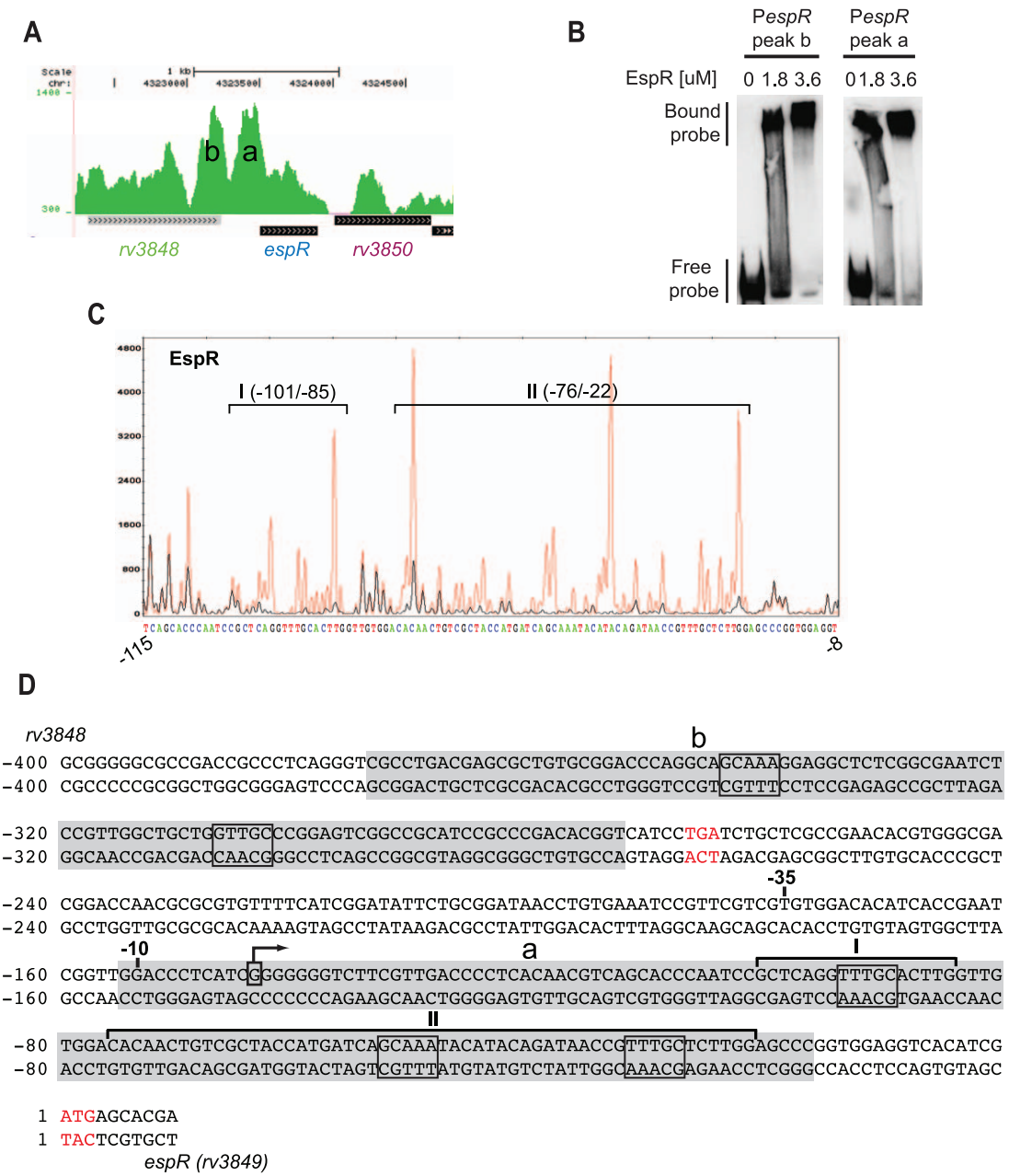


Figure 3. EspR autoregulation. (**A**) Pattern of EspR ChIP-Seq peaks (denominated peaks **a** and **b**) identified upstream of the *espR* start codon. (**B**) EMSA showing concentration-dependent EspR binding of DNA fragments covering the center sequences of peaks **a** and **b** shown in (A). (**C**) DNase I footprint from binding of EspR to peak **a** of *espR*. Red and black peaks represent DNA without or with 10 μM EspR, respectively. Both reactions were partially digested with DNase I and analysed by capillary electrophoresis in a genetic analyser (Applied Biosystems 3130xl). The corresponding nucleotide sequence is shown below. Regions I and II protected from DNase I digestion by EspR are denoted by black brackets and positions relative to the translational start are indicated. (**D**) Analysis of the *espR* promoter. The 5′ end of the transcript was determined by 5′ RACE using RNA extracted from *Mtb* H37Rv at mid-log phase. Translational stop of *rv3848* and translational start of *espR* are highlighted in red with +1 corresponding to the first base of the *espR* open reading frame. Transcriptional start mapped by 5′ RACE is boxed and indicated by a bent arrow. −10 and −35 positions for putative sigma factor binding sites are indicated. ChIP-Seq peak sequences are highlighted in gray, EspR consensus motifs framed and sites protected from DNase I cleavage bracketed as in (C).
doi:10.1371/journal.ppat.1002621.g003

early-log (day 2), mid-log (day 3) and stationary (days 4 and 5) phases of growth. Analysis of equivalent cell numbers showed that the intracellular concentration of EspR increases throughout the bacterial growth cycle, ranging from ~20,000 molecules at early log-phase (day 2) to ~100,000 molecules per cell at stationary phase (day 5) (Fig. 5A). Peak cell concentration of EspR in stationary phase is consistent with the growth arrest observed upon its premature induction (Fig. 4A). To determine if this peak was due to protein accumulation or to increased expression of the *espR* gene, we performed quantitative RT-PCR on RNA samples isolated from cells at different time points (Fig. 5B). Interestingly, throughout the growth cycle, the levels of *espR* mRNA varied in a manner inversely proportional to the amounts of EspR protein, suggesting that EspR stably accumulates in the bacteria while

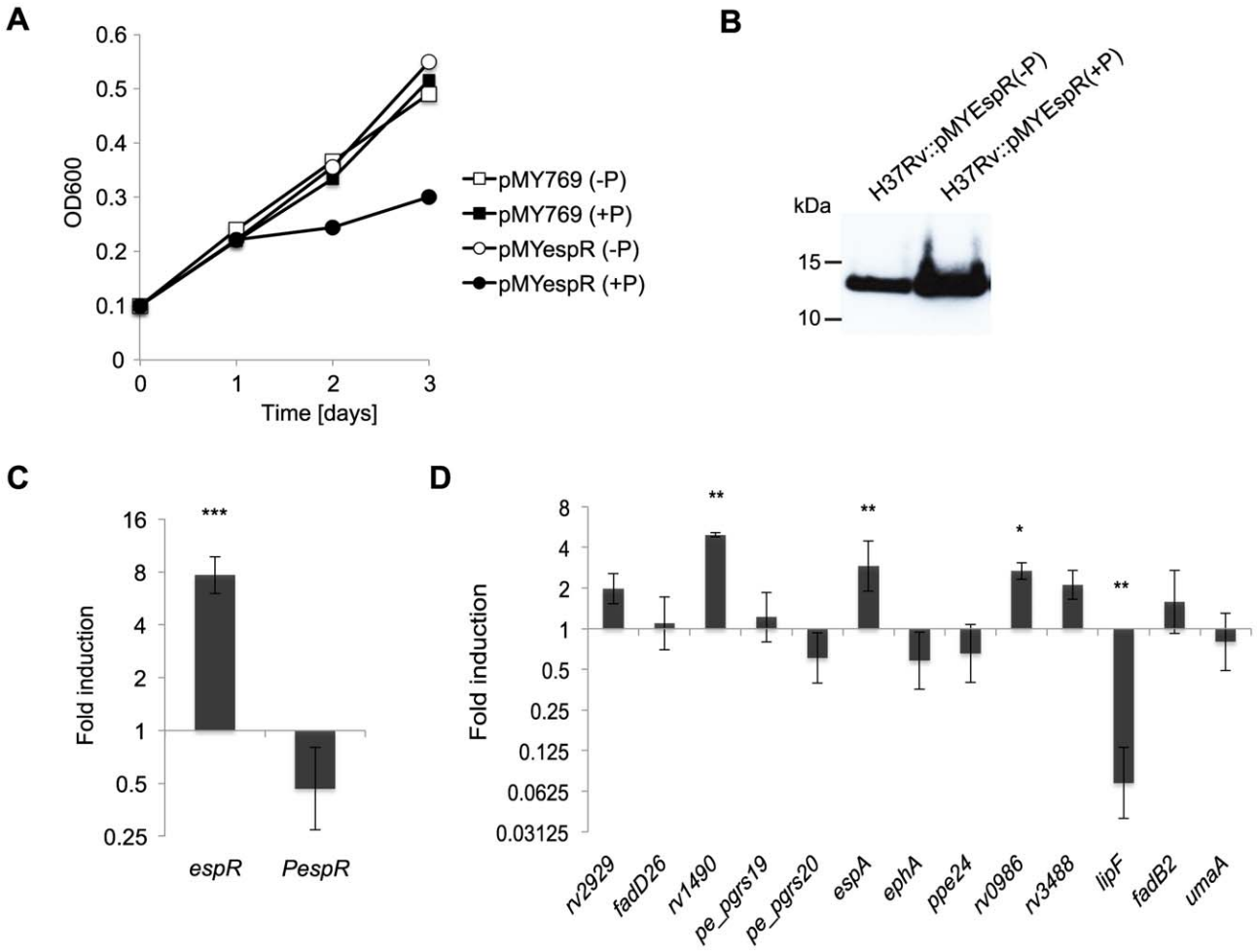


Figure 4. Gene expression associated with EspR binding. (**A**) Growth of H37Rv::pMYespR without (−P) or with (+P) pristinamycin IA. Growth was monitored at the designated time points measuring OD_{600}. (**B**) Immunoblot analysis of EspR expression in the absence (−P) or presence (+P) of pristinamycin IA after 3 days growth using rat polyclonal antibodies specific for EspR. Equivalent amounts of total protein lysates were loaded. (**C**, **D**) Quantitative RT-PCR analysis of relative mRNA levels extracted from H37Rv::pMYespR cells treated with 0 μg/ml or 2 μg/ml of pristinamycin using primers specific for the following regions: (**C**) the *espR* coding region (annealing to both endogenous and vector copies of *espR*) or to the 5′-UTR region of *espR* (*PespR*) (specific to endogenous *espR*). (**D**) The coding regions of genes related to the top 10 EspR binding peaks as obtained in ChIP-Seq experiments (see Table 1). Relative gene expression was normalized against *sigA* and displayed as fold-induction (log-2 scale) relative to the untreated sample (0 μg/ml pristinamycin). Shown are the mean ± s.d. of a minimum of duplicate measurements from the average of three independent experiments. Statistical significance was evaluated using Students T-test. * indicates $P<0.05$, ** $P<0.01$ and *** $P<0.001$.
doi:10.1371/journal.ppat.1002621.g004

autorepression may limit its own gene expression at late time points.

EspR is not secreted

To determine if the low intracellular levels of EspR observed at the early and mid-log phases of growth were due to intensive EspR secretion, we measured intra- and extra-cellular levels of EspR from strains *Mtb* H37Rv and *Mtb* H37RvΔRD1 cultured in Sauton's medium to mid-log phase. Under these conditions, we were unable to detect EspR among the culture filtrate (CF) proteins in either case, whereas EsxA was present in the CF of *Mtb* H37Rv, as expected, but not in CF from the ESX-1 mutant H37RvΔRD1 that lacks *esxA* among other genes (Fig. 6A).

To investigate whether EspR was exported from the cytosol but retained in the cell envelope, whole cell lysate (CL) was fractionated by ultracentrifugation into the cell wall/cell membrane (W/M) and cytosolic (CYT) components. Since the chromosome is known to be attached to the plasma membrane [22], half of the samples were treated with DNase I. EspR was detected in both of the untreated fractions but was mainly in the cytosol after DNase I treatment (Fig. 6B). Since previous studies were performed with the Erdman strain of *Mtb*, this provided a possible explanation for the localization discrepancy. Consequently, we repeated the experiment with the *Mtb* Erdman strain and the ESX-1 mutant *Mtb* Erdman 36–72 that fails to secrete EsxA [23]. Again, EspR was below the level of detection in the CF of either strain, whereas EsxA appeared in the CF of *Mtb* Erdman (Fig. S7). We then examined CF at different time points of Erdman cultures for the presence of EspR and the cytosolic marker GroEL2. EspR first appeared in the culture filtrate after 8 days of growth when it was accompanied by GroEL2, indicating that cell lysis had likely occurred (Fig. 6C).

Discussion

The EspR protein has attracted considerable interest because of its role in the regulation of virulence in *Mtb* and the remarkable,

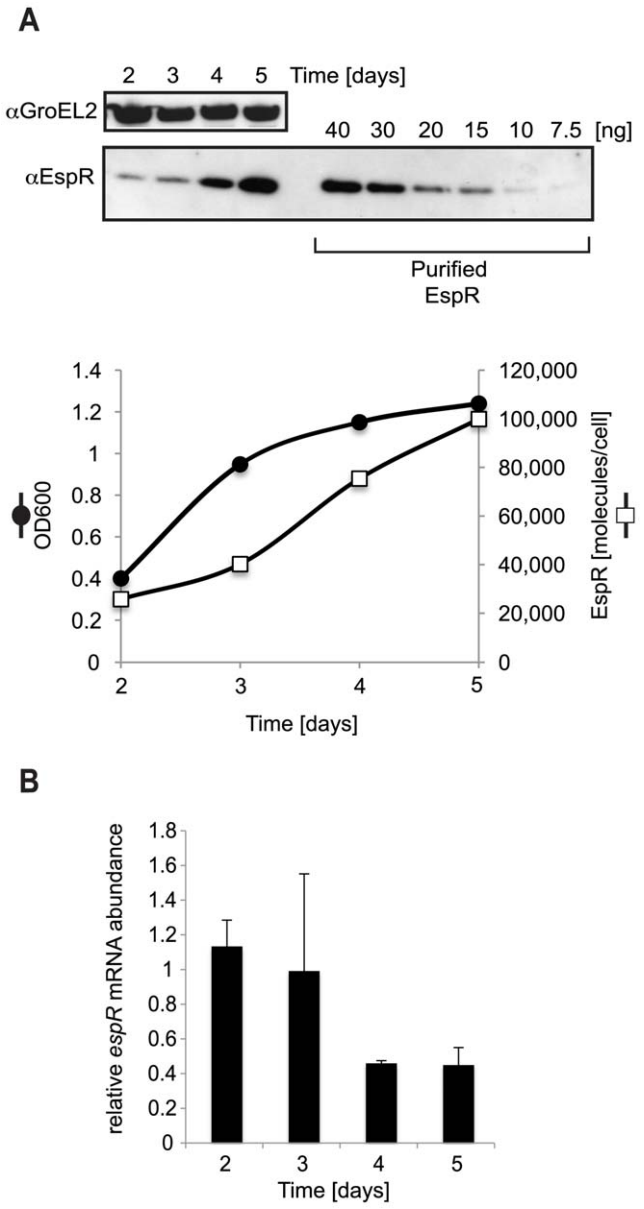


Figure 5. Growth phase-dependent variation of EspR intracellular levels. Time course analysis of *Mtb* H37Rv at the following phases of growth: early log (day 2), mid-log (day 3) and stationary (days 4 and 5). (**A**) Immunoblot analysis of cell lysates (CL) from equivalent amount of cells (approx. 2*10^7). For quantification purposes, the indicated amounts of purified EspR were electroblotted onto the same membrane as cell lysates. GroEL2 is used as a loading control and corresponding Western blot signals are used to normalize small loading discrepancies. The number of EspR molecules per cell was estimated from Western blot analysis correlated to OD_{600} measurements, assuming that OD_{600} $0.2 = 10^8$ cells [40]. The results are plotted in the bottom chart, with solid circles representing optical density at 600 nm (OD_{600}) while open squares represent the total number of EspR molecules per cell from one representative of four experiments. (**B**) Relative *espR* transcript amounts measured in total RNA harvested at the different time points by quantitative RT-PCR and normalized to *sigA* expression. The means and standard deviations of triplicate measurements are shown from two experiments.
doi:10.1371/journal.ppat.1002621.g005

and most unusual, property of being secreted by the very secretion system ESX-1 whose expression it controls [2]. This has led to the suggestion that a negative feedback loop modulates EspR secretion and is critical to successful infection. Studies of the three-dimensional structure of EspR and its truncated variant, EspRΔ10, together with models of DNA recognition and AFM analysis of single molecule EspR nucleoprotein complexes, indicate that EspR employs an atypical DNA recognition mechanism [10,11]. A dimer of dimers is thought to bind to DNA via one monomer of each dimer leaving the second monomer free to contact another site. Cooperative interactions then lead to multimerization and the formation of looped structures in which EspR acts as a bridge between two separated, or even remote, sites on DNA [10,11]. Such structures are typically formed by nucleoid-associated proteins (NAPs), like H-NS and Fis [12,24,25]. Collectively, these features led us to consider the possibility that EspR functions as a NAP rather than as a specific transcriptional activator of a limited number of genes required for pathogenesis.

To test this possibility, ChIP-Seq analysis was performed to assess the genome-wide distribution of EspR and to identify the sites and genes to which it binds. This resulted in the identification of at least 165 loci, often containing multiple peaks of EspR-binding, throughout the genome. Binding sites were distributed more or less evenly between intragenic and intergenic regions. The majority of the genes (~50%) were involved in cell envelope functions. Among them were genes that contribute to ESX-1, ESX-2 and ESX-5-related activities, and others that contribute to mycobacterial virulence, such as *lipF* and the PDIM locus. Surprisingly, the latter harbored the major peak of EspR binding in the genome whereas the previously reported site preceding the *espACD* operon was less prominent (Fig. 1 and Table 1).

Confirmation of EspR binding to a selection of major sites was obtained *in vitro*, from a combination of studies performed with highly purified EspR (Fig. 3), and *in vivo*, following overexpression of the protein (Fig. 4). This resulted in the definition of a consensus sequence, TTTGC[TC][GA], that agrees well with the motif predicted previously by molecular dynamic simulations, involving computation of the binding energy of the optimal interaction of EspR with the intergenic region upstream of *espA* [10]. These predictions were consistent with the findings of DNase footprinting or EMSA studies of the same locus. Using an *in silico* approach to scan the genome sequence >1,000 potential EspR sites were found, of which 163 had been detected experimentally by ChIP-Seq. While some of the *in silico* predictions may be fortuitous, this does raise the possibility that occupancy of EspR-binding sites may vary with growth phase or physiological conditions and that more sites will be uncovered. Furthermore, the EspR-binding motif deduced here should be considered as a core sequence for high affinity nucleation sites from where cooperative binding between EspR dimers can initiate and extend to form long oligomers and hence reach more distant sites [10]. The number of such sites and the distance between them most probably enables EspR to structure the chromosome.

The number of EspR molecules per cell was estimated by quantitative Western blotting at different stages of the growth cycle. There was a steady increase in concentration until ~100,000 molecules/cell were found at day 5. These levels are about 30-fold higher than those of well-characterized transcriptional activators, like Fnr in *E. coli*, but are comparable to levels of major NAPs, such as Fis or HU, during the exponential growth phase of *E. coli* [13,26]. The intracellular EspR concentration is clearly in excess of that required to occupy all the experimentally detected (582) or computer predicted (1026) binding sites.

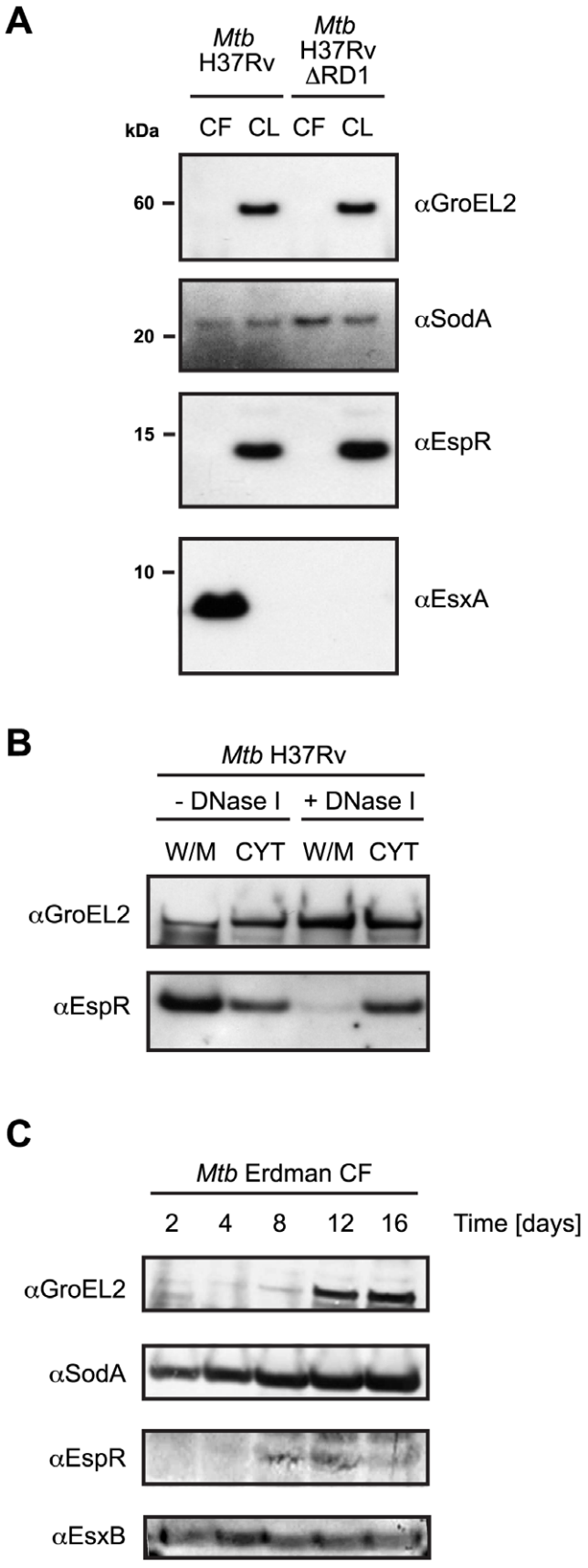


Figure 6. EspR is not a secreted protein. (**A**) Immunoblot analysis of 10 μg of culture filtrate (CF) and 5 μg of cell lysates (CL) of wild-type H37Rv (left) and H37RvΔRD1 (right) strains grown for 4 days after transfer into Sauton's medium without Tween-80. GroEL2 was used as a control for autolysis, SodA as a loading control for CF and CL samples and EsxA as a control for ESX-1-dependent secretion. (**B**) Immunoblot analysis of H37Rv CL shown in (**A**) fractionated into cell wall/cell membrane (W/M) and cytosolic (CYT) components by ultracentrifugation. GroEL2 was used as a loading control. (**C**) Immunoblot analysis of Erdman CF (20 μg) for each of the time points indicated.
doi:10.1371/journal.ppat.1002621.g006

The results of subcellular fractionation of *Mtb* H37Rv cells from mid-log phase indicate that EspR is predominantly a cytosolic protein although it can be found attached to cell membrane-bound DNA, a trait of the nucleoid. Prior treatment of this fraction with DNase I releases most of the EspR to the cytosol (Fig. 6B). In contrast to the findings of a previous report [2], we were unable to detect the secreted form of EspR in the culture filtrates of ESX-1 proficient and deficient strains of H37Rv nor in the Erdman strains at early time points. Since EspR is a relatively abundant protein and only found in the culture filtrate together with the cytosolic marker GroEL2, we conclude that it is released via cell lysis rather than secretion mediated by ESX-1.

EspR seemingly acts as an activator or a repressor depending on its binding position relative to the genes it controls. EspR production appears to be autoregulated as the protein binds to its own promoter region and downregulates expression in certain conditions (Fig. 4C). Interaction at the *espR* regulatory site occurs in a contrasting manner to that seen at the *espACD* locus where there are three prominent binding sites situated far upstream of the promoter (Fig. 1). Interestingly, some tubercle bacilli, including *M. bovis* and *M. microti*, have incurred the RD8 deletion in this region of the genome [27] that completely removes the three major EspR-binding sites. Two EspR-binding sites flank the *espR* promoter thus evoking an autoregulatory mechanism whereby EspR forms a loop at the promoter that either occludes RNA polymerase (RNAP) or traps RNAP that has already bound. In ChIP-Seq experiments performed with RNAP-antibodies, a major polymerase binding site was localized (data not shown) that partially overlaps EspR peak **a** (Fig. 3), consistent with the promoter prediction from 5′ RACE.

Loss of EspR strongly attenuates *Mtb* [2], suggesting that this is due to reduced functioning of the ESX-1 system as a result of insufficient EspA levels. However, in light of the present findings, this appears to be an oversimplification as expression of the genes for several other known virulence determinants are clearly subject to EspR regulation. Foremost among these is a major locus that encodes an enzyme system required for synthesis of PDIM, and in some strains PGL, both of which contribute extensively to virulence [16,17]. Another enzyme that has an important role in pathogenesis is the lipase, LipF [16], which has been implicated in modification of the mycobacterial cell wall as an adaptive response to acid damage [28]. LipF is also thought to degrade host lipids during infection [16]. The EspR binding site (Table 1) located far upstream of the *lipF* coding sequence overlaps the previously identified 59 bp acid-inducible promoter region, situated 515 bp from the start codon [28]. Occlusion of this site by EspR would therefore explain the observed repression of *lipF* transcription (Fig. 4D). The *lipF* gene, together with a number of other EspR gene targets like *fadD26* and *espA*, is also regulated by PhoP [29,30] and CRP [31], frequently with opposite effects on transcription.

Regulation of transcription orchestrated by EspR seems to occur at two levels. EspR binding at promoter regions, as in the case of *espR* or *lipF*, resembles global transcriptional regulators

where repression of transcription stems from occlusion of RNAP whereas activation of transcription occurs via favorable interaction with RNAP and/or other proteins. On the other hand, our genome-wide analysis revealed that more than half of the EspR-binding sites are intragenic and this refutes, at least partially, the hypothesis that EspR acts as a transcription factor *per se*. Moreover, EspR overexpression had little effect on some major EspR-bound genes (Fig. 4D), suggesting that EspR-binding does not necessarily affect transcription locally but rather serves as anchoring points to organize chromosome domains. NAPs with DNA-bridging activity, such as EspR, are often located at the boundaries of chromosomal domain loops [13] where they control gene expression in a temporal or spatial manner. In many bacteria, NAP expression levels are dependent on the growth phase [24]. This is true of *Mtb* since low EspR levels were detected at early- and mid-log phase compared to stationary phase, and premature conditional overexpression causes growth to slow down. The interplay between different NAPs alters chromosome structure and organization thereby influencing patterns of gene expression in a temporal manner.

Lsr2 is a DNA-bridging protein that also performs NAP functions in *Mtb* by recognizing AT-rich and xenogeneic regions. Binding sites of Lsr2 in *Mtb* have been mapped by Gordon *et al.* using ChIP-on-chip technology [32]. Comparison of the Lsr2 ChIP-on-chip and EspR ChIP-Seq results showed that 77% of the genes in the EspR regulon are also likely recognized by Lsr2 [32] owing to an extensive overlap between their repertoires (Fig. S2). For example, all three genes significantly upregulated upon EspR overexpression (Fig. 4D) also bind Lsr2 and major ChIP-Seq peaks of EspR are located close to Lsr2 binding sites in the ChIP-on-chip enriched regions (Fig. S2).

While Lsr2 and EspR are both subject to autoregulation there is no evidence for cross-regulation and Lsr2 seems to impact many more regions. Lsr2 has an N-terminal dimerization domain and a C-terminal DNA-binding domain whereas in EspR the opposite configuration exists. A further difference lies in the DNA recognition mechanisms since Lsr2 interacts with the minor groove while EspR binding is predicted to occur via the major groove. Together, this suggests that EspR and Lsr2 may control gene expression, including that of many cell wall functions, in a divergent manner with EspR possibly replacing Lsr2 at certain sites and vice-versa. It is striking that in all sequenced mycobacterial genomes *espR* is very close to the *hns* gene encoding another NAP [33] and this may also indicate functional interplay. These hypotheses can be tested experimentally by using the corresponding antibodies to perform ChIP-Seq experiments on *Mtb* strains at different stages in the growth cycle to localize their binding sites. The contribution of other global regulators that intersect with the EspR regulon, like PhoP and CRP, should also be examined. A regulatory scheme is emerging in which growth of *Mtb*, and hence pathogenesis, is controlled by chromosome remodeling, effected by different NAPs, thereby resulting in pleiotropic regulation of gene expression. EspR may thus play a central role in regulating virulence gene expression analagous to that of H-NS in enteropathogenic bacteria [12,13].

Materials and Methods

Bacterial strains and culture conditions

Mycobacterium tuberculosis (*Mtb*) H37Rv and Erdman were grown in 7H9 Broth (Difco) supplemented with 0.2% glycerol, Middlebrook albumin-dextrose-catalase (ADC) enrichment and 0.05% Tween 80 or on solid Middlebrook 7H11 medium (Difco) supplemented with oleic acid-albumin-dextrose-catalase (OADC). *Mtb* cells prepared for secretion analysis and cell fractionation were grown in Sauton liquid medium. *Escherichia coli* BL21(DE3) was used for expression of His_9-tagged EspR as described [10].

ChIP-Seq

Chromatin immunoprecipitation was performed essentially as previously described [34] but with some variations. Briefly, *Mtb* H37Rv cultures (50 ml) grown to OD_{600} 0.4–0.6 were treated with formaldehyde (final concentration 1%) for 10 min at 37°C. Cross-linking was quenched by addition of glycine to a final concentration of 125 mM. Harvested cells were washed twice with Tris-buffered saline (20 mM Tris-HCl pH 7.5, 150 mM NaCl), re-suspended in 600 µl Immunoprecipitation (IP) buffer (50 mM Hepes-KOH pH 7.5, 150 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.1% sodium deoxycholate, 0.1% SDS, mini-protease inhibitor cocktail (Roche)) and sonicated to shear DNA to an average size of 100–500 bp. Insoluble cellular debris was removed by centrifugation and the supernatant used as input sample in IP experiments. 300 µl of input was incubated with either no antibody (mock-IP) or 20 µl of serum containing rat anti-EspR polyclonal antibodies (kindly generated by Ida Rosenkrands) and then with 50 µl of Dynabeads sheep anti-rat IgG (Dynal Biotech) pre-saturated with 0.1 mg/ml salmon sperm DNA and 1 mg/ml Bovine Serum Albumin (BSA) in IP buffer. Washing and crosslink reversal of protein-DNA complexes, as well as purification of the resulting DNA was carried out as previously described [34]. Prior to sequencing, DNA fragment sizes were checked and enrichment was verified by gene-specific quantitative PCR (ChIP-qPCR).

EspR ChIP-Seq library construction and sequencing

DNA fragments (150 to 250 bp) were selected for library construction and sequencing libraries prepared using the ChIP-Seq Sample Preparation Kit (Illumina; San Diego, California, USA; Cat. No. IP-102-1001) according to the protocol supplied with the reagents. Prior and post library construction, chromatin immunoprecipitation products were quantified using the Qubit fluorometer (Invitrogen; Carlsbad, California, USA). One lane of each library was sequenced on the Illumina Genome Analyzer IIx using the Single-Read Cluster Generation Kit v4 and 36 Cycle Sequencing Kit v4. Data were processed using the Illumina Pipeline Software v1.60.

ChIP-Seq data analysis

ChIP-Seq experiments, with two independent mid-log phase cultures, generated 25.6 and 22 million reads, of which 95% could be successfully mapped to the *Mtb* H37Rv genome (NCBI accession NC_000962.2) using Bowtie [35] allowing up to 3 mismatches and up to 10 hits per read. As a control, input DNA was also sequenced and mapped to the *Mtb* H37Rv genome to identify sequencing artifacts and calculate enrichment values. Since comparison of the two ChIP-Seq datasets showed excellent correlation in binding signals (Fig. S1), sequenced DNA reads from both experiments were pooled together. In order to estimate the binding location in the enriched regions, we used a deconvolution algorithm that models the expected tag distribution on both strands [36]. Based on the deconvoluted profile, a score was calculated for each peak that was proportional to the read density in the peak. Considering the high genome coverage, only peaks with more than 600 reads per position were selected. Reads mapped to the forward and reverse strands were shifted (by 80 bp) and merged together to generate single peak profiles, which were visualized on the UCSC Genome Browser database [37]. Annotation of the peaks was performed using the *ChipPeakAnno*

package from Bioconductor [38]. The read count at each binding region (400 bp width) was determined as the total number of reads mapping to the region divided by the length of the region normalized to the total number of mapped reads across the whole genome. The enrichment at each locus was obtained as the ratio of the average read count from the ChIP sample (from two datasets) to the read count from the input DNA sample. Peaks with an enrichment ratio lower than 1.5 were filtered out. The ChIP-Seq data files have been deposited in NCBI's Gene Expression Omnibus [39] and can be accessed through GEO Series accession number GSE35149. Lsr2 data were retrieved from this database (accession number GSE18652).

Motif detection

Sequences of 102 bp covering the center of the 582 peaks that fulfilled the peak selection criteria were extracted and used for motif analysis using MEME [19] with motif occurrence set as "zero or one per sequence", minimum width as 5 and other parameters as default. Since many EspR binding sites are found within the highly repetitive sequences belonging to *pe* and *ppe* genes, the initial motif returned by MEME was biased and reflected a *pe-ppe* motif. Thus, from the 582 sequences, those belonging to the *pe* and *ppe* categories were excluded, leaving 416 sequences that were reanalyzed using MEME. This led to the identification of an unbiased overrepresented motif shown in Fig. 2. This motif was further used in FIMO (Find Individual Motif Occurences) from the MEME Suite web server [19] to search for motif occurrence in the entire set of EspR binding sequences setting the *p*-value threshold to 0.001.

Electrophoretic mobility shift assays (EMSA) and DNase I footprinting

EspR protein was produced and purified as previously described [10] and used in EMSA, DNase I footprinting and quantitative Western Blot experiments, as well as to immunize rats. EspR-DNA gel retardation assays were performed as recently described [10] using 5′-biotin-labeled forward primers and unlabeled reverse primers in PCR reactions designed to yield DNA probes of 99 to 120 bp encompassing selected ChIP-Seq peak sequences for analysis. DNase I footprinting assay was carried out using 6-FAM-labeled probes as described in [10]. 6-FAM-labeled forward primers and unlabeled reverse primers were used to synthesize a 361 bp DNA fragment starting 208 bp upstream of the *espR* translational start by PCR. The nucleotide sequences of protected regions were determined by DNA sequencing.

5′ RACE

RNA was extracted from *Mtb* H37Rv cells in the mid-log phase. 5′ RACE was performed as previously described [34] using the 5′/3′ RACE kit (2nd Generation, Roche) and primers specific for *espA* and *espR* for 5′-end mapping (see Table S1).

Inducible overexpression of EspR in *Mtb* H37Rv

Construction of the pMYespR vector for pristinamycin IA-inducible overexpression of *espR* was undertaken as previously described [21]. Plasmids pMYespR and the control pMY769 were transformed into H37Rv strain by electroporation, clones were selected on 7H11 agar plates containing 100 μg/ml spectinomycin and plasmid integration verified by PCR. The resulting H37Rv::pMYEspR strain was grown to mid-log phase (OD_{600} 0.4–0.6), diluted to $OD_{600} = 0.1$ (day 0) and split into two 30 ml volumes prior to addition of 0 or 2 μg/ml pristinamycin IA. Cells were harvested at day 3.

For whole cell Western blot analysis, bacteria were pelleted, washed twice in PBS containing 0.05% Tween 80 and resuspended in lysis buffer (10 mM Tris pH 7.9, 500 mM NaCl, 1 mM β-mercaptoethanol, 5% glycerol, 0.1 mM EDTA) prior to sonication (15 min at 4°C). Cell debris was then pelleted and the supernatant filter sterilized. After total protein quantification using Bradford reagent (Sigma), 10 μg of cell lysates were separated by SDS-PAGE, and specific proteins visualized by immunoblot as described below.

Immunoblot analysis

Proteins were separated on NuPAGE Novex 4–12% bis-Tris gels (Invitrogen) and transferred to nitrocellulose membranes. Membranes were incubated in TNT blocking buffer (25 mM Tris pH 7.5, 150 mM NaCl, 0.05% Tween 20) with 5% w/v skim milk powder for 2 h prior to incubation with primary antibodies diluted in TNT with 1% BSA overnight. Membranes were washed in TNT five times, then incubated with secondary antibodies for 1 h before washing. Immunoblotting were performed with rat polyclonal anti-EspR antibodies (kindly provided by Ida Rosenkrands), mouse monoclonal anti-EsxA antibodies (Hyb 76-8), mouse monoclonal anti-GroEL2 antibodies (IT-70), rabbit polyclonal anti-EsxB and anti-SodA antibodies. All antibodies were used at a dilution 1:1,000 except for anti-GroEL2 (1:5,000), anti-EsxB (1:500) and anti-SodA (1:50). Horseradish peroxidase (HRP) conjugated rabbit polyclonal anti-rat IgG, rabbit polyclonal anti-mouse IgG or goat polyclonal anti-rabbit IgG secondary antibodies (Sigma) were used at dilutions of 1:100,000. Secondary antibodies were visualized using chemiluminescent substrates (Sigma).

Quantitative RT-PCR

Quantitative RT-PCR reactions were performed with the 7900HT Fast Real-Time PCR System (Applied Biosystems) with the following parameters: 50°C for 2 min, 95°C for 10 min, followed by 40 cycles of 95°C for 15 s and 60°C for 60 s. Melt curve analysis was used to confirm specific amplification for each primer pair. Threshold cycle values were determined automatically using the SDS software.

Enrichment of ChIP DNA samples was measured by quantitative RT-PCR using peak-specific primer pairs (Table S1), IP or mock-IP DNA as template and SYBR-Green master mix (Applied Biosystems). To calculate the amount of amplified DNA, standard curves were generated for each primer pair using 10-fold dilutions of input DNA as template. All reactions were done in duplicate and averaged to determine the enrichment ratio calculated as [IP DNA]/[mock-IP DNA].

For gene expression analysis, RNA was extracted with Trizol reagent (Invitrogen) and treated with DNase I (Roche) prior to generation of the cDNA template. cDNA was synthesized using the RevertAid First Strand cDNA Synthesis Kit (Fermentas) according to the manufacturer's instructions using random hexamers primer. cDNA corresponding to 10 ng of input RNA was used in each RT-PCR reaction supplemented with specific primer pairs (200 nM each) listed in Table S1 and SYBR-Green master mix (Applied Biosystems). Relative mRNA levels were calculated using the ΔΔCt method, normalizing transcript levels to *sigA* signals and are average of three biological replicates.

Time course expression analysis

Mtb H37Rv cells were grown to OD_{600} 0.6 and then diluted in 100 ml with a starting OD_{600} of 0.05 (day 0). At various time points corresponding to the early-log (day 2), mid-log (day 3) and stationary (days 4 and 5) phase of growth, the number of cells was

estimated from OD_{600} and culture aliquots were pelleted by centrifugation for RNA and protein analysis. For protein analysis, 400 to 100 µl cell aliquots were resuspended in a 1× NuPAGE sample buffer (Invitrogen) and boiled for 45 min. For all time points, equivalent amounts of cells were subjected to SDS-PAGE together with purified EspR protein standards (40, 30, 20, 15, 10 and 7.5 ng) and proteins detected by immunoblotting. Quantification was performed by densitometry of scanned blots using the ImageJ software. A standard curve was made from the intensity of the purified EspR bands and was shown to be linear in the 7.5 to 40 ng range. Quantification of expression levels at each time point was estimated from the standard curve after correction for small sample loading discrepancies using GroEL2 band intensity for normalization. The experiment was repeated four times and a typical immunoblot pattern of EspR expression is shown in Fig. 5A. For transcript analysis, RNA isolation and quantitative RT-PCR was performed as described above.

EspR secretion analysis

Mtb starter cultures were first grown in complete 7H9 broth to late-logarithmic phase (OD_{600}~0.8–1). These were then used to inoculate Sauton's medium supplemented with 0.05% Tween-80, at starting OD_{600} of 0.1. Cells were grown to mid-log phase of growth (OD_{600}~0.6), centrifuged, washed once with phosphate buffered saline (PBS), resuspended in a volume of Sauton's medium without Tween-80 such that the OD_{600} was again ~0.6 then grown further at 37°C with shaking. Cultures were harvested at required times post-transfer by centrifugation to obtain culture filtrates and cell pellets. Culture filtrates (CF) were filtered sequentially through 0.4 and 0.2 micron filters to remove any residual cells and concentrated in Vivaspin columns with 5-kDa molecular weight cut-off membranes (Sartorius Stedim Biotech GmbH, Goettingen, Germany). Cell lysates (CL) were prepared by resuspending cell pellets in lysis buffer (PBS containing Roche mini-protease inhibitor cocktail tablets) followed by bead beating with 100 micron glass beads and centrifuged. Total protein concentration of all preparations was determined using the BCA assays (Pierce) with BSA as the standard. Equivalent total protein concentrations of CF and CL were analyzed by SDS-PAGE and immunoblotting. For cell fractionation experiments, CL samples were split into two equivalent volumes and treated with either 50 µl PBS or 50 µl of 1,000 units/ml DNase I overnight at 4°C. CL samples were centrifuged at 200,000 g for 4 h at 4°C in a Beckman-Coulter ultracentrifuge (TLA-100.3 rotor) to obtain the cell wall/cell membrane (W/M) fraction. Supernatants containing the cytosolic (CYT) fractions were carefully recovered and the W/M fractions resuspended in PBS with 0.1% Triton X-100 and mini-protease inhibitor cocktail. Equivalent amounts of the different cell fractions were separated by SDS-PAGE and specific proteins were visualized by immunoblotting.

Supporting Information

Figure S1 Correlation plot showing the reproducibility of EspR ChIP-Seq.
(PDF)

Figure S2 Comparison of EspR and Lsr2 binding sites on the *Mtb* genome. (A) Venn diagram showing significant overlap between the gene targets of Lsr2 [32] as obtained by ChIP-on-chip and EspR as obtained by ChIP-Seq (this study). (B–F) UCSC Genome Browser (http://genome.ucsc.edu) view of selected binding profiles as determined by ChIP-Seq for EspR (green, this study) and by ChIP-on-chip for Lsr2 (red, [32]). Shown are (B) the *rv1490* region; (C) the *rv0986-7-8* operon region; (D) the *espA-ephA* intergenic region; (E) the ESX-1 (*rv3864-3883c*) and ESX-2 (*rv3884c-rv3895c*) regions; and (F) the PDIM/PGL locus (*rv2928-rv2962c*). The GC content of *Mtb* H37Rv genome (in 20 bp windows) normalized to the median GC content (65.6%) and the gene positions are indicated below.
(PDF)

Figure S3 Analysis of the *espA* promoter. RNA extracted from *Mtb* H37Rv at mid-log phase was used for the 5′ RACE. Translational starts are highlighted in red with +1 corresponding to the first base of the *espA* open reading frame. Transcriptional start mapped by 5′ RACE is boxed and indicated by a bent arrow. −10 and −35 positions for putative sigma factor binding sites are indicated. ChIP-Seq peak sequences are highlighted in gray.
(PDF)

Figure S4 Validation of 11 EspR-binding sites selected among a wide range of scores by quantitative RT-PCR. The *sigA* and *rv0888* genes, showing no peak in EspR ChIP-Seq, were used as negative controls. Plot showing $\log_2$ enrichment calculated from ChIP-Seq and ChIP-qPCR experiments shows a good correlation between the output of the two experiments. The $\log_2$ enrichment values have also been indicated in the table below along with additional details about the peaks. The column "Peak feature" indicates where the EspR-binding site is located relative to the known gene annotation. "NA" denotes not applicable.
(PDF)

Figure S5 EMSA showing binding to the top five ChIP-Seq peak sequences related to the following genes: *fadD26*, *rv1490*, *rv2929*, *pe_pgrs19*, *espA*. A DNA fragment from within the *espA* coding region where no ChIP-Seq enrichment was observed was used as negative control.
(PDF)

Figure S6 DNase I footprint at peak "a" of the *espR* promoter. Red and black peaks represent DNA incubated without or with 10 µM EspR proteins, respectively. Both reactions were partially digested with DNase I and analysed by capillary electrophoresis in a genetic analyser (Applied Biosystems 3130xl). The corresponding sequencing reaction of the DNA fragment is shown at bottom. Regions I and II protected from DNase I digestion by EspR and regions IIaΔ10 and IIbΔ10 protected from DNase I digestion by EspRΔ10 are denoted by square brackets. Positions indicated are relative to the translational start.
(PDF)

Figure S7 Immunoblot analysis of 10 µg of culture filtrate (CF) and 5 µg of cell lysates (CL) of *Mtb* Erdman wild-type (left) and 36–72 (transposon insertion in the *pe35* promoter blocking *esxA* expression and therefore ESX-1 function [23]) (right); strains were grown for 4 days after transfer into Sauton's medium without Tween-80. GroEL2 was used as a control for autolysis, SodA as a loading control for CF and CL samples and EsxA as a control for ESX-1-dependent secretion.
(PDF)

Table S1 List of primers used in this study.
(PDF)

Table S2 List of the EspR ChIP-Seq peaks on the *Mtb* H37Rv genome sorted by score (highest to lowest). Peaks are numbered by order of appearance in the genome. The column "Feature" indicates where the EspR-binding site is located relative to the known gene annotation. The "Distance to feature" column designates the distance (in bp) from the midpoint position of the calculated peak to the closest translational start site. Values in the "Score" column are arbitrary units proportional to the peak

density. The "Enrichment" column represents the ratio between the average ChIP-Seq read count at a particular locus (400 bp) and the input DNA read count at that locus.
(XLS)

Acknowledgments

We thank Stefanie Boy and Philippe Busso for technical assistance. Antibodies against GroEL2 and SodA were received as part of the National Institutes of Health, National Institute of Allergy and Infectious Diseases contract (no. HHSN266200400091c) entitled "Tuberculosis Vaccine Testing and Research Materials", awarded to Colorado State University.

Author Contributions

Conceived and designed the experiments: BB JMC RH CS FP STC. Performed the experiments: BB JMC RH CS. Analyzed the data: BB JMC RH CS SU JR FP STC. Wrote the paper: BB STC.

References

1. Homolka S, Niemann S, Russell DG, Rohde KH (2010) Functional genetic diversity among Mycobacterium tuberculosis complex clinical isolates: delineation of conserved core and lineage-specific transcriptomes during intracellular survival. PLoS Pathog 6: e1000988.
2. Raghavan S, Manzanillo P, Chan K, Dovey C, Cox JS (2008) Secreted transcription factor controls Mycobacterium tuberculosis virulence. Nature 454: 717–721.
3. Fortune SM, Jaeger A, Sarracino DA, Chase MR, Sassetti CM, et al. (2005) Mutually dependent secretion of proteins required for mycobacterial virulence. Proc Natl Acad Sci U S A 102: 10676–10681.
4. Garces A, Atmakuri K, Chase MR, Woodworth JS, Krastins B, et al. (2010) EspA acts as a critical mediator of ESX1-dependent virulence in Mycobacterium tuberculosis by affecting bacterial cell wall integrity. PLoS Pathog 6: e1000957.
5. Abdallah AM, Gey van Pittius NC, Champion PA, Cox J, Luirink J, et al. (2007) Type VII secretion–mycobacteria show the way. Nat Rev Microbiol 5: 883–891.
6. Simeone R, Bottai D, Brosch R (2009) ESX/type VII secretion systems and their role in host-pathogen interaction. Curr Opin Microbiol 12: 4–10.
7. Bitter W, Houben EN, Bottai D, Brodin P, Brown EJ, et al. (2009) Systematic genetic nomenclature for type VII secretion systems. PLoS Pathog 5: e1000507.
8. Hsu T, Hingley-Wilson SM, Chen B, Chen M, Dai AZ, et al. (2003) The primary mechanism of attenuation of bacillus Calmette-Guerin is a loss of secreted lytic function required for invasion of lung interstitial tissue. Proc Natl Acad Sci U S A 100: 12420–12425.
9. van der Wel N, Hava D, Houben D, Fluitsma D, van Zon M, et al. (2007) M. tuberculosis and M. leprae translocate from the phagolysosome to the cytosol in myeloid cells. Cell 129: 1287–1298.
10. Blasco B, Stenta M, Alonso-Sarduy L, Dietler G, Peraro MD, et al. (2011) Atypical DNA recognition mechanism used by the EspR virulence regulator of Mycobacterium tuberculosis. Mol Microbiol 82: 251–264.
11. Rosenberg OS, Dovey C, Tempesta M, Robbins RA, Finer-Moore JS, et al. (2011) EspR, a key regulator of Mycobacterium tuberculosis virulence, adopts a unique dimeric structure among helix-turn-helix proteins. Proc Natl Acad Sci U S A 108: 13450–13455.
12. Browning DF, Grainger DC, Busby SJ (2010) Effects of nucleoid-associated proteins on bacterial chromosome structure and gene expression. Curr Opin Microbiol 13: 773–780.
13. Dillon SC, Dorman CJ (2010) Bacterial nucleoid-associated proteins, nucleoid structure and gene expression. Nat Rev Microbiol 8: 185–195.
14. Sala C, Grainger DC, Cole ST (2009) Dissecting regulatory networks in host-pathogen interaction using chIP-on-chip technology. Cell Host Microbe 5: 430–437.
15. Becq J, Gutierrez MC, Rosas-Magallanes V, Rauzier J, Gicquel B, et al. (2007) Contribution of horizontally acquired genomic islands to the evolution of the tubercle bacilli. Mol Biol Evol 24: 1861–1871.
16. Camacho LR, Ensergueix D, Perez E, Gicquel B, Guilhot C (1999) Identification of a virulence gene cluster of Mycobacterium tuberculosis by signature-tagged transposon mutagenesis. Mol Microbiol 34: 257–267.
17. Cox JS, Chen B, McNeil M, Jacobs WR, Jr. (1999) Complex lipid determines tissue-specific replication of Mycobacterium tuberculosis in mice. Nature 402: 79–83.
18. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, et al. (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature 393: 537–544.
19. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, et al. (2009) MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res 37: W202–208.
20. Forti F, Crosta A, Ghisotti D (2009) Pristinamycin-inducible gene regulation in mycobacteria. J Biotechnol 140: 270–277.
21. Hartkoorn RC, Sala C, Magnet SJ, Chen JM, Pojer F, et al. (2010) Sigma factor F does not prevent rifampin inhibition of RNA polymerase or cause rifampin tolerance in Mycobacterium tuberculosis. J Bacteriol 192: 5472–5479.
22. Shapiro L, McAdams HH, Losick R (2009) Why and how bacteria localize proteins. Science 326: 1225–1228.
23. Brodin P, Majlessi L, Marsollier L, de Jonge MI, Bottai D, et al. (2006) Dissection of ESAT-6 system 1 of Mycobacterium tuberculosis and impact on immunogenicity and virulence. Infect immun 74: 88–98.
24. Dame RT (2005) The role of nucleoid-associated proteins in the organization and compaction of bacterial chromatin. Mol Microbiol 56: 858–870.
25. Skoko D, Yoo D, Bai H, Schnurr B, Yan J, et al. (2006) Mechanism of chromosome compaction and looping by the Escherichia coli nucleoid protein Fis. J Mol Biol 364: 777–798.
26. Azam TA, Iwata A, Nishimura A, Ueda S, Ishihama A (1999) Growth phase-dependent variation in protein composition of the Escherichia coli nucleoid. J Bacteriol 181: 6361–6370.
27. Gordon SV, Brosch R, Billault A, Garnier T, Eiglmeier K, et al. (1999) Identification of variable regions in the genomes of tubercle bacilli using bacterial artificial chromosome arrays. Mol Microbiol 32: 643–655.
28. Richter L, Tai W, Felton J, Saviola B (2007) Determination of the minimal acid-inducible promoter region of the lipF gene from Mycobacterium tuberculosis. Gene 395: 22–28.
29. Gonzalo-Asensio J, Mostowy S, Harders-Westerveen J, Huygen K, Hernandez-Pando R, et al. (2008) PhoP: a missing piece in the intricate puzzle of Mycobacterium tuberculosis virulence. PLoS ONE 3: e3496.
30. Walters SB, Dubnau E, Kolesnikova I, Laval F, Daffe M, et al. (2006) The Mycobacterium tuberculosis PhoPR two-component system regulates genes essential for virulence and complex lipid biosynthesis. Mol Microbiol 60: 312–330.
31. Rickman L, Scott C, Hunt DM, Hutchinson T, Menendez MC, et al. (2005) A member of the cAMP receptor protein family of transcription regulators in Mycobacterium tuberculosis is required for virulence in mice and controls transcription of the rpfA gene coding for a resuscitation promoting factor. Mol Microbiol 56: 1274–1286.
32. Gordon BR, Li Y, Wang L, Sintsova A, van Bakel H, et al. (2010) Lsr2 is a nucleoid-associated protein that targets AT-rich sequences and virulence genes in Mycobacterium tuberculosis. Proc Natl Acad Sci U S A 107: 5154–5159.
33. Werlang IC, Schneider CZ, Mendonca JD, Palma MS, Basso LA, et al. (2009) Identification of Rv3852 as a nucleoid-associated protein in Mycobacterium tuberculosis. Microbiology 155: 2652–2663.
34. Sala C, Haouz A, Saul FA, Miras I, Rosenkrands I, et al. (2009) Genome-wide regulon and crystal structure of BlaI (Rv1846c) from Mycobacterium tuberculosis. Mol Microbiol 71: 1102–1116.
35. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10: R25.
36. Rey G, Cesbron F, Rougemont J, Reinke H, Brunner M, et al. (2011) Genome-wide and phase-specific DNA-binding rhythms of BMAL1 control circadian output functions in mouse liver. PLoS Biol 9: e1000595.
37. Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D, et al. (2011) The UCSC Genome Browser database: update 2011. Nucleic acids res 39: D876–882.
38. Zhu LJ, Gazin C, Lawson ND, Pages H, Lin SM, et al. (2010) ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. BMC bioinformatics 11: 237.
39. Edgar R, Domrachev M, Lash AE (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic acids res 30: 207–210.
40. Tufariello JM, Jacobs WR, Jr., Chan J (2004) Individual Mycobacterium tuberculosis resuscitation-promoting factor homologues are dispensable for growth in vitro and in vivo. Infect immun 72: 515–526.

4.3 Whole Genome Re-sequencing

Contribution: Mapping and assembly of the high-throughput sequencing data, detection of sequence variation in the form of SNPs (single nucleotide polymorphisms) and InDels (insertions and deletions), and characterization of the nature of the substitutions occurring in coding sequences, as synonymous or nonsynonymous, based on the M. tb *H37Rv genome annotation.*

4.3.1 Spontaneous phthiocerol dimycocerosate-deficient variants of *M. tb*

The cell wall of *M. tb* possesses a unique array of complex lipids and carbohydrates that play critical roles in host-pathogen interactions and virulence of *M. tb*. Glycolipids are major *M. tb* cell wall constituents, known for their toxic or immunological properties. The most abundant of these lipids, the phthiocerol dimycocerosates (PDIMs) have been linked to virulence. Previous studies have shown that spontaneously arising PDIM-deficient strains were attenuated in guinea pigs and mice. Whole-genome re-sequencing of a spontaneously arising PDIM-negative clone and comparison with a PDIM-positive clone led to the identification of a nonsynonymous SNP in the *ppsD* gene resulting in a glycine to cysteine substitution at position 44 (G44C) in the PpsD protein. The *ppsD* gene encodes a modular polyketide synthase involved in PDIM biosynthesis (34). Further experimental validation confirmed that the spontaneous point mutation in the *ppsD* gene is in fact responsible for both the defect in PDIM production and attenuation of virulence in mice.

INFECTION AND IMMUNITY, July 2011, p. 2829–2838 Vol. 79, No. 7
0019-9567/11/$12.00 doi:10.1128/IAI.00097-11

Spontaneous Phthiocerol Dimycocerosate-Deficient Variants of *Mycobacterium tuberculosis* Are Susceptible to Gamma Interferon-Mediated Immunity▽

Meghan A. Kirksey,[1]†‡ Anna D. Tischler,[3]*† Roxane Siméone,[2]§ Katherine B. Hisert,[1]¶ Swapna Uplekar,[3] Christophe Guilhot,[2] and John D. McKinney[1,3]

Laboratory of Infection Biology, The Rockefeller University, New York, New York 10021[1]; Institut de Pharmacologie et de Biologie Structurale, Centre National de la Recherche Scientifique and Université P. Sabatier (Unité Mixte de Recherche 5089), 31077 Toulouse Cedex, France[2]; and Global Health Institute, Swiss Federal Institute of Technology Lausanne (EPFL), CH-1015 Lausanne, Switzerland[3]

Received 28 January 2011/Returned for modification 17 March 2011/Accepted 3 May 2011

Onset of the adaptive immune response in mice infected with *Mycobacterium tuberculosis* is accompanied by slowing of bacterial replication and establishment of a chronic infection. Stabilization of bacterial numbers during the chronic phase of infection is dependent on the activity of the gamma interferon (IFN-γ)-inducible nitric oxide synthase (NOS2). Previously, we described a differential signature-tagged mutagenesis screen designed to identify *M. tuberculosis* "counterimmune" mechanisms and reported the isolation of three mutants in the H37Rv strain background containing transposon insertions in the *rv0072*, *rv0405*, and *rv2958c* genes. These mutants were impaired for replication and virulence in NOS2$^{-/-}$ mice but were growth-proficient and virulent in IFN-γ$^{-/-}$ mice, suggesting that the disrupted genes were required for bacterial resistance to an IFN-γ-dependent immune mechanism other than NOS2. Here, we report that the attenuation of these strains is attributable to an underlying transposon-independent deficiency in biosynthesis of phthiocerol dimycocerosate (PDIM), a cell wall lipid that is required for full virulence in mice. We performed whole-genome resequencing of a PDIM-deficient clone and identified a spontaneous point mutation in the putative polyketide synthase PpsD that results in a G44C amino acid substitution. We demonstrate by complementation with the wild-type *ppsD* gene and reversion of the *ppsD* gene to the wild-type sequence that the *ppsD*(G44C) point mutation is responsible for PDIM deficiency, virulence attenuation in NOS2$^{-/-}$ and wild-type C57BL/6 mice, and a growth advantage *in vitro* in liquid culture. We conclude that PDIM biosynthesis is required for *M. tuberculosis* resistance to an IFN-γ-mediated immune response that is independent of NOS2.

Pathogenic mycobacteria possess a unique array of complex cell wall-associated lipids. The most abundant of these lipids, the phthiocerol dimycocerosates (PDIMs) (Fig. 1), are among the best characterized (23). PDIMs contain long-chain diols esterified by methyl-branched fatty acid chains. As early as 1974, it was recognized that a spontaneously arising PDIM-deficient variant of the laboratory strain H37Rv was attenuated in a guinea pig model of infection (11). Shortly thereafter, it was shown that the *in vivo* survival of an avirulent *Mycobacterium tuberculosis* strain was enhanced by coating the bacteria with cholesterol oleate and purified PDIM (16). A genetic link between PDIM production and virulence was not established until a quarter century later, when a large chromosomal locus was identified as playing an essential role in the biosynthesis and export of PDIM (3, 4, 6). Transposon insertions within the *fadD26*, *fadD28*, *mmpL7*, and *drrC* genes, and in the putative transcriptional promoter region upstream of the *fadD26* gene, were identified in strains deficient in surface-localized PDIM. The *fadD26* and *fadD28* mutants apparently fail to synthesize PDIM, whereas the *mmpL7* and *drrC* mutants produce PDIM but accumulate it intracellularly, thus implicating these genes in transmembrane transport of PDIM to the cell surface. Strains deficient in production or surface localization of PDIM are markedly attenuated for growth in the lungs of intravenously (3, 6) or intranasally (28) infected mice, and in gamma interferon (IFN-γ)-activated but not in nonactivated macrophages (28).

Mutant strains that lack surface-localized PDIM display enhanced membrane permeability (3), but the precise role of PDIM in virulence of *M. tuberculosis* is unclear. The attenuated growth of PDIM-deficient *M. tuberculosis* in IFN-γ-activated macrophages is reversed by treatment of the infected macrophages with *N*ω-nitro-L-arginine methyl ester (L-NAME), a small-molecule inhibitor of the mammalian inducible nitric oxide synthase (NOS2), suggesting that PDIM might play a role in countering the impact of this important host antimicrobial mechanism (28). However, PDIM-deficient bacteria do not show increased sensitivity to reactive nitrogen

* Corresponding author. Mailing address: Global Health Institute, Swiss Federal Institute of Technology Lausanne (EPFL), Station 19, CH-1015 Lausanne, Switzerland. Phone: (41) (0)21 693 0626. Fax: (41) (0)21 693 1790. E-mail: anna.tischler@epfl.ch.

‡ Present address: Department of Anesthesiology, Weill Cornell Medical Center, New York, NY 10021.

§ Present address: Institute Pasteur, Pathogénomique Mycobactérienne Intégrée, 75724 Paris Cedex 15, France.

¶ Present address: Division of Pulmonary and Critical Care Medicine, University of Washington School of Medicine, Seattle, WA 98195-6522.

† These authors contributed equally to this work.

▽ Published ahead of print on 16 May 2011.

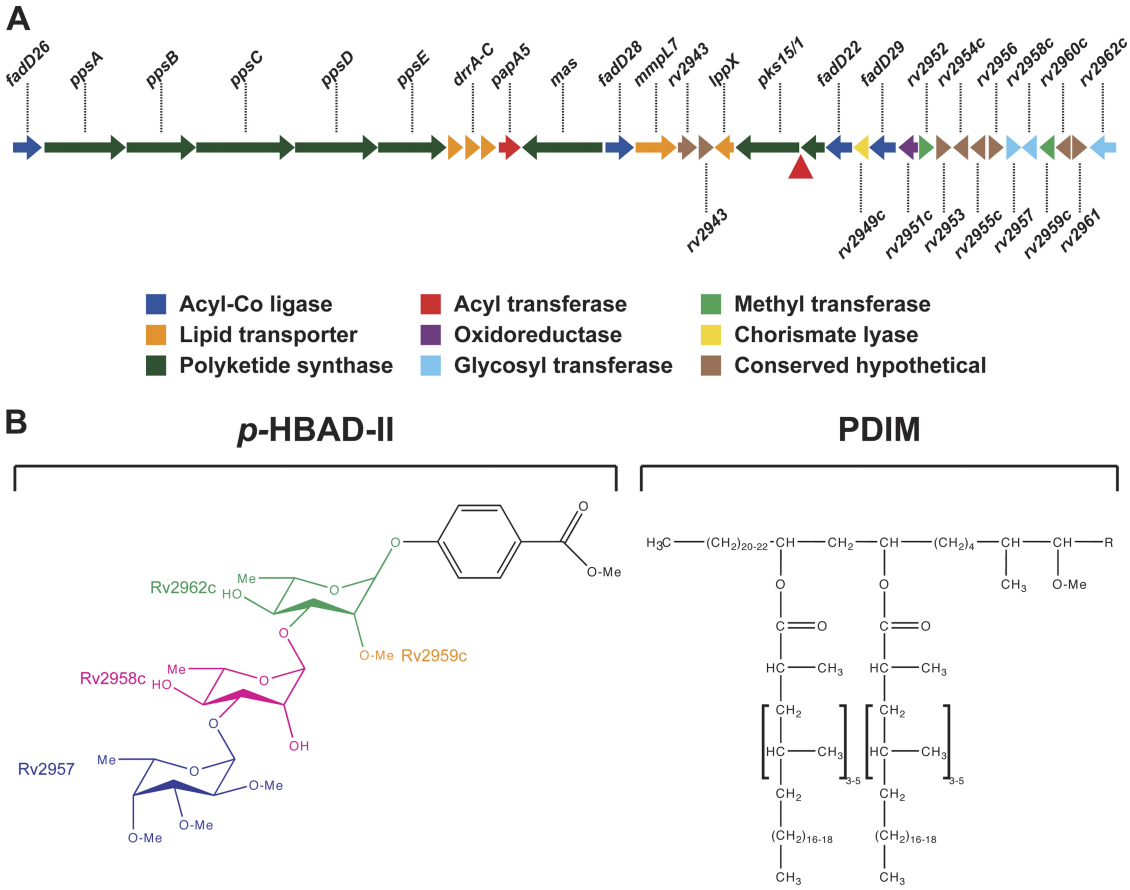


FIG. 1. PDIM and *p*-HBAD biosynthesis in *M. tuberculosis*. (A) Genomic locus responsible for PDIM and *p*-HBAD-II biosynthesis. In *M. tuberculosis* H37Rv and other members of the Euro-American lineage, a frameshift mutation (red triangle) disrupts the *pks15/1* open reading frame. Other *M. tuberculosis* lineages have an intact *pks15/1* locus that encodes a functional polyketide synthase. (B) Structures of *p*-HBAD-II and PDIM. The polyketide synthase Pks15/1 adds malonyl coenzyme A (malonyl-CoA) units to *p*-hydroxybenzoic acid to generate *p*-hydroxyphenylalkanoic acid derivatives, which are precursors of PGL biosynthesis (5). *M. tuberculosis* 37Rv lacks Pks15/1 activity and produces tri-glycosylated *p*-hydroxybenzoic acid (*p*-HBAD-II) and PDIM instead. The Rv2962c, Rv2958c, and Rv2957 glycosyl transferases add rhamnose → rhamnose → fucose to *p*-hydroxyphenylalkanoic acid (25), and the Rv2959c methyltransferase O-methylates the C2 ring position of the proximal rhamnosyl residue (24).

species (RNS) *in vitro* (3), suggesting that *M. tuberculosis* might be sensitized to RNS during intracellular growth or that the impact of L-NAME on intracellular bacteria might be indirect, possibly via modulation of expression of host (9) or pathogen (29, 35) genes that are RNS responsive. A detailed comparative analysis of PDIM-proficient and PDIM-deficient *M. tuberculosis* strains in macrophages revealed that PDIM participates in receptor-dependent phagocytosis and inhibition of phagosome acidification (2). PDIM insertion in model membranes caused alterations in membrane fluidity, suggesting that physical changes to the host cell membrane caused by interaction with PDIM may influence the uptake and ultimate cellular destination of *M. tuberculosis* (2).

Whereas all *M. tuberculosis* clinical isolates apparently produce PDIM, only a subset of isolates produce the closely related phenolic glycolipids (PGLs). The PGLs comprise a PDIM lipid core that is terminated by a glycosylated aromatic nucleus (26). A 7-bp deletion in the *pks15/1* polyketide synthase locus, resulting in a translational frameshift, is responsible for the lack of PGL production by the H37Rv laboratory strain and all other strains of the Euro-American *M. tuberculosis* lineage (5, 10). This frameshift abolishes production of the Pks15/1 polyketide synthase that is responsible for biosynthesis of the PGL precursor *p*-hydroxyphenylalkanoate from *p*-hydroxybenzoic acid (5). In the absence of Pks15/1 function, glycosylated *p*-hydroxybenzoic acid methyl esters (*p*-HBADs) accumulate and are released into the culture medium (5, 25). Three glycosyltransferases, encoded by the *rv2962c*, *rv2958c*, and *rv2957* genes, are thought to add, successively, two rhamnosyl residues and one fucosyl residue to the *p*-hydroxybenzoic acid moiety (Fig. 1B) (25). *M. tuberculosis* H37Rv produces the triglycosylated *p*-HBAD (*p*-HBAD-II). Disruption of *rv2962* abolishes *p*-HBAD production entirely; disruption of the *rv2958c* or *rv2957* gene results in the production of monoglycosylated *p*-HBAD (*p*-HBAD-I) (25).

In *M. tuberculosis* strains that produce PGL, a role for PGL in immune modulation and virulence has been reported (26). Whether the *p*-HBAD moieties secreted by PGL-negative *M. tuberculosis* strains might also play a role in immune modulation and virulence is not known. Previously, we described a

genetic screen that was designed to identify *M. tuberculosis* genes involved in countering IFN-γ-dependent host immune mechanisms other than NOS2 (12). Disruption of these "counterimmune" (*cim*) genes severely attenuated growth and virulence in NOS2$^{-/-}$ mice but had little or no impact on bacterial growth and virulence in IFN-γ$^{-/-}$ mice. One of the *cim* mutants identified in this study contained a transposon (Tn) insertion in the *rv2958c* gene, suggesting that secreted *p*-HBAD-II might, like full-length PGL, modulate the interaction of the bacterium and the host immune response. Here, we describe further studies to elucidate the role of the *rv2962c*, *rv2958c*, and *rv2957* glycosyltransferases in *M. tuberculosis* virulence and counterimmunity. In contrast to our previous report, we find that these genes are dispensable for bacterial growth and survival in wild-type (C57BL/6) and NOS2$^{-/-}$ mice. We demonstrate that the phenotypes we had previously ascribed to disruption of the *rv2958c* gene are attributable, instead, to the spontaneous loss of PDIM production in the *rv2958c*::Tn mutant. Whole-genome resequencing of a spontaneous PDIM-deficient variant that arose during *in vitro* cultivation of *M. tuberculosis* H37Rv identified a single-nucleotide polymorphism (SNP) in the *ppsD* gene, which encodes a modular polyketide synthase putatively involved in PDIM biosynthesis (34). We demonstrate by complementation and reversion that the spontaneous point mutation in the *ppsD* gene is responsible for both the defect in PDIM production and attenuation of virulence in mice. We additionally show that the spontaneous PDIM deficient variant has an *in vitro* growth advantage that allows it to replace the PDIM-proficient parental strain during subculture. We suggest that spontaneous loss of PDIM production is likely to be a common phenomenon that calls for a reexamination of published genetic studies of *M. tuberculosis* in which complementation analysis was not done or was unsuccessful.

MATERIALS AND METHODS

Bacteriology. *M. tuberculosis* strains from the McKinney lab were H37Rv (parental strain) and Tn*5370* (Tn) mutagenized derivatives *rv0072*::Tn, *rv0405*::Tn, and *rv2958c*::Tn, described previously (12). *M. tuberculosis* strains from the Guilhot lab were H37Rv (parental strain), Δ*rv2957*, Δ*rv2958c*, Δ*rv2959c*, and Δ*rv2962c*, described previously (24, 25). Bacteria were cultured at 37°C in Middlebrook 7H9 broth (Difco) containing 10% albumin-dextrose-saline, 0.5% glycerol, and 0.05% Tween 80 or on Middlebook 7H10 agar (Difco) containing 10% oleic acid-albumin-dextrose-catalase (BD Biosciences) and 0.5% glycerol. Cycloheximide was added at 10 μg ml^{-1} to prevent fungal contamination, as needed. Kanamycin (15 μg ml^{-1}), hygromycin (50 μg ml^{-1}), and sucrose (2%) were added to the growth media, as needed. Frozen stocks were prepared by growing liquid cultures in 7H9 broth to mid-log phase (optical density at 600 nm [OD_{600}] = 0.5) and freezing in aliquots at −80°C after the addition of glycerol (15% [vol/vol]).

Mouse infections. Male and female C57BL/6, NOS2$^{-/-}$, and IFN-γ$^{-/-}$ mice, 5 to 8 weeks of age, were purchased from Jackson Laboratories and housed in The Rockefeller University's Laboratory Animal Research Center or the EPFL Center of Phenogenomics under specific-pathogen-free conditions. Bacteria were grown to mid-log phase (OD_{600} = 0.5) in 7H9 broth, collected by centrifugation (2,500 × *g*, 15 min), resuspended in phosphate-buffered saline containing 0.05% Tween 80 (PBST), and centrifuged at a low speed (150 × *g*, 5 min) to remove clumps. The declumped supernatant was adjusted to an OD_{600} of 0.1 (~10^8 CFU ml^{-1}) and further diluted 2-fold before being loaded into the nebulizer. Mice were infected by the aerosol route with ~100 CFU using a custom-built aerosol exposure chamber (Department of Mechanical Engineering, University of Wisconsin, Madison, WI) and an exposure time of 15 min, as described previously (37). Infected mice were euthanized by CO_2 overdose. Bacterial CFU were enumerated by plating serially diluted lung homogenates on 7H10 agar and counting colonies after 3 to 4 weeks at 37°C. For survival experiments, infected mice were monitored twice daily, and animals that showed signs of illness (ruffled fur, immobility, hunched posture, labored breathing) were euthanized by CO_2 overdose and scored as "died." The animal protocols for these studies were reviewed and approved by the Institutional Animal Care and Use Committee (IACUC) of The Rockefeller University and by the chief veterinarian of the Swiss Federal Institute of Technology Lausanne (EPFL), the Service de la Consommation et des Affaires Vétérinaires of the Canton of Vaud, and the Swiss Office Vétérinaire Fédéral.

TABLE 1. Oligonucleotide primers[a]

Name	Sequence (5′ to 3′)
ppsAF	GCGAGGACCTGGTCGGTATC
ppsAR	GGCCTTGTTGAGGTTGGTC
ppsBF	GAACTCTGCCACGAGCTGG
ppsBR	GCACCGATGACGAGCTG
ppsDF	GTCTTAATTAAGGAAACCCTGGGACTCGAC
ppsDR	TAGGCGCGCCGCCAAGTGAATTGCCACCAG
ppsDF2	CTGGAATTC*TAAGAAGGAGATATACAT*ATGAC AAGTCTGGCGGAGC
ppsDR2	CTGGTTAACTCGGGGATGCTCACAGGTC
ppsDR4	AAGCTTGAGGGCGGATGTGAT
MVinsF	AGCGAGGACAACTTGAGC
ppsDF4	CGATTCGCGCTCAGCTAGAC
pJGR	AAATGCCGATATCCTATTGGC
pJGF	GTGGACCTCGACGACCTC

[a] Restriction enzyme sites are underlined. The strong ribosome-binding site in primer ppsDF2 is indicated in italics.

***p*-HBAD-II glycosylation and PDIM production.** *p*-HBAD-II biosynthesis by wild-type and mutant strains of *M. tuberculosis* was analyzed by thin-layer chromatography (TLC) of extracted and purified glycolipids, as described previously (25). PDIM biosynthesis was analyzed by growing bacteria to mid-log phase and labeling 10 ml of culture for 24 to 48 h with 10 μCi of [1-^{14}C]-propionate (specific activity of 54 Ci mol^{-1} [Amersham] or 55.9 Ci mol^{-1} [Campro Scientific]). Apolar lipids were extracted essentially as described previously (32). Labeled cells were collected by centrifugation (2,500 × *g*, 10 min), resuspended in 5 ml of 10:1 (vol/vol) methanol/0.3% NaCl, and 5 ml of petroleum ether was added. Samples were vortexed vigorously for 4 min and phase-separated by centrifugation (750 × *g*, 10 min). The upper layer was removed, and the extraction was repeated with an additional 5 ml of petroleum ether. Remaining bacteria in the combined petroleum ether fraction were killed by the addition of an equal volume of chloroform. Extracts were reduced to ~10 ml by overnight evaporation and spotted (25 to 30 μl) on a silica gel 60 F_{254} TLC plate (5 by 10 cm; Merck). TLC plates were developed in petroleum ether/diethyl ether (9:1 [vol/vol]), air-dried, and visualized using a Typhoon PhosphorImager (Amersham Biosciences).

Whole-genome resequencing. High-quality genomic DNA was prepared from PDIM-positive and PDIM-negative isolates of H37Rv by the cetyltrimethylammonium bromide (CTAB)-lysozyme method (17). Genomic DNA fragment sequencing libraries were prepared using a genomic DNA sample prep kit (Illumina) according to the manufacturer's instructions, with 5 μg of purified genomic DNA. Each genomic DNA fragment library was sequenced on one lane of an Illumina genome analyzer IIx sequencing chip using a single-read cluster generation kit v2 (Illumina) and a 36-cycle sequencing kit v2 (Illumina). Image analysis and base calling were done using the Illumina Pipeline software package, v1.32.

A total of 6.2 and 5.3 million reads 36 bases in length were obtained for the PDIM-positive and PDIM-negative H37Rv clones, respectively. These sequence reads were mapped to the reference *M. tuberculosis* H37Rv sequence using Maq v0.7.1 via ungapped alignments allowing up to two mismatches per read. This method mapped 95% and 89% of the PDIM-positive and PDIM-negative sequence reads, respectively, to the H37Rv genome. Maq was also used for SNP calling and filtering out low-quality SNPs using the SNP filter designed for single-end reads (18). SNPs identified in the *ppsA* and *ppsD* genes were further confirmed by PCR and sequencing using the primer pairs ppsAF/ppsAR and ppsDF/ppsDR (Table 1).

Plasmid construction. Oligonucleotide primers used for plasmid construction are listed in Table 1. For complementation of the *ppsD*(G44C) mutation, the full-length *ppsD* gene was amplified from PDIM-positive H37Rv genomic DNA using primers ppsDF2 and ppsDR2 that contain EcoRI and HpaI

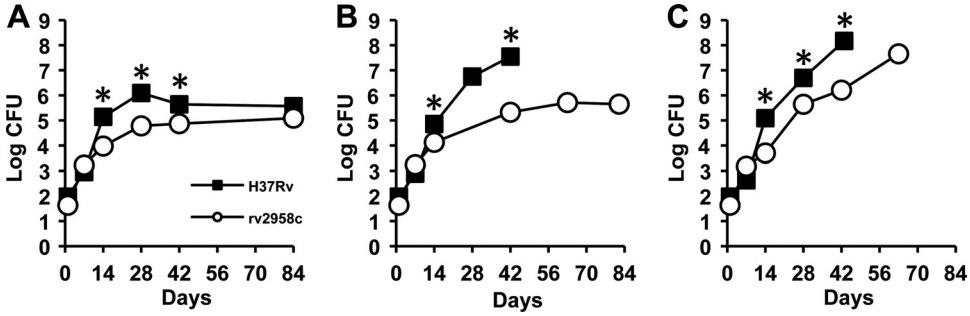


FIG. 2. Growth kinetics of *M. tuberculosis* strains H37Rv (wild type) and *rv2958c*::Tn in wild-type and immunodeficient mice. C57BL/6 (A), NOS2$^{-/-}$ (B), and IFN-$\gamma^{-/-}$ (C) mice were aerosol infected with *M. tuberculosis* strains H37Rv (squares) or *rv2958c*::Tn (circles). These strains were described previously (12). Groups of mice were sacrificed at the indicated time points, and bacterial CFU were enumerated by plating lung homogenates on 7H10 agar and scoring colonies after 3 to 4 weeks of incubation at 37°C. Symbols represent means ($n = 4$ or 5 mice per group per time point); error bars indicate standard errors. Asterisks indicate a statistically significant difference ($P < 0.05$) between the groups. Representative results of two independent experiments are shown.

restriction sites, respectively. The ppsDF2 primer additionally harbors a strong ribosome-binding site immediately upstream of the translational start site of the *ppsD* gene. The resulting PCR products were cloned in pCR2.1-TOPO (Invitrogen) and sequenced. Clones were identified that contained wild-type *ppsD* sequence between a unique NcoI restriction site and the 3′ end of the *ppsD* gene and between a unique HindIII site and the unique NcoI site. The first 979 bp of the *ppsD* gene 5′ to the unique HindIII site were amplified by PCR from PDIM-positive H37Rv genomic DNA using primers ppsDF2 and ppsDR4, cloned in pCR2.1-TOPO, and sequenced. The *ppsD* 5′-HindIII fragment was digested out of the pCR2.1 cloning vector with EcoRI and HindIII and cloned downstream of the strong constitutive *hsp60* promoter in pMV361, a vector that integrates at the *attB* site on the *M. tuberculosis* chromosome and contains a Kan^r marker for selection (33). The resulting pMV361-*ppsD* 5′-HindIII construct was digested with HindIII and HpaI, and the HindIII-NcoI and NcoI-3′ fragments of the *ppsD* gene were ligated together in the plasmid. The resulting full-length *ppsD* complementation construct, pAT223, was confirmed by sequencing.

For allelic exchange of the *ppsD*(G44C) point mutation, the wild-type *ppsD* sequence approximately 450 bp up- and downstream of the point mutation was amplified from PDIM-positive H37Rv genomic DNA using primers ppsDF and ppsDR, which contain PacI and AscI restriction sites, respectively. The resulting PCR product was digested with PacI and AscI and ligated to PacI/AscI-digested pJG1100, a suicide vector that contains Kan^r and Hyg^r markers for selection of recombinants and the *sacB* counterselectable marker that confers sucrose sensitivity, to generate pAT221. The construct was confirmed by sequencing.

Strain construction. Oligonucleotide primers used in PCR confirmation of strains are listed in Table 1. For complementation of the *ppsD*(G44C) point mutation with pAT223, the plasmid was electroporated into the PDIM-negative H37Rv *ppsD*(G44C) strain, and Kan^r colonies were selected. The presence of pAT223 was confirmed by PCR using primers MVinsF and ppsDR4. Reversion of the *ppsD*(G44C) point mutation to the wild-type sequence was accomplished by a two-step homologous recombination method. The pAT221 allelic exchange vector was electroporated into the PDIM-negative H37Rv *ppsD*(G44C) strain, and Kan^r Hyg^r colonies were selected. Isolates containing the pAT221 vector integrated at the *ppsD* gene were identified by PCR with the primer pairs ppsDF4/pJGR and pJGF/ppsDR4. These isolates were grown in 7H9 medium without antibiotic selection to mid-log phase and plated on 7H10 agar containing 2% sucrose to select isolates that had undergone a second recombination. Excision of the plasmid in sucrose-resistant isolates was confirmed by PCR using primers ppsDF4/ppsDR4, and the resulting PCR product was sequenced to determine whether the wild-type *ppsD* sequence or the *ppsD*(G44C) mutant sequence was present.

Statistics. Student's unpaired *t* test (one-tailed) was used to assess statistical significance of pairwise comparisons between groups of mice infected with PDIM-positive or PDIM-negative bacteria. The Mantel-Cox log-rank test was used for comparison of Kaplan-Meier plots of mouse survival. *P* values were calculated using GraphPad Prism 4.0 software (GraphPad Software, Inc.). *P* values of <0.05 were considered significant.

RESULTS

A strain of *M. tuberculosis* with a Tn insertion in the *rv2958c* gene is attenuated in wild-type, NOS2$^{-/-}$, and IFN-$\gamma^{-/-}$ mice. Previously, we reported the results of a pilot signature-tagged mutagenesis screen to identify *M. tuberculosis* genes involved in countering the impact of IFN-γ-dependent immune mechanisms other than NOS2 (12). This was accomplished by parallel screening of Tn-induced mutants in intravenously infected gene knockout mice to identify mutants that were attenuated for growth and virulence in NOS2$^{-/-}$ mice but unimpaired for growth and virulence in IFN-$\gamma^{-/-}$ mice. One of the mutants identified in this screen contained a Tn insertion in the *rv2958c* gene, encoding a putative rhamnosyl transferase involved in biosynthesis of *p*-HBAD-II (25). To confirm the phenotypes observed in the screen, we infected mice by the aerosol route with the H37Rv parental strain or with the *rv2958c*::Tn mutant derived from it. Consistent with the results of the screen, growth of the *rv2958c*::Tn mutant was attenuated in C57BL/6 (wild-type) mice (Fig. 2A) and in NOS2$^{-/-}$ mice (Fig. 2B). In contrast to our previous results, however, we found that growth of the *rv2958c*::Tn mutant was also somewhat attenuated in IFN-$\gamma^{-/-}$ mice (Fig. 2C). We do not know the reason for this discrepancy. A possible explanation is that screening and retesting of mutants in our previous report were done in mice infected by the intravenous route (12), whereas the experiments reported here were done in mice infected by the respiratory route, which is the natural route of infection.

Role of *p*-HBAD-II in *M. tuberculosis* growth in C57BL/6 and NOS2$^{-/-}$ mice. Concurrent with our report describing the results of our pilot screen to identify *M. tuberculosis* "counter-immune" (*cim*) mutants (12), another group reported the construction of deletion mutations in the *M. tuberculosis rv2962c*, *rv2958c*, and *rv2957* genes and the roles of the putative glycosyl transferases encoded by these genes in the biosynthesis of *p*-HBAD-II (Fig. 1) (25). In a separate report, the same group also described the construction of a deletion mutation in the *rv2959c* gene, encoding a methyltransferase responsible for O-methylation of the first rhamnosyl residue linked to the *p*-hydroxybenzoic acid moiety of *p*-HBAD-II (Fig. 1) (24). In order to elucidate the contributions of these genes to *M. tu-*

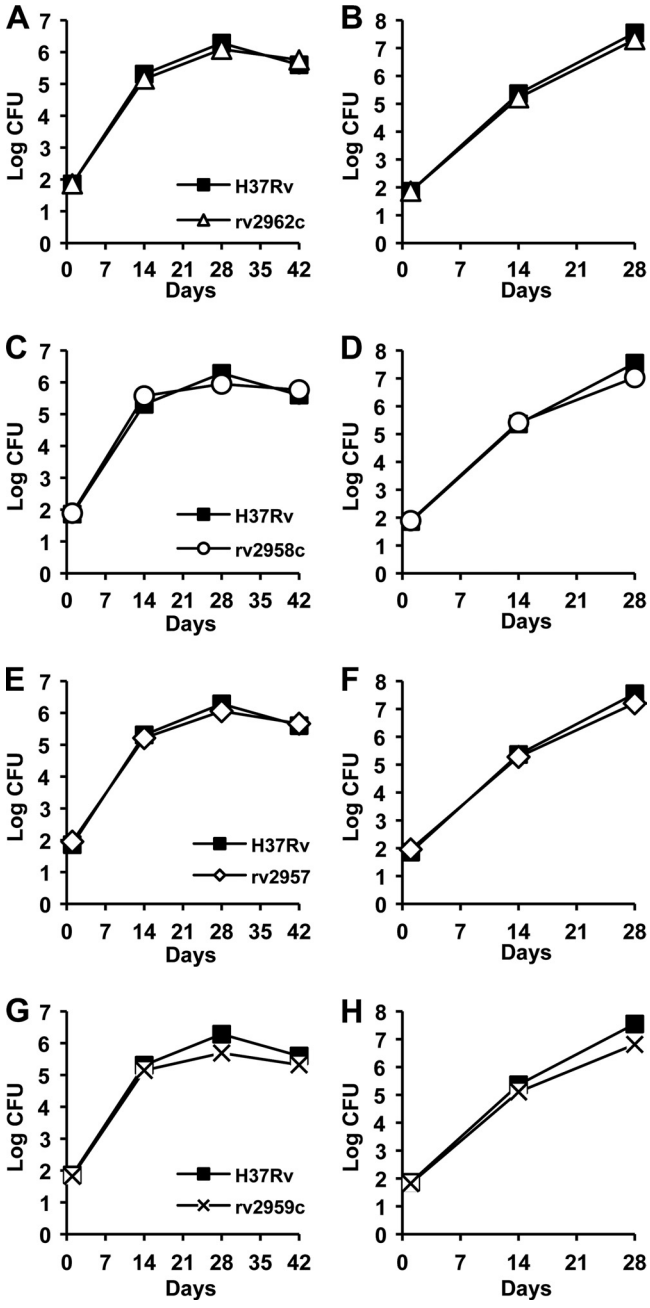


FIG. 3. *p*-HBAD-II biosynthesis is not required for *M. tuberculosis* growth and survival in wild-type or immunodeficient mice. C57BL/6 (A, C, E, and G) and NOS2$^{-/-}$ (B, D, F, and H) mice were aerosol infected with *M. tuberculosis* strains H37Rv (A to H, squares), Δ*rv2962c* (A and B, triangles), Δ*rv2958c* (C and D, circles), Δ*rv2958c* (E and F, diamonds), or Δ*rv2959c* (G and H, crosses). These strains were described previously (24, 25). Groups of mice were sacrificed at the indicated time points, and bacterial CFU were enumerated by plating lung homogenates on 7H10 agar and scoring colonies after 3 to 4 weeks of incubation at 37°C. Symbols represent means (n = 4 or 5 mice per group per time point); error bars indicate standard errors. This experiment was performed once.

berculosis virulence and counterimmunity, our groups jointly investigated the growth kinetics and virulence of the Δ*rv2962c* (Fig. 3A and B), Δ*rv2958c* (Fig. 3C and D), Δ*rv2957* (Fig. 3E and F), and Δ*rv2959c* (Fig. 3G and H) mutants in aerosol-infected mice. We found that none of these mutants was attenuated for growth and persistence in aerosol-infected C57BL/6 (Fig. 3A, C, E, and G) or NOS2$^{-/-}$ (Fig. 3B, D, F, and H) mice. These results were in sharp contrast to the phenotype of the *rv2958c*::Tn mutant isolated in our screen, which was markedly attenuated in both C57BL/6 and NOS2$^{-/-}$ mice (Fig. 2) (12).

Loss of PDIM production in the *M. tuberculosis rv2958c*::Tn strain. Analysis of bacterial cell wall lipid fractions by thin-layer chromatography revealed that the *rv2958c*::Tn mutant (12) was deficient in PDIM production (Fig. 4A, lane 3), compared to the H37Rv parental strain (Fig. 4A, lane 2). Loss of Rv2958c function *per se* was not responsible for loss of PDIM, because we found that the Δ*rv2958c* mutant (25) produced wild-type levels of PDIM (not shown). Two additional mutants identified in our counterimmune screen (12), containing Tn insertions in the *rv0072* (Fig. 4B, lane 4) and *rv0405* (Fig. 4B, lane 5) genes, were also found to be PDIM deficient. We speculated that loss of PDIM production might have occurred during the isolation of the *rv2958c*::Tn, *rv0072*::Tn, and *rv0405*::Tn mutants, which involves clonal selection and expansion of bacteria arising from a single transposition event. We therefore derived eight independent cell lines from our parental H37Rv frozen stock, which had been used to generate the Tn-induced mutants that were screened in mice (12), by plating for single colonies on nonselective media and analyzing the clonal cell lines derived from randomly selected individual colonies for PDIM production. Of the eight clonal cell lines that we analyzed, four were PDIM deficient (not shown), suggesting that our parental H37Rv stock (Fig. 4A, lane 2; Fig. 4B, lane 1) comprised a mixture of PDIM-positive and PDIM-negative variants.

Consistent with this interpretation, we found that PDIM production was dramatically reduced in the McKinney lab's parental H37Rv strain (Fig. 4A, lane 2) compared to that in the Guilhot lab's H37Rv strain (Fig. 4A, lane 1). We also found that PDIM production by the McKinney lab's parental H37Rv strain was undetectable after only four rounds of *in vitro* subculture from frozen stocks (Fig. 4B, lane 2). These observations suggest that spontaneously arising PDIM deficiency might confer a growth advantage *in vitro*, thus promoting replacement of the parental PDIM-positive strain with PDIM-negative variants during *in vitro* passage. To test this idea, we compared the *in vitro* growth kinetics of independently derived PDIM-positive and PDIM-negative H37Rv clones (see above) and found that the PDIM-negative bacteria had a significant and reproducible growth advantage over the PDIM-positive bacteria (Fig. 4C). We found that the *in vitro* growth advantage of PDIM-negative variants was particularly pronounced during the early outgrowth of cultures inoculated from frozen stocks (data not shown), which might reflect differential recovery of PDIM-positive and PDIM-negative strains from the stress caused by freezing/thawing.

Growth kinetics of PDIM-negative H37Rv, *rv2958c*::Tn, *rv0072*::Tn, and *rv0405*::Tn strains in C57BL/6, NOS2$^{-/-}$, and IFN-γ$^{-/-}$ mice. Previously, we reported that *M. tuberculosis* *rv0072*::Tn, *rv0405*::Tn, and *rv2958c*::Tn mutants were severely attenuated for growth and virulence in intravenously infected NOS2$^{-/-}$ mice but were only slightly attenuated in IFN-γ$^{-/-}$ mice (12). Our subsequent discovery that these strains were

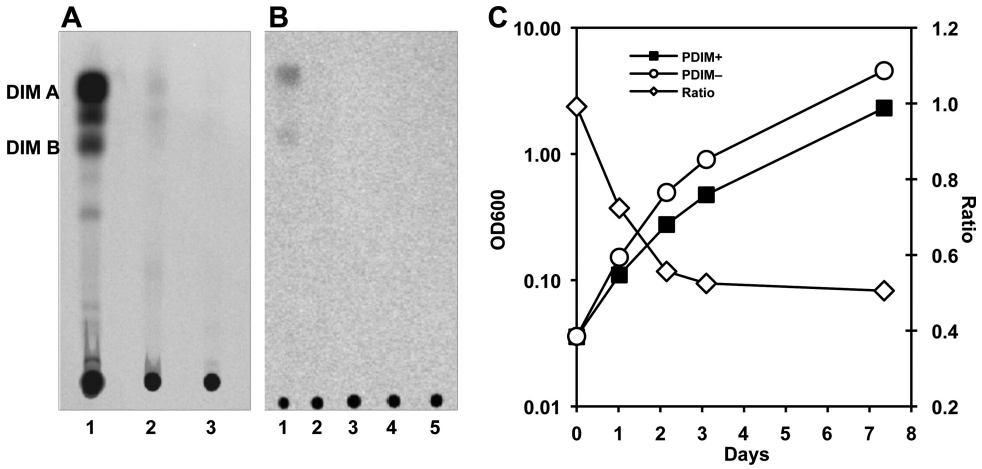


FIG. 4. PDIM deficiency confers an *in vitro* growth advantage in *M. tuberculosis* H37Rv. (A and B) Thin-layer chromatographic analysis of PDIM biosynthesis. Bacteria were labeled with [^{14}C]propionate, which preferentially labels PDIM (6), and cell wall lipids were extracted and separated by thin-layer chromatography. (A) *M. tuberculosis* strains. H37Rv, Guilhot lab (25) (lane 1); H37Rv, McKinney lab (12) (lane 2); *rv2958c*::Tn (12) (lane 3). (B) *M. tuberculosis* strains, McKinney lab (12). H37Rv (lane 1), H37Rv after subculture (lane 2), *rv2958c*::Tn (lane 3), *rv0072*::Tn (lane 4), *rv0405*::Tn (lane 5). (C) Independently derived subclones of PDIM-positive H37Rv (squares) and PDIM-negative H37Rv (circles) were grown in 7H9 broth with aeration at 37°C. Growth of the cultures was monitored by withdrawing aliquots and measuring the OD_{600} at the indicated time points (plotted on the primary *y* axis). The (PDIM-positive OD_{600})/(PDIM-negative OD_{600}) ratios at each time point are plotted on the secondary *y* axis (diamonds). Results are representative of three independent experiments.

PDIM deficient (Fig. 4A), presumably due to selection of preexisting PDIM-negative variants in our parental H37Rv stock during Tn mutagenesis, suggested that the attenuation of these mutants in mice might be due, at least in part, to their PDIM deficiency. To test this idea, we compared the growth kinetics of these mutants with a PDIM-negative H37Rv clone in aerosol-infected mice (Fig. 5). We found that the *rv0072*::Tn (Fig. 5A to C), *rv0405*::Tn (Fig. 5D to F), and *rv2958c*::Tn (Fig. 5G to I) mutants grew with kinetics similar to those of the PDIM-negative H37Rv clone in C57BL/6 (Fig. 5A, D, and G), NOS2$^{-/-}$ (Fig. 5B, E, and H), and IFN-$\gamma^{-/-}$ (Fig. 5C, F, and I) mice. These observations suggest that the *in vivo* phenotypes we reported previously for these mutants (12) were probably due to the spontaneous loss of PDIM production in these strains, rather than disruption of the *rv2958c*, *rv0072*, or *rv0405* genes *per se*.

Growth kinetics of PDIM-positive and PDIM-negative clones of H37Rv in C57BL/6, NOS2$^{-/-}$, and IFN-$\gamma^{-/-}$ mice. To further test the idea that PDIM deficiency might contribute to the *in vivo* attenuation of our *rv2958c*::Tn, *rv0072*::Tn, and *rv0405*::Tn mutants, we compared the growth kinetics of PDIM-positive and PDIM-negative clones of H37Rv in aerosol-infected mice. We found that, compared to the PDIM-positive H37Rv clone, the PDIM-negative H37Rv clone was markedly attenuated for growth in C57BL/6 (Fig. 6A), NOS2$^{-/-}$ (Fig. 6B), and IFN-$\gamma^{-/-}$ (Fig. 6C) mice. Virulence of the PDIM-negative H37Rv clone, measured in terms of survival time of infected mice, was also attenuated in NOS2$^{-/-}$ mice ($P = 0.0005$) and, to a lesser extent, in IFN-$\gamma^{-/-}$ mice ($P = 0.002$) (Fig. 6D). Median survival time (MST) of NOS2$^{-/-}$ mice was >200 days after infection with PDIM-negative H37Rv, compared to 64 days postinfection for NOS2$^{-/-}$ mice infected with PDIM-positive H37Rv (Fig. 6D). The MST of IFN-$\gamma^{-/-}$ mice was also longer for mice infected with PDIM-negative H37Rv (MST, 82 days) compared to mice infected with PDIM-positive H37Rv (MST, 59 days) (Fig. 6D). These data strongly suggest that the *in vivo* attenuation of the "counterimmune" mutants described in our previous report (12) can be attributed to the fortuitous loss of PDIM production in these strains.

Whole-genome resequencing of PDIM-negative H37Rv identifies a point mutation, *ppsD*(G44C), responsible for PDIM deficiency. Since our results demonstrated a correlation between spontaneous loss of the ability to produce PDIM and attenuation in the mouse model of infection, we sought to identify the mutation responsible for PDIM deficiency in a PDIM-negative clone. Analysis of the PDIM biosynthesis locus (Fig. 1) by Southern blotting did not identify any major insertions or genetic rearrangements (data not shown), suggesting that a point mutation might be the cause of PDIM deficiency. To identify putative point mutations that could contribute to PDIM deficiency, we performed whole-genome resequencing of PDIM-positive and PDIM-negative H37Rv clones using an Illumina genome analyzer platform. In comparison to the H37Rv reference sequence, our PDIM-positive H37Rv clone had 161 putative SNPs and 15 putative small sequence insertions or deletions (indels). The PDIM-negative H37Rv clone had 151 putative SNPs and 15 putative indels. These sequence alterations included 72 polymorphisms (57 SNPs and 15 indels) that were recently identified by whole-genome resequencing of six H37Rv isolates and that are likely to be errors within the H37Rv reference sequence (14). Analysis of the remaining sequence polymorphisms, excluding those in the highly repetitive GC-rich PE_PGRS coding regions, revealed 22 putative SNPs unique to the PDIM-negative H37Rv clone.

Among the SNPs present only in the PDIM-negative H37Rv isolate were two mutations in genes required for PDIM biosynthesis: a putative A to C at position 2544 in the *ppsA* gene and a G to T at position 130 in the *ppsD* gene. Only the G to T in the *ppsD* gene was predicted to be nonsynonymous, gen-

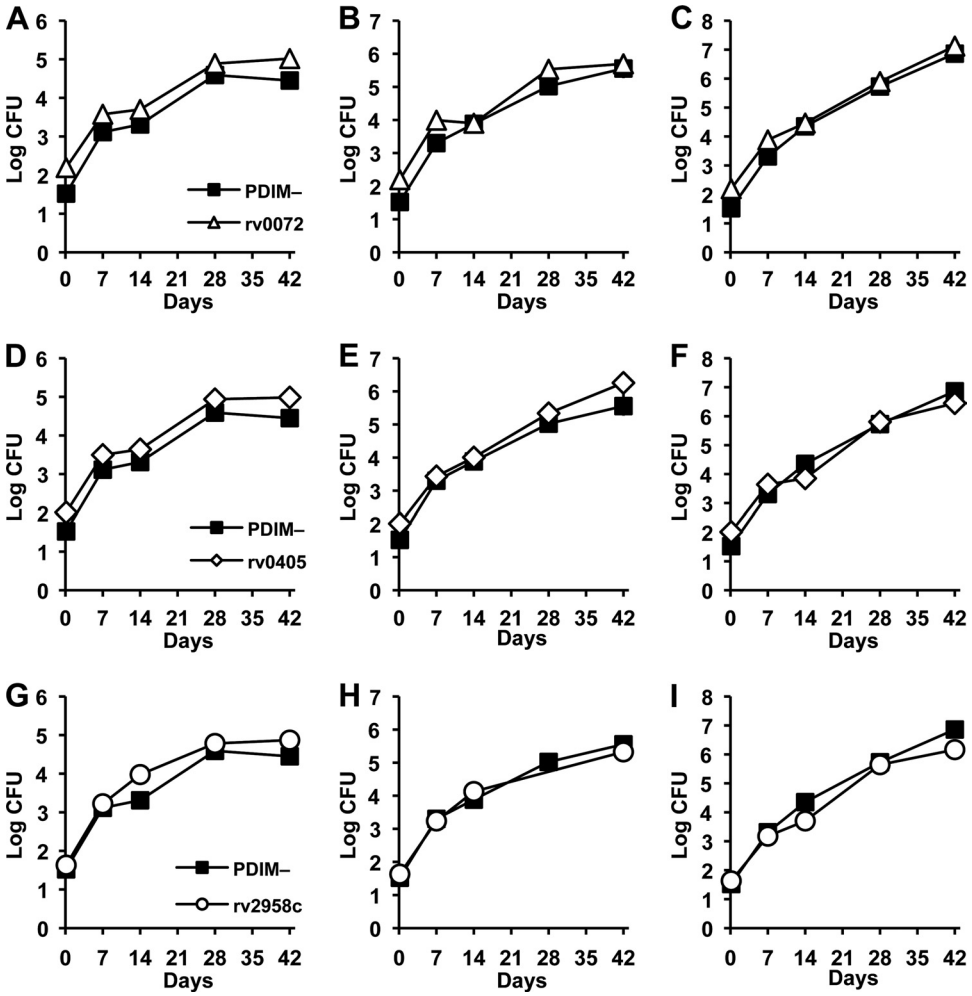


FIG. 5. Growth kinetics of PDIM-negative H37Rv and *M. tuberculosis rv0072*::Tn, *rv0405*::Tn, and *rv2958c*::Tn mutants in wild-type and immunodeficient mice. C57BL/6 (A, D, and G), NOS2$^{-/-}$ (B, E, and H), and IFN-$\gamma^{-/-}$ (C, F, and I) mice were aerosol infected with *M. tuberculosis* H37Rv PDIM-negative variant (A to I, squares), *rv0072*::Tn (A to C, triangles), *rv0405*::Tn (D to F, diamonds), or *rv2958c*::Tn (G to I, circles). The PDIM-negative variant of H37Rv is described herein; the Tn mutant strains were described previously (12). Groups of mice were sacrificed at the indicated time points, and bacterial CFU were enumerated by plating lung homogenates on 7H10 agar and scoring colonies after 3 to 4 weeks of incubation at 37°C. Symbols represent means (n = 4 or 5 mice per group per time point); error bars indicate standard errors. This experiment was performed once.

erating a glycine-to-cysteine transition at position 44 in the PpsD protein. PpsD is a modular polyketide synthase, and the G44C point mutation identified in our PDIM-negative clone is located within the putative β-ketoacyl acyl carrier protein synthase enzymatic domain (34). We reasoned that the G44C mutation might interfere with the activity of PpsD, since an additional cysteine residue could cause formation of aberrant disulfide bonds, thereby preventing proper protein folding or interfering with the function of the active site cysteine residue. We confirmed that the *ppsD*(G44C) point mutation was unique to the PDIM-negative clone by PCR and sequencing. We were not able to confirm the point mutation in the *ppsA* gene by sequencing of a PCR product, suggesting that there are likely to be some false positives among the SNPs identified by whole-genome resequencing.

We attempted to restore PDIM production to the PDIM-negative H37Rv *ppsD*(G44C) mutant using two approaches. We complemented the *ppsD*(G44C) mutation in *trans* by providing a wild-type copy of the *ppsD* gene under the control of a strong constitutive promoter on the vector pMV361, which integrates in the *M. tuberculosis* chromosome at the unique *attB* site. We also reverted the *ppsD*(G44C) mutation in the PDIM-negative clone to the wild-type *ppsD* sequence by a two-step homologous recombination method. Both strains were tested for the ability to produce PDIM *in vitro* by thin-layer chromatography of extractable cell wall lipids. Production of PDIM was restored by either complementation or reversion (*ppsD* rev) of the *ppsD*(G44C) point mutation (Fig. 7A), confirming that this spontaneous mutation is responsible for the deficiency in PDIM production of our PDIM-negative H37Rv clone.

The *ppsD*(G44C) point mutation enhances growth of H37Rv *in vitro*. Since the *ppsD*(G44C) spontaneous point mutation was responsible for PDIM deficiency, we sought to establish whether the *in vitro* and *in vivo* growth characteristics of our PDIM-negative H37Rv clone could also be attributed to this

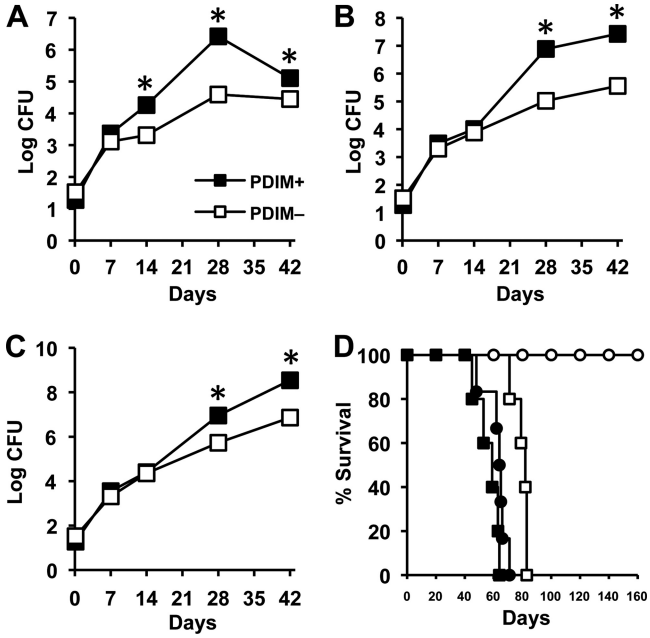


FIG. 6. Growth kinetics of PDIM-positive and PDIM-negative subclones of *M. tuberculosis* H37Rv in wild-type and immunodeficient mice. (A to C) C57BL/6 (A), NOS2$^{-/-}$ (B), and IFN-$\gamma^{-/-}$ (C) mice were aerosol-infected with PDIM-positive (filled squares) or PDIM-negative (open squares) subclones of *M. tuberculosis* H37Rv. Groups of mice were sacrificed at the indicated time points, and bacterial CFU were enumerated by plating lung homogenates on 7H10 agar and scoring colonies after 3 to 4 weeks of incubation at 37°C. Symbols represent means ($n = 4$ or 5 mice per group per time point); error bars indicate standard errors. Asterisks indicate statistically significant differences ($P < 0.05$) between the groups. (D) Survival of NOS2$^{-/-}$ (circles) and IFN-$\gamma^{-/-}$ (squares) mice ($n = 5$ or 6 per group) after aerosol infection with PDIM-positive (filled symbols) or PDIM-negative (open symbols) subclones of *M. tuberculosis* H37Rv. This experiment was performed once.

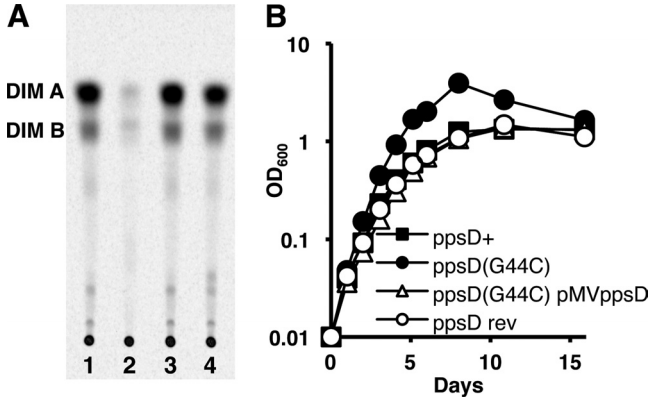


FIG. 7. The *ppsD*(G44C) point mutation is responsible for PDIM deficiency and the *in vitro* growth advantage of the H37Rv PDIM-negative subclone. (A) Thin-layer chromatographic analysis of PDIM biosynthesis. Bacteria were labeled with [^{14}C]propionate, which is preferentially incorporated into PDIM (6), and cell wall lipids were extracted and separated by thin-layer chromatography. *M. tuberculosis* H37Rv strains: *ppsD*+ (lane 1), *ppsD*(G44C) (lane 2), *ppsD* rev (lane 3), *ppsD*(G44C) pMV361-*ppsD* (lane 4). Results are representative of two independent experiments. (B) Bacteria were grown in 7H9 broth with aeration at 37°C. Growth was monitored by withdrawing aliquots and measuring the OD_{600} at the indicated time points. *M. tuberculosis* H37Rv strains: *ppsD*+ (filled squares), *ppsD*(G44C) (filled circles), *ppsD* rev (open circles), *ppsD*(G44C) pMV361-*ppsD* (open triangles). Results are representative of three independent experiments.

point mutation. We compared the *in vitro* growth kinetics of the PDIM-positive *ppsD*+ clone, the PDIM-negative *ppsD*(G44C) mutant, the *ppsD*(G44C) pMV-*ppsD* complemented mutant, and the strain in which the *ppsD* gene was reverted to the wild-type sequence (*ppsD* rev) in standard liquid culture conditions. The PDIM-negative *ppsD*(G44C) mutant consistently grew at a higher rate and to a higher cell density in liquid culture than the PDIM-positive *ppsD*+ clone (Fig. 7B). Either complementation or reversion of the *ppsD*(G44C) point mutation restored growth to a rate similar to the PDIM-positive *ppsD*+ clone (Fig. 7B). These data demonstrate that the *ppsD*(G44C) point mutation is responsible for the enhanced *in vitro* growth rate of the PDIM-negative H37Rv clone.

The *ppsD*(G44C) point mutation attenuates replication and virulence of H37Rv in mice. Next, we tested whether the spontaneous *ppsD*(G44C) mutation was also responsible for attenuation of the PDIM-negative H37Rv clone in the mouse model of infection. For these experiments, we focused on the strain in which *ppsD* was reverted to the wild-type sequence (*ppsD* rev), because the complemented strain exhibited a moderately reduced *in vitro* growth rate (Fig. 7B). The PDIM-positive *ppsD*+, PDIM-negative *ppsD*(G44C), and *ppsD* rev strains were introduced into the lungs of C57BL/6, NOS2$^{-/-}$, and IFN-$\gamma^{-/-}$ mice by the aerosol route, and replication of the bacteria in the lungs and survival of the mice were monitored. Consistent with previous experiments, the PDIM-negative *ppsD*(G44C) mutant was attenuated for growth in the lungs of C57BL/6 (Fig. 8A) and NOS2$^{-/-}$ (Fig. 8B) mice. Virulence of the PDIM-negative *ppsD*(G44C) mutant, measured by survival time of infected mice, was likewise modestly attenuated in IFN-$\gamma^{-/-}$ mice (Fig. 8C) ($P = 0.0018$) and dramatically attenuated in NOS2$^{-/-}$ mice (Fig. 8D) ($P = 0.0025$). Reversion of the *ppsD*(G44C) spontaneous point mutation to the wild-type *ppsD* sequence reversed each of these phenotypes. The *ppsD* rev strain replicated in the lungs of C57BL/6 (Fig. 8A) and NOS2$^{-/-}$ (Fig. 8B) mice with kinetics identical to those of the PDIM-positive *ppsD*+ clone. In addition, reversion to the wild-type *ppsD* sequence restored virulence in IFN-$\gamma^{-/-}$ (Fig. 8C) and NOS2$^{-/-}$ (Fig. 8D) mice. Survival times of IFN-$\gamma^{-/-}$ and NOS2$^{-/-}$ mice infected with PDIM-positive *ppsD*+ H37Rv or the *ppsD* rev strain were not significantly different (for IFN-$\gamma^{-/-}$ mice, $P = 0.159$; for NOS2$^{-/-}$ mice, $P = 0.398$). These data confirm that a spontaneous point mutation, *ppsD*(G44C), was responsible for both loss of PDIM production and attenuation of the PDIM-negative H37Rv clone in mice.

DISCUSSION

Long-term persistence in the face of innate and adaptive immune responses is a hallmark of tuberculosis, yet little is known about the "counterimmune" mechanisms that promote the persistence of *M. tuberculosis* in immunocompetent hosts. We previously identified the glycosyltransferase encoded by the *rv2958c* gene as a putative factor required to counter the

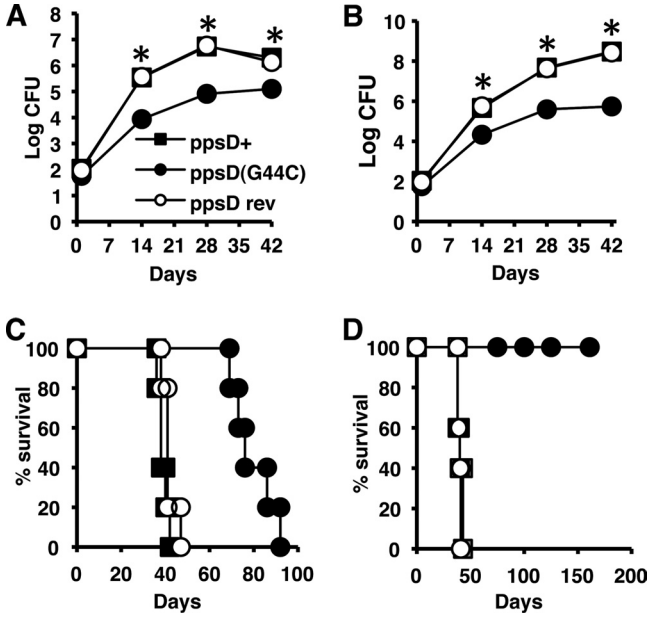


FIG. 8. Reversion of the *ppsD*(G44C) point mutation restores wild-type levels of growth and virulence in mice. (A to D) Mice were aerosol infected with *ppsD*+ (filled squares), *ppsD*(G44C) (filled circles), or *ppsD* rev (open circles) strains of *M. tuberculosis* H37Rv. (A and B) Bacterial growth in the lungs of C57BL/6 (A) and NOS2$^{-/-}$ (B) mice. Groups of mice were sacrificed at the indicated time points, and bacterial CFU were enumerated by plating lung homogenates on 7H10 and scoring colonies after 3 to 4 weeks of incubation at 37°C. Symbols represent means (n = 4 mice per group); error bars indicate standard errors. Asterisks indicate statistically significant differences ($P < 0.05$) in comparisons of *ppsD*(G44C) versus *ppsD*+ and *ppsD*(G44C) versus *ppsD* rev strains. This experiment was performed once. (C and D) Survival of IFN-$\gamma^{-/-}$ (C) and iNOS$^{-/-}$ (D) mice (n = 5 mice per group). This experiment was performed once.

impact of IFN-γ-dependent immune mechanisms other than NOS2 (12). Here, we tested mutants harboring targeted deletions of genes that are required for biosynthesis of secreted *p*-HBAD, including the *rv2958c* gene, for their ability to replicate in the lungs of mice. Although none of the Δ*rv2957*, Δ*rv2958c*, Δ*rv2959c*, or Δ*rv2962c* mutants were attenuated for growth, it remains possible that secreted *p*-HBADs play a role in virulence modulation that is not reflected in the bacterial growth kinetics, as shown previously for PGL (26).

We attribute the discrepancy between the phenotypes of the *rv2958c*::Tn and Δ*rv2958c* mutants in the mouse infection model to the spontaneous loss of PDIM production in our H37Rv parental strain and the *rv2958c*::Tn mutant derived from it. Our observations demonstrate a role for PDIM in countering the impact of an IFN-γ-dependent, NOS2-independent immune mechanism, in addition to the previously postulated role for PDIM in protecting the bacteria from the cidal activity of RNS (28). Consistent with this idea, we found that the survival time of IFN-$\gamma^{-/-}$ and NOS2$^{-/-}$ mice was not significantly different ($P = 0.07$) after infection with the PDIM-positive H37Rv clone. In contrast, following infection with the PDIM-negative H37Rv clone, NOS2$^{-/-}$ mice survived significantly longer than IFN-$\gamma^{-/-}$ mice ($P < 0.001$). Our results are also consistent with the results of a recent screen for mutants that specifically alter growth of *M. tuberculosis* in NOS2$^{-/-}$ mice, which identified two independent mutations in the *drrA* gene encoding a putative PDIM transporter (22). The immune mechanism responsible for this IFN-γ-dependent, NOS2-independent attenuation and the countermechanism by which PDIM confers protection are currently unknown.

We demonstrated that a point mutation, *ppsD*(G44C), in a PDIM-negative clone derived from our H37Rv parental strain was responsible for both a defect in PDIM production and attenuation in mice. To our knowledge, this is the first direct evidence that PpsD is required for PDIM biosynthesis and virulence. We consistently observed a low level of residual PDIM production by the *ppsD*(G44C) mutant. The *ppsD*(G44C) mutant was a clonal isolate from a single colony, suggesting that weak PDIM production is a property of this strain. It is possible that reversion of the G44C point mutation or compensatory mutations elsewhere in *ppsD* that restore PDIM synthesis occur spontaneously at some low frequency. It is unlikely that such strains would become a significant fraction of the population, however, since the *ppsD*(G44C) mutant grows at a higher rate. We therefore favor the possibility that the PpsDG44C mutant protein retains some residual activity that enables weak PDIM production. The low level of PDIM produced by the *ppsD*(G44C) mutant is apparently not sufficient, however, to support normal replication in the lungs of mice.

A correlation was previously observed between PDIM deficiency and a growth advantage in liquid culture (7). Similarly, our PDIM-negative *ppsD*(G44C) mutant had an enhanced *in vitro* growth rate and grew to a higher cell density than PDIM-positive clones. The *in vitro* growth advantage of such spontaneous PDIM-negative variants could explain why they are able to supplant the parental strain during repeated cycles of growth *in vitro*, for example, during mutant strain construction. Indeed PDIM-negative variants were isolated with significantly higher frequency in a culture that was serially passaged than in a nonpassaged control (7). It has also been suggested that spontaneous PDIM deficient variants might be selected by genetic manipulations involving electroporation or bacteriophage infections (11, 26). Since our results indicate that these spontaneous PDIM-deficient variants are attenuated for virulence, care should be taken to minimize the number of passages during genetic manipulation of *M. tuberculosis* and to ensure that strains used in animal infection studies are PDIM proficient.

Spontaneously arising PDIM-deficient variants have been described previously (1, 7, 8, 11, 15, 19–21) and are probably more common than has been reported or realized. Although spontaneous PDIM deficiency has most frequently been associated with the H37Rv strain, it has also recently been described for the *M. tuberculosis* Erdman strain and the clinical isolate HN878 (7, 20). Many of the attenuated *M. tuberculosis* mutants described in the scientific literature have not been complemented, and in other cases, complementation restored virulence only partially. Some of these mutants might have acquired unrecognized secondary mutations causing PDIM deficiency, which could be a factor contributing to their attenuation. Moreover, a number of genes have been implicated in PDIM synthesis, based on the PDIM-negative phenotypes of the corresponding mutants, in the absence of confirmatory complementation (for example, see references 13, 27, 28, 30,

31, and 36). Some of these mutant strains might contain unrecognized secondary mutations that are the true cause of their PDIM deficiency. This phenomenon might also explain the observation that *Mycobacterium leprae* produces PDIM despite lacking functional copies of certain polyketide biosynthesis genes that were reported to be essential for PDIM production in *M. tuberculosis* (discussed in reference 36).

Construction and characterization of random or targeted gene-disrupted mutants is a powerful technique to assign biological functions to biochemical pathways in mycobacteria. However, our findings underscore the idea that any functional assignment must be tentative in the absence of complementation analysis, even in cases where the mutation is not polar on the expression of neighboring genes.

ACKNOWLEDGMENTS

We thank Peter Giannakis and Laetitia Martin for expert technical assistance with animal experiments and Keith Harshman and Jérome Thomas at the University of Lausanne core facility for performing the whole-genome resequencing reactions.

This work was supported by a Robert D. Watkins Graduate Fellowship from the American Society for Microbiology (M.A.K.), the William Randolph Hearst Endowed Scholarship Fund (K.B.H.), NIH MSTP grant GM07739 (M.A.K. and K.B.H.) for the Tri-Institutional MD/PhD Program of Weill-Cornell Medical School, Rockefeller University, and the Sloan-Kettering Institute, an Irvington Institute Postdoctoral Fellowship of the Cancer Research Institute (A.D.T.), SystemsX, The Swiss Initiative in Systems Biology (S.U.), and National Institutes of Health grant AI046392 (J.D.M.).

REFERENCES

1. **Andreu, N., and I. Gibert.** 2008. Cell population heterogeneity in *Mycobacterium tuberculosis* H37Rv. Tuberculosis (Edinb.) **88:**553–559.
2. **Astarie-Dequeker, C., et al.** 2009. Phthiocerol dimycocerosates of *M. tuberculosis* participate in macrophage invasion by inducing changes in the organization of plasma membrane lipids. PLoS Pathog. **5:**e1000289.
3. **Camacho, L. R., et al.** 2001. Analysis of the phthiocerol dimycocerosate locus of *Mycobacterium tuberculosis*: evidence that this lipid is involved in the cell wall permeability barrier. J. Biol. Chem. **276:**19845–19854.
4. **Camacho, L. R., D. Ensergueix, E. Perez, B. Gicquel, and C. Guilhot.** 1999. Identification of a virulence gene cluster of *Mycobacterium tuberculosis* by signature-tagged transposon mutagenesis. Mol. Microbiol. **34:**257–267.
5. **Constant, P., et al.** 2002. Role of the pks15/1 gene in the biosynthesis of phenolglycolipids in the *Mycobacterium tuberculosis* complex: evidence that all strains synthesize glycosylated p-hydroxybenzoic acid methyl esters and that strains devoid of phenolglycolipids harbor a frameshift mutation in the pks15/1 gene. J. Biol. Chem. **277:**38148–38158.
6. **Cox, J. S., B. Chen, M. McNeil, and W. R. Jacobs, Jr.** 1999. Complex lipid determines tissue-specific replication of *Mycobacterium tuberculosis* in mice. Nature **402:**79–83.
7. **Domenech, P., and M. B. Reed.** 2009. Rapid and spontaneous loss of phthiocerol dimycocerosate (PDIM) from *Mycobacterium tuberculosis* grown *in vitro*: implications for virulence studies. Microbiol. **155:**3532–3543.
8. **Domenech, P., et al.** 2004. The role of MmpL8 in sulfatide biogenesis and virulence of *Mycobacterium tuberculosis*. J. Biol. Chem. **279:**21257–21265.
9. **Ehrt, S., et al.** 2001. Reprogramming of the macrophage transcriptome in response to interferon-γ and *Mycobacterium tuberculosis*: signaling roles of nitric oxide synthase-2 and phagocyte oxidase. J. Exp. Med. **194:**1123–1140.
10. **Gagneux, S., et al.** 2006. Variable host-pathogen compatibility in *Mycobacterium tuberculosis*. Proc. Natl. Acad. Sci. U. S. A. **103:**2869–2873.
11. **Goren, M. B., O. Brokl, and W. B. Schaefer.** 1974. Lipids of putative relevance to virulence in Mycobacterium tuberculosis: phthiocerol dimycocerosate and the attenuation indicator lipid. Infect. Immun. **9:**150–158.
12. **Hisert, K. B., et al.** 2004. Identification of *Mycobacterium tuberculosis* counterimmune (*cim*) mutants in immunodeficient mice by differential screening. Infect. Immun. **72:**5315–5321.
13. **Hotter, G. S., et al.** 2005. Transposon mutagenesis of Mb0100 at the *ppe1-nrp* locus in *Mycobacterium bovis* disrupts phthiocerol dimycocerosate (PDIM) and glycosylphenol-PDIM biosynthesis, producing an avirulent strain with vaccine properties at least equal to those of *M. bovis* BCG. J. Bacteriol. **187:**2267–2277.
14. **Ioerger, T. R., et al.** 2010. Variation among genome sequences of H37Rv strains of *Mycobacterium tuberculosis* from multiple laboratories. J. Bacteriol. **192:**3645–3653.
15. **Kana, B. D., et al.** 2008. The resuscitation-promoting factors of *Mycobacterium tuberculosis* are required for virulence and resuscitation from dormancy but are collectively dispensable for growth *in vitro*. Mol. Microbiol. **67:**672–684.
16. **Kondo, E., and K. Kanai.** 1976. A suggested role of a host-parasite lipid complex in mycobacterial infection. Jpn. J. Med. Sci. Biol. **29:**199–201.
17. **Larsen, M. H., K. Biermann, S. Tandberg, T. Hsu, and W. R. J. Jacobs.** 2007. Genetic manipulation of *Mycobacterium tuberculosis.* Curr. Protoc. Microbiol. **Chapter 10:**Unit 10A.2.
18. **Li, H., J. Ruan, and R. Durbin.** 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. **18:**1851–1858.
19. **Manjunatha, U. H., et al.** 2006. Identification of a nitroimidazo-oxazine-specific protein involved in PA-824 resistance in *Mycobacterium tuberculosis*. Proc. Natl. Acad. Sci., U. S. A. **103:**431–436.
20. **Marrero, J., K. Y. Rhee, D. Schnappinger, K. Pethe, and S. Ehrt.** 2010. Gluconeogenic carbon flow of tricarboxylic acid cycle intermediates is critical for *Mycobacterium tuberculosis* to establish and maintain infection. Proc. Natl. Acad. Sci., U. S. A. **107:**9819–9824.
21. **Matsunaga, I., et al.** 2004. *Mycobacterium tuberculosis pks12* produces a novel polyketide presented by CD1c to T cells. J. Exp. Med. **200:**1559–1569.
22. **Murry, J. P., A. K. Pandey, C. M. Sassetti, and E. J. Rubin.** 2009. Phthiocerol dimycocerosate transport is required for resisting interferon-γ-independent immunity. J. Infect. Dis. **200:**774–782.
23. **Onwueme, K. C., C. J. Vos, J. Zurita, J. A. Ferreras, and L. E. Quadri.** 2005. The dimycocerosate ester polyketide virulence factors of mycobacteria. Prog. Lipid Res. **44:**259–302.
24. **Perez, E., et al.** 2004. Molecular dissection of the role of two methyltransferases in the biosynthesis of phenolglycolipids and phthiocerol dimycoserosate in the *Mycobacterium tuberculosis* complex. J. Biol. Chem. **279:**42584–42592.
25. **Pérez, E., et al.** 2004. Characterization of three glycosyltransferases involved in the biosynthesis of the phenolic glycolipid antigens from the *Mycobacterium tuberculosis* complex. J. Biol. Chem. **279:**42574–42583.
26. **Reed, M. B., et al.** 2004. A glycolipid of hypervirulent tuberculosis strains that inhibits the innate immune response. Nature **431:**84–87.
27. **Rousseau, C., et al.** 2003. Virulence attenuation of two Mas-like polyketide synthase mutants of *Mycobacterium tuberculosis*. Microbiology **149:**1837–1847.
28. **Rousseau, C., et al.** 2004. Production of phthiocerol dimycocerosates protects *Mycobacterium tuberculosis* from the cidal activity of reactive nitrogen intermediates produced by macrophages and modulates the early immune response to infection. Cell. Microbiol. **6:**277–287.
29. **Schnappinger, D., et al.** 2003. Transcriptional adaptation of *Mycobacterium tuberculosis* within macrophages: insights into the phagosomal environment. J. Exp. Med. **198:**693–704.
30. **Sirakova, T. D., V. S. Dubey, M. H. Cynamon, and P. E. Kolattukudy.** 2003. Attenuation of *Mycobacterium tuberculosis* by disruption of a *mas*-like gene or a chalcone synthase-like gene, which causes deficiency in dimycocerosyl phthiocerol synthesis. J. Bacteriol. **185:**2999–3008.
31. **Sirakova, T. D., V. S. Dubey, H.-J. Kim, M. H. Cynamon, and P. E. Kolattukudy.** 2003. The largest open reading frame (*pks12*) in the *Mycobacterium tuberculosis* genome is involved in pathogenesis and dimycocerosyl phthiocerol synthesis. Infect. Immun. **71:**3794–3801.
32. **Slayden, R. A., and C. E. I. Barry.** 2001. Analysis of the lipids of *Mycobacterium tuberculosis*, p. 229–245. *In* T. Parish and N. G. Stoker (ed.), *Mycobacterium tuberculosis* protocols. Humana Press, Totowa, NJ.
33. **Stover, C. K., et al.** 1991. New use of BCG for recombinant vaccines. Nature **351:**456–460.
34. **Trivedi, O. A., et al.** 2005. Dissecting the mechanism and assembly of a complex virulence mycobacterial lipid. Mol. Cell **17:**631–643.
35. **Voskuil, M. I., et al.** 2003. Inhibition of respiration by nitric oxide induces a *Mycobacterium tuberculosis* dormancy program. J. Exp. Med. **198:**705–713.
36. **Waddell, S. J., et al.** 2005. Inactivation of polyketide synthase and related genes results in the loss of complex lipids in *Mycobacterium tuberculosis* H37Rv. Lett. Appl. Microbiol. **40:**201–206.
37. **Wiegeshaus, E. H., D. N. McMurray, A. A. Grover, G. E. Harding, and D. W. Smith.** 1970. Host-parasite relationships in experimental airborne tuberculosis. III. Relevance of microbial enumeration to acquired resistance in guinea pigs. Am. Rev. Respir. Dis. **102:**422–429.

Editor: J. L. Flynn

4.3.2 Sequencing of pyridomycin-resistant mutants of *M. tb*

Pyridomycin is a naturally occurring antibiotic produced by *Dactylosporangium fulvum* with specific cidal activity against mycobacteria. The aim of this study was to determine the target and mechanism of action of pyridomycin against *M. tb* and assess its potential as an anti-TB compound. The strategy used to identify the target of pyridomycin was based on analysis of the genome sequences of pyridomycin-resistant isolates of *M. tb* in order to identify genetic mutations responsible for this phenotype. It is expected that these mutations are likely to occur in genes encoding the drug target, as these changes would directly interfere with the activity of the drug. Whole-genome re-sequencing of pyridomycin-resistant colonies of *M. tb* H37Rv and comparison with the parental strain revealed a nonsynonymous SNP in the *inhA* gene resulting in the replacement of aspartic acid by glycine at position 148 (D148G). Further genetic validation, followed by biochemical and structural studies confirmed that pyridomycin inhibits InhA directly as a competitive inhibitor of the NADH-binding site and thus prevents synthesis of mycolic acids in *M. tb*. Incidentally, InhA is also the target of isoniazid, a first-line anti-TB drug but the two molecules have different binding sites. Coherently, the most frequently encountered isoniazid-resistant clinical isolates remain fully susceptible to pyridomycin, making it a promising alternative for the treatment of isoniazid-resistant tuberculosis. In addition to the identification of the target of pyridomycin, this study revealed a new druggable pocket in InhA that can be exploited in research on TB drug development.

TRANSPARENT PROCESS OPEN ACCESS
EMBO Molecular Medicine

Towards a new tuberculosis drug: pyridomycin – nature's isoniazid

*Ruben C. Hartkoorn[1], Claudia Sala[1], João Neres[1], Florence Pojer[1], Sophie Magnet[1], Raju Mukherjee[1], Swapna Uplekar[1], Stefanie Boy-Röttger[1], Karl-Heinz Altmann[2], Stewart T. Cole[1]**

Keywords: drug discovery; InhA; isoniazid; pyridomycin; tuberculosis

DOI 10.1002/emmm.201201689

Received June 29, 2012
Revised July 26, 2012
Accepted July 30, 2012

→See accompanying article
http://dx.doi.org/10.1002/emmm.201201811

Tuberculosis, a global threat to public health, is becoming untreatable due to widespread drug resistance to frontline drugs such as the InhA-inhibitor isoniazid. Historically, by inhibiting highly vulnerable targets, natural products have been an important source of antibiotics including potent anti-tuberculosis agents. Here, we describe pyridomycin, a compound produced by *Dactylosporangium fulvum* with specific cidal activity against mycobacteria. By selecting pyridomycin-resistant mutants of *Mycobacterium tuberculosis*, whole-genome sequencing and genetic validation, we identified the NADH-dependent enoyl-(Acyl-Carrier-Protein) reductase InhA as the principal target and demonstrate that pyridomycin inhibits mycolic acid synthesis in *M. tuberculosis*. Furthermore, biochemical and structural studies show that pyridomycin inhibits InhA directly as a competitive inhibitor of the NADH-binding site, thereby identifying a new, druggable pocket in InhA. Importantly, the most frequently encountered isoniazid-resistant clinical isolates remain fully susceptible to pyridomycin, thus opening new avenues for drug development.

INTRODUCTION

Today, infection with *Mycobacterium tuberculosis* accounts for up to two million deaths annually (Glaziou et al, 2009). Major confounding factors such as poverty, homelessness and the prevalence of HIV/AIDS (Harrington, 2010) mean that tuberculosis will indefinitely remain an important cause of morbidity and mortality throughout the world. Furthermore, despite the small, but growing number of drugs that are effective at killing *M. tuberculosis*, the current treatment is still burdened by its duration (typically 6 months for drug-sensitive strains) and the ever increasing number of multidrug (MDR) and extensively drug resistant (XDR) clinical isolates of *M. tuberculosis* (Cegielski, 2010). Together, this underlines the need for alternative therapeutic entities that can be used both to shorten the duration of therapy and to combat the growing problem of clinical drug resistance.

Natural products have long provided a rich source of effective anti-tuberculosis agents. The most active of these in current use, the rifamycins (rifampicin, rifabutin and rifapentine), inhibit RNA polymerase and are crucial for front-line treatment of the disease. Furthermore, several other natural products such as the aminoglycosides (streptomycin, amikacin and kanamycin) and the peptide antibiotic (capreomycin) are part of the current portfolio of anti-tuberculosis drugs. The rich diversity of natural products represents a powerful tool for drug discovery, firstly, in the form of leads for potential anti-microbial agents and secondly, as a means of identifying those targets that are most vulnerable in the bacterium.

In 1953, pyridomycin was first described as an antibiotic that exhibited specific activity against different mycobacteria including *M. tuberculosis* and *M. smegmatis* (Maeda et al, 1953). Pyridomycin (Fig 1A) is produced by *Streptomyces pyridomyceticus* (Maeda et al, 1953; Yagishita, 1954, 1955, 1957a,b) or *Dactylosporangium fulvum* (Shomura et al, 1986). Its biosynthesis was first studied in 1968 (Ogawara et al, 1968) and more recently in 2011 (Huang et al, 2011) when the involvement of both non-ribosomal peptide synthetases (NRPS) and polyketide synthases (PKS) was proposed. Despite this body of work, the mechanism of action of pyridomycin against *M. tuberculosis* is unknown, and its potential as an anti-tuberculosis compound has not been assessed.

(1) Ecole Polytechnique Fédérale de Lausanne, Global Health Institute, Lausanne, Switzerland
(2) Eidgenössische Technische Hochschule Zürich, Institut für Pharmazeutische Wissenschaften, HCI H 405, Zürich, Switzerland
***Corresponding author:** Tel: +41 21 693 1851; Fax +41 21 693 1790; E-mail: stewart.cole@epfl.ch

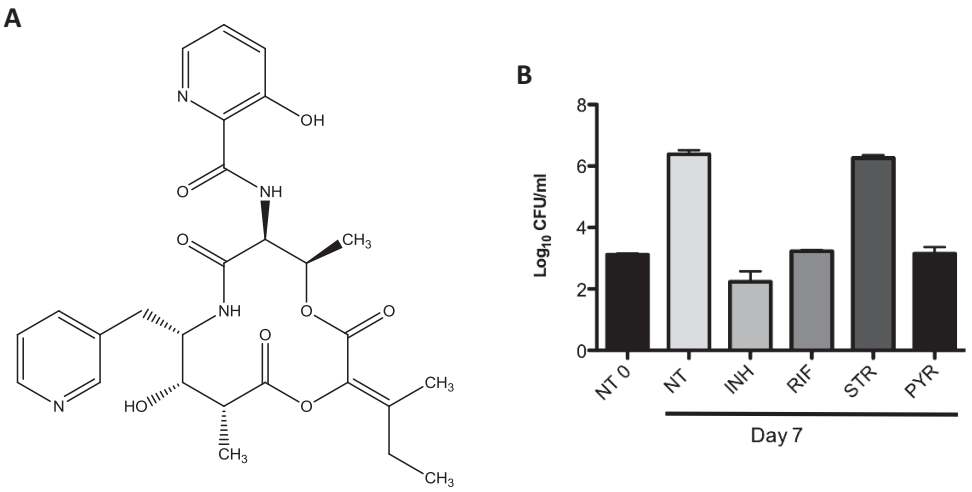


Figure 1. Chemical structure and intracellular activity of pyridomycin.

A. Chemical structure of pyridomycin.

B. The activity of pyridomycin on intracellular *M. tuberculosis* was tested in activated THP-1-derived macrophages. Cells were infected at an MOI of 1:1 with *M. tuberculosis* Erdman and treated with isoniazid (INH) at 1 μg/ml, rifampicin (RIF) at 1 μg/ml, streptomycin (STR) at 10 μg/ml or pyridomycin (PYR) at 10 μg/ml. Colony forming units (CFU) were determined after 7 days exposure to drugs. NT refers to the untreated sample and NT0 to untreated sample at time 0. The experiment was performed in duplicate and results are shown as mean values and standard errors.

The aim of this study was to determine how pyridomycin kills *M. tuberculosis* and to identify its target. To achieve this, a combination of approaches involving resistance mapping, genetic validation, biochemistry, enzyme inhibition and X-ray crystallographic analysis of the target are described. The combined results unambiguously indicate that pyridomycin is a competitive inhibitor of the NADH-binding site of InhA, NADH-dependent enoyl-[Acyl-Carrier-Protein] reductase, the target of the two anti-tuberculosis pro-drugs isoniazid and ethionamide (Banerjee et al, 1994; Vilcheze et al, 2006).

RESULTS

Purification of pyridomycin

Several strains of *Streptomyces pyridomyceticus* (NRRL B-2517, ISP-5024 and DSM40024) were initially tested for pyridomycin production with limited success, likely due to the presence of producing and non-producing populations in the same culture. Pyridomycin (Fig 1A) was, however, readily produced by and purified from *Dactylosporangium fulvum* (NRRL B-16292) with a yield of 20–40 mg/L at a purity >99% and with an NMR spectrum as previously reported (Kinoshita et al, 1989).

Anti-bacterial properties of pyridomycin

Pyridomycin has been described to act specifically against mycobacteria, with little or no activity against other Gram-positive and Gram-negative species (Maeda et al, 1953). In order to verify its spectrum of activity, the resazurin reduction microplate assay (REMA) was used to determine the minimum inhibitory concentration (MIC) for various bacteria. From Table 1, it can be clearly seen that pyridomycin is effective against all members of the *Mycobacterium* genus tested including *M. tuberculosis* (strain H37Rv, MIC = 0.31–0.63 μg/ml) and *M. smegmatis* (strain mc^2 155, MIC = 0.62–1.25 μg/ml). Pyridomycin, however, showed no detectable activity against other bacteria, including the close relative *C. glutamicum* (all MIC > 100 μg/ml). These data therefore agree with earlier observations (Maeda et al, 1953; Maeda, 1957) and suggest that pyridomycin targets a mycobacterial component that is either sufficiently divergent or absent in other genera.

To further understand the properties of pyridomycin against *M. tuberculosis*, its minimum bactericidal concentration (MBC) was determined and its activity against non-replicating and intracellular *M. tuberculosis* measured. MBC data demonstrated that pyridomycin is bactericidal against *M. tuberculosis* H37Rv at concentrations of 0.62–1.25 μg/ml. Evaluation of pyridomycin activity against non-replicating *M. tuberculosis* using the streptomycin-starved 18b (ss18b) model (Sala et al, 2010) revealed that pyridomycin is not effective, thereby implying that it may target a function involved in active growth. Finally, the intracellular killing activity of pyridomycin was assessed *ex vivo* after infection of activated THP1-derived macrophages. The results indicated that, when left untreated for a 7-day period, intracellular *M. tuberculosis* grew by at least 3 logs, whilst exposure to both pyridomycin (10 μg/ml) and rifampicin (1 μg/ml) prevented any multiplication within the macrophages (Fig 1B). Further controls showed that streptomycin (10 μg/ml) had no impact on the growth of intracellular bacteria while isoniazid (1 μg/ml) was able to reduce the intracellular *M. tuberculosis* load by 1 log (Fig 1B). Pyridomycin is therefore clearly able to enter macrophages and arrest bacterial replication.

Table 1. Bacterial susceptibility to pyridomycin as measured by resazurin reduction microtitre assay

Bacterium	Pyridomycin MIC (μg/ml)
Mycobacterium tuberculosis	0.39
Mycobacterium bovis BCG	0.39
Mycobacterium smegmatis	0.78
Mycobacterium marinum	3.13
Mycobacterium abscessus	6.25
Mycobacterium bolletii	6.25
Mycobacterium massiliense	6.25
Mycobacterium avium	12.5
Corynebacterium glutamicum	>100
Corynebacterium diphtheriae	>100
Micrococcus luteus	>100
Listeria monocytogenes	>100
Staphylococcus aureus	>100
Bacillus subtilis	>100
Enterococcus faecalis	>100
Escherichia coli	>100
Pseudomonas putida	>100
Pseudomonas aeruginosa	>100
Salmonella typhimurium	>100
Candida albicans	>100

Cytotoxicity of pyridomycin on human cell lines

To determine whether pyridomycin is cytotoxic to human cells, the concentration-dependent cytotoxicity of the compound was assessed on two human cell lines. Data indicated that the amount of pyridomycin needed to kill 50% of HepG2 cells (human hepatic cell line) or A549 cells (human lung epithelium cell line) was 100 and 50 μg/ml, respectively. Pyridomycin therefore shows higher selectivity for *M. tuberculosis* compared to the human cells tested (selectivity index >100-fold), in agreement with a previous finding that pyridomycin shows low toxicity in an acute murine model following 800 mg/kg intraperitoneal injection (Maeda et al, 1953).

Identification of the pyridomycin target

The strategy to identify the target and mechanism of action of pyridomycin was to raise pyridomycin-resistant mutants and to pinpoint the genetic mutations responsible for this phenotype, anticipating that these mutations would be in the gene for the drug target. Resistant mutants of strain H37Rv were selected on solid medium containing pyridomycin at 10× MIC (3 μg/ml) and arose at a frequency of around 10^{-6}. Of the 10 colonies selected for further analysis (PYR1 to 10), nine showed no change in the MIC to pyridomycin when re-tested by REMA, whereas mutant PYR7 retained a near 10-fold increase in its resistance level compared to the parent H37Rv (Fig 2). This phenotype was stably maintained and mutant PYR7 remained fully susceptible to isoniazid, moxifloxacin and rifampicin like wild-type H37Rv (Fig 2).

To identify the single nucleotide polymorphisms (SNPs) or insertion/deletions (INDELs) responsible for the pyridomycin resistance, the genomes of both PYR7 and the parental strain were sequenced to near completion by the Illumina protocol. Ninety-eight percent of the reads were successfully mapped to the H37Rv reference genome (Cole et al, 1998) resulting in an average 300-fold coverage. Comparison of the PYR7 and H37Rv assemblies revealed 63 SNPs of which 53 mapped to the repetitive PE and PPE gene families and were therefore discarded. Of the remaining 10 SNPs, nine were synonymous. The only non-synonymous mutation found was an a443g transition in *inhA* resulting in replacement of the aspartic acid at position 148 by a glycine (D148G). This missense mutation was subsequently confirmed by conventional Sanger sequencing. With reference to previously published structures of the NADH-dependent enoyl-ACP reductase InhA, Asp148 was found to be located near the NADH binding pocket (Dessen et al, 1995; Dias et al, 2007; Molle et al, 2010; Oliveira et al, 2006; Rozwarski et al, 1999; Vilcheze et al, 2006).

Genetic validation of InhA as the target of pyridomycin

To genetically validate InhA as the target of pyridomycin, we evaluated whether over expression of *inhA* caused an increase in resistance to the antibiotic in wild-type *M. tuberculosis*. For this purpose, we transformed strain H37Rv with a plasmid carrying the *inhA* gene under the control of the *hsp60* promoter (pMVinhA; Larsen et al, 2002). In Fig 2, it can be seen that H37Rv::pMVinhA displayed a 15-fold higher MIC for pyridomycin compared to the control strain H37Rv::pMV261 (from 0.31 to 5 μg/ml). When the same experiment was performed in the PYR7 background, no complementation of the resistant phenotype was observed, indicating that the associated mutation was dominant. Similar to the wild-type strain, we noticed a four-fold increase in resistance in PYR7::pMVinhA compared to the empty vector (PYR7::pMV261; from 2.5 to 10 μg/ml; Fig 2). In control experiments, overexpression of *inhA* also led to increased isoniazid resistance in both strains whilst not impacting the MIC of moxifloxacin (Fig 2). Together, these genetic data strongly suggest that InhA is the target of pyridomycin.

To further corroborate that the D148G mutation in *inhA* was indeed responsible for pyridomycin resistance, we overexpressed this allele in H37Rv and compared its effect with that of a well-characterized mutation associated with isoniazid resistance, InhA (S94A) (Vilcheze et al, 2006). Results presented in Fig 2 clearly show that, compared to overexpression of wild-type InhA (pMVinhA), overexpression of InhA (D148G) causes four-fold greater resistance to pyridomycin while overexpression of InhA (S94A) conferred only two-fold resistance. Furthermore, overexpression of InhA (D148G) had no impact on the MIC for isoniazid compared to overexpression of wild type InhA, while, as expected, overexpressing InhA (S94A) increased resistance around four-fold (Fig 2). None of the mutations affected the MIC for moxifloxacin (Fig 2). Collectively, the data prove that the D148G mutation in InhA is responsible for resistance to pyridomycin, whilst not noticeably affecting isoniazid activity.

In addition to isoniazid, InhA is also the target of ethionamide and triclosan. Susceptibility studies using these compounds on H37Rv and PYR7 indicate that both strains are equally sensitive with MICs of 2.0 and 12.5 μg/ml, respectively. This lack of cross-resistance indicates that D148G in InhA is likely

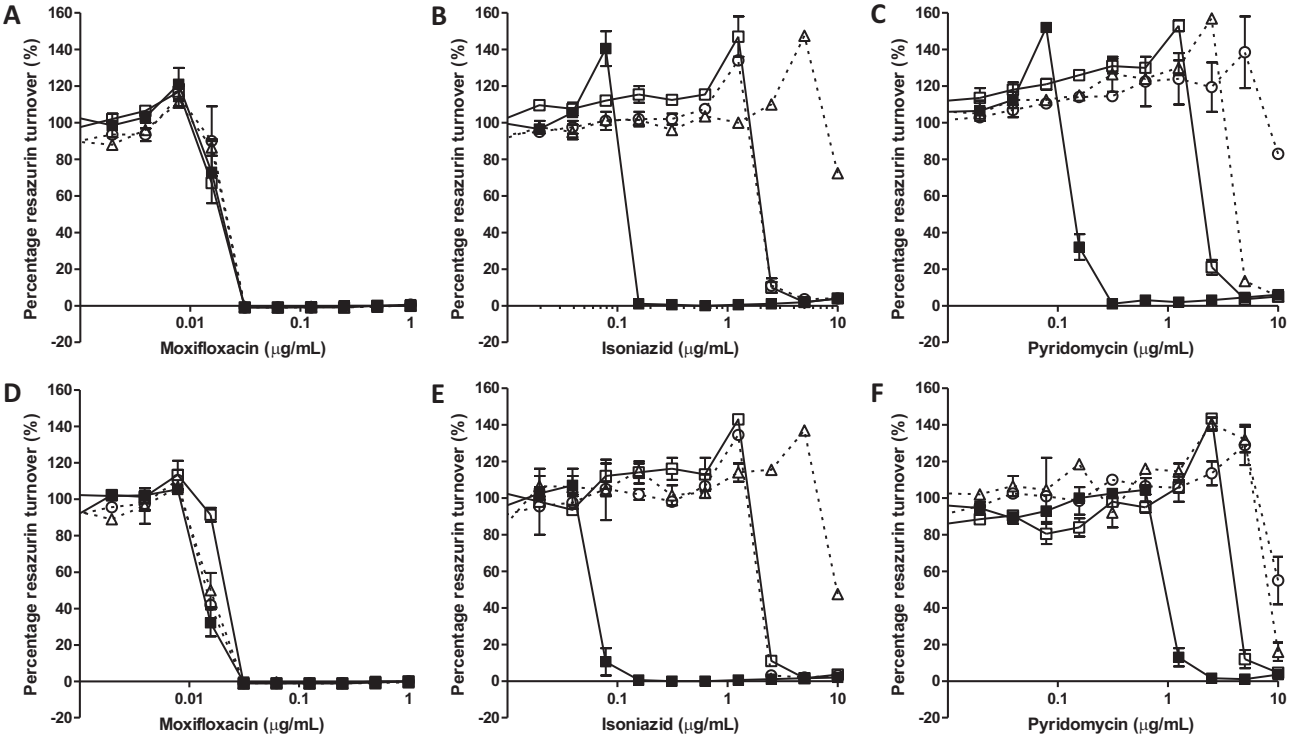


Figure 2. Genetic validation of InhA as the target of pyridomycin.

A-C. The compound susceptibility of wild-type H37Rv transformed with the control vector pMV261 (filled squares), pMVinhA (open squares), pMVinhA (S94A) (open triangle) or pMVinhA (D148G) (open circle) to: (**A**) moxifloxacin, (**B**) isoniazid or (**C**) pyridomycin.

D-F. The compound susceptibility of pyridomycin-resistant mutant PYR7 transformed with the control vector pMV261 (filled squares), pMVinhA (open squares), pMVinhA (S94A) (open triangle) or pMVinhA (D148G) (open circle) to: (**D**) moxifloxacin, (**E**) isoniazid or (**F**) pyridomycin.

to have no impact on the binding to InhA of either triclosan or the active metabolite of ethionamide, the ethionamide-NAD adduct.

Susceptibility of isoniazid-resistant clinical isolates to pyridomycin

Since our findings indicated that pyridomycin has the same target as isoniazid, we investigated whether isoniazid-resistant clinical isolates of *M. tuberculosis* retained susceptibility to pyridomycin. As isoniazid is a pro-drug, clinically relevant mutations that confer resistance are frequently found in the *katG* gene encoding the catalase–peroxidase required for isoniazid bio-activation or, less commonly, in the promoter region of *inhA*, which increases expression of the protein. Of the eight independent isoniazid-resistant clinical isolates analysed, four had mutations in *katG* (S315T), three in the promoter region of *inhA* [c (−15)t] and one isolate carried both mutations (Table 2). Analysis of the drug susceptibility of these isolates confirmed that all strains carrying the *katG* mutation displayed a high level of resistance to isoniazid (MIC >10 μg/ml) and those isolates carrying only the *inhA* promoter mutation showed intermediate isoniazid resistance (MIC = 1.25 μg/ml) compared to H37Rv (0.16 μg/ml; Table 2). On the contrary, isolates carrying the *katG* mutations showed no resistance to pyridomycin (MIC = 0.3–0.6 μg/ml), while a mutation in the *inhA* promoter resulted in increased pyridomycin resistance (MIC = 2.5–5 μg/ml; Table 2). For all clinical isolates tested, the susceptibility to moxifloxacin was similar to wild-type (MIC = 0.03–0.10 μg/ml). Thus, isoniazid-resistant clinical isolates carrying the *inhA* [*c* (−15)*t*] promoter mutation displayed cross-resistance with pyridomycin, whereas the more common *katG* (*S315T*) isoniazid-resistant mutants retained full sensitivity to the antibiotic.

Inhibition of mycolic acid synthesis by pyridomycin

It has been elegantly demonstrated that inhibition of InhA by isoniazid in *M. tuberculosis* leads to the specific depletion of mycolic acids from the bacterial cell wall without affecting fatty acid synthesis (Vilcheze et al, 2006). To show that pyridomycin inhibition of InhA also results in inhibition of mycolic acid synthesis, the mycolic and fatty acid content of *M. tuberculosis* was determined in the presence and absence of pyridomycin by radiometric thin layer chromatography (TLC). We found that pyridomycin caused a concentration-dependent reduction of mycolic acid synthesis (alpha-, methoxy- and keto-mycolic acids) whilst not affecting the fatty acid content (Fig 3). Furthermore, when performing the same experiment on PYR7, over five-fold higher pyridomycin concentrations were needed to inhibit mycolic acid biosynthesis consistent with the resistance level observed. Both H37Rv and PYR7 behaved similarly when

Table 2. Pyridomycin activity against isoniazid-resistant clinical isolates of *M. tuberculosis*

Isolate	KatG genotype[a]	*inhA* promoter genotype[b]	MIC (μg/ml)		
			Isoniazid	Pyridomycin	Moxifloxacin
H37Rv	wt	wt	0.16	0.31	0.03
1	S315T	wt	>10	0.63	0.03
2	S315T	wt	>10	0.50	0.05
3	S315T	wt	>10	0.50	0.05
4	S315T	wt	>10	0.31	0.03
5	S315T	c (−15)t	>10	2.5	0.02
6	wt	c (−15)t	1.25	3.75	0.10
7	wt	c (−15)t	1.25	3.75	0.06
8	wt	c (−15)t	1.25	5	0.06

wt, wild-type.

[a]Numbering refers to the KatG protein sequence.

[b]Numbering refers to the *inhA* coding sequence, with +1 corresponding to the first base of the ATG start codon.

the assay was repeated in the presence of isoniazid (Fig 3). Indeed, the latter caused a concentration-dependent decrease in the amount of mycolic acids in H37Rv and was equally effective at inhibiting mycolic acid synthesis in PYR7. Taken together, these data confirm that pyridomycin targets mycolic acid synthesis and demonstrate that the InhA (D148G) enzyme in PYR7 is much less susceptible to pyridomycin inhibition.

In vitro inhibition of InhA by pyridomycin

Inhibition of purified InhA by pyridomycin was studied to investigate if pyridomycin alone can inhibit the enzyme or whether *in vivo* bio-activation by an intracellular process is needed. InhA, InhA (S94A) and InhA (D148G) were successfully expressed and purified. All three enzymes were catalytically active and oxidized NADH in the presence of the substrate 2-trans-octenoyl-CoA (OcCoA). Initial experiments determined the NADH-binding constant (K_m) and confirmed that for InhA (S94A) it was around 6.5 times higher than for wild-type InhA (Table 3) as reported previously (Quemard et al, 1995; Rawat et al, 2003; Vilcheze et al, 2006). Surprisingly, we found that the K_m of InhA (D148G) for NADH was 14-fold greater than for wild-type InhA (Table 3) suggesting a lower affinity for NADH in the D148G mutant. All the enzymes had a similar V_{max} (around 0.52 μmol/min/mg) (Table 3). Enzyme inhibition studies showed that pyridomycin was able to inhibit both wild-type InhA and InhA (S94A) at a similar K_i (6.5 and 4.55 μM, respectively) (Table 3). InhA (D148G) could not be inhibited at all by pyridomycin at concentrations below 18.6 μM. Statistical analysis of inhibition of both wild-type InhA and InhA (S94A) by pyridomycin favours a model of competitive inhibition with NADH as indicated by similar *y*-axis intercepts on Lineweaver–Burk plots (Fig 4). These data prove that pyridomycin is the pharmacophore that inhibits InhA, and this activity is achieved by competitive inhibition of the NADH-binding site. Additionally, these biochemical and enzymological results confirm that InhA (D148G) is more resistant to pyridomycin while InhA (S94A) is as susceptible as the wild-type enzyme.

Crystal structures of InhA (D148G), wild-type InhA and InhA (S94A)

To further investigate the mode of binding of pyridomycin to InhA at the atomic level, we crystallized and solved the

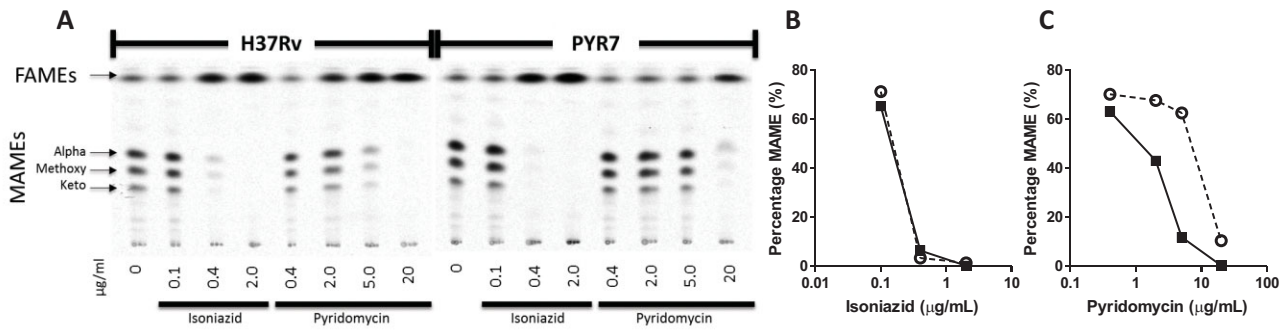


Figure 3. Inhibition of mycolic acid synthesis by pyridomycin. The fatty acid methyl ester (FAMEs) and mycolic acid methyl ester (MAMEs) profiles of wild-type H37Rv and pyridomycin-resistant mutant PYR7 were evaluated by thin-layer chromatography. Both strains were treated with different concentrations of isoniazid and pyridomycin for 3 h and labeled with [1,2-^{14}C]-acetate.

A. ^{14}C-labeled FAMEs and MAMEs were separated by thin-layer chromatography and detected by autoradiography.

B,C. Quantification of the MAME band intensity relative to the density of the FAMEs illustrates the inhibition of MAME synthesis by pyridomycin (**B**) and isoniazid (**C**) in H37Rv (black squares) and PYR7 (open circles).

Table 3. *In vitro* kinetic parameters of *M. tuberculosis* InhA and its inhibition by pyridomycin

	Wild-type InhA	InhA (S94A)	InhA (D148G)
NADH K_m (μM)	13.5 ± 2.3	83.5 ± 9.5	190 ± 16
NADH V_{max} (μmol/min/mg)	0.52 ± 0.03	0.50 ± 0.02	0.54 ± 0.3
Pyridomycin K_i (μM)	6.5 ± 1.2	5.0 ± 0.4	No inhibition at 18.6 μM

structures of wild-type InhA and InhA (S94A) mutant in complex with NADH as previously published (Dessen et al, 1995; Dias et al, 2007; Molle et al, 2010; Oliveira et al, 2006; Rozwarski et al, 1999; Vilcheze et al, 2006). In an attempt to obtain crystal structures of InhA in complex with pyridomycin, pre-crystallized InhA:NADH or InhA (S94A):NADH crystals were soaked in a pyridomycin solution. On penetration of pyridomicin, the crystals turned yellow but lost their ability to diffract. This suggests that pyridomycin may induce major conformational changes upon binding to the NADH co-factor pocket of InhA. For control purposes, InhA:NADH crystals were also soaked with triclosan, an inhibitor of the enoyl-ACP substrate binding site of InhA, and the structure successfully solved, thereby ruling out technical issues with soaking (data not shown). As an alternative strategy to define the pyridomycin binding site in InhA attempts were made to co-crystallize InhA or InhA (S94A) in the presence of pyridomycin alone or with the octenoyl CoA substrate; however, despite testing over 1000 conditions, no diffracting crystals have been obtained to date.

The D148G mutation leads to pyridomycin resistance as well as to a decrease in NADH affinity (Table 3). To determine the molecular basis for this resistance, we crystallized InhA (D148G) in the presence or absence of NADH. As with InhA and InhA (S94A), we obtained crystals only in presence of NADH and solved the InhA (D148G):NADH structure to 2.54 Å (Supporting Information Table S3). By comparing the structure of InhA (D148G):NADH with that of wild-type InhA:NADH and InhA (S94A):NADH obtained in this study and elsewhere (Dessen et al, 1995; Dias et al, 2007; Molle et al, 2010; Oliveira et al, 2006; Rozwarski et al, 1999; Vilcheze et al, 2006), we observed only one major effect, namely rotation of the phenylalanine 149 (Phe149) side chain by 90° (Fig 5). Otherwise, the positions of other active site residues as well as the overall structure of the protein were identical (Supporting Information Fig 1).

Interestingly, Phe149 in the InhA (D148G) mutant adopts the same orientation as InhA and InhA (S94A) in complex with INH-NADH (Dias et al, 2007; Rozwarski et al, 1999; Vilcheze et al, 2006). Thus, in the InhA (D148G) mutant, Phe149 is naturally placed for Pi-stacking with the INH moiety and this explains the unchanged MIC observed for isoniazid against wild-type InhA and the InhA (D148G) mutant *in vivo* (Fig 5). Furthermore, rotation of the Phe149 side chain in the InhA (D148G) mutant also results in opening of a water channel allowing the addition of two water molecules in the active site as well as increasing the distance between the nicotinamide ring of NADH and the ring of Phe149, leading to a decrease in NADH affinity as confirmed by kinetic studies (Fig 5). The same structural changes most probably also explain the decreased sensitivity of InhA (D148G) to pyridomycin compared to wild-type as NADH and pyridomycin share a similar binding pocket.

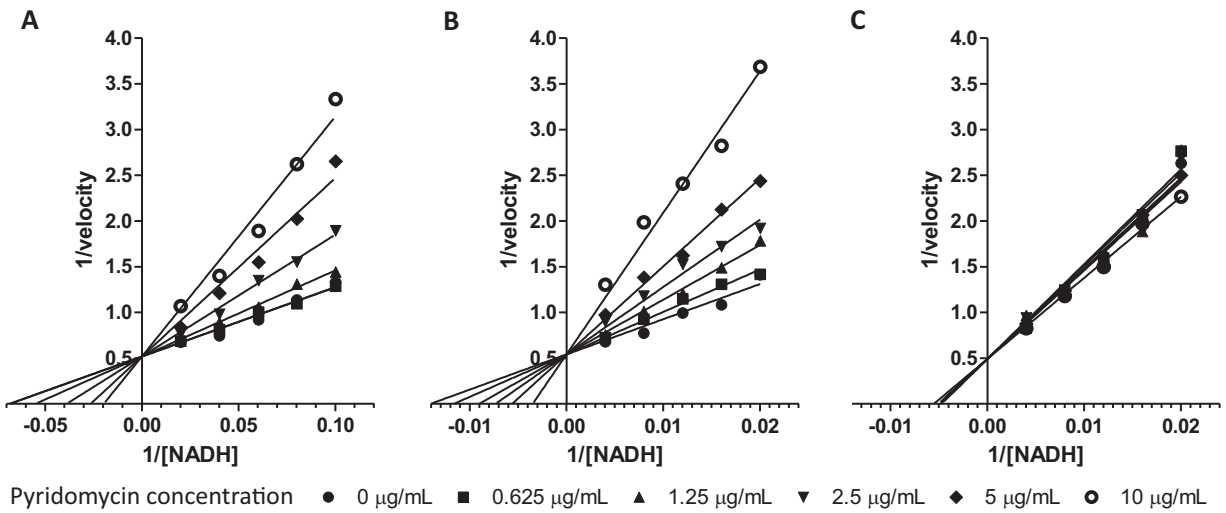


Figure 4. Inhibition of purified InhA by pyridomycin.

A-C. Lineweaver–Burk plot showing the competitive inhibition of wild-type InhA (**A**), InhA (S94A) (**B**) and InhA (D148G) (**C**) by pyridomycin in the presence of NADH.

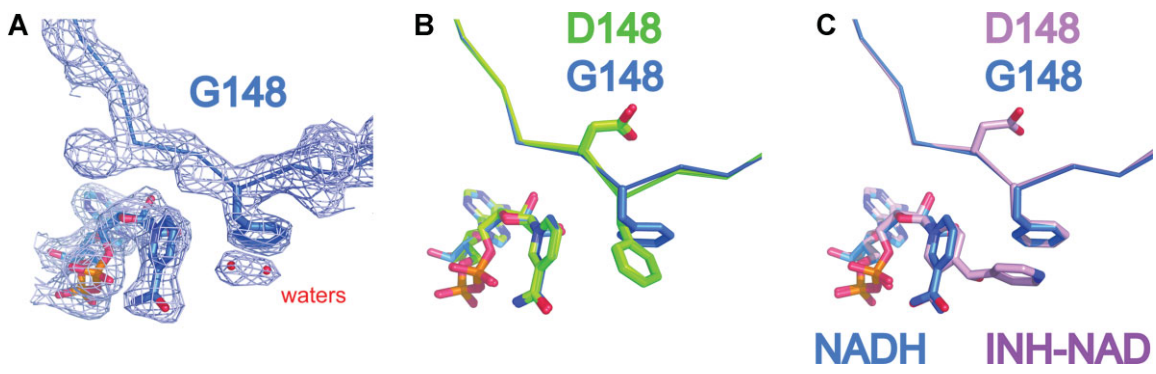


Figure 5. Crystal structures of InhA (D148G) compared to wild-type InhA and InhA (S94A).

A. The 2fo-fc electron density map at 1 sigma of InhA (D148G):NADH structure (at 2.45 Å) with Phe149 and NADH represented as sticks.

B. Superposition of InhA (D148G):NADH (blue) on InhA:NADH and InhA (S94A):NADH (both in green) clearly shows the D148G mutation and the resulting 90° rotation of Phe149.

C. Overlay of InhA (D148G):NADH (Blue) and InhA:NAD-INH (PDB code 1Z1D: Pink) shows that, in this case, Phe149 occupies the same position and explains why InhA (D148G) is still sensitive to inhibition by isoniazid.

DISCUSSION

To inhibit or kill competing organisms, numerous microbes produce and secrete natural products with antibiotic activity. This rich source of chemically diverse compounds was successfully exploited for the production of many anti-microbial and anti-cancer drugs in the first decades of antibiotic development but was then abandoned (Fischbach & Walsh, 2009). Since the selective pressure imparted by some antibiotics with broad-spectrum activity may have led to naturally occurring resistance in other bacteria that share the same ecological niche, many potentially bio-active molecules may have been overlooked (Smith et al, 2010). The ecological and evolutionary forces that have shaped natural products, particularly the selection and inhibition of vulnerable targets, remain unknown but understanding this relationship is important for drug discovery.

The treatment of tuberculosis is increasingly menaced by the emergence of drug-resistant strains of *M. tuberculosis* and this has led to renewed interest in finding new bactericidal inhibitors. As in other anti-infective areas, target-based screens have been unsuccessful prompting investigators to adopt whole cell screening once more (Cole & Riccardi, 2011; Payne et al, 2007). Consequently, we are reinvestigating diverse natural products for hit generation as these compounds have been optimized by evolution as antibacterial agents. In this study, we have used pyridomycin, a natural product with anti-mycobacterial activity, to identify the intracellular target and characterized its mechanism of action. Pyridomycin was first described in 1953 (Maeda et al, 1953), shortly after the introduction of isoniazid into clinical practice, but was then apparently neglected. In a remarkable coincidence, we show that, like isoniazid, pyridomycin directly targets the NADH-dependent enoyl (ACP)-reductase InhA and causes inhibition by competing for the NADH-binding pocket.

InhA (also known as FabI) is an essential component of the type II fatty acid synthase system (FasII) involved in fatty acid elongation and is required for mycolate production in *M. tuberculosis*. The FasII system is highly conserved in bacteria but absent from humans making it an attractive drug target (Heath et al, 2002; McMurry et al, 1998; Vilcheze et al, 2006, 2011). Although isoniazid is certainly the most effective known inhibitor of InhA in *M. tuberculosis*, it is a pro-drug requiring activation by the KatG catalase–peroxidase to form an adduct with NAD. Clinically significant resistance to isoniazid is mainly attributed to loss or alteration of KatG activity. The INH-NAD adduct acts as a slow, tightly binding competitive inhibitor of the NADH-binding site of InhA (Rawat et al, 2003). Interestingly, mycobacterial InhA is also targeted by the small molecules ethionamide (Morlock et al, 2003) (also a pro-drug) and triclosan (Parikh et al, 2000) whilst in other bacteria, FabI is inhibited by the natural products vinaxanthone and cephalochromin (Zhang & Rock, 2008). Other natural products with broad-spectrum activity, such as thiolactomycin, cerulenin and platensimycin, have been shown to target different components of the FasII system in other bacteria (Zhang & Rock, 2008). Our results suggest that pyridomycin is the most potent natural product to inhibit FasII specifically in *M. tuberculosis* and resistance arises due to remodelling of the NAD-binding site in InhA, notably through mutation of D148G.

Pharmacologically validated drug targets are scarce in *M. tuberculosis* with InhA being among the best (Lamichhane, 2011). For this reason, several attempts have been made to develop heterocyclic inhibitors that differ from isoniazid and ethionamide in their structure and activity. Examples include heterocyclic boron containing compounds (diazaborine) that react with NAD^+ ribose to form diazyborine-NAD adducts that inhibit InhA similarly to INH-NAD (Baldock et al, 1996). Triclosan analogues (Freundlich et al, 2009; Vilcheze et al, 2011) and di-phenyl ether compounds (Sullivan et al, 2006) have been shown to inhibit InhA with nanomolar K_i and micromolar MICs and are both promising candidates as anti-tuberculosis compounds. Further screening studies for *Escherichia coli* FabI inhibitors have also revealed effective novel structures, many of

which however do not have good MIC against *M. tuberculosis* (Lu & Tonge, 2008; Payne et al, 2002).

It is noteworthy that, while the FasII system is present in most bacteria, pyridomycin is a specific inhibitor of mycobacterial species (Table 1). Amongst the mycobacteria tested here, the InhA proteins share a high level of sequence identity (77%) and both Asp148 and Phe149 (as well as other active site residues) are strictly conserved, which may explain why all mycobacteria were susceptible to the antibiotic. Additionally, other pathogenic mycobacteria such as *M. leprae* and *M. ulcerans* also share near identical InhA proteins (91 and 93% identity to the *M. tuberculosis* ortholog, respectively) and are expected to be susceptible to pyridomycin (Supporting Information Fig 1). The level of sequence identity between *M. tuberculosis* InhA and FabI from the different Gram-positive and Gram-negative bacteria tested here is considerably lower, ranging from 27 to 33%. Neither Asp148 nor Phe149 are conserved in these enzymes, which probably accounts for their pyridomycin resistance. It is possible that through cohabitating with producers of pyridomycin, or a related natural product, the ancestors of these bacteria may have acquired resistance to the antibiotic as has been proposed for arylomycin, a natural product that inhibits signal peptidase I (Smith et al, 2010).

The increasing emergence of isoniazid resistance in clinical isolates of *M. tuberculosis* is an important problem for tuberculosis therapy and seriously compromises the effectiveness of current treatment. In 50–95% of the cases, resistance to isoniazid is caused by mutations in *katG* (Zhang & Yew, 2009). Low-level resistance to isoniazid is also associated with upregulation of *inhA* but mutations in the *inhA* gene itself are much less common (8–43%) (Zhang & Yew, 2009). Our data clearly show that pyridomycin does not require activation. This is of particular significance because it means that pyridomycin can effectively kill isoniazid-resistant *M. tuberculosis* carrying *katG* mutations as demonstrated by our susceptibility testing of isoniazid-resistant clinical isolates.

The pyridomycin resistant strain PYR7, that carries the D148G mutation in InhA, is not cross-resistant to isoniazid and ethionamide while, conversely, the S94A variant that displays isoniazid resistance remains susceptible to pyridomycin (Fig 2). This suggests that, while both pharmacophores are competitive inhibitors of NADH-binding, they bind to the pocket in different ways. Additionally, the lack of cross resistance with triclosan, the scaffold for other InhA inhibitor programs, is promising as it demonstrates that there are multiple ways of inhibiting the same protein without cross-resistance occurring. These are important findings for rational drug design and could lead to the development of pyridomycin derivatives that kill multiple mycobacteria unlike isoniazid, which is effective solely against *M. tuberculosis*.

MATERIALS AND METHODS

General information

Bacterial strains, culture conditions, pyridomycin production and purification, expression and purification of proteins and details of materials are described in Supporting Information Materials and Methods.

Determination of pyridomycin MIC

The drug susceptibility of all bacteria was determined using the resazurin microtitre assay (REMA; Palomino et al, 2002). Briefly, log-phase bacteria were diluted to an OD_{600} of 0.0001, and grown in a 96-well plate in the presence of serial compound dilutions. After 10 generations (7 days for *M. tuberculosis*) bacterial viability was determined using 10 μl of resazurin (0.025% w/v), and calculated as a percentage of resazurin turnover in the absence of compound. Methodology for determining the minimum bactericidal activity (MBC) and intracellular activity are described in the Supporting Information Materials and Methods.

Isolation and characterization of pyridomycin-resistant H37Rv clones

Pyridomycin-resistant H37Rv mutants were isolated by plating 10^9 CFU on solid 7H10 medium containing 3 μg/ml of pyridomycin (10 × MIC). Following 4 weeks of incubation (37 °C), colonies were picked and grown in 7H9 medium without pyridomycin. Colonies were then retested for susceptibility to pyridomycin, moxifloxicin, isoniazid and rifampicin. Pyridomycin-resistant clone PYR7 was selected for whole genome sequencing using Illumina technology and reads aligned to the genome sequence of the parent H37Rv genome to identify SNPs (protocols described in Supporting Information Materials and Methods). The SNP in *inhA* was validated by conventional Sanger sequencing using an ABI3130XL genetic analyser (Applied Biosystems).

Determination of MIC on isoniazid-resistant clinical isolates

Nine isoniazid-resistant clinical isolates that were 'genotyped' using the Line probe assay (1st line drugs, GenoType MTBDRplus) were kindly supplied by the Centre Hospitalier Universitaire Vaudois (CHUV) and the Hôpitaux Universitaires de Genève (HUG). Isolates were grown to mid log phase in liquid medium and tested for their pyridomycin, isoniazid and moxifloxacin susceptibility by REMA.

Over-expression of *inhA* in H37Rv and PYR7

To genetically validate that InhA was the target of pyridomycin, the impact of over-expressing either wild-type InhA or the mutant forms in strains H37Rv and PYR7 was determined. Briefly, Quick change site-directed mutagenesis was used with pMVInhA [where *inhA* is expressed from the *hsp60* promoter in pMV261 (Larsen et al, 2002)] to introduce t280g or a443g mutations thus generating pMVInhA (S94A) and pMVInhA (D148G), respectively (primers in Supporting Information Table 1). Plasmids pMVInhA, pMVInhA (S94A) and pMVInhA (D148G) were transformed into H37Rv and PYR7. Transformants were selected using kanamycin, then verified by colony PCR for the presence of the kanamycin-resistance cassette before determining their MIC for pyridomycin, isoniazid and moxifloxacin by REMA.

Inhibition of mycolic acid synthesis by pyridomycin

The inhibition of mycolic acid production by pyridomycin and isoniazid was determined as previously described (Vilcheze et al, 2006). Briefly, early log-phase cultures of H37Rv and PYR7 ($OD_{600} = 0.3$, 4 ml) were treated for 3 h with isoniazid (0.1, 0.4 and

The paper explained

PROBLEM:

Even today, infection with *Mycobacterium tuberculosis* accounts for up to two million deaths annually. The effectiveness of current anti-tuberculosis drugs to combat these infections is increasingly compromised by the escalating prevalence of multi- and extensively drug-resistant tuberculosis. For these cases, the most effective anti-tubercular compounds such as isoniazid and rifampicin are no longer effective and this can result in mortality rates approaching 100% for patients with extensively drug-resistant tuberculosis. For these reasons, it is imperative to ensure that the pipeline of drug candidates to treat tuberculosis is well filled.

RESULTS:

We show here that the natural product pyridomycin is a very selective bactericidal compound against mycobacteria including *Mycobacterium tuberculosis*, the causative bacterium of tuberculosis in humans. By selecting and isolating *M. tuberculosis* mutants resistant to pyridomycin and sequencing their genome, we found that a single mutation in a gene named *inhA* is responsible for the resistance. InhA is already the target of the current frontline anti-tuberculosis agent isoniazid. However, most interestingly, no cross resistance was observed between pyridomycin and isoniazid, both in laboratory strains containing mutations in InhA or in the most frequently encountered isoniazid-resistant clinical isolates that contain mutations in *katG* (a gene required to activate isoniazid). We then present detailed genetic and biochemical studies to confirm that pyridomycin itself inhibits InhA and that in live bacteria, this inhibition leads to the depletion of mycolic acids, an essential cell wall component. Finally, studies of the crystal structure of the InhA protein and the pyridomycin-resistant form give valuable insight into the binding pocket of pyridomycin.

IMPACT:

Inhibition of InhA is one of the most effective means of killing *Mycobacterium tuberculosis*, and this is the mechanism behind one of the most potent anti-tubercular agents currently used: isoniazid. The increasing emergence of multi- and extensively drug-resistant tuberculosis (both of which are resistant to isoniazid) means that for these cases, this target can no longer be effectively inhibited by current therapy. Our finding that pyridomycin kills *M. tuberculosis* by inhibiting InhA (even in isoniazid-resistant clinical isolates) provides a promising basis for the development of pyridomycin or a related agent for the treatment of isoniazid-resistant tuberculosis.

2 μg/ml) or pyridomycin (0.4, 2, 5, and 20 μg/ml) prior to labelling for 20 h with [1,2-^{14}C] acetate (1 μCi/ml; Perkin Elmer). Bacteria were harvested and washed with double-distilled H_2O to remove excess [1,2-^{14}C]-acetate. The bacterial pellet was then treated with 2 ml of 20% tetrabutyl ammonium hydroxide overnight at 100 °C to extract the mycolic acids from the cell wall. Mycolic acids were subsequently methylated and extracted by incubation with an equal volume of methylene chloride and 100 μl of methyl iodide for 1 h at room temperature with mixing. The organic phase containing the mycolic acid methyl esters (MAMEs) was washed once with 3N HCl and once with H_2O, dried under nitrogen and resuspended in a smaller volume of methylene chloride. Radiolabelled MAMEs were analysed by TLC. For each condition 10,000 dpm were loaded on a HPTLC Silica Gel 60 F254 plate and run three times with 95:5 v/v hexane/ethyl acetate. The MAMEs were visualized using a Typhoon™ scanner (GE Healthcare) and quantified with the software ImageQuant™ (GE Healthcare).

Steady state kinetics and inhibition of InhA

Inhibition of InhA activity was investigated using InhA, InhA (S94A) and InhA (D148G) as described previously (Dessen et al, 1995; Quemard et al, 1995). Briefly, kinetic parameters were determined by following NADH oxidation at 340 nm using a TECAN FL200 spectrophotometer. All reactions were performed at 25 °C with 100 nM InhA (or mutants) in 30 mM PIPES (pH 6.8), 150 mM NaCl and 10% glycerol. After addition of variable concentrations of NADH, reactions were initiated by adding 2-trans-octenoyl-CoA (synthesized as described in Supporting Information Materials and Methods) to a final concentration of 250 μM. Steady state K_m for NADH was determined by measuring enzyme kinetics at different NADH concentrations (0–800 μM). NADH K_m and pyridomycin K_i were determined by measuring enzyme kinetics with both different NADH (wild type InhA – 10, 12.5, 16.7, 25 and 50 μM, for both InhA (S94A) and InhA (D148G) —50, 62.5, 83.3, 125, 250 μM NAHD) and different pyridomycin concentrations (0, 0.625, 1.25, 2.5, 5 and 10 μg/ml). Kinetic parameters were analysed and calculated using GraphPad Prism 5.

Crystallization, data collection and structure determination

Crystals of the InhA:NADH complex, InhA (D148G):NADH complex and InhA (S94A):NADH complex were obtained by vapour diffusion at 18°C by equilibrating 2 μl hanging drops containing a 1:1 mixture of approximately 10 mg/ml protein incubated with 5 mM NADH and crystallization buffer (10% v/v MPD and 0.1 M TRIS, pH 8.0 for InhA:NADH complex; 8% v/v MPD and 0.1 M Bicine, pH 9.0 for InhA (D148G):NADH complex; and 8% v/v MPD, 50 mM Sodium Citrate, pH 6.5 and 0.1 M Hepes, pH 7.5 for InhA (S94A):NADH complex) over a 500 μl reservoir of the same crystallization solution. Crystals were stabilized by soaking briefly in a cryoprotectant solution (25% glycerol w/v in crystallization buffer) and flash frozen in liquid nitrogen before data collection. Diffraction data were collected on X06DA of the Swiss Light Source (SLS, PSI, Villigen, Switzerland) and on ID29 at the European Synchrotron Radiation Facility (ESRF, Grenoble, France). Data were indexed, integrated and scaled with XDS (Kabsch, 2010). Both

wild-type InhA and its mutants crystallized in the P6222 space group with unit cell dimensions of approximately $a = b = 98$ Å, $c = 139$ Å, $\alpha = \beta = 90^\circ$ and $\gamma = 120^\circ$, with one molecule per asymmetric unit (Supporting Information Table 2). Phase determinations were carried out by molecular replacement using Phaser (McCoy et al, 2007), part of the CCP4 Suite, using as a search model the published structure of the InhA:NADH complex from *M. tuberculosis* (PDB code 3OEW). The initial molecular replacement models were manually adjusted in COOT, part of the CCP4 Suite (Winn et al, 2011) and refined with REFMAC5 (Murshudov et al, 1997). The refined structures were evaluated with PROCHECK (Laskowski et al, 1996). Structure figures were prepared with PyMOL (Molecular Graphics System, Version 1.5.0.1 Schrödinger, LLC). All crystallographic statistics are listed in Supporting Information Table 2. Coordinates and structure factors for the above complexes structures have been deposited in the Protein Data Bank (PDB accession codes 4DRE, 4DTI and 4DQU, respectively).

Author contributions

RCH, CS, JN, FP, SM, STC designed the experiments; RCH, CS, JN, FP, SM, SB, performed the experiments; RCH, CS, JN, FP, SM, RM, SU, KHA, STC analysed data; KHA contributed reagents; RCH, CS, FP, STC wrote the paper.

Acknowledgements

The authors would like to thank Drs. Catherine Vilchèze, Katia Jaton and Beatrice Ninet for kindly providing plasmids and clinical isolates, Prof. Paul J. Dyson for providing resources and support in pyridomycin purification, Dr. Keith Harshman for the Illumina sequencing, and Patricia Schneider, Philippe Busso and Anthony Vocat for technical assistance. The research leading to these results received funding from the European Community's Seventh Framework Programme (Grant 260872) and Systems X. ch. F.P. is a Marie Heim-Voegtlin fellow of the Swiss National Science Foundation. João Neres is the recipient of a Marie Curie fellowship from the European Commission (Grant 252802).

Supporting Information is available at EMBO Molecular Medicine Online.

The authors declare that they have no conflict of interest.

References

Baldock C, Rafferty JB, Sedelnikova SE, Baker PJ, Stuitje AR, Slabas AR, Hawkes TR, Rice DW (1996) A mechanism of drug action revealed by structural studies of enoyl reductase. Science 274: 2107-2110

Banerjee A, Dubnau E, Quemard A, Balasubramanian V, Um KS, Wilson T, Collins D, de Lisle G, Jacobs WR, Jr (1994) inhA, a gene encoding a target for isoniazid and ethionamide in *Mycobacterium tuberculosis*. Science 263: 227-230

Cegielski JP (2010) Extensively drug-resistant tuberculosis: "There must be some kind of way out of here." Clin Infect Dis 50: S195-S200

Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry CE III, *et al* (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature 393: 537-544

Cole ST, Riccardi G (2011) New tuberculosis drugs on the horizon. Curr Opin Microbiol 14: 570-576

Dessen A, Quemard A, Blanchard JS, Jacobs WR, Jr, Sacchettini JC (1995) Crystal structure and function of the isoniazid target of *Mycobacterium tuberculosis*. Science 267: 1638-1641

Dias MV, Vasconcelos IB, Prado AM, Fadel V, Basso LA, de Azevedo WF, Jr, Santos DS (2007) Crystallographic studies on the binding of isonicotinyl-NAD adduct to wild-type and isoniazid resistant 2-trans-enoyl-ACP (CoA) reductase from *Mycobacterium tuberculosis*. J Struct Biol 159: 369-380

Fischbach MA, Walsh CT (2009) Antibiotics for emerging pathogens. Science 325: 1089-1093

Freundlich JS, Wang F, Vilcheze C, Gulten G, Langley R, Schiehser GA, Jacobus DP, Jacobs WR, Jr, Sacchettini JC (2009) Triclosan derivatives: towards potent inhibitors of drug-sensitive and drug-resistant *Mycobacterium tuberculosis*. ChemMedChem 4: 241-248

Glaziou P, Floyd K, Raviglione M (2009) Global burden and epidemiology of tuberculosis. Clin Chest Med 30: 621-636 vii

Harrington M (2010) From HIV to tuberculosis and back again: a tale of activism in 2 pandemics. Clin Infect Dis 50: S260-S266

Heath RJ, White SW, Rock CO (2002) Inhibitors of fatty acid synthesis as antimicrobial chemotherapeutics. Appl Microbiol Biotechnol 58: 695-703

Huang TT, Wang YM, Yin J, Du YH, Tao MF, Xu J, Chen WQ, Lin SJ, Deng ZX (2011) Identification and characterization of the pyridomycin biosynthetic gene cluster of *Streptomyces pyridomyceticus* NRRL B-2517. J Biol Chem 286: 20648-20657

Kabsch W (2010) Xds. Acta Crystallogr D Biol Crystallogr 66: 125-132

Kinoshita M, Nakata M, Takarada K, Tatsuta K (1989) Synthetic studies of pyridomycin 5. Total synthesis of pyridomycin. Tetrahedron Lett 30: 7419-7422

Lamichhane G (2011) Novel targets in M. tuberculosis: search for new drugs. Trends Mol Med 17: 25-33

Larsen MH, Vilcheze C, Kremer L, Besra GS, Parsons L, Salfinger M, Heifets L, Hazbon MH, Alland D, Sacchettini JC, *et al* (2002) Overexpression of inhA, but not kasA, confers resistance to isoniazid and ethionamide in *Mycobacterium smegmatis*, *M. bovis* BCG and *M. tuberculosis*. Mol Microbiol 46: 453-466

Laskowski RA, Rullmannn JA, MacArthur MW, Kaptein R, Thornton JM (1996) AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. J Biomol NMR 8: 477-486

Lu H, Tonge PJ (2008) Inhibitors of FabI, an enzyme drug target in the bacterial fatty acid biosynthesis pathway. Acc Chem Res 41: 11-20

Maeda K (1957) Degradation studies on pyridomycin; chemical studies on antibiotics of Streptomyces. V. J Antibiot (Tokyo) 10: 94-106

Maeda K, Kosaka H, Okami Y, Umezawa H (1953) A new antibiotic, pyridomycin. J Antibiot (Tokyo) 6: 140

McCoy AJ, Grosse-Kunstleve RW, Adams PD, Winn MD, Storoni LC, Read RJ (2007) Phaser crystallographic software. J Appl Crystallogr 40: 658-674

McMurry LM, Oethinger M, Levy SB (1998) Triclosan targets lipid synthesis. Nature 394: 531-532

Molle V, Gulten G, Vilcheze C, Veyron-Churlet R, Zanella-Cleon I, Sacchettini JC, Jacobs WR, Jr, Kremer L (2010) Phosphorylation of InhA inhibits mycolic acid biosynthesis and growth of *Mycobacterium tuberculosis*. Mol Microbiol 78: 1591-1605

Morlock GP, Metchock B, Sikes D, Crawford JT, Cooksey RC (2003) ethA, inhA, and katG loci of ethionamide-resistant clinical *Mycobacterium tuberculosis* isolates. Antimicrob Agents Chemother 47: 3799-3805

Murshudov GN, Vagin AA, Dodson EJ (1997) Refinement of macromolecular structures by the maximum-likelihood method. Acta Crystallogr D Biol Crystallogr 53: 240-255

Ogawara H, Maeda K, Umezawa H (1968) The biosynthesis of pyridomycin. I. Biochemistry 7: 3296-3302

Oliveira JS, Pereira JH, Canduri F, Rodrigues NC, de Souza ON, de Azevedo WF, Jr, Basso LA, Santos DS (2006) Crystallographic and pre-steady-state

kinetics studies on binding of NADH to wild-type and isoniazid-resistant enoyl-ACP (CoA) reductase enzymes from *Mycobacterium tuberculosis*. J Mol Biol 359: 646-666

Palomino JC, Martin A, Camacho M, Guerra H, Swings J, Portaels F (2002) Resazurin microtiter assay plate: simple and inexpensive method for detection of drug resistance in *Mycobacterium tuberculosis*. Antimicrob Agents Chemother 46: 2720-2722

Parikh SL, Xiao G, Tonge PJ (2000) Inhibition of InhA, the enoyl reductase from *Mycobacterium tuberculosis*, by triclosan and isoniazid. Biochemistry 39: 7645-7650

Payne DJ, Gwynn MN, Holmes DJ, Pompliano DL (2007) Drugs for bad bugs: confronting the challenges of antibacterial discovery. Nat Rev Drug Discov 6: 29-40

Payne DJ, Miller WH, Berry V, Brosky J, Burgess WJ, Chen E, DeWolf WE, Jr, Fosberry AP, Greenwood R, Head MS, *et al* (2002) Discovery of a novel and potent class of FabI-directed antibacterial agents. Antimicrob Agents Chemother 46: 3118-3124

Quemard A, Sacchettini JC, Dessen A, Vilcheze C, Bittman R, Jacobs WR, Jr, Blanchard JS (1995) Enzymatic characterization of the target for isoniazid in *Mycobacterium tuberculosis*. Biochemistry 34: 8235-8241

Rawat R, Whitty A, Tonge PJ (2003) The isoniazid-NAD adduct is a slow, tight-binding inhibitor of InhA, the *Mycobacterium tuberculosis* enoyl reductase: adduct affinity and drug resistance. Proc Natl Acad Sci USA 100: 13881-13886

Rozwarski DA, Vilcheze C, Sugantino M, Bittman R, Sacchettini JC (1999) Crystal structure of the *Mycobacterium tuberculosis* enoyl-ACP reductase, InhA, in complex with NAD+ and a C16 fatty acyl substrate. J Biol Chem 274: 15582-15589

Sala C, Dhar N, Hartkoorn RC, Zhang M, Ha YH, Schneider P, Cole ST (2010) Simple model for testing drugs against nonreplicating *Mycobacterium tuberculosis*. Antimicrob Agents Chemother 54: 4150-4158

Shomura T, Amano S, Yoshida J, Kojima M (1986) *Dactylosporangium-fulvum* Sp-Nov. Int J Syst Bacteriol 36: 166-169

Smith PA, Roberts TC, Romesberg FE (2010) Broad-spectrum antibiotic activity of the arylomycin natural products is masked by natural target mutations. Chem Biol 17: 1223-1231

Sullivan TJ, Truglio JJ, Boyne ME, Novichenok P, Zhang X, Stratton CF, Li HJ, Kaur T, Amin A, Johnson F, *et al* (2006) High affinity InhA inhibitors with activity against drug-resistant strains of *Mycobacterium tuberculosis*. ACS Chem Biol 1: 43-53

Vilcheze C, Baughn AD, Tufariello J, Leung LW, Kuo M, Basler CF, Alland D, Sacchettini JC, Freundlich JS, Jacobs WR, Jr (2011) Novel inhibitors of InhA efficiently kill *Mycobacterium tuberculosis* under aerobic and anaerobic conditions. Antimicrob Agents Chemother 55: 3889-3898

Vilcheze C, Wang F, Arai M, Hazbon MH, Colangeli R, Kremer L, Weisbrod TR, Alland D, Sacchettini JC, Jacobs WR, Jr (2006) Transfer of a point mutation in *Mycobacterium tuberculosis* inhA resolves the target of isoniazid. Nat Med 12: 1027-1029

Winn MD, Ballard CC, Cowtan KD, Dodson EJ, Emsley P, Evans PR, Keegan RM, Krissinel EB, Leslie AG, McCoy A, *et al* (2011) Overview of the CCP4 suite and current developments. Acta Crystallogr D Biol Crystallogr 67: 235-242

Yagishita K (1954) Studies on the pyridomycin production by *Streptomyces albidofuscus*. I. On pyridomycin production of a lactose-utilizing mutant. J Antibiot (Tokyo) 7: 143-148

Yagishita K (1955) Studies on the pyridomycin production. II. X-ray irradiation on the pyridomycin-producing strain. J Antibiot (Tokyo) 8: 201-204

Yagishita K (1957a) Studies on the pyridomycin production. III. Medium selection and a device of a method of detecting the precursor. J Antibiot (Tokyo) 10: 5-14

Yagishita K (1957b) Studies on the pyridomycin production. IV. Metabolic studies on *Streptomyces pyridomyceticus*. J Antibiot (Tokyo) 10: 15-20

Zhang Y, Yew WW (2009) Mechanisms of drug resistance in *Mycobacterium tuberculosis*. Int J Tuberc Lung Dis 13: 1320-1330

Zhang YM, Rock CO (2008) Membrane lipid homeostasis in bacteria. Nat Rev Microbiol 6: 222-233

5. Conclusion and Perspectives

The work carried out in this thesis presents the successful application of microarray and next-generation sequencing (NGS) technologies in comparative and functional genomics research of *Mycobacterium tuberculosis* (*M. tb*), the etiologic agent of tuberculosis. We have established and employed computational strategies for the analysis, integration, and interpretation of genomic and transcriptomic datasets to improve our understanding of the evolution, physiology and virulence of the MTBC. We performed comparative genomic analysis to assess the genetic diversity in MTBC strains by means of SNP genotyping, automated Sanger sequencing and whole genome re-sequencing in order to reveal genomic differences that could have functional consequences. Our studies in functional genomics investigated the role of several transcription regulators in *M. tb* using ChIP-on-chip or ChIP-seq, and determined the transcriptome profiles of *M. tb* in different conditions with the use of cDNA microarrays or high-throughput cDNA sequencing (RNA-seq). Some of our findings may have direct implications in the development of diagnostics, vaccines and new drugs for TB control.

This chapter will recapitulate the major findings of the studies presented in this thesis and comment on other areas of TB research that could benefit from the applications of NGS (in addition to those summarized in section **2.2.4**). When applicable, there will be a discussion of cutting edge NGS methodologies that could further our efforts to uncover novel aspects of *M. tb* biology and pathogenesis. Finally, the potential utility of NGS in the clinical setting will be discussed along with the bioinformatics challenges associated with the interpretation of data generated by these technologies.

In order to promote the development of TB vaccines, it is necessary to improve our understanding of the genetic mechanisms contributing to BCG attenuation and to the variable efficacy among individual vaccine strains. Using high-throughput SNP genotyping we were able to identify a minimal set of 115 nonsynonymous SNPs that distinguish attenuated *M. bovis* BCG vaccine strains and virulent *M. bovis* strains. Two follow up publications by . have characterized the functional effect of some of the nsSNPs identified in our study (Mendoza Lopez *et al.* 2010; Chen *et al.* 2012). We also identified strain-specific SNPs among BCG strains that could be studied further to determine their functional consequences and possible contribution to the variation in protective efficacy of different BCG vaccines.

Identification of immunodominant *M. tb* antigens and studying their interaction with the host immune cells is crucial for the development of an effective TB control strategy. In order to evade the host immune system, variants of pathogens emerge, with alterations in epitope regions recognized by critical host immune cells. We characterized the sequence diversity in the highly immunodominant *esx* gene family among clinical isolates of *M. tb* in order to identify sequence polymorphisms that may affect their immunogenicity. Our work has led to the identification of non-synonymous mutations affecting known T-cell epitopes in EsxB (CFP-10) and EsxH (TB10.4). EsxB peptides are used in commercially available diagnostic tests for TB and the nonsynonymous substitutions may influence responses to EsxB peptides in such diagnostic tests. As EsxH has been used to obtain a fusion protein-based subunit vaccine failure to consider mycobacterial strain diversity could have a significant negative impact on vaccine efficacy.

We exploited the power of NGS by performing whole genome re-sequencing to detect sequence variation among *M. tb* strains, which could result in phenotypic differences. In one study, genome comparison of drug-resistant mutants with the sensitive wild-type strain we identify InhA as the target of the anti-TB drug, pyridomycin. Our finding that pyridomycin kills *M. tb* by inhibiting InhA (even in isoniazid-resistant clinical isolates) provides a promising basis for the development of pyridomycin or a related agent for the treatment of isoniazid-resistant tuberculosis. Using a similar approach we identified a spontaneous mutation that makes *M. tb* strains incapable of producing PDIMs (phthiocerol dimycocerosates), which are cell-wall associated lipids involved in *M. tb* virulence.

We applied two high-throughput approaches for detection of sequence variation - SNP-genotyping arrays and whole genome re-sequencing. NGS offers unprecedented coverage and resolution for detection of rare genetic variants, which is especially useful in case of genetically monomorphic species such as *M. tb*. As NGS becomes more affordable it is likely that whole genome re-sequencing will replace array-based SNP detection. In *M. tb* research, whole genome re-sequencing has been mostly used to study pathogen evolution and population structure. The ability to discriminate between closely related strains using NGS has provided us with an enhanced view of pathogen biology and allowed reconstruction and refinement of the existing MTBC phylogeny (Comas *et al.* 2010; Schürch 2011). NGS has also been valuable in TB drug discovery; re-sequencing of drug resistant mutants has been successfully used for identification of the targets for two anti-TB drugs, namely, TMC207 (Andries *et al.* 2005), and pyridomycin (Hartkoorn *et al.* 2012). Whole-genome comparisons surpass traditional typing methods in assessment of genetic diversity. Based on NGS data, Niemann *et al.* (2009)

uncovered significant genomic differences in drug sensitive and MDR-isolates of *M. tb* exhibiting identical IS6110-RFLP patterns, which may have important clinical implications.

It has been established that different *M. tb* strains have distinctive epidemiological and clinical characteristics (Lopez *et al.* 2003). The impact of differences in virulence and immunogenicity between *M. tb* on the outcome of disease is yet unclear. *M. tb* specific $CD4^{+}$ T-cell signatures have been shown to vary significantly among patients with active disease, latent infection, and upon initiation of anti-TB therapy (Harari *et al.* 2011). Our laboratory has started a collaborative project that involves detailed profiling of *M. tb* specific host cellular immune response associated with different disease outcomes, and in response to therapy. The immunity profiles will be correlated with genotypes of *M. tb* clinical isolates, which will be determined by whole genome re-sequencing.

In addition to the characterization of individual isolates, NGS can also be used to reveal the taxonomic and functional diversity of microbial communities. This will help us understand the impact of genetic diversity of *M. tb* diversity on immune response before and after anti-TB therapy. Metagenomics has been widely applied for profiling microbial communities from different habitats, such as aquatic ecosystems, soil, bioreactors, etc. The human microbiome project is an international effort that aims to characterize microbial communities at five different sites on the human body (gut, nasal, oral, vaginal, skin) and to correlate changes in microbial communities with disease states (Gevers *et al.* 2012). Metagenomic data analysis involves mapping of the reads directly against an entire database of complete or draft genomes. The results are represented as the relative abundance of different

species in the sample. Recent metagenomic studies of the lung microbiome revealed alterations in the lung flora of people suffering from respiratory diseases such as asthma and cystic fibrosis, compared to healthy individuals. A similar approach could be applied to understand the interaction of *M. tb* with the lung flora, and compare the lung microbiome during active disease and in response to therapy. The resulting information may promote the development of microbiome-based biomarkers for diagnosis and follow-up of anti-TB treatment.

In order to determine their biological significance, genomic differences need to be studied in a functional context. Gene expression and its regulation by transcription factors contribute to the complexity of biological function. Within transcriptional regulation the interactions of transcription factors (TFs) with DNA are of central importance. Using a combination of the two high-throughput approaches, we characterized the transcriptome of *M. tb* and studied the genome-wide dynamics of two key components of the transcription machinery, RNA polymerase (RNAP) and NusA, in exponential and stationary phases of growth. We demonstrated that NusA interacts with RNAP ubiquitously and that its profile mirrors RNAP distribution in both the exponential and stationary phases of growth. Differential binding of RNAP and NusA in the two growth conditions correlated with transcriptional activity as reflected by RNA abundance. By systematic integration of the ChIP-seq and RNA-seq data, we identified a set of transcription units (TU) in the *M. tb* genome, and mapped their putative promoters. Analysis of RNAP and NusA binding across the promoter and the body of TUs and their correlation with transcription offered new insights on the biological roles of RNAP and NusA in promoting transcription. The high resolution and accuracy offered by RNA-seq helped us uncover novel anti-sense and

intergenic transcripts and enabled refinement of the *M. tb* genome annotation. Our catalogue of the gene expression and the genome-wide dynamics of the RNA polymerase and NusA in exponential and stationary phases of growth is a valuable resource to the TB research community.

We also exploited the ChIP-on-chip and ChIP-seq technologies to define the regulon of a specific sigma factor, SigF, and gain a better understanding of the role of virulence regulator EspR. In the case of SigF, integration of the genome wide binding profile with transcriptomic data provided a complete set of genes whose expression was directly regulated by SigF. Our results confirmed that SigF controls its own expression, and regulates transcription of genes involved in lipid and intermediary metabolism, virulence, and even a small RNA. Together with the 3D-structure of EspR (Blasco *et al.* 2011), ChIP-seq data support its role as a novel nucleoid-associated protein (NAP) that plays architectural and regulatory roles that impact *M. tb* pathogenesis through multiple genes . The EspR regulon comprised genes encoding a host of diverse functions such as production of the complex PDIM lipids, the virulence-mediating ESX-1 cluster, PE/PPE cell-surface proteins, etc.

M. tb possesses a vast repertoire of regulatory elements as it encodes 13 sigma factors directing the specificity of the RNAP core enzyme to a defined subset of promoters and over 200 potential transcriptional regulators. In addition, a number of NAPs contribute to shaping the chromosome and impact transcriptional activity in *M. tb*. The studies mentioned above have been a major step towards our ultimate goal of generating a complete regulatory map of the *M. tb* genome to fully understand transcriptional control and uncover the role of diverse regulatory elements in different conditions.

As a continuation of our efforts, generation of ChIP-seq datasets using specific antibodies against several regulatory proteins, including the two-component response regulator PhoP, termination factor Rho, topology enzymes TopA and GyrA, and various sigma factors (SigA, SigB, SigE) will be carried out in the near future. When available, isogenic mutants will be used as a negative control in ChIP-seq and for comparison in RNA-seq. The ChIP-seq results can be integrated with the RNAP, NusA and RNA-seq datasets in different growth conditions in order to correlate the presence of specific regulatory proteins with transcriptional activity. Integration of binding profiles for sigma factors and RNAP with transcriptional activity will allow definition of promoter features and correlation of promoter strength with utilization of specific sigma factors in both exponential and stationary phase.

NAPs are the bacterial equivalent of histones and act by stabilizing long-range structures in the genome through cooperative binding to multiple sites. New technologies based on chromosome confirmation capture (3C) can be used to probe this higher level of regulation by NAPs (Dekker *et al.* 2002). The carbon-copy chromosome confirmation capture (5C) technique allows identification of physical interactions between distant DNA regions and can be used to generate a conformational map of the chromosome (Dostie *et al.* 2006). More recently, chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) has been developed and applied in eukaryotes to capture associations among genomic regions mediated by a particular protein through a ChIP step (Fullwood & Ruan 2009). As these technologies mature, they will be adapted to *M. tb* to elucidate spatial and temporal aspects of gene regulation by NAPs (EspR, H-NS, HupB, etc).

Generation of a regulatory map of the genome will require integration of the numerous NGS datasets using sophisticated data-mining, clustering and network biology approaches to identify interactions and dependencies within the transcription complex. Machine learning techniques, notably Bayesian networks can be used to integrate data from different kinds of experiments, and to find associations among diverse kinds of data. A better understanding of the molecular mechanisms at the basis of *M. tb* multiplication, survival and control of gene expression will underpin future translational research.

It is clear that genomic, transcriptomic and metagenomic data sets produced by NGS can provide us with unprecedented opportunities to address fundamental questions related to genome function and disease. NGS applications also hold great promise in the clinical and public health domain, particularly in areas of diagnosis, molecular epidemiology, and treatment control. Researchers have exploited the discriminatory power of whole genome sequencing in a number of recent outbreaks to investigate the real-time evolution of disease-associated clonal isolates, tracing person-to-person transmission and identifying sources of outbreaks. Some examples include, the cholera epidemic in Haiti (Chin *et al.* 2011)
Chin *et al.* 2011), the *E. coli* O104:H4 outbreak in Germany (Grad *et al.* 2012), MRSA (methicillin-resistant *Staphylococcus aureus*) cases in a hospital in Thailand (Harris *et al.* 2010), and a TB outbreak in British Columbia (Gardy *et al.* 2011). Whole genome sequencing offers a resolution superior to all existing molecular typing tools.

Drug susceptibility testing for bacterial pathogens can provide clues to the potential for disease progression and aid further investigation and clinical management of the disease. The use of whole genome re-sequencing as a

diagnostic method is appropriate only in organisms that exhibit a high correlation between phenotype and genotype. In this respect, *M. tb* represents an ideal candidate due to lack of heterogeneity between MTBC strains and the fact that many of the drug-resistance traits arise through point mutations or indels in single genes. Moreover, due to the slow growth rate of *M. tb*, the turnaround time for phenotypic susceptibility is too long. At present, the WHO endorsed genotypic test, namely the GeneXpert MTB/RIF assay, can rapidly distinguish MTBC from other acid-fast bacteria and detect rifampicin resistance (Helb *et al.* 2010). This test could be complemented by whole genome sequencing of the organism which could, over time, replace the remaining diagnostic functions including the precise species identification, susceptibility testing for remaining antibiotics, and epidemiological typing, which are currently achieved using a number of different techniques at reference laboratories.

NGS has the potential to transform clinical and public health microbiology but there are certain challenges that need to be addressed. The capacity to generate NGS data greatly outpaces our ability to analyze it. The lack of standardized and automated data interpretation represents the major hurdle to clinical implementation. Since the introduction of the first NGS platform in 2004, the field of NGS has been developing at a tremendous rate. Major advances include, reduction in the sequencing costs and turnaround time, and improvements in sequencing read length, depth and data accuracy offered by various NGS platforms.

The massive data produced by NGS presents a significant challenge for data storage, analyses and management. The development of a customizable data analysis pipeline for interpretation of NGS data is time and labour intensive,

and depends on a complex set of choices from among the many bioinformatics and statistical tools that are available. A variety of open-source and commercial data analysis tools for NGS are being developed. NGS data analysis is a multi-step process and the tools generally fit into the following broad categories: (1) quality control of raw data, (2) alignment of reads to a reference or *de novo* assembly in absence of a reference, (3) base-calling and detection of variation (i.e. SNPs, insertions and deletions), (4) method-specific data analysis (i.e. peak calling for ChIP-seq, analysis of differential gene expression for RNA-seq, etc.), (5) visualization and analysis in terms of the genome annotation.

The challenge for bioinformaticians is to create tools that are easy to use and provide output that can be readily interpreted by bench scientists and clinicians with no specialist knowledge of NGS analysis techniques. An ideal software would be platform and organism independent, able to support analysis of different types of NGS data and perform downstream tasks ranging from epidemiological tracking to drug susceptibility testing. This needs to be supported by comprehensive databases that would enable retrieval and comparison of sequence data from clinical isolate(s) with existing strains in the database. One must stress the importance of genome annotation for interpretation of how genetic sequences translate to resistance phenotypes. The reliability of NGS as a means of predicting antimicrobial susceptibility is critically dependent upon the availability of a current and curated database of reference sequences. In order to address the present challenges, future advances in NGS should focus on improvements not limited to but including the following areas: (1) reducing the quantity of starting material required in order to save time and effort needed for culturing, (2) decreasing costs of DNA library preparation prior to sequencing, (3)

improving our understanding of variation between different sequencing platforms, especially in terms of error profiles, (4) developing easy-to-use, standardized and replicable data assembly and analysis workflows, (5) constructing and maintaining accurate and readily accessible annotation databases for reference genomes, and, (6) maintaining public data repositories to allow easy retrieval of raw and pre-processed NGS data.

In conclusion, high-throughput approaches for comparative and functional genomics based on NGS offer new hope in the fight against *M. tb*. On the one hand, we have been able to obtain novel insights into bacterial evolution and disease etiology that will promote identification of new drug targets and antigens for vaccine development, and on the other hand we have the potential to develop tools to detect, monitor, and control pathogen threats in real time.

6. Acknowledgements

As I look back to the five years of study and research at EPFL, I feel deeply indebted to my advisors, colleagues, friends and family for their support and encouragement along this long but fulfilling journey. Despite the insufficiency of words to adequately express thanks to them, I must insist on mentioning their names in print!

I would like to start by expressing my gratitude to my thesis advisors Stewart Cole and Jacques Rougemont. I feel so fortunate to have worked under the guidance of Stewart Cole. His vision, immense knowledge, patience and generosity, have been my inspiration in the completion of this research work. I am grateful to Jacques Rougemont for his valuable insights and sound advice. He has taught me how to think like a bioinformatician!

I am thankful to my external collaborators Dr Stephen Gordon, Dr Beate Heym, and Dr Anna Tischler, and my colleagues in the Cole Lab, Dr Claudia Sala, Dr Ruben Hartkoorn, and Benjamin Blasco for their hard work and cooperation which has been instrumental in the success of my PhD.

I am extremely grateful to Claudia Sala, Jocelyne Lew, Rasika Uplekar, Chhavi Jain, and Sophie Magnet for their help and support in the final stages of my PhD.

I would like to thank the members of my thesis committee, Prof. Roland Brosch, Dr Keith Harshman, Prof. Bart Deplancke and Prof. John McKinney for their constructive feedback and encouragement.

I feel lucky to have been in the company of extremely talented and enthusiastic researchers in the Cole Lab and the Global Health Institute. I have shared so many wonderful memories with them, both inside and outside the lab. I will always cherish their friendship.

Finally, I would like to thank my parents, my sister, my family members and my friends for their unconditional love and encouragement. I hope I have made you proud!

7. Bibliography

Aaron, L. *et al.*, 2004. Tuberculosis in HIV-infected patients: a comprehensive review. *Clinical Microbiology and Infection*, 10(5), pp.388–398.

Abdallah, A.M. *et al.*, 2008. The ESX-5 secretion system of *Mycobacterium marinum* modulates the macrophage response. *Journal of immunology (Baltimore, MD: 1950)*, 181(10), pp.7166–7175.

Abdallah, A.M. *et al.*, 2007. Type VII secretion - mycobacteria show the way. *Nature reviews. Microbiology*, 5(11), pp.883–891.

Alderson, M.R. *et al.*, 2000. Expression cloning of an immunodominant family of *Mycobacterium tuberculosis* antigens using human CD4(+) T cells. *The Journal of experimental medicine*, 191(3), pp.551–560.

Allison, D.B. *et al.*, 2006. Microarray data analysis: from disarray to consolidation and consensus. *Nature Reviews Genetics*, 7(1), pp.55–65.

Allix-Béguec, C. *et al.*, 2008. Evaluation and strategy for use of MIRU-VNTRplus, a multifunctional database for online analysis of genotyping data and phylogenetic identification of *Mycobacterium tuberculosis* complex isolates. *Journal of clinical microbiology*, 46(8), pp.2692–2699.

Anders, S. & Huber, W., 2010. Differential expression analysis for sequence count data. *Genome biology*, 11(10), p.R106.

Andersen, P. *et al.*, 1995. Recall of long-lived immunity to *Mycobacterium tuberculosis* infection in mice. *Journal of immunology (Baltimore, MD: 1950)*, 154(7), pp.3359–3372.

Andersen, P. *et al.*, 1991. T-cell proliferative response to antigens secreted by *Mycobacterium tuberculosis*.

Andries, K. *et al.*, 2005. A diarylquinoline drug active on the ATP synthase of Mycobacterium tuberculosis. *Science (New York, N.Y.)*, 307(5707), pp.223–227.

GLOBAL PLAN TO STOP TB 2006-2015, Stop TB Partnership, 2006. Available at: http://www.stoptb.org/assets/documents/global/plan/GlobalPlanFinal.pdf [Accessed July 13, 2012].

Global TB Vaccine Pipeline, Stop TB Partnership, 2011. Available at: http://www.stoptb.org/wg/new_vaccines/assets/documents/Global%20TB%20Vaccine%20Pipeline_Mar%202012.ppt [Accessed July 13, 2012a].

Global Tuberculosis Control 2011, World Health Organization, 2011. Available at: http://whqlibdoc.who.int/publications/2011/9789241564380_eng.pdf [Accessed July 13, 2012b].

International Roadmap for Tuberculosis Research, Stop TB Partnership, 2011.[Accessed July 10, 2012c].

MDR and XDR-TB progress report 2011, WHO Stop TB Department, 2011. Available at: http://www.who.int/entity/tb/challenges/mdr/factsheet_mdr_progress_march2011.pdf [Accessed July 10, 2012d].

Working Group on New TB Drugs, Stop TB Partnership, 2012. Available at: http://www.newtbdrugs.org/pipeline.php [Accessed July 10, 2012].

Arnvig, K. & Young, D., 2012. Non-coding RNA and its potential role in *Mycobacterium tuberculosis* pathogenesis. *RNA biology*, 9(4).

Arnvig, K.B. & Young, D.B., 2009. Identification of small RNAs in *Mycobacterium tuberculosis*. *Molecular microbiology*, 73(3), pp.397–408.

Bailey, T.L. *et al.*, 2009. MEME SUITE: tools for motif discovery and searching. *Nucleic acids research*, 37(Web Server issue), pp.W202–8.

Baker, L. *et al.*, 2004. Silent nucleotide polymorphisms and a phylogeny for *Mycobacterium tuberculosis*. *Emerging Infectious Diseases*, 10(9), pp.1568–1577.

Behr, M.A. *et al.*, 1999. Comparative genomics of BCG vaccines by whole-genome DNA microarray. *Science (New York, N.Y.)*, 284(5419), pp.1520–1523.

Benjamini, Y. & Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, 57, pp.289–300.

Bercovier, H., Kafri, O. & Sela, S., 1986. Mycobacteria possess a surprisingly small number of ribosomal RNA genes in relation to the size of their genome. *Biochemical and Biophysical Research Communications*, 136(3), pp.1136–1141.

Betts, J.C. *et al.*, 2002. Evaluation of a nutrient starvation model of *Mycobacterium tuberculosis* persistence by gene and protein expression profiling. *Molecular microbiology*, 43(3), pp.717–731.

Betts, J.C. *et al.*, 2003. Signature gene expression profiles discriminate between isoniazid-, thiolactomycin-, and triclosan-treated *Mycobacterium tuberculosis*. *Antimicrobial agents and chemotherapy*, 47(9), pp.2903–2913.

Blasco, B. *et al.*, 2011. Atypical DNA recognition mechanism used by the EspR virulence regulator of *Mycobacterium tuberculosis*. *Molecular microbiology*, 82(1), pp.251–264.

Blasco, B. *et al.*, 2012. Virulence Regulator EspR of *Mycobacterium tuberculosis* Is a Nucleoid-Associated Protein E. J. Rubin, ed. *PLoS pathogens*, 8(3), p.e1002621.

Boehme, C.C. *et al.*, 2010. Rapid molecular detection of tuberculosis and rifampin resistance. *The New England journal of medicine*, 363(11), pp.1005–1015.

Boshoff, H.I.M. *et al.*, 2004. The transcriptional responses of *Mycobacterium tuberculosis* to inhibitors of metabolism: novel insights into drug mechanisms of action. *The Journal of biological chemistry*, 279(38), pp.40174–40184.

Bottai, D. *et al.*, 2012. Disruption of the ESX-5 system of *Mycobacterium tuberculosis* causes loss of PPE protein secretion, reduction of cell wall integrity and strong attenuation. *Molecular microbiology*, 83(6), pp.1195–1209.

Brennan, P.J., 2012. A new TB vaccine blueprint. *Tuberculosis (Edinburgh, Scotland)*, 92 Suppl 1, pp.S1–S1.

Brennan, P.J., 2003. Structure, function, and biogenesis of the cell wall of *Mycobacterium tuberculosis*. In Tuberculosis. pp. 91–97.

Brosch, R. *et al.*, 2002. A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. *Proceedings of the National Academy of Sciences of the United States of America*, 99(6), pp.3684–3689.

Brosch, R. *et al.*, 2007. Genome plasticity of BCG and impact on vaccine efficacy. *Proceedings of the National Academy of Sciences*, 104(13), pp.5596–5601.

Brosch, R. *et al.*, 1999. Genomic analysis reveals variation between *Mycobacterium tuberculosis* H37Rv and the attenuated *M. tuberculosis* H37Ra strain. *Infection and immunity*, 67(11), pp.5768–5774.

Brosch, R. *et al.*, 1998. Use of a *Mycobacterium tuberculosis* H37Rv Bacterial Artificial Chromosome Library for Genome Mapping, Sequencing, and Comparative Genomics. *Infection and immunity*, 66(5), pp.2221–2229.

Brudey, K. *et al.*, 2006. *Mycobacterium tuberculosis* complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology. *BMC microbiology*, 6, p.23.

Buck, M.J. & Lieb, J.D., 2004. ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics*, 83(3), pp.349–360.

Calmette, A., 1931. Preventive Vaccination Against Tuberculosis with BCG. *Proceedings of the Royal Society of Medicine*, 24(11), pp.1481–1490.

Camus, J. *et al.*, 2002a. TubercuList: a Regularly Updated Database Dedicated to the *Mycobacterium tuberculosis* Genome.

Camus, J.-C. *et al.*, 2002b. Re-annotation of the genome sequence of *Mycobacterium tuberculosis* H37Rv.

Carver, T. *et al.*, 2012. Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics (Oxford, England)*, 28(4), pp.464–469.

Cavusoglu, C. *et al.*, 2002. Characterization of rpoB mutations in rifampin-resistant clinical isolates of *Mycobacterium tuberculosis* from Turkey by DNA sequencing and line probe assay. *Journal of clinical microbiology*, 40(12), pp.4435–4438.

Charlet, D. *et al.*, 2005. Reduced expression of antigenic proteins MPB70 and MPB83 in *Mycobacterium bovis* BCG strains due to a start codon mutation in sigK. *Molecular microbiology*, 56(5), pp.1302–1313.

Chen, J.M. *et al.*, 2012. A Point Mutation in cycA Partially Contributes to the D-cycloserine Resistance Trait of *Mycobacterium bovis* BCG Vaccine Strains. *PloS one*.

Chin, C.-S. *et al.*, 2011. The origin of the Haitian cholera outbreak strain. *The New England journal of medicine*, 364(1), pp.33–42.

Cole, S. *et al.*, 1998. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature*, 393(6685), pp.537–.

Cole, S.T., 2002a. Comparative and functional genomics of the *Mycobacterium tuberculosis* complex. *Microbiology (Reading, England)*, 148(Pt 10), pp.2919–2928.

Cole, S.T., 2002b. Comparative mycobacterial genomics as a tool for drug target and antigen discovery. *The European respiratory journal. Supplement*, 36, pp.78s–86s.

Cole, S.T. & Barrell, B.G., 1998. *Genetics and Tuberculosis: Novartis Foundation Symposium 217* Novartis Foundation, D. J. Chadwick, & G. Cardew, eds., Chichester, UK: John Wiley & Sons, Ltd.

Cole, S.T. & Riccardi, G., 2011. New tuberculosis drugs on the horizon. *Current opinion in microbiology*, 14(5), pp.570–576.

Cole, S.T. *et al.*, 2001. Massive gene decay in the leprosy bacillus. *Nature*, 409(6823), pp.1007–1011.

Collins, D.M., 2003. Different susceptibility of two animal species infected with isogenic mutants of *Mycobacterium bovis* identifies phoT as having roles in tuberculosis virulence and phosphate transport. *Microbiology (Reading, England)*, 149(11), pp.3203–3212.

Comas, I. *et al.*, 2009. Genotyping of genetically monomorphic bacteria: DNA sequencing in *Mycobacterium tuberculosis* highlights the limitations of current methodologies. *PloS one*, 4(11), p.e7815.

Comas, I. *et al.*, 2010. Human T cell epitopes of *Mycobacterium tuberculosis* are evolutionarily hyperconserved. *Nature genetics*, 42(6), pp.498–503.

Corbett, E.L. *et al.*, 2006. Tuberculosis in sub-Saharan Africa: opportunities, challenges, and change in the era of antiretroviral treatment. *Lancet*, 367(9514), pp.926–937.

Daniel, T.M., 2006. The history of tuberculosis. *Respiratory Medicine*, 100(11), pp.1862–1870.

Davids, V. *et al.*, 2006. The Effect of Bacille Calmette-Guérin Vaccine Strain and Route of Administration on Induced Immune Responses in Vaccinated Infants. *Journal of Infectious Diseases*, 193(4), pp.531–536.

Dekker, J. *et al.*, 2002. Capturing chromosome conformation. *Science (New York, N.Y.)*, 295(5558), pp.1306–1311.

DiChiara, J.M. *et al.*, 2010. Multiple small RNAs identified in *Mycobacterium bovis* BCG are also expressed in *Mycobacterium tuberculosis* and *Mycobacterium smegmatis*. *Nucleic acids research*, 38(12), pp.4067–4078.

Dillon, S.C. & Dorman, C.J., 2010. Bacterial nucleoid-associated proteins, nucleoid structure and gene expression. *Nature reviews. Microbiology*, 8(3), pp.185–195.

Dostie, J. *et al.*, 2006. Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome research*, 16(10), pp.1299–1309.

Dreszer, T.R. *et al.*, 2012. The UCSC Genome Browser database: extensions and updates 2011. *Nucleic acids research*, 40(Database issue), pp.D918–23.

Dubos, R., 1952. *The white plague: tuberculosis, man, and society*,

Dunne, W.M., Westblade, L.F. & Ford, B., 2012. Next-generation and whole-genome sequencing in the diagnostic clinical microbiology laboratory. *European journal of clinical microbiology & infectious diseases : official publication of the European Society of Clinical Microbiology*.

Dye, C. *et al.*, 1999. Consensus statement. Global burden of tuberculosis: estimated incidence, prevalence, and mortality by country. WHO Global Surveillance and Monitoring Project. *JAMA : the journal of the American Medical Association*, 282(7), pp.677–686.

Ernst, J.D., Trevejo-Nuñez, G. & Banaiee, N., 2007. Genomics and the evolution, pathogenesis, and diagnosis of tuberculosis. *Journal of Clinical Investigation*, 117(7), pp.1738–1745.

Espinal, M.A. *et al.*, 1999. Rational "DOTS plus" for the control of MDR-TB. *The international journal of tuberculosis and lung disease: the official journal of the International Union against Tuberculosis and Lung Disease*, 3(7), pp.561–563.

Ewing, B. & Green, P., 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome research*, 8(3), pp.186–194.

Fang, Z., Martin, J.A. & Wang, Z., 2012. Statistical methods for identifying differentially expressed genes in RNA-Seq experiments. *Cell & bioscience*, 2(1), p.26.

Filliol, I. *et al.*, 2006. Global phylogeny of *Mycobacterium tuberculosis* based on single nucleotide polymorphism (SNP) analysis: insights into tuberculosis evolution, phylogenetic accuracy other DNA fingerprinting systems, and recommendations for a minimal standard SNP set. *Journal of bacteriology*, 88(2), pp.759–772.

Fine, P.E., 1995. Variation in protection by BCG: implications of and for heterologous immunity. *The Lancet*, 346(8986), pp.1339–1345.

Fleischmann, R.D. *et al.*, 2002. Whole-genome comparison of *Mycobacterium tuberculosis* clinical and laboratory strains. *Journal of bacteriology*, 184(19), pp.5479–5490.

Fortune, S.M. *et al.*, 2005. Mutually dependent secretion of proteins required for mycobacterial virulence. *Proceedings of the National Academy of Sciences of the United States of America*, 102(30), pp.10676–10681.

Frigui, W. *et al.*, 2008. Control of *M. tuberculosis* ESAT-6 secretion and specific T cell recognition by PhoP. *PLoS pathogens*, 4(2), p.e33.

Fullwood, M.J. & Ruan, Y., 2009. ChIP-based methods for the identification of long-range chromatin interactions. *Journal of cellular biochemistry*, 107(1), pp.30–39.

Gagneux, S. *et al.*, 2006. Variable host-pathogen compatibility in *Mycobacterium tuberculosis*. *Proceedings of the National Academy of Sciences of the United States of America*, 103(8), pp.2869–2873.

Gardy, J.L. *et al.*, 2011. Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. *The New England journal of medicine*, 364(8), pp.730–739.

Garnier, T., 2003. The complete genome sequence of *Mycobacterium bovis*. *Proceedings of the National Academy of Sciences*, 100(13), pp.7877–7882.

Gentleman, R.C. *et al.*, 2004. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, 5(10), p.R80.

Gerdes, K., Christensen, S.K. & Løbner-Olesen, A., 2005. Prokaryotic toxin–antitoxin stress response loci. *Nature reviews. Microbiology*, 3(5), pp.371–382.

Gevers, D. *et al.*, 2012. The human microbiome project: a community resource for the healthy human microbiome. *PLoS biology*, 10(8), p.e1001377.

Gey Van Pittius, N.C. *et al.*, 2001. The ESAT-6 gene cluster of *Mycobacterium*

tuberculosis and other high G+C Gram-positive bacteria. *Genome biology*, 2(10).

Gioffré, A. *et al.*, 2005. Mutation in mce operons attenuates *Mycobacterium tuberculosis* virulence. *Microbes and Infection*, 7(3), pp.325–334.

Girling, D. *et al.*, 1976. SHORT-COURSE CHEMOTHERAPY OF TUBERCULOSIS. *The Lancet*, 307(7955), pp.358–359.

Glickman, M.S. & Jacobs, W.R., 2001. Microbial pathogenesis of *Mycobacterium tuberculosis*: dawn of a discipline. *Cell*, 104(4), pp.477–485.

Gordon, B.R.G. *et al.*, 2010. Lsr2 is a nucleoid-associated protein that targets AT-rich sequences and virulence genes in *Mycobacterium tuberculosis*. *Proceedings of the National Academy of Sciences*, 107(11), pp.5154–5159.

Gordon, S.V. *et al.*, 1999. Identification of variable regions in the genomes of tubercle bacilli using bacterial artificial chromosome arrays. *Molecular microbiology*, 32(3), pp.643–655.

Grad, Y.H. *et al.*, 2012. Genomic epidemiology of the *Escherichia coli* O104:H4 outbreaks in Europe, 2011. *Proceedings of the National Academy of Sciences*, 109(8), pp.3065–3070.

Groenen, P.M. *et al.*, 1993. Nature of DNA polymorphism in the direct repeat cluster of *Mycobacterium tuberculosis*; application for strain differentiation by a novel typing method. *Molecular microbiology*, 10(5), pp.1057–1065.

Gutacker MM, Smoot JC, Migliaccio CAL, *et al.* Genome-wide analysis of synonymous single nucleotide polymorphisms in *Mycobacterium tuberculosis* complex organisms: resolution of genetic relationships among closely related microbial strains. *Genetics* 2002; 162: 1533–43.

Harari, A. *et al.*, 2011. Dominant TNF-α+ *Mycobacterium tuberculosis*-specific CD4+ T cell responses discriminate between latent infection and active disease. *Nature medicine*, 17(3), pp.372–376.

Harrington, C.A., Rosenow, C. & Retief, J., 2000. Monitoring gene expression using DNA microarrays. *Current opinion in microbiology*, 3(3), pp.285–291.

Harris, S.R. *et al.*, 2010. Evolution of MRSA during hospital transmission and intercontinental spread. *Science (New York, N.Y.)*, 327(5964), pp.469–474.

Harshman, K. & Martínez-A, C., 2002. DNA microarrays: a bridge between genome sequence information and biological understanding. *European Review*, 10(03), pp.389–408.

Hartkoorn, R., 2012. Towards a new tuberculosis drug: pyridomycin – nature's isoniazid. *EMBO Molecular Medicine*, pp.1–33.

Hartkoorn, R.C. *et al.*, 2012. Genome-wide Definition of the SigF Regulon in *Mycobacterium tuberculosis*. *Journal of bacteriology*.

Helb, D. *et al.*, 2010. Rapid detection of *Mycobacterium tuberculosis* and rifampin resistance by use of on-demand, near-patient technology. *Journal of clinical microbiology*, 48(1), pp.229–237.

Hirsh, A.E. *et al.*, 2004. Stable association between strains of *Mycobacterium tuberculosis* and their human host populations. *Proceedings of the National Academy of Sciences of the United States of America*, 101(14), pp.4871–4876.

Horsburgh, C.R., 1991. *Mycobacterium avium* complex infection in the acquired immunodeficiency syndrome. *The New England journal of medicine*, 324(19), pp.1332–1338.

Ioerger, T.R. *et al.*, 2009. Genome Analysis of Multi- and Extensively-Drug-Resistant Tuberculosis from KwaZulu-Natal, South Africa. *PloS one*, 4(11), p.e7778.

Ishikawa, S. *et al.*, 2010. RNA polymerase trafficking in *Bacillus subtilis* cells. *Journal of bacteriology*, 192(21), pp.5778–5787.

Kamerbeek, J. *et al.*, 1997. Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. *Journal of clinical microbiology*, 35(4), pp.907–914.

Kaufmann, S. & Hussey, G., 2010. New vaccines for tuberculosis. *The Lancet*.

Kumar, P. *et al.*, 2009. The *Mycobacterium tuberculosis* protein kinase K modulates activation of transcription from the promoter of mycobacterial monooxygenase operon through phosphorylation of the transcriptional regulator VirS. *The Journal of biological chemistry*, 284(17), pp.11090–11099.

Lagranderie, M.R. *et al.*, 1996. Comparison of immune responses of mice immunized with five different *Mycobacterium bovis* BCG vaccine strains. *Infection and immunity*, 64(1), pp.1–9.

Langmead, B. *et al.*, 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*, 10(3), p.R25.

Lawn, S.D.S. & Zumla, A.I.A., 2011. Tuberculosis. *Audio, Transactions of the IRE Professional Group on*, 378(9785), pp.57–72.

Lee, J.-H., Karakousis, P.C. & Bishai, W.R., 2008. Roles of SigB and SigF in the *Mycobacterium tuberculosis* sigma factor network. *Journal of bacteriology*, 190(2), pp.699–707.

Legrand, E. *et al.*, 2001. Use of spoligotyping to study the evolution of the direct repeat locus by IS6110 transposition in *Mycobacterium tuberculosis*. *Journal of clinical*

microbiology, 39(4), pp.1595–1599.

Lew, J.M. *et al.*, 2011. TubercuList-10 years after. *Tuberculosis (Edinburgh, Scotland)*, 91(1), pp.1–7.

Lewis, K.N. *et al.*, 2003. Deletion of RD1 from *Mycobacterium tuberculosis* Mimics Bacille Calmette-Guérin Attenuation. *Journal of Infectious Diseases*, 187(1), pp.117–123.

Li, H. *et al.*, 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16), pp.2078–2079.

Li, H., Ruan, J. & Durbin, R., 2008a. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome research*, 18(11), pp.1851–1858.

Li, R. *et al.*, 2008b. SOAP: short oligonucleotide alignment program. *Bioinformatics (Oxford, England)*, 24(5), pp.713–714.

Ling, D.I., Flores, L.L., *et al.*, 2008a. Commercial nucleic-acid amplification tests for diagnosis of pulmonary tuberculosis in respiratory specimens: meta-analysis and meta-regression. *PloS one*, 3(2), p.e1536.

Ling, D.I., Zwerling, A.A. & Pai, M., 2008b. GenoType MTBDR assays for the diagnosis of multidrug-resistant tuberculosis: a meta-analysis. *European Respiratory Journal*, 32(5), pp.1165–1174.

Lopez, B. *et al.*, 2003. A marked difference in pathogenesis and immune response induced by different *Mycobacterium tuberculosis* genotypes. *Clinical & Experimental Immunology*, 133(1), pp.30–37.

López-Campos, G. *et al.*, 2012. *Microarray Detection and Characterization of Bacterial Foodborne Pathogens*, Boston, MA: Springer US.

Lönnroth, K. *et al.*, 2009. Drivers of tuberculosis epidemics: the role of risk factors and social determinants. *Social science & medicine (1982)*, 68(12), pp.2240–2246.

Lun, D.S. *et al.*, 2009. A blind deconvolution approach to high-resolution mapping of transcription factor binding sites from ChIP-seq data. *Genome biology*, 10(12).

MacGurn, J.A. *et al.*, 2005. A non-RD1 gene cluster is required for Snm secretion in *Mycobacterium tuberculosis*. *Molecular microbiology*, 57(6), pp.1653–1663.

Mahairas, G.G. *et al.*, 1996. Molecular analysis of genetic differences between *Mycobacterium bovis* BCG and virulent *M. bovis*. *Journal of bacteriology*, 178(5), pp.1274–1282.

Manca, C. *et al.*, 1999. *Mycobacterium tuberculosis* catalase and peroxidase activities and resistance to oxidative killing in human monocytes in vitro. *Infection and*

immunity, 67(1), pp.74–79.

Manca, C. *et al.*, 2001. Virulence of a *Mycobacterium tuberculosis* clinical isolate in mice is determined by failure to induce Th1 type immunity and is associated with induction of IFN-alpha /beta. *Proceedings of the National Academy of Sciences of the United States of America*, 98(10), pp.5752–5757.

Manganelli, R. *et al.*, 2004. Sigma factors and global gene regulation in *Mycobacterium tuberculosis*. *Journal of bacteriology*, 186(4), pp.895–902.

Manganelli, R. *et al.*, 2001. The *Mycobacterium tuberculosis* ECF sigma factor sigmaE: role in global gene expression and survival in macrophages. *Molecular microbiology*, 41(2), pp.423–437.

Mardis, E.R., 2008. The impact of next-generation sequencing technology on genetics. *Trends in genetics*, 24(3), pp.133–141.

Marmiesse, M. *et al.*, 2004. Macro-array and bioinformatic analyses reveal mycobacterial "core" genes, variation in the ESAT-6 gene family and new phylogenetic markers for the *Mycobacterium tuberculosis* complex. *Microbiology (Reading, England)*, 150(Pt 2), pp.483–496.

Mazurek, G.H. *et al.*, 2001. Comparison of a whole-blood interferon gamma assay with tuberculin skin testing for detecting latent *Mycobacterium tuberculosis* infection. *JAMA : the journal of the American Medical Association*, 286(14), pp.1740–1747.

Meier, T. *et al.*, 2005. Sensitivity of a new commercial enzyme-linked immunospot assay (T SPOT-TB) for diagnosis of tuberculosis in clinical practice. *European journal of clinical microbiology & infectious diseases : official publication of the European Society of Clinical Microbiology*, 24(8), pp.529–536.

Mendoza Lopez, P. *et al.*, 2010. Characterization of the transcriptional regulator Rv3124 of *Mycobacterium tuberculosis* identifies it as a positive regulator of molybdopterin biosynthesis and defines the functional consequences of a non-synonymous SNP in the *Mycobacterium bovis* BCG orthologue. *Microbiology (Reading, England)*, 156(Pt 7), pp.2112–2123.

Metzker, M.L., 2009. Sequencing technologies — the next generation. *Nature Reviews Genetics*, 11(1), pp.31–46.

Migliori, G.B. *et al.*, 2007. First tuberculosis cases in Italy resistant to all tested drugs. *Euro surveillance : bulletin européen sur les maladies transmissibles European communicable disease bulletin*, 12(5), p.E070517.1.

Mishra, A.K. *et al.*, 2011. Lipoarabinomannan and related glycoconjugates: structure, biogenesis and role in *Mycobacterium tuberculosis* physiology and host-pathogen interaction. *FEMS microbiology reviews*, 35(6), pp.1126–1157.

Mooney, R.A. *et al.*, 2009. Regulator trafficking on bacterial transcription units in vivo. *Molecular cell*, 33(1), pp.97–108.

Morgan, M. *et al.*, 2005. A commercial line probe assay for the rapid detection of rifampicin resistance in *Mycobacterium tuberculosis*: a systematic review and meta-analysis. *BMC infectious diseases*, 5, pp.62–.

Mostowy, S., 2003. The in vitro evolution of BCG vaccines. *Vaccine*, 21(27-30), pp.4270–4274.

Mostowy, S. *et al.*, 2002. Genomic deletions suggest a phylogeny for the *Mycobacterium tuberculosis* complex. *Journal of Infectious Diseases*, 186(1), pp.74–80.

Muñoz-Elías, E.J. *et al.*, 2006. Role of the methylcitrate cycle in *Mycobacterium tuberculosis* metabolism, intracellular growth, and virulence. *Molecular microbiology*, 60(5), pp.1109–1122.

Murray, J.F., 2004. A Century of Tuberculosis. *American Journal of Respiratory and Critical Care Medicine*, 169(11), pp.1181–1186.

N Sarita Shah, *et al.*, 2011. Increasing Drug Resistance in Extensively Drug-Resistant Tuberculosis, South Africa. *Emerging Infectious Diseases*, 17(3), p.510.

Nicol, M.P.M. & Wilkinson, R.J.R., 2008. The clinical consequences of strain diversity in *Mycobacterium tuberculosis*. *Journal of Cell Biology*, 102(10), pp.955–965.

Oettinger, T. *et al.*, 1999. Development of the *Mycobacterium bovis* BCG vaccine: review of the historical and biochemical evidence for a genealogical tree. *Tubercle and lung disease: the official journal of the International Union against Tuberculosis and Lung Disease*, 79(4), pp.243–250.

Pai, M., Zwerling, A. & Menzies, D., 2008. Systematic review: T-cell-based assays for the diagnosis of latent tuberculosis infection: an update. *Annals of internal medicine*, 149(3), pp.177–184.

Pan, W., 2002. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments.

Pareek, C.S., Smoczynski, R. & Tretyn, A., 2011. Sequencing technologies and genome sequencing. *Journal of applied genetics*, 52(4), pp.413–435.

Park, H.-D. *et al.*, 2003. Rv3133c/dosR is a transcription factor that mediates the hypoxic response of *Mycobacterium tuberculosis*. *Molecular microbiology*, 48(3), pp.833–843.

Park, P.J., 2009. ChIP–seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics*, 10(10), pp.669–680.

Pellin, D. *et al.*, 2012. A genome-wide identification analysis of small regulatory RNAs in *Mycobacterium tuberculosis* by RNA-Seq and conservation analysis. *PloS one*, 7(3), p.e32723.

Prod'hom, G. *et al.*, 1997. Rapid discrimination of *Mycobacterium tuberculosis* complex strains by ligation-mediated PCR fingerprint analysis. *Journal of clinical microbiology*, 35(12), pp.3331–3334.

Pym, A.S. *et al.*, 2002. Loss of RD1 contributed to the attenuation of the live tuberculosis vaccines *Mycobacterium bovis* BCG and *Mycobacterium microti*. *Molecular microbiology*, 46(3), pp.709–717.

Quackenbush, J., 2001. Computational analysis of microarray data. *Nature Reviews Genetics*.

Raghavan, S. *et al.*, 2008. Secreted transcription factor controls *Mycobacterium tuberculosis* virulence. *Nature*, 454(7205), pp.717–721.

Ramage, H.R., Connolly, L.E. & Cox, J.S., 2009. Comprehensive functional analysis of *Mycobacterium tuberculosis* toxin-antitoxin systems: implications for pathogenesis, stress responses, and evolution. *PLoS genetics*, 5(12), p.e1000767.

Ramakrishnan, L., Federspiel, N.A. & Falkow, S., 2000. Granuloma-specific expression of Mycobacterium virulence proteins from the glycine-rich PE-PGRS family. *Science (New York, N.Y.)*, 288(5470), pp.1436–1439.

Ramaswamy, S.V. *et al.*, 2000. Molecular genetic analysis of nucleotide polymorphisms associated with ethambutol resistance in human isolates of *Mycobacterium tuberculosis*. *Antimicrobial agents and chemotherapy*, 44(2), pp.326–336.

Raviglione, M. *et al.*, 2012. Scaling up interventions to achieve global tuberculosis control: progress and new developments. *Lancet*, 379(9829), pp.1902–1913.

Roberts, D.M. *et al.*, 2004. Two sensor kinases contribute to the hypoxic response of *Mycobacterium tuberculosis*. *The Journal of biological chemistry*, 279(22), pp.23082–23087.

Rodriguez, G.M. *et al.*, 2002. ideR, An essential gene in mycobacterium tuberculosis: role of IdeR in iron-dependent gene expression, iron metabolism, and oxidative stress response. *Infection and immunity*, 70(7), pp.3371–3381.

Rougemont, J. & Naef, F., 2012. Computational analysis of protein-DNA interactions from ChIP-seq data. *Methods in molecular biology (Clifton, N.J.)*, 786, pp.263–273.

Saini, D.K. *et al.*, 2004. DevR-DevS is a bona fide two-component system of *Mycobacterium tuberculosis* that is hypoxia-responsive in the absence of the DNA-

binding domain of DevR. *Microbiology (Reading, England)*, 150(Pt 4), pp.865–875.

Sala, C. & Hartkoorn, R.C., 2011. Tuberculosis drugs: new candidates and how to find more. *Future microbiology*, 6(6), pp.617–633.

Sala, C. *et al.*, 2009. Genome-wide regulon and crystal structure of BlaI (Rv1846c) from *Mycobacterium tuberculosis*. *Molecular microbiology*, 71(5), pp.1102–1116.

Sassetti, C.M., Boyd, D.H. & Rubin, E.J., 2003. Genes required for mycobacterial growth defined by high-density mutagenesis. *Molecular microbiology*, 48(1), pp.77–84.

Schaaf, H.S. & Zumla, A., 2009. *Tuberculosis: A Comprehensive Clinical Reference* 1st ed. eds., Saunders.

Schnappinger, D. *et al.*, 2003. Transcriptional Adaptation of *Mycobacterium tuberculosis* within Macrophages: Insights into the Phagosomal Environment. *The Journal of experimental medicine*, 198(5), pp.693–704.

Schürch, A., 2011. DNA fingerprinting of *Mycobacterium tuberculosis*: from phage typing to whole-genome sequencing. *Infection*.

Scorpio, A. *et al.*, 1997. Rapid differentiation of bovine and human tubercle bacilli based on a characteristic mutation in the bovine pyrazinamidase gene. *Journal of clinical microbiology*, 35(1), pp.106–110.

Shabbeer, A. *et al.*, 2012. Web tools for molecular epidemiology of tuberculosis. *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases*, 12(4), pp.767–781.

Sharma, K. *et al.*, 2006. Transcriptional control of the mycobacterial embCAB operon by PknH through a regulatory protein, EmbR, in vivo. *Journal of bacteriology*, 188(8), pp.2936–2944.

Sherman, D.R. *et al.*, 2001. Regulation of the *Mycobacterium tuberculosis* hypoxic response gene encoding alpha -crystallin. *Proceedings of the National Academy of Sciences of the United States of America*, 98(13), pp.7534–7539.

Siegrist, M.S. *et al.*, 2009. Mycobacterial Esx-3 is required for mycobactin-mediated iron acquisition. *Proceedings of the National Academy of Sciences*, 106(44), pp.18792–18797.

Siméone, R., Bottai, D. & Brosch, R., 2009. ESX/type VII secretion systems and their role in host–pathogen interaction. *Current opinion in microbiology*, 12(1), pp.4–10.

Simpson, J.T. *et al.*, 2009. ABySS: a parallel assembler for short read sequence data. *Genome research*, 19(6), pp.1117–1123.

Skjøt, R.L.V. *et al.*, 2002. Epitope mapping of the immunodominant antigen TB10.4 and the two homologous proteins TB10.3 and TB12.9, which constitute a subfamily of the esat-6 gene family. *Infection and immunity*, 70(10), pp.5446–5453.

Small, P.M. *et al.*, 1994. The Epidemiology of Tuberculosis in San-Francisco - a Population-Based Study Using Conventional and Molecular Methods. *New England Journal of Medicine*, 330(24), pp.1703–1709.

Smith, N.H. *et al.*, 2006. Ecotypes of the *Mycobacterium tuberculosis* complex. *Journal of theoretical biology*, 239(2), pp.220–225.

Smith, N.H. *et al.*, 2003. The population structure of *Mycobacterium bovis* in Great Britain: clonal expansion. *Proceedings of the National Academy of Sciences of the United States of America*, 100(25), pp.15271–15275.

Smyth, G.K. & Speed, T., 2003. Normalization of cDNA microarray data. *Methods (San Diego, Calif.)*, 31(4), pp.265–273.

Sorek, R., Kunin, V. & Hugenholtz, P., 2008. CRISPR - a widespread system that provides acquired resistance against phages in bacteria and archaea. *Nature reviews. Microbiology*, 6(3), pp.181–186.

Sreevatsan, S. *et al.*, 1996. Identification of a polymorphic nucleotide in oxyR specific for *Mycobacterium bovis*. *Journal of clinical microbiology*, 34(8), pp.2007–2010.

Sreevatsan, S. *et al.*, 1997. Restricted structural gene polymorphism in the *Mycobacterium tuberculosis* complex indicates evolutionarily recent global dissemination. *Proceedings of the National Academy of Sciences of the United States of America*, 94(18), pp.9869–9874.

Sterling, T.R. *et al.*, 2011. Three months of rifapentine and isoniazid for latent tuberculosis infection. *The New England journal of medicine*, 365(23), pp.2155–2166.

Stermann, M. *et al.*, 2004. A promoter mutation causes differential nitrate reductase activity of *Mycobacterium tuberculosis* and *Mycobacterium bovis*. *Journal of bacteriology*, 186(9), pp.2856–2861.

Supply, P. *et al.*, 1997. Identification of novel intergenic repetitive units in a mycobacterial two-component system operon. *Molecular microbiology*, 26(5), pp.991–1003.

Supply, P. *et al.*, 2003. Linkage disequilibrium between minisatellite loci supports clonal evolution of *Mycobacterium tuberculosis* in a high tuberculosis incidence area. *Molecular microbiology*, 47(2), pp.529–538.

Supply, P. *et al.*, 2006. Proposal for standardization of optimized mycobacterial interspersed repetitive unit-variable-number tandem repeat typing of *Mycobacterium*

tuberculosis. *Journal of clinical microbiology*, 44(12), pp.4498–4510.

Supply, P. *et al.*, 2000. Variable human minisatellite-like regions in the *Mycobacterium tuberculosis* genome. *Molecular microbiology*, 36(3), pp.762–771.

Tekaia, F. *et al.*, 1999. Analysis of the proteome of *Mycobacterium tuberculosis* in silico. *Tubercle and lung disease : the official journal of the International Union against Tuberculosis and Lung Disease*, 79(6), pp.329–342.

Thompson, W., McCue, L.A. & Lawrence, C.E., 2005. Using the Gibbs motif sampler to find conserved domains in DNA and protein sequences. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]*, Chapter 2, p.Unit 2.8.

Thorvaldsdóttir, H., Robinson, J.T. & Mesirov, J.P., 2012. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics*.

Tibshirani, R., 2006. A simple method for assessing sample sizes in microarray experiments. *BMC bioinformatics*, 7, p.106.

Trachtenberg, A.J. *et al.*, 2012. A primer on the current state of microarray technologies. *Methods in molecular biology (Clifton, N.J.)*, 802, pp.3–17.

Trunz, B.B., Fine, P. & Dye, C., 2006. Effect of BCG vaccination on childhood tuberculous meningitis and miliary tuberculosis worldwide: a meta-analysis and assessment of cost-effectiveness. *Lancet*, 367(9517), pp.1173–1180.

Tsolaki, A.G. *et al.*, 2004. Functional and evolutionary genomics of *Mycobacterium tuberculosis*: insights from genomic deletions in 100 strains.

Valouev, A. *et al.*, 2008. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nature methods*, 5(9), pp.829–834.

Valway, S.E. *et al.*, 1998. An outbreak involving extensive transmission of a virulent strain of *Mycobacterium tuberculosis*. *The New England journal of medicine*, 338(10), pp.633–639.

van Embden, J.D. *et al.*, 1993. Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology. *Journal of clinical microbiology*, 31(2), pp.406–409.

van Rie, A. *et al.*, 1999. Exogenous reinfection as a cause of recurrent tuberculosis after curative treatment. *The New England journal of medicine*, 341(16), pp.1174–1179.

Voelkerding, K.V., Dames, S.A. & Durtschi, J.D., 2009. Next-Generation Sequencing: From Basic Research to Diagnostics. *Clinical Chemistry*, 55(4).

Voskuil, M.I. *et al.*, 2003. Inhibition of respiration by nitric oxide induces a *Mycobacterium tuberculosis* dormancy program. *The Journal of experimental medicine*, 198(5), pp.705–713.

Waddell, S.J. *et al.*, 2004. The use of microarray analysis to determine the gene expression profiles of *Mycobacterium tuberculosis* in response to anti-bacterial compounds. *Tuberculosis*, 84(3-4), pp.263–274.

Wallis, R.S. *et al.*, 2010. Biomarkers and diagnostics for tuberculosis: progress, needs, and translation into practice. *Lancet*, 375(9729), pp.1920–1937.

Wang, L. *et al.*, 2009. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics (Oxford, England)*, 26(1), pp.136–138.

Warren, R.L. *et al.*, 2007. Assembling millions of short DNA sequences using SSAKE. *Bioinformatics (Oxford, England)*, 23(4), pp.500–501.

Wilson, D.L. *et al.*, 2003. New normalization methods for cDNA microarray data. *Bioinformatics (Oxford, England)*, 19(11), pp.1325–1332.

Wilson, M. *et al.*, 1999. Exploring drug-induced alterations in gene expression in Mycobacterium tuberculosis by microarray hybridization.

Young, D.B., Gideon, H.P. & Wilkinson, R.J., 2009. Eliminating latent tuberculosis. *Trends in microbiology*, 17(5), pp.183–188.

Zerbino, D.R. & Birney, E., 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research*, 18(5), pp.821–829.

Zhang, F. & Xie, J.-P., 2011. Mammalian cell entry gene family of *Mycobacterium tuberculosis*. *Molecular and cellular biochemistry*, 352(1-2), pp.1–10.

Zhang, Y. *et al.*, 1992. The catalase—peroxidase gene and isoniazid resistance of *Mycobacterium tuberculosis*. *Nature*, 358(6387), pp.591–593.

Zheng, H. *et al.*, 2008. Genetic basis of virulence attenuation revealed by comparative genomic analysis of *Mycobacterium tuberculosis* strain H37Ra versus H37Rv. *PloS one*, 3(6), p.e2375.

Zumla, A. *et al.*, 2009. Reflections on the white plague. *The Lancet infectious diseases*, 9(3), pp.197–202.

Zumla, A., Hafner, R., *et al.*, 2012a. Advancing the development of tuberculosis therapy. *Nature Reviews Drug Discovery*, 11(3), pp.171–172.

Zumla, A.A., Abubakar, I.I., *et al.*, 2012b. Drug-resistant tuberculosis-current dilemmas, unanswered questions, challenges, and priority needs. *Journal of Infectious Diseases*, 205 Suppl 2, pp.S228–S240.

Miss Swapna Uplekar

Rue de Crissier 3, 1020 Renens VD, Switzerland
+ 41 78 920 1983 • uplekar@gmail.com

Bioinformatics graduate with undergraduate training in molecular biology and biochemistry. Interested in applications of genomics and bioinformatics for the study of infectious diseases.

EDUCATION AND WORK EXPERIENCE

Since Nov 2007 — **Doctoral Program in Biotechnology and Bioengineering**
École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
Advisors: Prof. Stewart Cole and Dr. Jacques Rougemont

Thesis: Comparative and functional genomics of *Mycobacterium tuberculosis*.
Analysis of data generated using ChIP-chip, ChIP-seq, RNA-seq and whole genome resequencing of *M. tuberculosis* isolates.

2006 – 2007 (June – Feb) — **Interim Course Director and Lecturer**
Welingkar Institute of Management, Mumbai, India

Role: Lecturer in Perl programming and academic coordinator for the MSc Bioinformatics program held in collaboration with Nottingham Trent University, UK.

2004 – 2005 (Sept – Sept) — **M Res (Master of Research) in Bioinformatics**
University of Newcastle, Newcastle-upon-Tyne, England, UK

Projects: Development of a sequence annotation workflow.
Metabolic reconstruction and comparison of the Genus *Bacillus*.

2001 – 2004 (Sept – June) — **BSc in Biochemistry with Molecular Biology & Biotechnology**
University of Bristol, Bristol, England, UK

Project: Role of Neuroplastin in monocarboxylate transport.

2003 – 2003 (May – July) — **Summer Internship in Bioinformatics**
Centre for DNA Fingerprinting and Diagnostics, Hyderabad, India

Project: Phylogenetic characterization of rice tungro bacilliform virus.

1999 – 2001 (June – June) — **Higher Secondary School Certificate (Baccalaureate)**
Ramnivas Ruia College, Mumbai, India

Subjects: Biology, Chemistry, Physics, Mathematics, English, Sanskrit

TECHNICAL SKILLS

Bioinformatics Packages: Comparative genomics and phylogenetics, structural biology, high-throughput sequencing data assembly, analysis and visualization tools.

Programming: Perl, Python Basics, R (Statistics), Microsoft Excel.

Informatics: Mac OS, Unix and Windows operating systems.

Well versed with commonly used Molecular Biology and Biochemistry techniques.

CONFERENCES AND WORKSHOPS

Genome Informatics meeting at Cold Spring Harbor Laboratory, New York.
"Towards a regulatory map of the genome of *M. tb*", poster presentation.

Eighth International Conference on Pathogenesis of Mycobacterial Infections, Sweden.
"Comparative genomics of *esx* genes from clinical isolates of *M. tb*", poster presentation.

AFFILIATIONS

- Systems-X iPhD (interdisciplinary) candidate, Swiss Initiative in Systems Biology
- Member of Swiss Institute of Bioinformatics (since 2008)
- Student representative, PhD in Biotechnology and Bioengineering (2007-08)

EXTRA-CURRICULAR

- President of Yuva, Indian Student Association at EPFL (since 2011)
- Trained extensively in keyboards and musical theory
- Enthusiastic traveller and photographer

LANGUAGES

English (Fluent), French (Elementary), Hindi (Fluent), Marathi (Fluent)

PUBLICATIONS

Chen JM, **Uplekar S**, Gordon SV, Cole ST. **A Point Mutation in** *cycA* **Partially Contributes to the D-cycloserine Resistance Trait of** *Mycobacterium bovis* **BCG Vaccine Strains**. *PLoS ONE* (2012).

Blasco B, Chen JM, Hartkoorn R, Sala C, **Uplekar S**, Rougemont J, Pojer F & Cole ST. **Virulence Regulator EspR of** *Mycobacterium tuberculosis* **is a Nucleoid-Associated Protein.** *PLoS Pathogens.* **8**, e1002621 (2012).

Hartkoorn RC, Sala C, **Uplekar S**, Busso P, Rougemont J & Cole ST. **Genome-wide Definition of the SigF Regulon in** *Mycobacterium tuberculosis***.** *J Bacteriology* (2012).

Uplekar S, Heym B, Friocourt V, Rougemont, J & Cole, ST. **Comparative genomics of Esx genes from clinical isolates of** *Mycobacterium tuberculosis* **provides evidence for gene conversion and epitope variation.** *Infection. Immunity.* **79**, 4042–4049 (2011).

Kirksey MA, Tischler AD, Siméone R, Hisert KB, **Uplekar S**, Guilhot C & McKinney JD. **Spontaneous phthiocerol dimycocerosate-deficient variants of** *Mycobacterium tuberculosis* **are susceptible to gamma interferon-mediated immunity.** *Infection. Immunity.* **79**, 2829–2838 (2011).

Massouras A, Hens K, Gubelmann C, **Uplekar S**, Decouttere F, Rougemont J, Cole ST & Deplancke B. **Primer-initiated sequence synthesis to detect and assemble structural variants.** *Nature Methods* **7**, 485–486 (2010).

Garcia Pelayo MC*, **Uplekar S*,** Keniry A, Mendoza Lopez P, Garnier T, Nunez Garcia J, Boschiroli L, Zhou X, Parkhill J, Smith N, Hewinson RG, Cole ST & Gordon SV. **A comprehensive survey of single nucleotide polymorphisms (SNPs) across** *Mycobacterium bovis* **strains and** *M. bovis* **BCG vaccine strains refines the genealogy and defines a minimal set of SNPs that separate virulent** *M. bovis* **strains and** *M. bovis* **BCG strains.** *Infection. Immunity.* **77**, 2230–2238 (2009).