Universidade do Minho
Escola de Ciências da Saúde

Hanna Raquel Nebenzahl Guimaraes

A genomic exploration of transmissibility in *Mycobacterium tuberculosis*

julho de 2016

Universidade do Minho
Escola de Ciências da Saúde

Hanna Raquel Nebenzahl Guimaraes

A genomic exploration of transmissibility in *Mycobacterium tuberculosis*

Tese de Doutoramento em Ciências da Saúde

Trabalho efetuado sob a orientação da
Prof. Doutora Margarida Correia-Neves
e da
Prof. Doutora Megan Murray

julho de 2016

DECLARAÇÃO

Nome: Hanna Raquel Nebenzahl Guimaraes

Endereço Electrónico: hanna.guimaraes@gmail.com

Número do Bilhete de Identidade: 12670480

Título da tese:

A genomic exploration of transmissibility in *Mycobacterium tuberculosis*

Orientadoras:

Prof. Doutora Margarida Correia-Neves

Prof. Doutora Megan Murray

Ano de conclusão: 2016

Designação do Doutoramento: Ciências da Saúde

Universidade do Minho, 28 de Julho de 2016

Assinatura: _____

STATEMENT OF INTEGRITY

I hereby declare having conducted my thesis with integrity. I confirm that I have not used plagiarism or any form of falsification of results in the process of the thesis elaboration.

I further declare that I have fully acknowledged the Code of Ethical Conduct of the University of Minho.

University of Minho, 28 de Julho de 2016

Full name:

Hanna Raquel Nebenzahl Guimaraes

Signature: _____

Agradecimentos

À Escola de Ciências da Saúde agradeço a grandissima oportunidade de realizar o doutoramento em parceria com mais do que uma instituição de renome no estrangeiro.

À Prof. Margarida Correia Neves e à Prof. Joana Palha agradeço a confiança no projeto proposto, e pela flexibilidade para com as várias modificacoes logisticas e geográficas.

À minha co-orientadora Prof. Megan Murray agradeço todos os ensinamentos transmitidos, para não mencionar a oportunidade de fazer parte de uma equipe estimulante e de alto rigor científico. Ao Prof. Dick van Soolingen agradeço o acolhimento simpático e abertura para usufruir da sua magnífica coleção de estirpes.

Aos colaboradores pelo caminho, entre os quais o Prof. Martien Borgdorff e Prof. Reinout van Crevel, agradeço as discussões proveitosas.

Ao meu marido, por toda a ajuda na minha aprendizagem em programação, e muito mais. Ao meu filho, pela compania in utero, e a Golda, pela compania cá fora. E aos meus pais, que puseram a minha educaçao à frente de tudo mais, e que por muitos anos sobreviveram as saudades.

Abstract

The ability of *Mycobacterium tuberculosis* (Mtb) to be transmitted from host to host is not well understood. Previous molecular epidemiology studies have shown that while some clinical strains of Mtb are able to cause infection and disease in a large number of individuals exposed to them, others are confined in their transmission, despite the ample chance for the spread of the infection. Since preventing transmission of Mtb is the key to a continued decline in tuberculosis cases, understanding the host and bacterial factors that are associated with transmissibility could be useful in developing strategies to prevent transmission.

Previous work has focused on cluster size as a measurable proxy for transmissibility, and several studies have found that host risk factors are associated with clustering and cluster size. This thesis set out to explore if and what bacterial factors, such as phylogenetic lineage and genomic markers, lie behind an increased transmissibility phenotype. We describe a novel approach, called the Propensity to Propagate (PPP), with which to adjust for host risk factors when quantifying transmissibility. Using this method, we found no significant differences to propagate between four different lineages within the Netherlands, as measured by molecular-typing defined cluster sizes. When looking more specifically at infectivity (as defined by mean number of positive contacts around each patient) and number of secondary cases within two years after diagnosis of an index case sharing the same fingerprint, we found evidence of phylogenetic lineage influencing these two indicators, namely, a decreased ability to infect and a lower secondary case rate in ancient phylogenetic lineages (*Mycobacterium africanum* and EAI) compared to their modern counterparts (Euro-American, Beijing, and CAS).

One simple approach to discovering more specific genetic regions behind transmissibility involves checking the absence/presence of mutations in the genes of interest between transmissible and non-transmissible phenotypes. In one of our studies, a multivariate logistic regression-based analysis of patient-, microorganism- and disease-related factors failed to reveal any significant association between frameshift-causing indels in Mycobacterium cyclase/LuxR-like genes (mclxs) and transmissibility.

Finally, using a large, well-characterized, complete data set of typed strains to identify strains found in large clusters as a proxy for a transmission phenotype as well as related strains that have not been transmitted, we selected 100 bacterial isolates after controlling for epidemiologic and host factors that may influence transmission. After whole genome sequencing, we subjected them to evolutionary convergence analysis. We identified six bacterial DNA regions - espE, PE-PGRS33, PE-PGRS56, Rv0197, Rv2813-14c and Rv2815-16c - to be associated with Mtb transmission and validated these regions by studying the response of human white blood cells to extracts from a subset of the tuberculosis bacteria that carried or did not carry mutations in these DNA regions. We show that there are differences in the immune response – as reflected by in vitro monocyte and T-cell cytokine production, reactive oxygen species release and neutrophil apoptosis - that associate with these genetic changes.

These findings not only contribute to our understanding of the interplay of bacterial factors in creating more successful strains at transmitting, but also have implications in the future of disease surveillance and curbing of transmission, by providing for instance tools with which to flag patients carrying particularly transmissible strains.

Resumo

A forma com que a bacteria *Mycobacterium tuberculosis* (Mtb) é transmitida de um hospedeiro para outro não está ainda bem estudada. Estudos de epidemiologia molecular têm demonstrado que, enquanto que algumas estirpes de Mtb tendem a causar infecção e doença num grande número de indivíduos, sugerindo uma grande capacidade de transmissão entre estes, outras apresentam uma propagação restrita, independentemente de terem elevadas oportunidades de disseminação. Uma vez que a prevenção da transmissão da Mtb é fundamental para o declínio continuado da tuberculose, o estudo dos fatores bacterianos que estão associados à transmissibilidade poderá ser útil para o desenvolvimento de novas estratégias de controlo da tuberculose.

Trabalhos anteriores focaram-se no tamanho dos clusters como medida de transmissibilidade, e vários estudos demonstraram uma associação entre os fatores de risco do hospedeiro e o agrupamento e tamanho dos clusters. Nesta foi explorada a existência de fatores bacterianos, tais como linhagem filogenética ou marcadores genéticos, responsáveis por um fenótipo de maior ou menor transmissibilidade. Descrevemos assim uma nova abordagem, chamada propensão para propagar (PPP), com a qual é possível corrigir o viés dos factores de risco do hospedeiro na quantificação da transmissibilidade de uma estirpe. Ao aplicar este método não foi possível detectar diferenças significativas de propagação - considerando o tamanho de clusters definidos por tipagem molecular - entre quatro linhagens diferentes presentes na Holanda. Mas uma análise específica da capacidade de infetar (definida pelo número médio de contatos positivos de cada doente) e do número de casos secundários, no espaço de dois anos após o diagnóstico de um caso índice com o mesmo perfil genético, determinou que a linhagem filogenética influencia estes dois indicadores. Concretamente, as linhagens filogenéticas mais antigas (*Mycobacterium africanum* e EAI) apresentam uma menor capacidade de infectar e um menor número de casos secundários quando comparadas com as suas equivalentes modernas (Euro-Americano, Beijing, e CAS).

Uma abordagem simples para identificar regiões genéticas específicas responsáveis pelas diferenças na transmissibilidade das estirpes envolve a análise da distribuição de mutações nos genes de interesse entre o fenótipo transmissível e o não-transmissível. Neste sentido uma análise baseada em regressão logística multivariada de fatores relacionados com o doente, o microorganismo ou a doença, não revelou qualquer associação significativa entre frameshift-causing indels (a presença de inserções ou deleções nucleotídicas causando a interrupção precoce da grelha de leitura dos genes) em genes Mycobacterium ciclase / LuxR-like (mclxs) e a transmissibilidade.

Finalmente, uma grande coleção de isolados bem caracterizados e sujeitos a tipagem molecular foi usada para identificar estirpes pertencentes a clusters grandes - representativas de um fenótipo de transmissão elevada - bem como estirpes com propagação limitada. Selecionamos 100 estirpes tendo em consideração os factores epidemiológicos do hospedeiro com influência na transmissibilidade da estirpe. Após sequenciação do genoma, estas estirpes foram submetidas a uma análise de convergência evolutiva. Identificamos seis genes/regiões intergénicas - Espe, PE PGRS33, PE PGRS56, Rv0197, e Rv2815-16c Rv2813-14c - como estando associados com a transmissão de Mtb, e validamos estes resultados através da análise da resposta de leucócitos a extractos de bactérias com ou sem as mutações nos seis genes/regiões intergénicas mencionados. Demostramos que existem diferenças na resposta imunitária – em termos da produção in vitro de citoquinas pelos monócitos e células T, espécies reativas de oxigénio e apoptose de neutrófilos - associadas às alterações genéticas estudadas.

As conclusões desta tese não só contribuem para a melhor compreensão da interação de fatores bacterianos no estabelecimento de linhagens com uma maior transmissibilidade, como têm implicações para o futuro da vigilância e contenção da transmissão, providenciando, por exemplo, as ferramentas necessárias para a identificação de doentes portadores de estirpes particularmente transmissíveis.

The work presented in this thesis was performed in the Department of Epidemiology of the Harvard School of Public Health, Boston, United States, at the Tuberculosis Reference Laboratory of the National Institute for Public Health and the Environment, Bilthoven, the Netherlands, and at the Microbiology and Infection Research Domain in the Life and Health Sciences Research Institute (ICVS), School of Health Sciences, University of Minho, Braga, Portugal (ICVS/3B's – PT Government Associate Laboratory, Braga/ Guimarães, Portugal). The financial support was given by the Fundação para a Ciência e Tecnologia (FCT) by means of a PhD grant (SFRH/BD/69390/2010).

Table of Contents

Chapter I

General Introduction

1.1 OVERALL INTRODUCTION

Although a largely curable disease, tuberculosis (TB) remains a major cause of morbidity and mortality worldwide, with over 8.7 million new cases and 1.4 million deaths in 2011 (World Health Organization 2015). Caused by *Mycobacterium tuberculosis* (Mtb), an acid-fast, intracellular bacillus, it is a disease predominantly of the lungs (approximately 70% of cases), although Mtb can disseminate to other organs, including lymph nodes, bone and meninges (Harisinghani et al. 2000). Following infection, one can either spontaneously clear the infection, progress to disease, or remain asymptomatic (latent), a result of the inability of the host to eliminate the bacteria but to at least control it's growth (Stewart et al. 2003). Only 5-10% of such latent cases will develop active TB disease in their lifetimes.

The control of the global TB epidemic has been thwarted by the lack of sensitive and rapid diagnostics, given that clinical data, albeit important, is insufficient for diagnoses. X-rays cannot exclude extrapulmonary TB and may not even be available in countries where resources are limited. Smear microscopy of sputum is inexpensive, simple, and results are available within hours, but the sensitivity is only about 50-60%, or even lower in countries with a high prevalence of both pulmonary TB and HIV infection (Siddiqi et al. 2003). The Tuberculosis Skin Test (TST), in use since 1910, is based on a protein–purified derivate resulting from a culture filtrate of tubercle bacilli containing over 200 antigens common both in bacilli Chalmette-Guerin vaccine (BCG) and in most non TB bacteria. As such, the test specificity is low and also the ability to distinguish latent infections is limited, decreasing its usefulness in a high prevalence setting. Newer tests, such as the Interferon gamma release assays (IGRAs), work by measuring the interferon gamma cytokine, a proxy of the person's immune response to the bacteria. Even though results are available within 24 hours, and prior BCG vaccination does not cause false positive results, IGRAs are most often used in low prevalence resource rich settings due to its high cost and necessary laboratory facilities. Discordance between TST and IGRA occurs in 10-20% of individuals, but the underlying mechanisms are poorly understood (Jones-Lopez et al. 2015; Ribeiro-Rodrigues et al. 2014).

Treatment of TB spans four to nine months on a cocktail of various first- and second-line antibiotics, depending on whether the disease is active or latent, and on the drug-resistance profile of the strain. Treating drug-resistant TB, which does not respond to the main drugs used for TB, requires people to endure a longer treatment course of up to twenty pills a day, and in the early stages of treatment, a daily injection. The side effects of such treatment are severe and some of the drugs are very expensive.

1.2 THE BIOLOGY OF *MYCOBACTERIUM TUBERCULOSIS* TRANSMISSION

Mtb is transmitted almost exclusively by inhalation of droplet nuclei bearing Mtb particles released from the lungs of patients with pulmonary or laryngeal disease. Bacteria that traverse the mouth or nasal passages, upper respiratory tract, and bronchi to reach the alveoli of the lung are phagocytised by macrophages, where they can initiate rounds of intracellular replication and cell lysis (O'Garra et al. 2013).

Macrophages are key effector cells in mycobacterial killing, evoking a vigorous host cellular immune response involving cytokines and a large number of chemokines. But, they can also provide a niche for bacterial multiplication. Dendritic cells then engulf bacteria, or bacterial components, circulate to the draining lymph nodes and prime T cells, which then return to the lungs to orchestrate control of the infection (Orme et al. 2014). T cells enhance the antibacterial activity of macrophages by releasing cytokines, such as interferon-γ, which generally results in arrest or clearance of the infection. If the immune response is insufficient to control the initial infection, clinical symptoms and associated pathology, including tissue necrosis and cavitation, will develop within ~1 year in the form of primary progressive disease. Individuals with cavitary TB are especially infectious (Rodrigo et al. 1997), since lung tissue destruction leads to the formation of macroscopic open spaces that contain numerous Mtb bacilli and connect to large airways, facilitating efficient expectoration of the bacteria (Kaplan et al. 2003).

Several lines of evidence indicate that — in addition to their widely known roles in protecting an infected individual from rapidly lethal TB — human T cell responses actually contribute to the lung tissue

destruction underlying cavitary TB, thereby enhancing to host-to-host TB transmission. Multiple studies have revealed that individuals with TB who are co-infected with HIV have a lower frequency of cavitary TB, and a recent systematic review revealed a linear correlation between the number of circulating CD4+ T cells and the frequency of cavitary TB (Kwan & Ernst 2011). Indeed, HIV-infected people transmit TB less efficiently than do HIV-uninfected people (Kwan & Ernst 2011). It is unclear whether the effect of CD4+ T cells on the promotion of cavitary TB is direct or indirect, and the mechanisms by which CD4+ T cells contribute to lung tissue damage and cavitary TB are not well characterized.

Compared with many other diseases, the timescales involved in TB are long and there is large variation between different individuals. Most individuals in fact develop a T-cell response in the absence of any clinical symptoms, which is defined as a latent infection, indicated by a positive TST or IGRA. In latent patients, bacteria can persist within infected tissues, such as granulomas, as well as in other sites that function to contain bacterial spread (Ramakrishnan 2012). Transmission from latent patients is only possible if there is reactivation of the initial infection (Silva Miranda et al. 2012). Isoniazid preventive therapy – the use of isoniazid monodrug therapy to interrupt progression from latent infection to active TB – helps in avoiding patients from transmitting in the future (Churchyard et al. 2014; World Health Organization 1982).

In most instances, patients respond to antibiotic treatment by clearance of the bacilli from tissues, partial reversal of the granulomatous process, and clinical cure. As such, starting appropriate therapy as early as possible not only contributes towards more timely curing of the patient, but also makes it less likely for drug resistance to develop and for the bacteria to be transmitted.

Prior vaccination with BCG, a live attenuated strain that is closely related to Mtb, establishes a primed population of T cells and protects against severe childhood forms of disease, including milliary and extrapulmonary TB and the often fatal TB meningitis (Roy et al. 2014). It also confers protection against leprosy. With a long-established safety profile and inexpensive cost, it remains the most widely

used vaccine in the world, currently compulsory in ≥64 countries and administered in >167 countries (Zwerling et al. 2011). However, the level of protection conferred by BCG is extremely variable, differing by target population given to, form of pulmonary TB and whether there is HIV co-infection. The reason for this remains a topic of active research — one of the hypotheses gleaned from newer genomic evidence points to differences between the BCG strains themselves, which have evolved from the original variant used in 1921 (Behr 2002). Meanwhile, a vaccine with reliable efficacy in preventing transmission of the infection does not yet exist (Franco-Paredes et al. 2006).

1.3 DETERMINANTS OF TRANSMISSION

An extensive crosstalk between internal (bacterial) and external (host and environmental) factors may influence the success of the bacteria in transmitting.

1.3.1 ENVIRONMENTAL

For centuries, TB has been linked anecdotally with environmental risk factors that go hand-in-hand with poverty, such as crowded housing and inadequate ventilation (Beggs et al. 2003; Schmidt 2008), associations that have been confirmed even in more developed settings (Baker et al. 2008; Wanyeki et al. 2006). Both factors increase the likelihood of transmission by either increasing exposure to the bacteria or by insufficient dilution and removal of infectious droplet nuclei from the air. Factors reflective of the level of healthcare accessibility, such as delay in diagnosis or ineffective treatment (either due to non-compliance or sub-optimal drug regimen), are also responsible for prolonging the period of infectiousness of the host, thus increasing potential exposure to others.

1.3.2 HOST

Comorbid conditions that dampen the host immune system, such as HIV co-infection, anti-TNF treatment, diabetes mellitus and malnutrition, have all been identified as risk factors for transmission (Kwan & Ernst 2011; Ali 2013; Qu et al. 2012; Cegielski & McMurray 2004). Demographic factors and risky social behaviors engaged by hosts, such as smoking or alcohol/drug addiction, have also been shown to contribute towards increasing the likelihood of transmission (Godoy et al. 2013; Nava-Aguilera

et al. 2009; Boum et al. 2014). Even in the absence of any of the above, various lines of evidence indicate that genetic factors partly determine differences in host susceptibility to mycobacterial infection (Schurr 2011). A family-based study in a hyperendemic area for TB with a sub-population that shows persistent lack of TST reactivity (thus appearing to be naturally resistant to infection by Mtb) has identified major loci in different chromosomal regions purported to influence T cell-dependent responses to tuberculin (Cobat et al. 2009). Gene expression analyses performed in TB patients versus uninfected healthy controls have also defined biomarkers predictive of susceptibility (Bellamy et al. 1998; Milano et al. 2016; Maertzdorf et al. 2011). Furthermore, gene expression levels in ex vivo Mtb-stimulated macrophages have revealed two cytokine genes (IL17 and IL6) associated to pulmonary manifestation of disease, rather than the meningeal or latent forms (Thuong et al. 2008).

1.3.3 BACTERIAL

Some strains may be inherently more transmissible than others, perhaps because they are particularly likely to give rise to sputum smear–positive disease, they are associated with a more insidious onset of clinical symptoms (so patients are infectious for longer), or the strains are more virulent and are therefore more likely to give rise to secondary cases within the period studied. A strain's overall propensity to transmit can be broken down into further components, such as infectivity (the ability of the bacteria to survive its aerosolized stage and reach the alveoli of the host) and breakdown to disease (also referred to as pathogenicity or virulence).

There have been studies suggesting that certain differences in the apparent virulence of specific Mtb strains can be explained by the genetic variability of the organism, such as by Rhee et al., which looked for associations between three *katG*-463 and *gyrA*-95 genotypes and the epidemiologically and clinically measured properties of infectivity and pathogenicity in a population-based sample of TB patients (Rhee et al. 1999). There have only been a couple of other studies aimed at identifying genomic markers for overall increased transmissibility. In 2007, Talarico et al. showed that PE_PGRS33 alleles that would result in a significant change to the PE_PGRS33 protein due to large insertions/ deletions or frameshift mutations were significantly associated with clustering based on genotype

and absence of cavitations in the lungs, compared to isolates having PE_PGRS33 alleles that would result in no or minimal change to the PE_PGRS33 protein. A later study by the same author comparing the genomic content of one strain from a large cluster to that of a non-clustered strain from the same community identified 25 genes that differed between the two strains, potentially contributing to the observed differences in transmission (Talarico et al. 2007; Talarico et al. 2011).

Phylogenetic lineages, a form of bacterial variation reflecting adaption to populations from different parts of the world, have also been shown to have epidemiological and clinical implications. Based on the genotype, Mtb has seven lineages: three 'ancient' (lineage-1 and two *Mycobacterium africanum* lineages), and three 'modern' (lineages-2, 3, 4) (Comas et al. 2009) and one intermediate (lineage-7), recently described in Ethiopia (Firdessa et al. 2013). Several population-based studies have used genotypic clustering as a proxy for transmissibility between different phylogenetic lineages (Buu et al. 2012; J. Anderson et al. 2013; Toungoussova et al. 2003; Hanekom et al. 2007). As an example selected from many others, particular sublineages of lineage 4 (Euro-American) have been shown to have an increased ability to cause secondary cases as determined by genotypic clustering (J Anderson et al. 2013). Other lineages have been associated to a particular site of disease, such as the East African Indian and Indo-Oceanic lineages, both associated to extrapulmonary manifestation (Click et al. 2012). Eighty percent of strains from modern Beijing sublineages, but not from ancient sublineages, have been found to synthesize relatively high quantities of phenolic glycolipid (PGL), which suppresses proinflammatory cytokines. While this is suggestive that modern sublineages are more pathogenic, molecular epidemiologic studies on the transmissibility of the Beijing lineage have been contradictory so far.

1.4 IMPLICATIONS ON TUBERCULOSIS CONTROL EFFORTS

Continued investigations of the association between bacterial genotypes (be it phylogenetic lineages, or presence/absence of genetic variations) and epidemiologically defined phenotypes (i.e. transmissibility, infectivity, pathogenicity) may:

1. Contribute to our understanding of the interplay of bacterial factors in creating more successful strains at transmitting.

2. Provide tools for disease surveillance and curbing of transmission in a population by flagging patients carrying particularly transmissible strains:
    - Intensifying contact tracing efforts around such patients.
    - Offering specific preventive measures to such patients i.e. patients with cavitary pulmonary TB receiving anti-TB medications supplemented with nebulized interferon-gamma have been found to have fewer bacilli in the lungs and less inflammation, thereby reducing the transmissibility of Mtb in the early phase of treatment (Dawson et al. 2009).

1.5 METHODS USED TO MEASURE AND MAP TRANSMISSION

1.5.1 MOLECULAR TYPING METHODS

Molecular epidemiology of TB emerged in the early 1990s thanks to the development of DNA fingerprinting techniques, such as restriction fragment length polymorphism (RFLP) of the insertion sequence (IS) 6110 and variable number of tandem repeats (VNTR). In the former, the number of IS6110 copies, a repetitive, mobile insertion sequence element of 1.35 kb (McAdam et al. 1990), varies from zero to about 25 per strain. This variation in insertion sites means IS6110 typing yields thousand of different banding patterns. VNTR typing utilizes locus-specific primers to amplify an unknown number of tandem repeats at 24 different sites, resulting in varying amplicon lengths for each. Since specificity seems to be comparable to IS6110 RFLP (de Beer et al. 2013), and it is faster and cheaper to perform (Allix-Béguec et al. 2008), VNTR typing has become the new standard in public

health applications of Mtb in the USA, Europe, and other parts of the world. Spoligotyping, a rapid, polymerase chain reaction (PCR)-based method that detects the presence of 43 unique DNA sequences interspaced between 36-base pair repeats in the DR (direct repeat) locus of Mtb complex isolates, is also widely used and highly reproducible (Driscoll 2009).

Based on comparisons of these markers of Mtb isolates from different patients, the resultant fingerprints can be classified as being either 'clustered' or 'unique'. Individuals with identical or similar fingerprint patterns (i.e. up to 1 site with a different amplicon length in VNTR) are considered to be 'clustered'. Clustered TB patients are believed to be involved in recent transmission chains, while patients with unique isolates are more likely to have reactivated TB acquired in the past.

Genotypic tools can therefore provide novel insight into Mtb transmission by identifying clusters of active cases and getting a sense via clustering rates (% of isolates that are clustered out of the total number of isolates) of how much transmission is occurring in a particular setting and over time (Adams et al. 2012; Tuite et al. 2013; Ferdinand et al. 2013). In the Netherlands, genotypic clustering has been shown to have steadily decreased over the years, from above 40% in 1993 to around 30% in 2008 (unpublished data). Clustering rates can also highlight increased transmission in specific higher –risk or immigrant sub-populations (Rossi et al. 2012; Iñigo et al. 2007).

Several population-based studies have used genotypic clustering as a proxy for transmissibility between different phylogenetic lineages (Buu et al. 2012; J. Anderson et al. 2013; Toungoussova et al. 2003; Hanekom et al. 2007). Findings from these studies have been varied and non-conclusive so far, most likely reflecting differences in other factors affecting transmission between different settings.

1.5.2 CONTACT TRACING

Genotyping of Mtb is also used to support contact tracing and source case finding, a form of active case detection (via interviews) entailing the systematic evaluation of the contacts of known TB patients to identify active disease or latent TB infection. Traditional contact tracing focuses on smear-positive

(defined by the presence of at least one acid fast bacilli (AFB+) in at least one sputum sample) adult index TB cases, as these are the most infectious. The primary goal is the early diagnosis and treatment of contacts with disease, both interrupting ongoing transmission and reducing morbidity and mortality in affected individuals. This strategy may be worthwhile in contacts (i.e. household) of patients with TB because they are at higher risk of TB (having had prolonged exposure and shared environmental risk factors with the index case) than members of the general population (Greenaway et al. 2003; Fox et al. 2013). Contact investigation has been widely implemented in high-income countries for decades (Erkens et al. 2010). Recently there has been a growing interest in it also being performed in resource-limited settings, as national programs seek new methods for improving case detection (Fox et al. 2012; Shapiro et al. 2012). Used as a parallel tool to fingerprinting data by confirming or disputing the epidemiologic link between two patients (Lambregts-Van Weezenbeek et al. 2003), it helps by pointing towards potential secondary cases whose recent infection may be subsequently confirmed via the application of immunological tests, such as TST and IGRAs.

1.5.3 IMMUNOLOGICAL MARKERS OF TRANSMISSION

TST and IGRAs are 'indirect tests' designed to detect latent TB infection, that is, they do not detect the actual bacilli but instead an immune response that suggest past or present exposure to Mtb bacilli. TST conversion – where a result changes from "negative" (typically 0-4mm diameter induration) to "positive" (typically equal to or >10mm diameter induration) within a 24 month period - may therefore be used as a proxy measure of infection, by calculating the proportion of contacts that started with a negative result and converted to positive during a determined follow-up period. This method has been used to find out whether transmission has taken place in certain settings i.e. health-care associated workplaces (Reynolds et al. 2006; Lee et al. 2005) as well at to identify populations that are at highest risk for being infected (Sherman et al. 2011). Both tests are however better suited to low prevalence settings where BCG vaccination, which may give a false positive result on the TST, is not routinely used.

1.5.4 WHOLE GENOME SEQUENCING

More recently, the exponentially falling cost of whole genome sequencing (WGS) has turned it into an increasingly accessible tool in epidemiologic studies. Since it monitors all variation in a bacterial genome it therefore has the highest level of discriminatory power that is possible at the DNA level. For example, in an investigation of a crack cocaine related outbreak in Vancouver, Canada, Gardy et al. demonstrated the value of performing WGS on all the Mtb isolates in the outbreak. While MIRU-VNTR grouped the isolates into a single genotype, WGS showed that there were really two concomitant outbreaks with two distinct strains that had evolved from a common ancestor and kept the same MIRU-VNTR genotype (Gardy et al. 2011). More recently, a threshold number (i.e. below five) of single nucleotide polymorphisms (SNPs) between strains has been shown to translate into an epidemiological link, whilst stains differing by more than 12 SNPs reflect unlinked cases (Walker et al. 2013). In a subsequent study of transmission epidemiology, Walker et al used WGS to analyze all available isolates from Mtb cases in Oxfordshire, UK from 2007 – 2012. Although they used a cut-off of 12 or fewer SNP differences to define clustered strains, the differences within the clusters ranged from zero to just 7 SNPs, with a median of 1 SNP difference (Walker et al. 2014). Tracing outbreaks by genome analysis of SNPs is therefore more discriminative than any of the more traditional methods (IS6110 RFLP, spoligotyping and MIRU-VNTR), but the process is still poorly standardized, and because each lab can use one of several available bioinformatics program, it is hard to compare results from studies performed in different labs.

1.6 THE NETHERLANDS: A SETTING CONDUCIVE TO STUDYING TRANSMISSION

Since 1993, patient epidemiological data and DNA fingerprints of virtually all Mtb complex isolates have been stored in a database at the National Institute for Public Health and the Environment (RIVM; Bilthoven) and the National Tuberculosis Register (NTR) of the Netherlands. Both IS6110 restriction fragment length polymorphism (RFLP) and 24-locus variable-number tandem-repeat (VNTR) typing are routinely performed, the discriminatory power and agreement of findings between both methods of which were evaluated in 2013 (de Beer et al. 2013). Systematic contact investigations around source cases are also routinely performed by the TB Public Health Services, and the TST test used to investigate presumably exposed contacts.

Given this wealth of data, multiple studies on the epidemiology of TB in the Netherlands have been generated. The observed declining incidence over the past decade, for example, has been attributed to older birth cohorts with high infection prevalence being replaced by those with lower infection prevalence (Borgdorff et al. 2005). Prevalence among immigrants has been associated with immigration figures (numbers of incoming people), (Borgdorff et al. 2010) and host risk factors, such as young age (<35 years) and geographic origin of first patients in a DNA fingerprint cluster, have been shown to be predictors of outbreaks (Kik et al. 2008).

WGS has also featured in more recent publications, such as in an outbreak investigation of over 100 cases beginning in the Dutch city of Harlingen in 1992 and extending through 2008. Most of the strains had identical IS6110 RFLP patterns, making it impossible to identify the sources of infection or routes of transmission. WGS of three of these isolates identified eight polymorphic SNPs specific for the Harlingen strains. By tracing the evolution of the nucleotide changes in these eight positions, the isolates of the Harlingen cluster could be divided into five SNP clusters, with the earliest isolate in each cluster defined as the index case, defined as the earliest isolate in each SNP cluster, from which the subsequent chains of transmission events could be delineated (Schürch et al. 2010).

The low TB prevalence setting of the Netherlands presents itself as an advantage in the study of the role of genotype on transmission, since it is less likely to be confounded by a high background infection pressure, where a TST result is more likely to fail at distinguishing recent from past infection. Furthermore, in the Netherlands there is no routine BCG vaccination program that could affect the interpretation of TST results, making TST a suitable tool for the detection of Mtb infection in contact investigations, particularly amongst the native population. The importance of a bacteriological component in TB transmission was first addressed in a study from 2011, where it was found that large clusters were independently associated with an increased number of TST positive contacts, suggesting that the spread of Mtb also depends on bacteriological factors (Verhagen et al. 2011).

1.7 RESEARCH QUESTIONS

The discovery of genotypic markers associated with increased transmissibility in Mtb would represent an important step in advancing mycobacterial virulence studies. In contrast to our understanding of host and environmental influences on infection and disease, there is a lack of systematic information on the influence of the genetic diversity of the pathogen itself that may provide important clues for basic and applied research on TB and its relation with the human host. This thesis focuses on addressing some of the gaps in our knowledge of bacterial factors behind transmissibility, such as:

- How can we control for host-related factors in measuring the "transmissibility" of Mtb, and are there any differences between phylogenetic lineages once these factors are adjusted for?
- Are there any differences in the infectivity and pathogenicity of different phylogenetic lineages of Mtb?
- Do variations in Mycobacterium cyclase/LuxR-like (mclx) genes play a role in virulence-related fitness and host adaptation ability to the host?
- Can we identify SNPs and genes associated to transmissibility? And if so, do we find biological correlates of their interaction with the human immune system?

1.8 THESIS OVERVIEW

Chapter 2 describes a new method to improve proxy measures of "transmissibility" by adjusting for patient-related factors, thus strengthening the causal association found with bacterial factors.

Chapter 3 looks more closely at the "transmissibility" phenotype, breaking it down into the bacteria's ability to spread or progress to disease, and explores any differences in these by phylogenetic lineage.

The study in **chapter 4** aims at correlating genetic variation (in mclx genes) with phylogenetic and epidemiological characteristics, including transmissibility, of Mtb strains.

Chapter 5 describes the application of evolutionary convergence analysis to identify SNPs/genes associated with "transmissibility" in Mtb.

In **chapter 6**, the general discussion chapter, the implications of these studies are discussed.

1.9 REFERENCES

Adams, L. V. et al., 2012. Molecular epidemiology of HIV-associated tuberculosis in Dar es Salaam, Tanzania: Strain predominance, clustering, and polyclonal disease. *Journal of Clinical Microbiology*, 50(8), p.2645–2650.

Ali, T., 2013. Clinical use of anti-TNF therapy and increased risk of infections. *Drug, Healthcare and Patient Safety*, 5, p.79.

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), p.2692–2699.

Anderson, J. et al., 2013. Sublineages of lineage 4 (Euro-American) *Mycobacterium tuberculosis* differ in genotypic clustering. *International Journal of Tuberculosis and Lung Disease*, 17(7), p.885–891.

Baker, M. et al., 2008. Tuberculosis associated with household crowding in a developed country. *Journal of Epidemiology and Community Health*, 62(8), p.715–721.

de Beer, J.L. et al., 2013. Comparative study of IS6110 restriction fragment length polymorphism and variable-number tandem-repeat typing of *Mycobacterium tuberculosis* isolates in the Netherlands, based on a 5-year nationwide survey. *Journal of Clinical Microbiology*, 51(4), p.1193–1198.

Beggs, C.B. et al., 2003. The transmission of tuberculosis in confined spaces: An analytical review of alternative epidemiological models. *International Journal of Tuberculosis and Lung Disease*, 7(11), p.1015–1026.

Behr, M.A., 2002. BCG–different strains, different vaccines? *The Lancet Infectious Diseases*, 2(2), p.86–92.

Bellamy, R. et al., 1998. Variations in the NRAMP1 gene and susceptibility to tuberculosis in West Africans. *The New England Journal of Medicine*, 338(10), p.640–644.

Borgdorff, M.W. et al., 2010. Progress towards tuberculosis elimination: secular trend, immigration and transmission. *The European Respiratory Journal : official journal of the European Society for Clinical Respiratory Physiology*, 36(2), p.339–347.

Borgdorff, M.W. et al., 2005. Tuberculosis elimination in the Netherlands. *Emerging Infectious Diseases*, 11(4), p.597–602.

Boum, Y. et al., 2014. Male Gender is independently associated with pulmonary tuberculosis among sputum and non-sputum producers people with presumptive tuberculosis in Southwestern Uganda. *BMC Infectious Diseases*, 14(1), p.638.

Buu, T.N. et al., 2012. Increased transmission of *Mycobacterium tuberculosis* Beijing genotype strains associated with resistance to streptomycin: a population-based study. *PloS ONE*, 7(8), p.e42323.

Cegielski, J.P. & McMurray, D.N., 2004. The relationship between malnutrition and tuberculosis: Evidence from studies in humans and experimental animals. *International Journal of Tuberculosis and Lung Disease*, 8(3), p.286–298.

Churchyard, G.J. et al., 2014. A trial of mass isoniazid preventive therapy for tuberculosis control. *The New England Journal of Medicine*, 370, p.301–310.

Click, E.S. et al., 2012. Relationship between *Mycobacterium tuberculosis* phylogenetic lineage and clinical site of tuberculosis. *Clinical Infectious Diseases*, 54(2), p.211–219.

Cobat, A. et al., 2009. Two loci control tuberculin skin test reactivity in an area hyperendemic for tuberculosis. *The Journal of Experimental Medicine*, 206(12), p.2583–2591.

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).

Dawson, R. et al., 2009. Immunomodulation with recombinant interferon-gamma1b in pulmonary tuberculosis. *PloS ONE*, 4(9), p.e6984.

Driscoll, J.R., 2009. Spoligotyping for molecular epidemiology of the *Mycobacterium tuberculosis* complex. *Methods in Molecular Biology (Clifton, N.J.)*, 551, p.117–128.

Erkens, C.G.M. et al., 2010. Tuberculosis contact investigation in low prevalence countries: a European consensus. *The European Respiratory Journal*, 36(4), p.925–949.

Ferdinand, S. et al., 2013. Use of genotyping based clustering to quantify recent tuberculosis transmission in Guadeloupe during a seven years period: analysis of risk factors and access to health care. *BMC Infectious Diseases*, 13(1), p.364.

Firdessa, R. et al., 2013. Lineages causing pulmonary and extrapulmonary tuberculosis, Ethiopia. *Emerging Infectious Diseases*, 19(3), p.460–463.

Fox, G.J. et al., 2013. Contact investigation for tuberculosis: A systematic review and meta-analysis. *European Respiratory Journal*, 41(1), p.140–156.

Fox, G.J. et al., 2012. Contact Investigation in Households of Patients with Tuberculosis in Hanoi, Vietnam: A Prospective Cohort Study. *PLoS ONE*, 7(11).

Franco-Paredes, C. et al., 2006. Vaccination strategies to prevent tuberculosis in the new millennium: from BCG to new vaccine candidates. *International journal of infectious diseases : IJID : official publication of the International Society for Infectious Diseases*, 10(2), p.93–102.

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), p.730–739.

Godoy, P. et al., 2013. Smoking in tuberculosis patients increases the risk of infection in their contacts. *International Journal of Tuberculosis and Lung Disease*, 17(6), p.771–776.

Greenaway, C. et al., 2003. Yield of casual contact investigation by the hour. *International Journal of Tuberculosis and Lung Disease*, 7(12 SUPPL. 3).

Hanekom, M. et al., 2007. A recently evolved sublineage of the *Mycobacterium tuberculosis* Beijing strain family is associated with an increased ability to spread and cause disease. *Journal of Clinical Microbiology*, 45(5), p.1483–1490.

Harisinghani, M.G. et al., 2000. Tuberculosis from head to toe. *Radiographics : a review publication of the Radiological Society of North America, Inc*, 20(2), p.449–470.

Iñigo, J. et al., 2007. Analysis of changes in recent tuberculosis transmission patterns after a sharp increase in immigration. *Journal of Clinical Microbiology*, 45(1), p.63–69.

Jones-Lopez, E.C. et al., 2015. Cough aerosol cultures of *Mycobacterium tuberculosis*: Insights on TST / IGRA discordance and transmission dynamics. *PLoS ONE*, 10(9), p.1–18.

Kaplan, G. et al., 2003. *Mycobacterium tuberculosis* Growth at the Cavity Surface: A Microenvironment with Failed Immunity. *Infection and Immunity*, 71(12), p.7099–7108.

Kik, S. V et al., 2008. Tuberculosis outbreaks predicted by characteristics of first patients in a DNA fingerprint cluster. *American Journal of Respiratory and Critical Care Medicine*, 178(1), p.96–104.

Kwan, C. & Ernst, J.D., 2011. HIV and tuberculosis: A deadly human syndemic. *Clinical Microbiology Reviews*, 24(2), p.351–376.

Lambregts-Van Weezenbeek, C.S.B. et al., 2003. Tuberculosis contact investigation and DNA fingerprint surveillance in The Netherlands: 6 Years' experience with nation-wide cluster feedback and cluster monitoring. *International Journal of Tuberculosis and Lung Disease*, 7(12 SUPPL. 3).

Lee, E.H. et al., 2005. Nosocomial transmission of *Mycobacterium tuberculosis* in a children's hospital. *International Journal of Tuberculosis and Lung Disease*, 9(6), p.689–692.

Maertzdorf, J. et al., 2011. Human gene expression profiles of susceptibility and resistance in tuberculosis. *Genes and Immunity*, 12(1), p.15–22.

McAdam, R.A. et al., 1990. Characterization of a *Mycobacterium tuberculosis* insertion sequence belonging to the IS3 family. *Molecular Microbiology*, 4(9), p.1607–1613.

Milano, M. et al., 2016. Single Nucleotide Polymorphisms in IL17A and IL6 Are Associated with Decreased Risk for Pulmonary Tuberculosis in Southern Brazilian Population. *Plos ONE*, 11(2), p.e0147814.

Nava-Aguilera, E. et al., 2009. Risk factors associated with recent transmission of tuberculosis: Systematic review and meta-analysis. *International Journal of Tuberculosis and Lung Disease*, 13(1), p.17–26.

O'Garra, A. et al., 2013. The immune response in tuberculosis. *Annual Review Immunology, 31, p.475-527*.

Orme, I.M., Robinson, et al., 2014. The balance between protective and pathogenic immune responses in the TB-infected lung. *Nature Immunology*, 16(1), p.57–63.

Qu, H.Q. et al., 2012. Host susceptibility to tuberculosis: Insights from a longitudinal study of gene expression in diabetes. *International Journal of Tuberculosis and Lung Disease*, 16(3), p.370–372.

Ramakrishnan, L., 2012. Revisiting the role of the granuloma in tuberculosis. *Nature Reviews Immunology*, 12(5), p.352–366.

Reynolds, D.L. et al., 2006. Transmission of *Mycobacterium tuberculosis* from an infant. *International Journal of Tuberculosis and Lung Disease*, 10(9), p.1051–1056.

Rhee, J.T. et al., 1999. Molecular epidemiologic evaluation of transmissibility and virulence of *Mycobacterium tuberculosis*. *Journal of Clinical Microbiology,* 37(6), p.1764-70.

Ribeiro-Rodrigues, R. et al., 2014. Discordance of tuberculin skin test and interferon gamma release assay in recently exposed household contacts of pulmonary TB cases in Brazil. *PLoS ONE*, 9(5).

Rodrigo, T. et al., 1997. Characteristics of tuberculosis patients who generate secondary cases. *International Journal of Tuberculosis and Lung Disease*, 1(4), p.352–357.

Rossi, C. et al., 2012. *Mycobacterium tuberculosis* transmission over an 11-year period in a low-incidence, urban setting. setting. *International Journal of Tuberculosis and Lung Disease*, 16(3), p.312–318.

Roy, A. et al., 2014. Effect of BCG vaccination against *Mycobacterium tuberculosis* infection in children: systematic review and meta-analysis. *BMJ (Clinical research ed.)*, 349, p.g4643.

Schmidt, C.W., 2008. Linking TB and the environment: An overlooked mitigation strategy. *Environmental Health Perspectives*, 116(11).

Schürch, A.C. et al., 2010. High-resolution typing by integration of genome sequencing data in a large tuberculosis cluster. *Journal of Clinical Microbiology*, 48(9), p.3403–3406.

Schurr, E., 2011. The contribution of host genetics to tuberculosis pathogenesis. *Kekkaku*, 86(1), p.17–28.

Shapiro, A.E. et al., 2012. Community-based targeted case finding for tuberculosis and HIV in household contacts of patients with tuberculosis in South Africa. *American Journal of Respiratory and Critical Care Medicine*, 185(10), p.1110–1116.

Sherman, H.A. et al., 2011. Housekeeping health care workers have the highest risk for tuberculin skin test conversion. *International Journal of Tuberculosis and Lung Disease*, 15(8), p.1050–1055.

Siddiqi, K., Lambert, M.L. & Walley, J., 2003. Clinical diagnosis of smear-negative pulmonary tuberculosis in low-income countries: The current evidence. *Lancet Infectious Diseases*, 3(5), p.288–296.

Silva Miranda, M. et al., 2012. The tuberculous granuloma: An unsuccessful host defence mechanism providing a safety shelter for the bacteria? *Clinical and Developmental Immunology*, 2012:139127.

Stewart, G.R., et al., 2003. Tuberculosis: a problem with persistence. *Nature Reviews Microbiology*, 1(2), p.97–105.

Talarico, S. et al., 2007. Association of *Mycobacterium tuberculosis* PE_PGRS33 polymorphism with clinical and epidemiological characteristics. *Tuberculosis*, 87(4), p.338–346.

Talarico, S. et al., 2011. Identification of factors for tuberculosis transmission via an integrated multidisciplinary approach. *Tuberculosis (Edinburgh, Scotland)*, 91(3), p.244–249.

Thuong, N.T.T. et al., 2008. Identification of tuberculosis susceptibility genes with human macrophage gene expression profiles. *PLoS Pathogens*, 4(12).

Toungoussova, O.S. et al., 2003. Molecular epidemiology and drug resistance of *Mycobacterium tuberculosis* isolates in the Archangel prison in Russia: predominance of the W-Beijing clone family. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, 37(5), p.665–672.

Tuite, A.R. et al., 2013. Epidemiological evaluation of spatiotemporal and genotypic clustering of *mycobacterium tuberculosis* in Ontario, Canada. *International Journal of Tuberculosis and Lung Disease*, 17(10), p.1322–1327.

Verhagen, L.M. et al., 2011. Mycobacterial factors relevant for transmission of tuberculosis. *The Journal of Infectious Diseases*, 203(9), p.1249–1255.

Walker, T.M. et al., 2014. Assessment of *Mycobacterium tuberculosis* transmission in Oxfordshire, UK, 2007–12, with whole pathogen genome sequences: an observational study. *The Lancet Respiratory Medicine*, p.285–292.

Walker, T.M. et al., 2013. Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study. *The Lancet Infectious Diseases*, 13(2), p.137–146.

Wanyeki, I. et al., 2006. Dwellings, crowding, and tuberculosis in Montreal. *Social Science and Medicine*, 63(2), p.501–511.

World Health Organization, 1982. Efficacy of various durations of isoniazid preventive therapy for tuberculosis: five years of follow-up in the IUAT trial. *Bulletin of the World Health Organization*, 60, p.555–564.

World Health Organization, 2015. Global tuberculosis report 2015.

Zwerling, A. et al., 2011. The BCG world atlas: A database of global BCG vaccination policies and practices. *PLoS Medicine*, 8(3).

Chapter II

A novel approach - the Propensity to Propagate (PTP) method for controlling for host factors in studying the transmission of *Mycobacterium tuberculosis*

OPEN ACCESS Freely available online

PLOS | ONE

A Novel Approach - The Propensity to Propagate (PTP) Method for Controlling for Host Factors in Studying the Transmission of Mycobacterium Tuberculosis

Hanna Nebenzahl-Guimaraes[1,2,3]*, Martien W. Borgdorff[4,5], Megan B. Murray[6], Dick van Soolingen[1,7]

1 National Institute for Public Health and the Environment (RIVM), Bilthoven, The Netherlands, **2** Life and Health Sciences Research Institute (ICVS), School of Health Sciences, University of Minho, Braga, Portugal, **3** ICVS/3B's, PT Government Associate Laboratory, Braga/Guimarães, Portugal, **4** Public Health Service, Amsterdam, Netherlands, **5** Department of Clinical Epidemiology, Academic Medical Centre, University of Amsterdam, Amsterdam, Netherlands, **6** Department of Global Health and Social Medicine, Harvard Medical School, Boston, United States of America, **7** Department of Medical Microbiology, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands

Abstract

Rationale: Understanding the genetic variations among Mycobacterium tuberculosis (MTB) strains with differential ability to transmit would be a major step forward in preventing transmission.

Objectives: To describe a method to extend conventional proxy measures of transmissibility by adjusting for patient-related factors, thus strengthening the causal association found with bacterial factors.

Methods: Clinical, demographic and molecular fingerprinting data were obtained during routine surveillance of verified MTB cases reported in the Netherlands between 1993 and 2011, and the phylogenetic lineages of the isolates were inferred. Odds ratios for host risk factors for clustering were used to obtain a measure of each patient's and cluster's propensity to propagate (CPP). Mean and median cluster sizes across different categories of CPP were compared amongst four different phylogenetic lineages.

Results: Both mean and median cluster size grew with increasing CPP category. On average, CPP values from Euro-American lineage strains were higher than Beijing and EAI strains. There were no significant differences between the mean and median cluster sizes among the four phylogenetic lineages within each CPP category.

Conclusions: Our finding that the distribution of CPP scores was unequal across four different phylogenetic lineages supports the notion that host-related factors should be controlled for to attain comparability in measuring the different phylogenetic lineages' ability to propagate. Although Euro-American strains were more likely to be in clusters in an unadjusted analysis, no significant differences among the four lineages persisted after we controlled for host factors.

Citation: Nebenzahl-Guimaraes H, Borgdorff MW, Murray MB, van Soolingen D (2014) A Novel Approach - The Propensity to Propagate (PTP) Method for Controlling for Host Factors in Studying the Transmission of Mycobacterium Tuberculosis. PLoS ONE 9(5): e97816. doi:10.1371/journal.pone.0097816

Editor: Olivier Neyrolles, Institut de Pharmacologie et de Biologie Structurale, France

Received February 24, 2014; **Accepted** April 24, 2014; **Published** May 21, 2014

Funding: Portuguese Foundation for Science and Technology (FCT) (SFRH/BD/33902/2009 to HN-G). 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: hanna.guimaraes@gmail.com

Introduction

Transmission of *Mycobacterium tuberculosis* (MTB) occurs through aerosol droplets. Subsequent cases in transmission chains result in "clusters" of patients who share Mtb strains of the same genotype or molecular fingerprint [1]. Cluster sizes vary widely, which may reflect the fact that strains do not spread equally or that they differ in their rate of progression to active TB disease. The identification of strains that cause large tuberculosis (TB) outbreaks, such as CDC1551 or Harlingen [2,3], has lead to studies on the virulence of such strains. Indeed molecular epidemiologic studies have suggested that some strains are more successfully transmitted than others [4–6]. The mechanisms however governing this variability remain largely unknown, with much research focused on the contribution of host risk factors. In the Netherlands for example, age, sex, homelessness, alcohol or drug abuse, living in an urban area and smear positivity have all been associated to increased transmissibility [7]. There is substantial evidence however to suggest that bacterial factors also contribute to variability in cluster size and the extent of transmission of TB. For example, Verhagen and colleagues showed that newly diagnosed index cases in a larger cluster infected more people than did newly diagnosed cases in smaller clusters [8]. This implies that clusters not only grow over time because of well-known patient risk factors for TB transmission, but also because the strain itself generates an increased number of tuberculin skin test-positive contacts, and spreads more effectively than other strains.

Phylogenetic lineages reflect evolutionary divergence associated with different geographical regions [9]. Beijing lineage strains, for example, are predominantly found in Asia, yet are widely

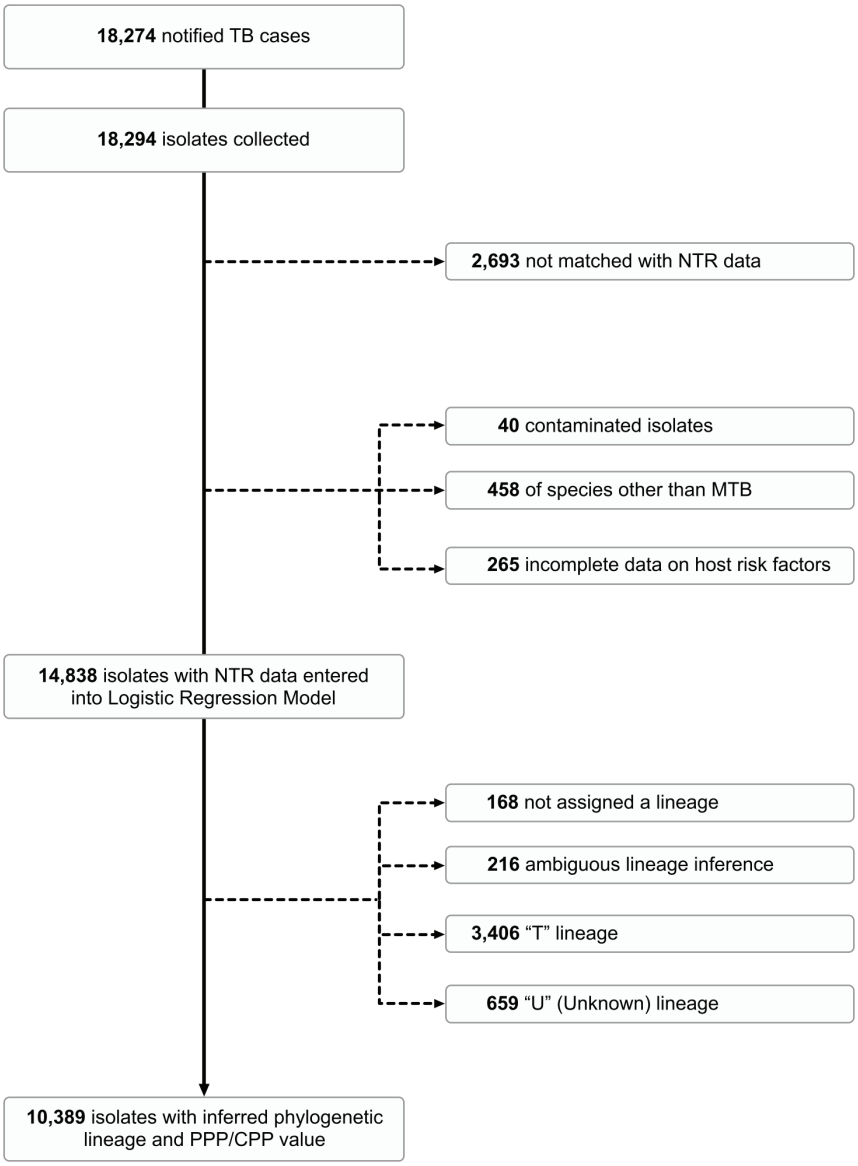


Figure 1. Flow-diagram of exclusion criteria applied to dataset.
doi:10.1371/journal.pone.0097816.g001

disseminated and present in more countries than any other lineage strain. This suggests that this evolutionary lineage may have evolved unique properties leading to its successful clonal expansion [10,11]. To date, studies examining the association between phylogenetic lineages of MTB and transmissibility have typically used DNA fingerprinting clustering rates as measures of transmissibility, with very few adjusting for host-related factors.

Since preventing transmission of MTB is key to a sustained decline in TB incidence, understanding the genetic variations between strains with differential ability to transmit would be a major step forward. In order to distinguish bacterial factors associated with transmission from those that pertain to the host however, the influence of host-related factors needs to be addressed. In the Netherlands, a nationwide surveillance of TB including structural DNA fingerprinting of all *M. tuberculosis* isolates has been in place since 1993. Patient information is available for all registered TB cases, of which there are approximately one thousand per year. Here, we describe a method to complement and extend the conventional use of cluster size and proportion of cases in a cluster as proxy measures of transmissibility by adjusting for patient related factors, thus strengthening the causal association found with bacterial factors. Since cluster size may reflect both the propensity of a strain to be transmitted and to cause disease given an infection, we have chosen to use the term "propagation" instead of transmissibility as a more accurate description of cluster growth.

Methods

Data Collection and DNA Fingerprinting

The National Institute for Public Health and the Environment (RIVM) in Bilthoven, The Netherlands, serves as a reference laboratory for the secondary laboratory diagnosis of all TB cases in The Netherlands, offering identification, drug susceptibility testing, and molecular typing. DNA fingerprints of all nationwide MTB complex isolates and their cluster status have been stored in

Table 1. Host risk factors for clustering of MTB in the Netherlands, 1993–2011.

		No. (%) in clustering state:			
Category and case group		**Clustered**	**Non-clustered**	**OR (95% CI)**	**Adjusted OR (95% CI)**
Sex	Males	5385 (60.8)	3474 (39.2)	1 (Ref)	1 (Ref)
	Females	3200 (53.5)	2779 (46.5)	**0.74 (0.70–0.79)**	**0.87 (0.81–0.93)**
Age at diagnosis (years)	0–15	366 (69.2)	163 (30.8)	**1.28 (1.11–1.56)**	1.05 (0.86–1.29)
	16–30	3383 (63.7)	1929 (36.3)	1 (Ref)	1 (Ref)
	31–45	2485 (60.9)	1593 (39.1)	**0.89 (0.82–0.97)**	**0.86 (0.78–0.94)**
	46–60	1254 (59.5)	852 (40.5)	**0.84 (0.76–0.93)**	**0.77 (0.69–0.86)**
	61–75	715 (45.1)	871 (54.9)	**0.47 (0.42–0.52)**	**0.40 (0.35–0.45)**
	76–90	370 (31.6)	800 (68.4)	**0.26 (0.23–0.30)**	**0.19 (0.17–0.23)**
	>90	12 (21.1)	45 (78.9)	**0.15 (0.08–0.29)**	**0.12 (0.06–0.22)**
Disease Classification	Pulmonary	5107 (61.6)	3187 (38.4)	1 (Ref)	1 (Ref)
	Extrapulmonary	2465 (51.2)	2347 (48.8)	**0.66 (0.61–0.70)**	**0.76 (0.69–0.83)**
	Pulmonary-extrapulmonary	1013 (58.5)	719 (41.5)	**0.88 (0.79–0.98)**	**0.90 (0.80–1.01)**
Smear-positivity	No	5068 (54.7)	4205 (45.3)	1 (Ref)	1 (Ref)
	Yes	3517 (63.2)	2048 (36.8)	**1.43 (1.33–1.53)**	**1.17 (1.07–1.27)**
Alcohol consumption	No	8426 (57.6)	6200 (42.4)	1 (Ref)	1 (Ref)
	Yes	159 (75.0)	53 (25.0)	**2.20 (1.61–3.01)**	1.29 (0.92–1.80)
Drug-use	No	8213 (57.0)	6193 (43.0)	1 (Ref)	1 (Ref)
	Yes	372 (86.1)	60 (13.9)	**4.67(3.55–6.15)**	**2.75 (2.05–3.67)**
Homelessness	No	8362 (57.4)	6198 (42.6)	1 (Ref)	1 (Ref)
	Yes	223 (80.2)	55 (19.8)	**3.0 (2.23–4.04)**	**1.58 (1.15–2.18)**
Health-care worker	No	8463 (57.8)	6187 (42.2)	1 (Ref)	1 (Ref)
	Yes	122 (64.9)	66 (35.1)	**1.35 (1.00–1.83)**	**1.00 (0.73–1.38)**
Traveler to endemic areas	No	8423 (58.0)	6090 (42.0)	1 (Ref)	1 (Ref)
	Yes	162 (49.8)	163 (50.2)	**0.72 (0.58–0.90)**	**0.58 (0.46–0.73)**
Origin	Native Dutch	2343 (58.4)	1667 (41.6)	1 (Ref)	1 (Ref)
	Foreign-born (Asia)	1247 (39.0)	1949 (61.0)	**0.41 (0.37–0.45)**	**0.28 (0.25–0.31)**
	Foreign-born (Africa)	3253 (65.0)	1749 (35.0)	**1.18 (1.09–1.29)**	**0.76 (0.69–0.84)**
	Foreign-born (America)	704 (72.1)	273 (27.9)	**1.64 (1.41–1.91)**	1.06 (0.90–1.25)
	Foreign-born (Europe)	415 (53.1)	366 (46.9)	**0.72 (0.62–0.84)**	**0.43 (0.37–0.51)**
Total cases		8585 (57.9)	6253 (42.1)		

OR, odds ratio; CI, confidence interval; Statistically significant OR are highlighted in bold.
doi:10.1371/journal.pone.0097816.t001

a RFLP database since 1993. The Registration Committee of the Netherlands Tuberculosis Register (NTR) approved this retrospective study and provided demographic and clinical information for patients. Because these data are de-identified by name, DNA fingerprinting results from the RIVM were linked on the basis of sex, date of birth, year of diagnosis and postal code. All notified MTB culture-positive cases between 1993 and 2011 were included in the study. In case of patients with multiple isolates, only the isolate with the earliest date of diagnosis was included. Contaminated isolates were excluded from the database.

Isolates recovered from patients between 1993 and 2009 underwent IS*6110* and polymorphic GC-rich sequence (PGRS) restriction fragment length polymorphism (RFLP) typing (n = 15,073), and those from 2004 onward to variable number of

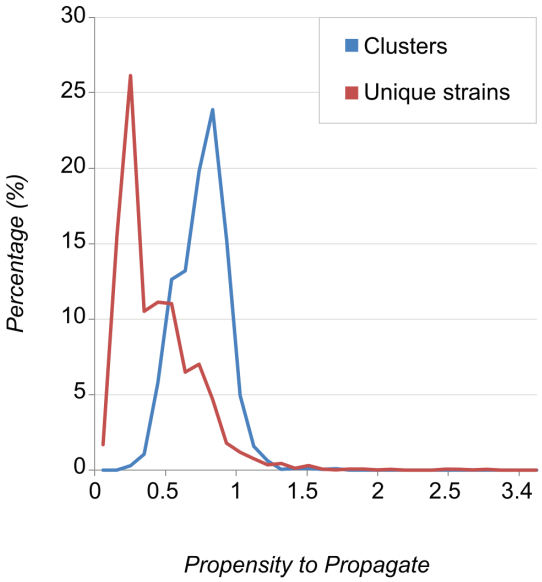


Figure 2. Distribution of Propensity to Propagate values.
doi:10.1371/journal.pone.0097816.g002

tandem repeat (VNTR) typing (n = 5,870) [12,13], In the period of 2004–2008 both RFLP and VNTR typing were performed [14]. In addition, 4,433 randomly selected isolates were spoligotyped. We defined a cluster as groups of patients who shared TB isolates with identical RFLP or VNTR patterns or, if strains had fewer than five IS*6110* copies, identical PGRS RFLP patterns [15].

Classification into phylogenetic lineages

The phylogenetic label of a spoligotyped isolate was used to infer the lineage of isolates belonging to the same RFLP or VNTR cluster as the spoligotyped isolate. Following this, the MIRU-VNTRplus online tool was used to perform MIRU Best Match Analysis (stringent cut-off of 0·17) followed by MIRU Tree-based identification to identify the phylogenetic lineages of strains with MIRU patterns [16]. Resulting matched phylogenetic lineages from clustered isolates were extrapolated to the remaining isolates of the respective clusters. Remaining strains without an inferred lineage were assigned one on the basis of RFLP similarity (≥80%) to a reference dataset of pre-identified strains with RFLP patterns in a tree generated by the neighbor-joining method with the Kimura 2 parameter on BioNumerics software for Windows (version 6·6, Applied Maths). The same procedure was repeated for strains with RFLP PGRS patterns. Finally, any remaining MIRU-typed strains without an inferred lineage were subjected to MIRU Best Match Analysis (relaxed cut-off of 0·3). This was purposely left as the last in the series of steps for the classification of lineages as it is the least optimized for minimizing fine-tuned mismatching that can occur as an exception among strains belonging to the Euro-American family [17].

Four major phylogenetic lineages were identified: Euro-American, Central Asian Strain (CAS), East-African-Indian (EAI) and Beijing (Table S1). Strains not assigned a phylogenetic lineage or assigned more than one major phylogenetic family per cluster were excluded from analysis. Strains classified as either "T" or "U" (Unknown) were also excluded due to the ambiguity of these classifications (Figure 1).

We considered the possibility that the use of spoligotyping, MIRU- or RFLP-typing for inferring phylogenetic lineages in this study may have resulted in misclassification of lineage, due to the propensity of these markers for convergent evolution and resulting homoplasies [18]. To assess this, we compared the inferred phylogenetic lineages with those determined using single nucleotide polymorphisms (SNP) markers in a subset of strains (n = 248) that were also whole-genome sequenced [19].

Statistical Analysis

We used a logistic regression model to determine independent host risk factors including demographic, behavioral, and sputum smear status, for clustering. Variables with p-values <0·20 were entered into a multivariate model. Crude and adjusted odds ratios (OR) are presented with 95% confidence intervals (CI). Estimates for the adjusted ORs were each multiplied as weights to calculate each patient's propensity to propagate (PPP). The geometric mean of PPP values belonging to a cluster was taken as the overall measure of a cluster's propensity to propagate (CPP). Confidence intervals for the median CPP by phylogenetic lineage were calculated using nonparametric bootstrapping methods based on 10,000 replicates. An analysis of variance (ANOVA) with Bonferroni correction was performed to determine CPP comparability among the four phylogenetic lineages. We repeated this step on a validation subset of strains (n = 2,136) whose lineages were determined using the highly reliable MIRU Best Match Analysis (stringent cut-off) and SNP markers [9]. We also explored the variability of CPP by phylogenetic lineage stratified by host region of origin, by repeating the ANOVA on a subset of clusters composed of patients of a particular region only (Europe versus Asia). In a sensitivity analysis, we checked the consequences of excluding extra-pulmonary cases from the dataset. Finally, the proportion of clustered isolates was calculated for each phylogenetic lineage. Mean and median cluster size (plus interquartile ranges and 95% CI, respectively) were calculated for three increasing CPP categories (<0·5, 0·5–0·8 and >0·8) for each of the four phylogenetic lineages. SAS software for Windows, version 9·3, was used for statistical analyses.

Results

During the period January 1993 to December 2011, 18,294 isolates were collected from 18,274 notified TB cases in the Netherlands and their clustering status ascertained, of which 15,601 (85%) were successfully matched with the NTR data. Of these, 14,838 (94%) were non-contaminated MTB cultures with completely ascertained information on host risk factors (Figure 1). The mean age of MTB positive TB cases was 41 years (SD, 20); 8,859 (60%) were male; and 10,005 (67%) were foreign-born.

Host-related factors for clustering

Of the 14,838 strains with both DNA fingerprinting and host-related data, 8,585 were clustered (57·9%) and 6,253 were non-clustered (42·1%). Table 1 shows that patients were more likely to be in a cluster if they were smear-positive, had a pulmonary manifestation and were younger, male, alcohol or IV drug users, homeless, a health-care worker, native Dutch or foreign-born from Africa or the Americas. Patients were less likely to be in a cluster if they had travelled to an endemic area in the past two months or were foreign-born from Asia or Europe. In the multivariate model, all risk factors for clustering remained significant with the exception of alcohol consumption, being a health-care worker or being a foreign-born from the Americas. Being a foreign-born from Africa turned into a protective factor against clustering after adjustment in multivariate analysis. Resulting values for PPP and CPP ranged from 0 (a low risk profile for clustering, i.e. an elderly female patient with extra-pulmonary, smear-negative TB and no

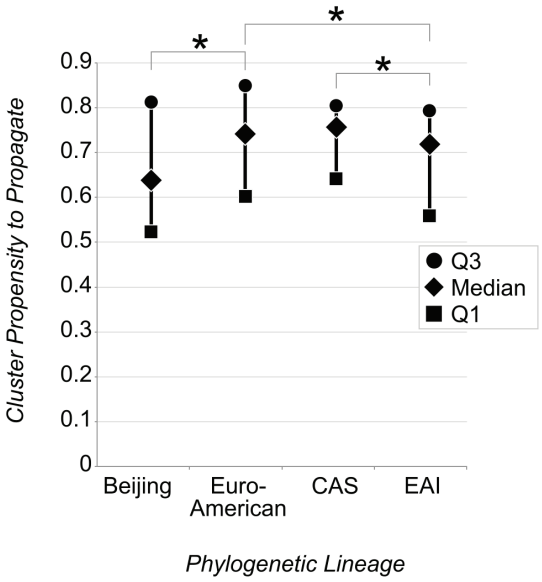


Figure 3. Distribution of Cluster Propensity to Propagate by 4 Phylogenetic Lineages. * 0.05 Level of Significance. Q1 – Lower Quartile; Q3 – Upper Quartile.
doi:10.1371/journal.pone.0097816.g003

behavioral risk factors) to 3·9 (a high risk profile i.e. a young (<30 years) male patient with pulmonary, smear-positive TB and at least one behavioral risk factor), with the distribution of CPP values skewed to the right of PPP values from patients with unique isolates (Figure 2).

Host-related factors by phylogenetic lineage

Of the 10,389 *M. tuberculosis* isolates which had both a CPP and/or PPP value and an assigned phylogenetic lineage, 6,595 were classified as Euro-American (63·5%), 1,327 as CAS (12·8%), 1,422 as EAI (13·7%) and 1,045 as Beijing (10·0%). The excluded 15% of strains that were not matched with the NTR data fall into a similar lineage distribution. Lineage misclassification was estimated at 19%, with 200 out of 248 strains in this study having concordant lineage classifications to SNP-based inferences. Of the 10,389 strains, 4,491 (43·2%) were non-clustered and the remainder consisted of 1,505 clusters, representing 175 CAS clusters, 972 Euro-American, 202 EAI and 156 Beijing. Median values for CPP were 0·64 (95% CI: 0·57–0·67), 0·76 (95% CI: 0·73–0·77), 0·75 (95% CI: 0·71–0·76) and 0·72 (95% CI: 0·70–0·73) for Beijing, Euro-American, CAS and EAI strains. CPP values from strains of the Euro-American lineage were on average higher than those of Beijing and EAI strains, and CAS strains were also on average higher than EAI strains at a 0·05 level of significance (Figure 3). CPP values of strains belonging to the validation subset of strains classified using high reliability markers showed a similar trend, with the median CPP of strains of the Euro-American lineage remaining on average higher than those of Beijing and EAI strains at a 0·05 level of significance. Repeating the ANOVA on clusters composed of patients of European origin only (n = 277) showed a significantly lower mean CPP in clusters of the Euro-American strain (0·73; 95% CI: 0·70–0·76) compared to that of in clusters of Beijing (0·89; 95% CI: 0·80–0·98) or CAS (0·86; 95% CI: 0·77–0·96) strains, at a 0·05 level of significance. In the subset of clusters composed of patients of Asian origin only (n = 57), mean CPP values were 0·46 (95% CI: 0·43–0·49), 0.40 (95% CI: 0·32–0·48), 0·40 and 0·43 (95% CI: 0·39–0·47), for Beijing, Euro-American, CAS (n = 1) and EAI strains, respectively. Excluding extra-pulmonary cases (n = 4,812) from the dataset and logistic regression model resulted in Beijing strains maintaining the lowest median CPP (0·71, 95% CI: 0·65–0·78) compared to that of Euro-American (0·85, 95% CI: 0·82–0·85), CAS (0·84, 95% CI: 0·80–0·85) and EAI (0·83, 95% CI: 0·75–0·86) strains.

Propagation by phylogenetic lineage

The proportion of clustered isolates was 60·7% (95% CI, 59·5–61·9) for Euro-American strains, 49·2% (95% CI, 46·5–51·9) for CAS strains, 51·1% (95% CI, 48·5–53·7) for EAI strains and 49·4% (95% CI, 46·4–52·4) for Beijing strains. Both minimum and average PPP/CPP per cluster size increased with rising cluster size (Figure 4). Likewise, mean and median cluster size grew with increasing CPP category (Figure 5). There were no significant differences between the mean and median values of cluster size between the four phylogenetic lineages within each CPP category.

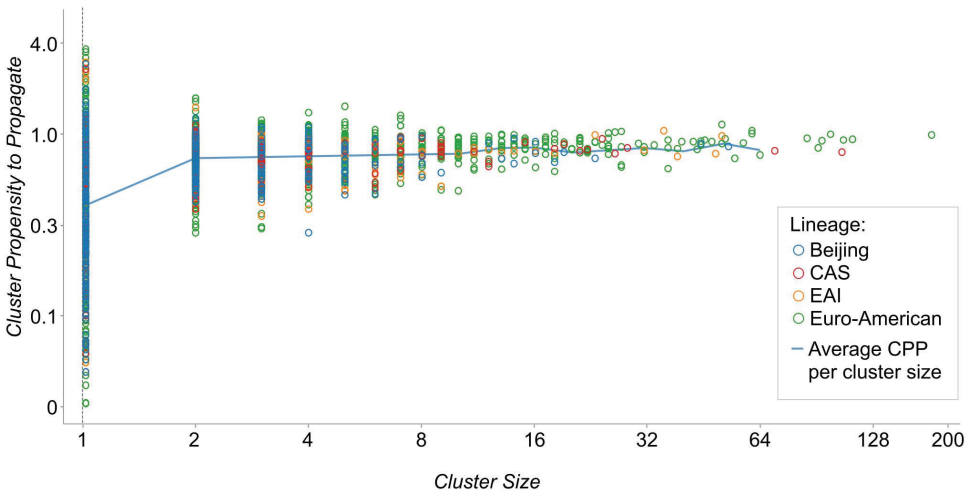


Figure 4. Distribution of Propensity to Propagate by Cluster Size.
doi:10.1371/journal.pone.0097816.g004

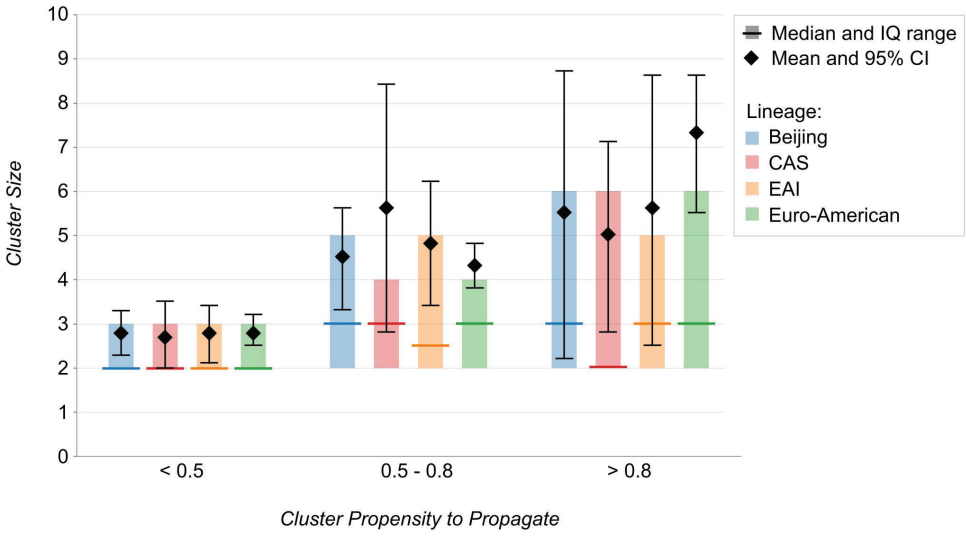


Figure 5. Distribution of Cluster Propensity to Propagate by Cluster Size and 4 Phylogenetic Lineages.
doi:10.1371/journal.pone.0097816.g005

Discussion

In this long-term Netherlands-based study, we compared the propensity to propagate of four major MTB lineages using a novel method designed to differentiate host and bacterial factors associated with strain transmissibility and progression. We found that although Euro-American strains were more likely to be found in clusters in an unadjusted analysis, no significant differences among the four different lineages persisted after we controlled for host factors.

The range of host factors associated with clustering that we identified in this study include demographic (age, gender and geographic origin), clinical (pulmonary manifestation and smear-positivity) and behavioral (drug-use, homelessness) determinants that have been identified in previous studies in this (and other) settings [7,20]. The clearly skewed distribution of cluster CPP values to the right of PPP values from patients with unique isolates confirms the role of host-related factors in propagation. The method described in this study to correct for host-related factors in transmission enables the identification of highly propagating strains (i.e. belonging to a larger than average cluster size for its CPP score) from non-propagating ones (i.e. non-clustered isolates with a high PPP). This selection process is useful to hone in on a crisp phenotype that is necessary to study bacterial factors associated with transmission, by means of genomic comparison in future whole-genome sequencing studies [21]. It would for example be interesting to subject the CDC1551 outbreak to our new approach in order to separate host risk factors from the true bacteriological component.

Our finding that the distribution of Cluster Propensity to Propagate scores was not equal across the Euro-American, Beijing, CAS and EAI lineages supports the notion that host-related factors need to be controlled for in order to attain comparability in measuring the ability of different phylogenetic lineages to propagate. Other previous studies in low prevalence settings such as Montreal and San Francisco have found the EAI lineage to be associated with lower rates of transmission [22], and the Euro-American lineage three times more likely to cause a secondary case [23]. The former adjusted their OR for clustering for age, whilst the rate measure used in the latter did not adjust for host-related factors. Discrepancies between results from studies measuring transmissibility between phylogenetic lineages may therefore partly be due to differences in how and if host-related factors are controlled for at all. This also seems important in the light of studies on co-evolution between bacteria and hosts; to facilitate a meaningful interpretation such studies should take patient risk factors for transmission and breakdown to disease into consideration [24]. In high prevalence settings this may be especially challenging.

A major strength of our study was the use of a large sample size over a long time period to accurately quantify the contribution of host-related factors in clustering within this setting. With 69% of patients being foreign-born from 159 different countries, our study sample is also globally representative; given the phylogeographic diversity of the major MTB lineages this is crucial to perform comparative analyses to identify associations between strain lineages and transmissibility. There is also an advantage in conducting this analysis in a low prevalence setting such as the Netherlands where the majority of people are susceptible and not vaccinated with BCG. This means that cluster sizes more closely reflect the biological underpinning of increased transmissibility rather than the proportion of the population that is still susceptible to MTB. Finally, our use of mean and median cluster size (therefore excluding non-clustered strains) across CPP categories instead of clustering rates decreases possible bias from the over-representation of foreign-born patients, associated with non-clustered strains from reactivation of latent TB infections acquired before immigration, among non-Euro-American strains (74·3%, 95% CI: 73·0–75·6) versus Euro-American strains (53·3%, 95% CI: 52·2–54·4).

Although our results contrast with those from studies carried out in other populations where Beijing has been associated with greater virulence and transmissibility [25–27], they are consistent with those from other low incidence immigrant-receiving settings such as the United States and Canada where it was concluded that Beijing strains do not pose more of a public health threat than non-Beijing strains [23,28]. The successful spread of this genotype in Asia and other parts of the world may therefore be related to a higher ability to withstand exposure to antituberculosis drugs and BCG vaccination, rather than a higher ability to propagate [11,29].

It has also been hypothesized that lineages that are rare in a specific human population are not adapted to transmit and cause secondary cases [23]. In Sweden for example, despite the close proximity to Russia and the Baltic states, Beijing was found to have a lower clustering rate, no absolute increase in number over time and very little observed transmission from immigrants to indigenous population [30]. In our study, there was no statistically significant difference between the median and mean cluster sizes of Beijing versus Euro-American strains after taking host propensity to propagate factors into account. This was also found to be the case for EAI and CAS strains, which suggests that imported strains in the Netherlands are not necessarily less adapted to the native host population and are just as likely to propagate as locally occurring strains of the Euro-American lineage. A lower mean CPP of Euro-American versus non-Euro-American strains found in clusters of European origin only suggests the possibility of co-evolution between phylogenetic lineages to their sympatric host population, as has been previously reported [23]. No significant differences were found however between CPP of phylogenetic lineages in clusters of Asian origin only, which may reflect the smaller sample size and reduced power to detect such an association.

The inclusion of *M. africanum* isolates, which have been associated with a lower rate of disease transmission compared to other MTB strains [31], for comparison in our study was not possible due to the very small number of patients infected with this strain. A differential representation of lineages amongst the native Dutch (who are not BCG vaccinated or previously exposed) versus the foreign-born population also represents a possible source of bias. In this dataset, the percentage of lineages circulating in the native Dutch were 7.6%, 10.4%, 25.3% and 36.7% in the CAS, EAI, Beijing and Euro-American lineages, respectively. It should also be noted that the weights used to calculate each patient's propensity to propagate (PPP) in this study depended on the clustering status given by molecular epidemiology data (RFLP- and VNTR-typing) alone, whose accuracy is limited.

In sum, this study demonstrates the importance of controlling for host-related factors in measuring the transmissibility of strains and describes a method to do so in order to identify bacterial factors in future studies. It also shows that there are no significant differences in the ability to propagate of four main phylogenetic lineages in the Netherlands, which is indicative that the spread of imported strains (most often of the EAI, CAS and Beijing lineages) is not necessarily curbed by a lack of adaptation to the native host population.

Supporting Information

Table S1 Classification of MIRU and spoligotypes into four lineage groups.
(DOC)

Acknowledgments

Herre Heersma is acknowledged for developing the RFLP/PGRS pattern comparison algorithm in BioNumerics, and Philip Supply for his advice on phylogenetic lineage classification using the MIRU-VNTRplus online tool. We thank Margarida Correia-Neves for her critical review of the manuscript, the staff of the RIVM mycobacteriological laboratory for their work on the RFLP and VNTR typing of *M. tuberculosis* isolates, and the Municipal Health Services for their voluntary collaboration in the nationwide tuberculosis surveillance.

Author Contributions

Conceived and designed the experiments: HNG MWB MBM DvS. Performed the experiments: HNG. Analyzed the data: HNG. Wrote the paper: HNG. Interpreted the data: HNG MWB MBM DvS. Reviewed the manuscript: MWB MBM DvS.

References

1. World Health Organization (2012) Global tuberculosis control: WHO Report 2012. Geneva: WHO Press.
2. Valway SE, Sanchez MP, Shinnick TF, Orme I, Agerton T et al. (1998) An outbreak involving extensive transmission of virulent strain of Mycobacterium tuberculosis. N Engl J Med 338: 633–639.
3. Kiers A, Drost AP, van Soolingen D, Veen J (1987). Use of DNA fingerprinting in international source case finding during a large outbreak of tuberculosis in the Netherlands. Int J Tuberc Lung Dis 1: 239–245.
4. Yang Z, Barnes PF, Chaves F, Eisenach KD, Weis SE et al. (1998) Diversity of DNA fingerprints of Mycobacterium tuberculosis isolates in the United States. J Clin Microbiol 36: 1003–1007.
5. Dillaha J, Yang Z, Ijaz K, Eisenach K, Cave M, et al. (2000) Transmission over a 54-year time span of an endemic strain of Mycobacterium tuberculosis in hospital, nursing home, jail and community settings. Presented at ATS International Conference; May 5-10, 2000; Toronto, Canada.
6. Ijaz K, Yang Z, Templeton G, Stead WW, Bates JH, et al. (2004) Persistence of a strain of Mycobacterium tuberculosis in a prison system. Int J Tuberc Lung Dis 8: 994–1000.
7. Borgdorff MW, van den Hof S, Kremer K, Verhagen L, Kalisvaart N et al. (2010) Progress towards tuberculosis elimination: secular trend, immigration and transmission. Eur Respir J 36: 339–347.
8. Verhagen LM, van den Hof S, van Deutekom H, Hermans PW, Kremer K et al. (2011) Mycobacterial factors relevant for transmission of Tuberculosis. J Infect Dis 203: 1249–1255.
9. Gagneux S, Small PM. (2007) Global phylogeography of Mycobacterium tuberculosis and implications for tuberculosis product development. Lancet Infect Dis 7: 328–37.
10. Glynn JR, Whiteley J, Bifani PJ, Kremer K, van Soolingen D. (2002) Worldwide occurrence of Beijing/W strains of Mycobacterium tuberculosis: a systematic review. Emerg Infect Dis 8: 843–849.
11. Parwati I, van Crevel R, van Soolingen D. (2010) Possible underlying mechanisms for successful emergence of the Mycobacterium tuberculosis Beijing genotype strains. Lancet Infect Dis 2: 103–111.
12. Van Embden JD, Cave MD, Crawford JT, Dale JW, Eisenach KD et al. (1993) Strain identification of Mycobacterium tuberculosis by DNA fingerprinting: recommendations for a standardized methodology. J Clin Microbiol 31: 406–409.
13. Supply P, Allix C, Lesejean S, Cardoso-Oelemann M, Rüsch-Gerdes S et al. (2006) Proposal for standardization of optimized mycobacterial interspersed repetitive unit-variable-number tandem repeat typing of Mycobacterium tuberculosis. J Clin Microbiol 44: 4498–4510.
14. De Beer JL, van Ingen J, de Vries G, Erkens C, Sebek M et al. (2013) Comparative study of IS6110 restriction fragment length polymorphism and variable-number tandem-repeat typing of Mycobacterium tuberculosis isolates in the Netherlands, based on a 5-year nationwide survey. J Clin Microbiol 51: 1193–1198.
15. Van Soolingen D, De Haas PE, Hermans PW, Groenen PM, Van Embden JD. (1993) Comparison of various repetitive DNA elements as genetic markers for strain differentiation and epidemiology of Mycobacterium tuberculosis. J Clin Microbiol 31: 1987–1995.
16. Weniger T, Krawczyk J, Supply P, Niemann S, Harmsen D. (2010) MIRU-VNTRplus: a web tool for polyphasic genotyping of Mycobacterium tuberculosis complex bacteria. Nucleic Acids Res 38(Web Server Issue).
17. Allix-Béguec C, Harmsen D, Weniger T, Supply P, Niemann S. (2008) Evaluation and user-strategy of MIRU-VNTRplus, a multifunctional database for online analysis of genotyping data and phylogenetic identification of Mycobacterium tuberculosis complex isolates. J Clin Microbiol 46: 2692–2699.
18. Comas I, Homolka S, Niemann S, Gagneux S. (2009) Genotyping of genetically monomorphic bacteria: DNA sequencing in Mycobacterium tuberculosis highlights the limitations of current methodologies. PLoS ONE 4(11): e7815.
19. Bryant JM, Schurch AC, van Deutekom H, Harris SR, de Beer J et al. (2013) Inferring patient to patient transmission of Mycobacterium tuberculosis from whole genome sequencing data. BMC Infectious Diseases 13: 110.
20. Kik SV, Verver S, van Soolingen D, de Haas PE, Cobelens FG et al. (2008) Tuberculosis outbreaks predicted by characteristics of first patients in a DNA fingerprint cluster. Am J Respir Crit Care Med 178: 96–104.
21. Talarico S., Ijaz K, Zhang X, Mukasa LN, Zhang L et al. (2011) Identification of factors for tuberculosis transmission via an integrated multidisciplinary approach. Tuberculosis 91: 244–249.

22. Albanna AS, Reed MB, Kotar KV, Fallow A, McIntosh FA et al. (2011) Reduced transmissibility of East African Indian strains of Mycobacterium tuberculosis. PLoS ONE 6: e25075.
23. Gagneux S, DeRiemer K, Van T, Kato-Maeda M, de Jong BC et al. (2006) Variable host-pathogen compatibility in Mycobacterium tuberculosis. PNAS 103: 2869–2873.
24. Tho DQ, Torok ME, Yen NT, Bang ND, Lan NT et al. (2012) Influence of antituberculosis drug resistance and Mycobacterium tuberculosis lineage on outcome in HIV-associated tuberculous meningitis. Antimicrob Agents Chemother 56: 3074–3079.
25. Yang C, Luo T, Sun G, Qiao K, Sun G et al. (2012) Mycobacterium tuberculosis Beijing strains favor transmission but not drug resistance in China. Clin Infect Dis 55: 1179–1187.
26. Toungoussova OS, Mariandyshev A, Bjune G, Sandven P, Caugant DA. (2003) Molecular epidemiology and drug resistance of Mycobacterium tuberculosis isolates in the Archangel prison in Russia: predominance of the W-Beijing clone family. Clin Infect Dis 37: 665–672.
27. Wada T, Fujihara S, Shimouchi A, Harada M, Ogura H et al. (2009) High transmissibility of the modern Beijing Mycobacterium tuberculosis in homeless patients of Japan. Tuberculosis 89: 252–255.
28. Langlois-Klassen D, Senthilselvan A, Chui L, Kunimoto D, Saunders LD et al. (2013) Transmission of Mycobacterium tuberculosis Beijing strains, Alberta, Canada, 1991–2007. Emerg Infect Dis 19: 701–711.
29. De Steenwinkel JE, ten Kate MT, de Knegt GJ, Kremer K, Aarnoutse RE et al. (2012) Drug susceptibility of Mycobacterium tuberculosis Beijing genotype and association with MDR TB. Emerg Infect Dis 18: 660–663.
30. Ghebremichael S, Groenheit R, Pennhag A, Koivula T, Andersson E et al. (2010) Drug resistant Mycobacterium tuberculosis of the Beijing genotype does not spread in Sweden. PLoS ONE 5: 210893.
31. De Jong B, Hill PC, Aiken A, Awine T, Antonio M et al. (2008) Progression to active tuberculosis, but not transmission, varies by M. tuberculosis lineage in The Gambia. J Infect Dis 198: 1037–1043.

Chapter III

Transmission and progression to disease of *Mycobacterium tuberculosis* phylogenetic lineages in the Netherlands

JCM
Journals.ASM.org

Transmission and Progression to Disease of *Mycobacterium tuberculosis* Phylogenetic Lineages in The Netherlands

Hanna Nebenzahl-Guimaraes,[a,b,c] Lilly M. Verhagen,[d,e] Martien W. Borgdorff,[f,g] Dick van Soolingen[a,h]

National Institute for Public Health and the Environment (RIVM), Bilthoven, The Netherlands[a]; Life and Health Sciences Research Institute (ICVS), School of Health Sciences, University of Minho, Braga, Portugal[b]; ICVS/3Bs, PT Government Associate Laboratory, Braga/Guimarães, Portugal[c]; Wilhelmina Children's Hospital Utrecht, Utrecht, The Netherlands[d]; Laboratory of Pediatric Infectious Diseases, Radboud University Medical Centre, Nijmegen, The Netherlands[e]; Public Health Service, Amsterdam, The Netherlands[f]; Department of Clinical Epidemiology, Academic Medical Centre, University of Amsterdam, Amsterdam, The Netherlands[g]; Department of Medical Microbiology, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands[h]

The aim of this study was to determine if mycobacterial lineages affect infection risk, clustering, and disease progression among *Mycobacterium tuberculosis* cases in The Netherlands. Multivariate negative binomial regression models adjusted for patient-related factors and stratified by patient ethnicity were used to determine the association between phylogenetic lineages and infectivity (mean number of positive contacts around each patient) and clustering (as defined by number of secondary cases within 2 years after diagnosis of an index case sharing the same fingerprint) indices. An estimate of progression to disease by each risk factor was calculated as a bootstrapped risk ratio of the clustering index by the infectivity index. Compared to the Euro-American reference, *Mycobacterium africanum* showed significantly lower infectivity and clustering indices in the foreign-born population, while *Mycobacterium bovis* showed significantly lower infectivity and clustering indices in the native population. Significantly lower infectivity was also observed for the East African Indian lineage in the foreign-born population. Smear positivity was a significant risk factor for increased infectivity and increased clustering. Estimates of progression to disease were significantly associated with age, sputum-smear status, and behavioral risk factors, such as alcohol and intravenous drug abuse, but not with phylogenetic lineages. In conclusion, we found evidence of a bacteriological factor influencing indicators of a strain's transmissibility, namely, a decreased ability to infect and a lower clustering index in ancient phylogenetic lineages compared to their modern counterparts. Confirmation of these findings via follow-up studies using tuberculin skin test conversion data should have important implications on *M. tuberculosis* control efforts.

Curbing tuberculosis (TB) transmission is a challenge in high-burden countries. However, even in low-prevalence settings, controlling TB is an important requirement due to human migration from higher-incidence areas to Western countries (1). In Western countries, studies on transmission are more feasible, as all cases undergo extended diagnostic algorithms and all clinical and demographic data are recorded. Current molecular typing methods, such as variable number of tandem repeat (VNTR) typing and restriction fragment length polymorphism (RFLP) typing, allow identification of clusters of *Mycobacterium tuberculosis* isolates with identical genotypes that, in population-based studies, reveal recent transmission (2, 3). Spoligotyping and VNTR typing can identify the genotype family of the isolate, revealing bacterial variation via the identification of phylogenetic lineages (4, 5).

While many studies have elucidated the variation in the disease's spread and outcome attributable to host and environmental factors, there is also evidence that bacterial factors may affect the spread of tuberculosis (6). In The Netherlands, for example, one study showed that the number of positive contacts around a case increases with growing cluster size (7). In a subsequent study in the same setting, cluster size growth was not different between phylogenetic lineages after controlling for host risk factors (8). However, this study could not distinguish between transmission rates and progression to disease. There are, however, indications that progression to disease is partly dependent on bacterial variation. It has, e.g., been postulated that some *Mycobacterium africanum* strains might transmit equally well as other *M. tuberculosis* complex strains but might be less associated with progression to disease (9). We will refer to these two properties that affect the degree of clustering as infectivity (the bacterium's ability to establish an initial infection in the human host) and progression to disease (the bacterium's capacity to produce disease) (10).

In the low-incidence context of The Netherlands, with a globally representative cohort of patients, we aim to determine differences in indices of infectivity, clustering, and estimated progression to disease of different mycobacterial lineages using fingerprinting data and contact investigation. This will provide insights into the role of bacteriological factors in TB transmission, which itself may affect future TB control measures.

MATERIALS AND METHODS

Data collection and DNA fingerprinting. The National Institute for Public Health and the Environment (RIVM) is a reference laboratory for secondary laboratory diagnosis of all TB cases in The Netherlands, offer-

Received 20 May 2015 Returned for modification 19 June 2015
Accepted 23 July 2015
Accepted manuscript posted online 29 July 2015
Citation Nebenzahl-Guimaraes H, Verhagen LM, Borgdorff MW, van Soolingen D. 2015. Transmission and progression to disease of *Mycobacterium tuberculosis* phylogenetic lineages in The Netherlands. J Clin Microbiol 53:3264–3271. doi:10.1128/JCM.01370-15.
Editor: G. A. Land
Address correspondence to Hanna N. Guimaraes, hanna.guimaraes@gmail.com.

doi:10.1128/JCM.01370-15

ing identification, drug susceptibility testing, and molecular typing for each TB case. DNA fingerprints of all nationwide *M. tuberculosis* complex isolates and their cluster statuses have been stored in an RFLP/VNTR database since 1993. The registration committee of The Netherlands Tuberculosis Register (NTR) approved this retrospective study and provided anonymized demographic and clinical information for patients. Because these data are deidentified by name, DNA fingerprinting results were matched by sex, date of birth, year of diagnosis, and postal code. All notified culture-positive cases of *M. tuberculosis* between 1993 and 2011 were included in the study. For patients with multiple isolates sharing identical fingerprints, only the isolate with the earliest diagnosis date was included. Contaminating isolates were excluded.

Isolates recovered from patients between 1993 and 2009 underwent IS*6110* typing and polymorphic GC-rich sequence (PGRS) RFLP typing (n = 15,073), and those from 2004 onward were subjected to VNTR typing (n = 5,870) (11, 12). In the period of 2004 to 2008, both RFLP and 24-locus VNTR typing were performed to obtain a smooth transition in typing methods and to evaluate VNTR typing performance (3). In addition, 4,433 randomly selected isolates were spoligotyped (n = 4,433). We defined a cluster as a group of patients who shared *M. tuberculosis* isolates with identical RFLP or VNTR patterns or, if strains had fewer than five IS*6110* copies, identical PGRS RFLP patterns.

Conventional contact investigation. Systematic contact investigation by TB Public Health Services in The Netherlands is conducted per the stone-in-the-pond principle, in which the decision to extend conventional contact investigation to the next ring of contacts is based on the prevalence of infection in the investigated ring (13). Contacts are defined by the frequency and intimacy of their contacts with the TB index case. The tuberculin skin test (TST) is used to investigate presumably exposed contacts. If the number of TST-positive contacts in the first ring suggests a high spread of tuberculosis, a larger ring of contacts is investigated. We have defined positive contacts as contacts with a TST induration ≥10 mm and/or contacts who received a diagnosis of TB disease. If contact investigations become very large, identified TB infections and secondary cases are less likely to be related to the index case. To minimize the probability that positive contacts in our research were unrelated to the defined index case, we only included contacts in the first ring around the index patient. First-ring contacts are defined as contacts that are physically close to the index patient, considering environmental factors, such as room size, ventilation, air purification, and air circulation. In addition, the patient and the contact must be able to indicate where they met and must have a long-standing relationship to qualify as a first-ring contact. Examples of first-ring contacts are household members, close work colleagues, and close friends.

Classification into phylogenetic lineages. The phylogenetic lineages of isolates were determined using a combination of spoligotyping, the MIRU (mycobacterial interspersed repetitive unit) best match analysis offered by the MIRU-VNTR*plus* online tool, and RFLP similarity, as described in a previous study using the same data set (8, 14). Three species (*M. africanum*, *Mycobacterium bovis*, and *M. tuberculosis*) and four major phylogenetic lineages of *M. tuberculosis* were identified: the Euro-American, Central Asian strain (CAS), East African Indian (EAI), and Beijing genotypes. Strains not assigned a phylogenetic lineage or assigned to multiple major phylogenetic families per cluster were not analyzed. Strains classified as either T or U (unknown) also were excluded due to the ambiguity of these classifications.

Definitions. For our infectivity index, we took the mean number of positive contacts around each patient who underwent contact investigation. We excluded patients with missing data on contact investigation or those who had zero contacts investigated, as well as those for whom we lacked ethnicity information. Because TB transmission almost exclusively results from patients with pulmonary TB, we also excluded patients with extrapulmonary TB, leaving us with a total of 2,809 cases (Fig. 1).

For our clustering index, we used the number of secondary cases occurring within 2 years of the index case diagnosis. The 2-year cutoff has been shown to best reflect recent transmission as opposed to disease reactivation (1, 15). We defined index cases as patients who had strains with RFLP or VNTR patterns not seen in other patients in the previous 2 years. We searched for index cases based on RFLP-typing data from 1995 to 2007 and for index cases based on VNTR typing from 2007 to 2009. We excluded RFLP-defined index cases from 1993 and 1994 and VNTR-defined index cases from 2005 and 2006 (n = 2,684), because we could not determine whether the strains of these index cases were unobserved in the previous 2 years. Similarly, we excluded RFLP-defined index cases occurring after 2007 and VNTR-defined index cases occurring after 2009 (n = 950), because we could not follow these index cases for a full 2 years. Secondary cases from these index cases (included in the counts) were also excluded. Finally, we excluded cases between 1995 and 2007 occurring <2 years after a previous patient with the same RFLP fingerprint yet diagnosed >2 years after a cluster's start (n = 722) and cases occurring between 2007 and 2009 that occurred <2 years after a previous patient with the same VNTR fingerprint yet >2 years after a cluster's start (n = 40). After excluding extrapulmonary cases, 4,432 patients remained: 2,881 nonclustered index patients, 607 index patients who were the first patient of a cluster, and 944 secondary cases within 2 years of a cluster's start (Fig. 1).

Finally, estimates of progression to disease were calculated as risk ratios (RR) of the population risk of disease given exposure to a risk factor by the population risk of infection given exposure to the same risk factor (dividing the clustering odds ratios [ORs] by the infectivity ORs).

Statistical analysis. We used a multivariate negative binomial regression model to determine the association between phylogenetic lineages and the infectivity and clustering indices. Since TST is poorly specific among *Mycobacterium bovis* BCG contacts and positive TSTs may represent old infections, we divided our data sets into native and foreign-born (FB) cohorts in order to address important differences between the two: FB patients are often BCG vaccinated (in contrast to native Dutch patients, who are not), while the prevalence of infection is higher among FB patients. Second-generation patients (born to FB patients) were included in the native cohorts, given that, like native patients, they are not BCG vaccinated and they have already been born in a setting of lower prevalence of infection. Studies carried out in The Netherlands have also previously demonstrated that contact investigation practices vary by demographic characteristics of the index patient (16). As such, in both analyses, we adjusted for index patient-related factors, including demographic, behavioral, and sputum smear status. In addition, the logarithm of the number of investigated contacts around a source case was used as an offset in the multivariate model assessing the association between phylogenetic lineages and the spread index, since the greater number of contacts around a source are investigated, the likelier it is to detect TST positive contacts. Variables with P values of ≤0.20 were entered into the multivariate model. Crude and adjusted ORs are presented with 95% confidence intervals (CIs). Estimates of TB progression were calculated for any risk factor that was significant in either multivariate regression model. To calculate the variance for the estimate of TB progression, we performed a bootstrapping procedure, running our multivariate negative binomial regression models 10,000 times on bootstrapped data sets. The median of the resulting 10,000 RRs was used as the estimate of TB progression, while the 2.5th and 97.5th percentiles were used as the 95% cutoffs for the estimate CI. All analyses were conducted using SAS (Windows version 9.3), SPSS program for Windows version 20.0 (SPSS Inc., Chicago, IL, USA) and R (version 3.1.2 for Windows).

RESULTS

Between January 1993 and December 2011, 18,294 isolates were collected from the same number of notified TB cases in The Netherlands, and their clustering statuses were ascertained, of which 15,601 (85%) were successfully matched with the NTR data. Of these, 15,224 (98%) were noncontaminated *M. tuberculosis* cul-

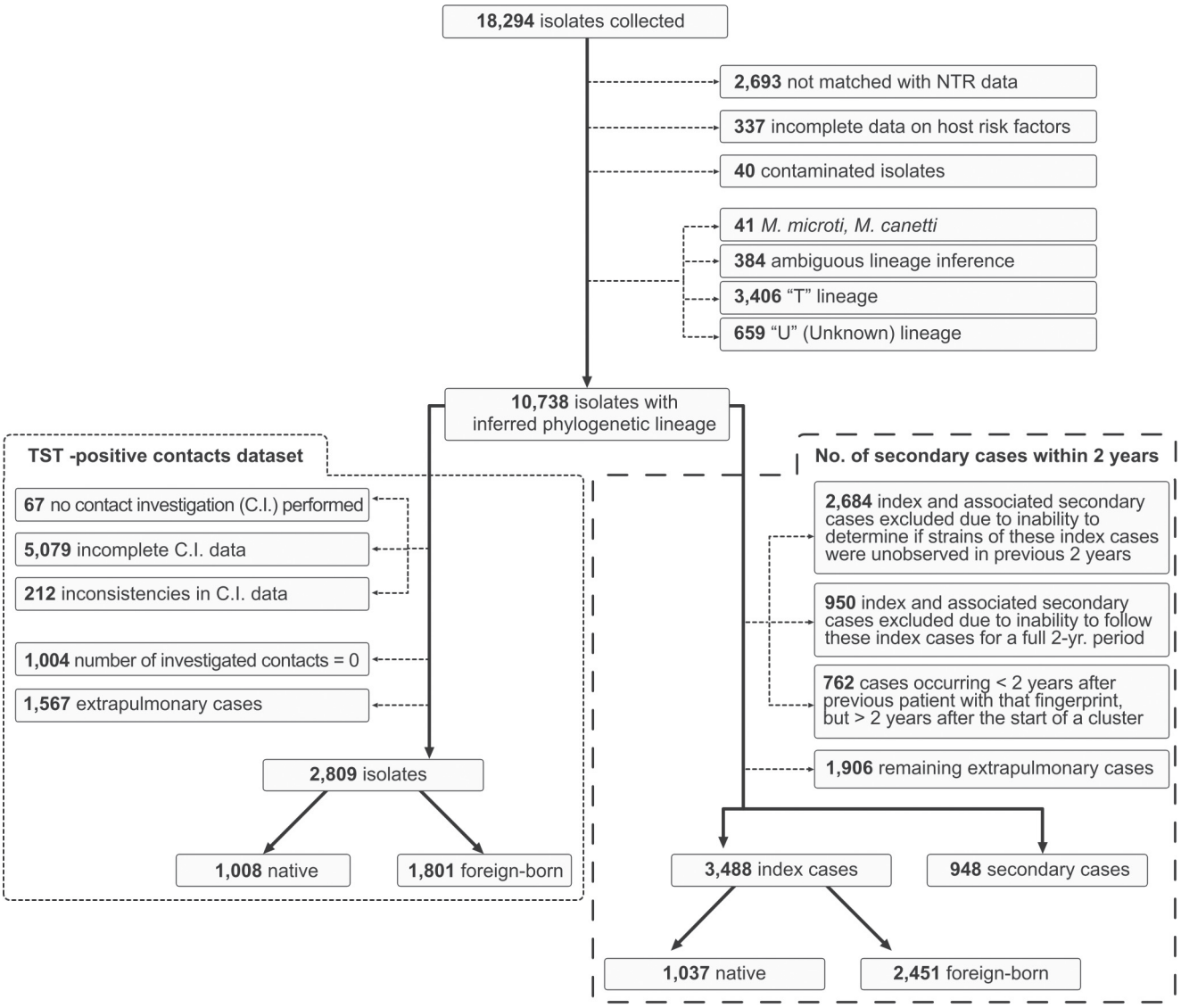


FIG 1 Flow-diagram of exclusion criteria applied to data set.

tures with completely ascertained information on host risk factors. After phylogenetic lineage assignment, there were 10,738 isolates that were *M. bovis*, *M. africanum*, or *M. tuberculosis* of the Euro-American, Beijing, CAS, or EAI lineages (Fig. 1). The mean age of the patients carrying these strains was 41 years (standard deviation, 20 years); 6,394 (60%) were male; and 7,762 (72%) were foreign born.

Mycobacterial genotypes. The Euro-American lineage was predominant in both the infectivity (78% in native cohort; 56% in FB cohort) and clustering (79% in native cohort; 64% in FB cohort) data sets. In contrast, both *M. africanum* and *M. bovis* represented less than 1% of all cases in the infectivity data set. In the clustering data set, both *M. africanum* and *M. bovis* represented only 2% of all cases (Tables 1 and 2).

Infectivity by mycobacterial lineage. The proportion of cases in which a contact investigation was performed in The Netherlands was approximately equal between lineages, though slightly lower in the FB cohort for Beijing and EAI compared to the native counterpart (Fig. 2). The average number of TST-positive contacts declined significantly in the >65 years age category in the native cohort and in the <20 years age category in the FB cohort. Smear positivity was associated with an increased average number of TST-positive contacts in both native and FB cohorts. There were no significant differences in infectivity by gender, homelessness, and alcohol use in the two cohorts, although use of intravenous drugs in the native populations and rural residence in the FB population were associated with a decreased average number of TST-positive contacts. The mean number of TST-positive contacts around an index case was significantly lower for *M. bovis* than for the Euro-American reference lineage in the native population in multivariable analysis. In the FB population, *M. africanum* and EAI presented a significantly lower number of TST-positive contacts (Table 1).

Clustering by mycobacterial lineage. The number of secondary cases declined significantly with increasing age (>65 years) in both the native and FB cohorts. Smear positivity was also associ-

TABLE 1 Risk factors among native and foreign-born index cases for infectivity (number of TST-positive contacts per index case)

Characteristic	Native cohort					Foreign-born cohort				
	No. of index cases	Mean no. of TST-positive contacts/ index case	Univariate analysis		Multivariate analysis, relative no. (95% CI)	No. of index cases	Mean no. of TST-positive contacts/ index case	Univariate analysis		Multivariate analysis, relative no. (95% CI)
			Relative no. (95% CI)	*P*				Relative no. (95% CI)	*P*	
Age, yr				0.00					0.033	
0–19	99	1.02	1.0 (0.68–1.6)		0.96 (0.65–1.4)	200	0.83	1.11 (0.83–1.5)		0.69 (0.53–0.89)
20–39	363	0.99	1 (Ref[a])		1 (Ref)	1024	0.75	1 (Ref)		1 (Ref)
40–64	339	0.82	0.83 (0.63–1.1)		0.91 (0.69–1.2)	459	0.60	0.81 (0.65–1.0)		0.96 (0.79–1.2)
≥65	207	0.51	0.51 (0.37–0.72)		0.50 (0.36–0.70)	118	0.49	0.66 (0.45–0.97)		0.80 (0.57–1.1)
Sex				0.15					0.32	
Male	635	0.78	1 (Ref)		1 (Ref)	1,125	0.73	1 (Ref)		
Female	373	0.93	1.2 (0.93–1.5)		1.0 (0.82–1.3)	676	0.66	0.73 (0.65–0.82)		
Smear positivity				<0.001					<0.001	
Negative	395	0.52	0.50 (0.39–0.65)		0.48 (0.38–0.62)	754	0.47	0.54 (0.45–0.65)		0.55 (0.46–0.65)
Positive	613	1.04	1 (Ref)		1 (Ref)	1,047	0.87	1 (Ref)		1 (Ref)
Lineage				0.00					0.006	
Euro-American	786	0.91	1 (Ref)		1 (Ref)	1,157	0.75	1 (Ref)		1 (Ref)
Beijing	112	0.67	0.73 (0.50–1.1)		1.1 (0.77–1.6)	182	0.55	0.74 (0.54–1.0)		0.94 (0.71–1.2)
CAS	30	0.23	0.26 (0.11–0.61)		0.64 (0.27–1.5)	198	0.86	1.2 (0.86–1.5)		1.0 (0.79–1.3)
EAI	62	0.56	0.62 (0.37–1.04)		0.78 (0.47–1.3)	237	0.53	0.71 (0.54–0.94)		0.64 (0.49–0.83)
M. Africanum	5	0.40	0.44 (0.067–2.9)		0.44 (0.055–3.5)	15	0.27	0.36 (0.11–1.1)		0.30 (0.10–0.89)
M. bovis	13	0.23	0.25 (0.068–0.94)		0.23 (0.059–0.94)	12	0.17	0.22 (0.052–0.96)		0.51 (0.11–2.4)
Residency				0.07					0.002	
Urban	690	0.90	1 (Ref)		1 (Ref)	1,067	0.79	1 (Ref)		1 (Ref)
Rural	318	0.70	0.78 (0.60–1.0)		0.78 (0.60–1.0)	734	0.58	0.74 (0.62–0.89)		0.71 (0.60–0.84)
Alcohol abuse				0.19					0.69	
No	969	0.85	1 (Ref)		1 (Ref)	1,779	0.70	1 (Ref)		
Yes	39	0.54	0.64 (0.33–1.2)		0.59 (0.30–1.2)	22	0.59	0.84 (0.36–1.9)		
Drug abuse				0.01					0.97	
No	953	0.86	1 (Ref)		1 (Ref)	1,722	0.70	1 (Ref)		
Yes	55	0.40	0.47 (0.26–0.83)		0.43 (0.24–0.78)	79	0.71	1.0 (0.65–1.6)		
Traveler to country of endemicity				0.17					0.17	
No	973	0.85	1 (Ref)		1 (Ref)	1,759	0.69	1 (Ref)		1 (Ref)
Yes	35	0.51	0.61 (0.30–1.2)		0.53 (0.25–1.1)	42	1.02	1.5 (0.83–2.6)		1.5 (0.88–2.4)
Homeless				0.72					0.94	
No	978	0.84	1 (Ref)			1,753	0.70	1 (Ref)		
Yes	30	0.73	0.88 (0.43–1.8)			48	0.69	0.98 (0.56–1.7)		
Site of disease				0.21					0.099	
Pulmonary	891	0.86	1 (Ref)		1 (Ref)	1,430	0.73	1 (Ref)		1 (Ref)
Pulmonary + extrapulmonary	117	0.67	0.78 (0.53–1.1)		1.0 (0.70–1.5)	371	0.60	0.83 (0.66–1.0)		0.88 (0.71–1.1)

[a] Ref, reference.

ated with an increased number of secondary cases in both cohorts, and female gender was associated with an increased number of secondary cases only among the FB. Rural residence was associated with a decreased number of secondary cases only in the FB cohort. Relative to the Euro-American reference in the multivariable analysis, the number of secondary cases was significantly lower for *M. bovis* in the native-born population and for *M. africanum* in the FB population (Table 2).

Estimates of progression to disease by mycobacterial lineage. Estimates of progression to disease were significantly lower in the >65 years age category in both ethnic cohorts and significantly higher in the 0- to 19-years age category in the FB cohort. Additionally, in the FB-born population, estimates of progression to disease were significantly lower in smear-negative patients. Both alcohol and drug abuse were significantly associated with higher estimates in the native population. No

TABLE 2 Risk factors among native and foreign-born index cases for clustering (number of secondary cases within 2 years of an index case)

Characteristic	Native cohort					Foreign-born cohort				
	No. of index cases	Mean no. of second cases per index case	Univariate analysis		Multivariate analysis, relative no. (95% CI)	No. index cases	Mean no. of second cases per index case	Univariate analysis		Multivariate analysis, relative no. (95% CI)
			Relative no. (95% CI)	*P*				Relative no. (95% CI)	*P*	
Age, yr										
0–19	59	0.46	1.02 (0.53–1.90)	0.001	1.04 (0.56–1.93)	276	0.36	1.32 (0.92–1.92)	0.716	1.42 (0.99–2.03)
20–39	216	0.45	1 (Ref[a])		1 (Ref)	1,411	0.27	1 (Ref)		1 (Ref)
40–64	267	0.34	0.65 (0.41–0.98)		0.68 (0.46–1.02)	569	0.31	1.14 (0.86–1.51)		1.06 (0.80–1.42)
≥65	495	0.1	0.22 (0.12–0.33)		0.21 (0.13–0.32)	195	0.08	0.30 (0.16–0.55)		0.30 (0.16–0.55)
Sex										
Female	387	0.23	0.87 (0.61–1.23)	0.4255		975	0.23	0.51 (0.29–0.90)	0.019	1.29 (1.01–1.66)
Male	650	0.27	1 (Ref)			1,476	0.31	1 (Ref)		1 (Ref)
Smear positivity										
Negative	393	0.18	0.58 (0.40–0.83)	0.0027	0.62 (0.44–0.88)	1,076	0.17	0.48 (0.38–0.61)	<.0001	0.50 (0.39–0.65)
Positive	644	0.3	1 (Ref)		1 (Ref)	1,375	0.36	1 (Ref)		1 (Ref)
Lineage										
Euro-American	756	0.25	1 (Ref)	0.1971	1 (Ref)	1,385	0.3	1 (Ref)	0.081	1 (Ref)
Beijing	73	0.36	1.44 (0.79–2.62)		1.20 (0.68–2.11)	354	0.28	0.93 (0.66–1.32)		1.07 (0.76–1.52)
CAS	28	0.29	1.15 (0.41–3.09)		0.79 (0.30–2.10)	290	0.28	0.93 (0.64–1.36)		0.87 (0.59–1.27)
EAI	118	0.33	1.34 (0.82–2.19)		0.90 (0.54–1.50)	335	0.22	0.74 (0.51–1.07)		0.80 (0.57–1.17)
M. Africanum	8	0.13	0.50 (0.04–5.22)		0.26 (0.03–2.53)	65	0.15	0.52 (0.22–1.21)		0.47 (0.31–0.94)
M. bovis	54	0.06	0.22 (0.06–0.78)		0.11 (0.006–0.56)	22	0.09	0.30 (0.06–1.67)		0.31 (0.06–1.67)
Residency										
Urban	785	0.35	1 (Ref)	0.02	1 (Ref)	1,547	0.34	1 (Ref)	0.004	1 (Ref)
Rural	252	0.23	0.65 (0.45–0.94)		0.99 (0.69–1.43)	904	0.24	0.70 (0.55–0.89)		0.77 (0.61–0.99)
Alcohol abuse										
No	1,009	0.24	1 (Ref)	0.02	1 (Ref)	2,430	0.28	1 (Ref)	0.964	
Yes	28	0.68	2.8 (1.22–6.42)		1.88 (0.89–3.99)	21	0.29	1.93 (0.28–3.74)		
Drug abuse										
No	1,017	0.25	1 (Ref)	0.05	1 (Ref)	2,398	0.28	1 (Ref)	0.905	
Yes	20	0.65	2.64 (0.98–7.03)		1.42 (0.58–3.49)	53	0.26	1.53 (0.42–2.18)		
Traveler to country of endemicity										
No	988	0.26	1 (Ref)	0.07	1 (Ref)	2,391	0.28	1 (Ref)	0.167	1 (Ref)
Yes	49	0.10	0.39 (0.14–1.09)		0.26 (0.09–0.70)	60	0.15	0.53 (0.22–1.30)		0.47 (0.20–1.14)
Homeless										
No	1,028	0.25	1 (Ref)	0.75		2,401	0.27	1 (Ref)	0.233	1 (Ref)
Yes	9	0.33	1.31 (0.25–6.99)			50	0.44	1.61 (0.74–3.49)		1.19 (0.5–2.56)
Site of disease										
Pulmonary	883	0.26	1 (Ref)	0.22	1 (Ref)	1,949	0.29	1 (Ref)	0.129	1 (Ref)
Pulmonary + extrapulmonary	154	0.19			0.94 (0.58–1.53)	502	0.23	0.79 (0.58–1.07)		0.92 (0.67–0.99)

[a] Ref, reference.

significant differences were found across phylogenetic lineages (Table 3).

DISCUSSION

In this study, we observed variations between the infectivity and clustering indices of different phylogenetic subgroups of *M. tuberculosis*, *M. bovis*, and *M. africanum* after controlling for clinical and demographic index host factors. *M. africanum* and *M. bovis* showed both significantly lower infectivity and clustering indices in the FB and native populations, respectively. A significantly lower infectivity was also observed for the EAI lineage in the larger FB population.

Our findings around *M. africanum* are consistent with previous experiments characterizing its reduced ESAT-6 (early secretory antigenic target-6) immunogenicity and candidate genes behind its attenuated phenotype (17). However, they are only

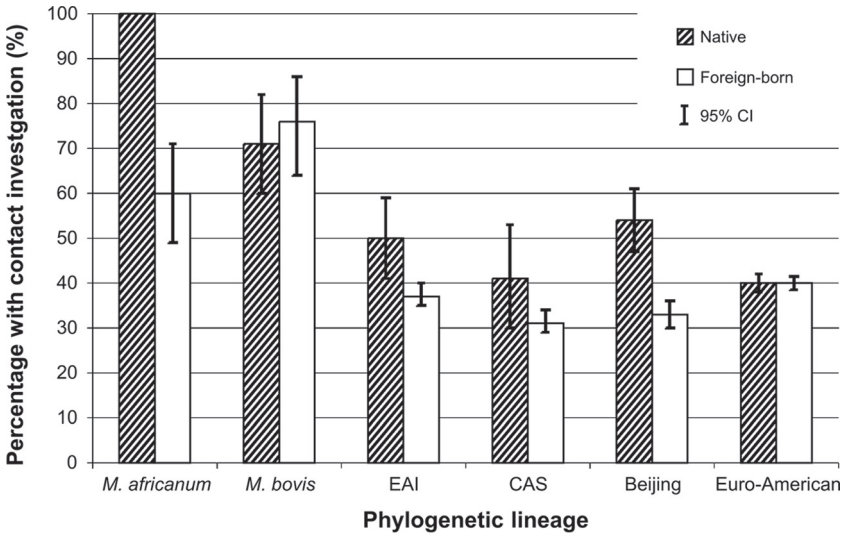


FIG 2 Proportion of cases in which contact investigation was performed by phylogenetic lineage.

TABLE 3 Estimates of progression to disease by risk factor

Characteristic	Median of bootstrapped progression-to-disease RR (95% CI)	
	Native cohort	Foreign-born cohort
Age, yr		
0–19	1.09 (0.54–2.39)	2.05 (1.25–3.74)
20–39	1 (Ref[a])	1 (Ref)
40–64	0.83 (0.56–1.25)	1.15 (0.79–1.74)
≥65	0.76 (0.58–0.94)	0.60 (0.40–0.87)
Smear positivity		
Negative	0.62 (0.27–1.52)	0.26 (0.17–0.38)
Positive	1 (Ref)	1 (Ref)
Lineage		
Euro-American	1 (Ref)	1 (Ref)
Beijing	1.33 (0.60–3.48)	1.32 (0.64–2.35)
CAS	2.47 (0.73–11.26)	0.86 (0.54–1.43)
EAI	1.30 (0.57–3.30)	1.16 (0.82–1.68)
M. africanum	2.07 (0.96–2.11)	1.14 (0.70–2.11)
M. bovis	0.89 (0.66–1.31)	0.92 (0.45–2.00)
Residency		
Urban	1 (Ref)	1 (Ref)
Rural	1.67 (1.01–3.01)	1.65 (0.91–2.42)
Alcohol abuse		
No	1 (Ref)	
Yes	9.78 (1.52–159.85)	
Drug abuse		
No	1 (Ref)	
Yes	3.79 (1.20–22.68)	
Traveler to country with endemicity		
No	1 (Ref)	1 (Ref)
Yes	0.77 (0.41–1.30)	0.39 (0.13–0.95)

[a] Ref, reference.

partially consistent with those from a study conducted in the Gambia, where *M. africanum* was shown to transmit equally well to household contacts but less likely than *M. tuberculosis* to progress to disease (9). While numbers in our native population were too low to detect any associations in both indices, in the larger FB cohort, our findings suggest that lower infectivity might also be a component of the overall lower transmissibility of *M. africanum*. Perhaps because of this lower infectivity, we did not observe the previously reported lower estimate of progression to disease in *M. africanum*. Possible explanations for this disparity may lie in the slightly different definition of infectivity used in the Gambia, where they used the incidence of TST conversion (using a follow-up period of 3 months) specifically within households as the outcome. In addition, we may not be comparing exactly the same genotype; in our FB cohort, only 3 of 183 (1.7%) *M. africanum* strains with a known birth country came from the Gambia.

In a cohort of native and FB TB cases in Montreal, the EAI lineage was also significantly associated with lower number of TST-positive contacts around index cases and with less clustering (lower proportion of patients clustering, as defined by identical RFLP or spoligotypes) in multivariable analysis (18). It is interesting to observe this trend in our study, which includes only pulmonary cases of EAI, given the association this lineage has with the extrapulmonary site of disease (16). In a secondary cohort of only FB cases in the Montreal study, the EAI lineage was significantly associated with less TST positivity but not with less clustering (18). This again agrees with our study, where we observed a significant association of EAI with lower infectivity but not with lower clustering. These findings on the EAI genotype are hard to explain using the molecular epidemiological data from Vietnam, where approximately 40% of cases are caused by EAI strains and another 40%, by the Beijing genotype strains (19). If EAI strains are less successful at infecting, one would expect them to disappear in a few generations and be replaced by other, more fit, strains. This shift is perhaps occurring at the very moment, as Beijing genotype isolates have been associated with a lower age of patients and, hence, with active transmission.

Although *M. bovis* was spread significantly less in the native

population and although the estimates of average number of secondary cases were lower than other lineages, the fact that there were three documented secondary cases (from three different index cases with unique fingerprints) does not rule out the possible occurrence of transmission of *M. bovis* in The Netherlands, where pasteurization practices have been in place for decades. Ingestion of unpasteurized dairy products has been suggested as the likely route of infection in extrapulmonary cases in second-generation immigrants in The Netherlands who may have traveled back to their country of origin (20). Yet, all three *M. bovis* index cases with secondary cases in our clustering cohort also had pulmonary manifestations; two of these index cases were FB but had no indication of recent travel to a country of *M. bovis* endemicity. Indeed, instances of human-to-human transmission of *M. bovis* have been documented in other settings (21, 22). Together these observations suggest that, from a public health perspective, contact investigation and treatment of pulmonary *M. bovis* patients should not altogether differ from those of *M. tuberculosis* patients.

Unlike studies conducted in other populations, where the Beijing strain was associated with greater virulence and transmissibility, we did not find that the Beijing strain had higher indices of infectivity, clustering, or progression to disease in The Netherlands (23, 24). This is concordant with other recent studies conducted in similarly low-incidence, immigrant-receiving settings, such as the United States and Canada, which concluded that Beijing strains are no more of a public health threat than non-Beijing strains (25, 26). The observed higher success rate of Beijing strains may therefore result from circumstances characteristic of high-prevalence settings, such as mass use of BCG vaccination, development of resistance, crowding of the human population, and other unknown factors.

Other clinical and demographic factors positively associated with either infectivity or clustering indices, such as smear positivity, a lower age, and residing in an urban area, have been similarly described in previous studies (27–29). The significantly lower estimate of progression to disease given an elderly source likely reflects a lower dose of infection (due to a less close contact) and propensity for older patients to have older contacts themselves, as well as the higher proportion of long, latent infections (possibly associated with lower virulence) in this age category (30). Likewise, the significant association between alcohol and drug abuse with higher estimates of progression to disease can be linked to the direct effects of both substances on immunity, the indirect effects of substance-related disorders (i.e., malnutrition), and other potential confounding factors, such as homelessness (31, 32). There are two possible reasons behind the less-expected association between use of intravenous drugs and the lower average number of TST-positive contacts in the native cohort. Contacts of drug abusers are often intravenous drug users themselves, a scenario in which the accurate definition of a first-ring contact is prone to misclassification (contacts could be misclassified as first-ring contacts while they actually do not have much contact with an index case and, therefore, do not become TST positive). It has also been described that drug use can comprise cellular immunity (even in the absence of HIV infection) so that TST sensitivity in drug users is lower (33, 34).

The low prevalence setting of this study means that the investigation of the role of the *M. tuberculosis* genotype on transmission is less likely to be confounded by a high background infection pressure, where a TST result is more likely to fail at distinguishing recent from past infection. Furthermore, in The Netherlands there is no routine BCG vaccination program that could affect the interpretation of TST results, making TST a suitable tool for the detection of recent *M. tuberculosis* infection in contact investigations. This advantage applies solely to the native cohort, however, as patients in the FB cohort are far more likely to have been BCG vaccinated than native patients (40% versus 8%, respectively) and have had higher exposure to TB in their country of origin; both of these factors might lead to an overestimation of infectivity. It is encouraging, however, to observe the same trend of lower infectivity in EAI result in another study which did adjust for the probability of previous latent TB (18). On the other hand, the facts that FB patients often have FB contacts and that contact tracing in this group is less efficient imply that we might have also underestimated infectivity (and, by implication, biased the progression to disease index upward) in this group. The same reasoning applies to cases of addiction to alcohol and drugs, where an increased likelihood of homelessness means infected contacts are less likely to be found.

It is important to remember the potential shortcomings from the molecular epidemiology data underpinning these findings. A lack of clinical follow-up data of infected contacts meant that we were unable to link infected contacts to secondary cases and, thus, to estimate the proportion of secondary cases infected by a specific index case. In this low-burden country, however, there is likely a large overlap in the number of infected contacts around an index case and the number of secondary cases occurring within 2 years of that index case. It nevertheless meant that we could not control for risk factors across the transmission chain, such as rates of latent TB treatment and existing medical risk factors in secondary cases, which could influence the likelihood of progression to disease or the susceptibility to infection of the host, respectively. Studies using a prospective cohort approach (i.e., with access to household contacts and TST conversion data) that can bypass some of these issues are warranted to confirm these findings.

In sum, the lower infectivity or overall transmissibility observed in this study for *M. bovis*, *M. africanum*, and EAI—all, ancient lineages—matches the hypothesis that modern strains, as a consequence of their access to rapidly increasing numbers of susceptible hosts, have been selected for more rapid disease progression and transmission (35). Validation of this scenario via future experimental studies could have important implications on how TB control efforts may be determined not only by index case host characteristics, but also by a bacterial signature, such as phylogenetic lineage.

ACKNOWLEDGMENTS

This study was supported by the Portuguese Foundation for Science and Technology (FCT) (reference SFRH/BD/33902/2009 to H.N.-G).

We thank Rogier Donders, Megan Murray, and Maha Farhat for their statistical support and discussion of the methodology used. We thank the staff of the RIVM mycobacteriological laboratory for their work on the RFLP and VNTR typing of *M. tuberculosis* isolates, and we thank the Municipal Health Services for their voluntary collaboration in the nationwide tuberculosis surveillance.

REFERENCES

1. **Borgdorff MW, van den Hof S, Kremer K, Verhagen L, Kalisvaart N, Erkens C, van Soolingen D.** 2010. Progress towards tuberculosis elimination: secular trend, immigration and transmission. Eur Respir J **36:**339–347. http://dx.doi.org/10.1183/09031936.00155409.

2. **Goguet de la Salmoniere YO, Li HM, Torrea G, Bunschoten A, van Embden J, Gicquel B.** 1997. Evaluation of spoligotyping in a study of the transmission of *Mycobacterium tuberculosis*. J Clin Microbiol **35:**2210–2214.
3. **de Beer JL, van Ingen J, de Vries G, Erkens C, Sebek M, Mulder A, Sloot R, van den Brandt AM, Enaimi M, Kremer K, Supply P, van Soolingen D.** 2013. Comparative study of IS*6110* restriction fragment length polymorphism and variable-number tandem-repeat typing of *Mycobacterium tuberculosis* isolates in The Netherlands, based on a 5-year nationwide survey. J Clin Microbiol **51:**1193–1198. http://dx.doi.org/10.1128/JCM.03061-12.
4. **Kato-Maeda M, Gagneux S, Flores LL, Kim EY, Small PM, Desmond EP, Hopewell PC.** 2011. Strain classification of *Mycobacterium tuberculosis*: congruence between large sequence polymorphisms and spoligotypes. Int J Tuberc Lung Dis **15:**131–133.
5. **Allix-Béguec C, Harmsen D, Weniger T, Supply P, Niemann S.** 2008. Evaluation and strategy for use of MIRU-VNTR*plus*, a multifunctional database for online analysis of genotyping data and phylogenetic identification of *Mycobacterium tuberculosis* complex isolates. J Clin Microbiol **46:**2692–2699. http://dx.doi.org/10.1128/JCM.00540-08.
6. **Kik SV, Verver S, van Soolingen D, de Haas PE, Cobelens FG, Kremer K, van Deutekom H, Borgdorff MW.** 2008. Tuberculosis outbreaks predicted by characteristics of first patients in a DNA fingerprint cluster. Am J Respir Crit Care Med **178:**96–104. http://dx.doi.org/10.1164/rccm.200708-1256OC.
7. **Verhagen LM, van den Hof S, van Deutekom H, Hermans PW, Kremer K, Borgdorff MW, van Soolingen D.** 2011. Mycobacterial factors relevant for transmission of tuberculosis. J Infect Dis **203:**1249–1255. http://dx.doi.org/10.1093/infdis/jir013.
8. **Nebenzahl-Guimaraes H, Borgdorff MW, Murray MB, van Soolingen D.** 2014. A novel approach: the propensity to propagate (PTP) method for controlling for host factors in studying the transmission of *Mycobacterium tuberculosis*. PLoS One **9:**e97816. http://dx.doi.org/10.1371/journal.pone.0097816.
9. **de Jong B, Hill PC, Aiken A, Awine T, Awine T, Antonio M, Adetifa IM, Jackson-Sillah DJ, Fox A, Deriemer K, Gagneux S, Borgdorff MW, McAdam KP, Corrah T, Small PM, Adegbola RA.** 2008. Progression to active tuberculosis, but not transmission, varies by *M. tuberculosis* lineage in The Gambia. J Infect Dis **198:**1037–1043. http://dx.doi.org/10.1086/591504.
10. **Rhee JT, Piatek AS, Small PM, Harris LM, Chaparro SV, Kramer FR, Alland D.** 1999. Molecular epidemiologic evaluation of transmissibility and virulence of *Mycobacterium tuberculosis*. J Clin Microbiol **37:**1764–1770.
11. **van Embden JD, Cave MD, Crawford JT, Dale JW, Eisenach KD, Gicquel B, Hermans P, Martin C, McAdam R, Shinnick TM.** 1993. Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology. J Clin Microbiol **31:**406–409.
12. **Supply P, Allix C, Lesejean S, Cardoso-Oelemann M, Rüsch-Gerdes S, Willery E, Savine E, de Haas P, van Deutekom H, Roring S, Bifani P, Kurepina N, Kreiswirth B, Sola C, Rastogi N, Vatin V, Gutierrez MC, Fauville M, Niemann S, Skuce R, Kremer K, Locht C, van Soolingen D.** 2006. Proposal for standardization of optimized mycobacterial interspersed repetitive unit-variable-number tandem repeat typing of *Mycobacterium tuberculosis*. J Clin Microbiol **44:**4498–4510. http://dx.doi.org/10.1128/JCM.01392-06.
13. **van Soolingen D, Borgdorff MW, de Haas PEW, Sebek MM, Veen J, Dessens M, Kremer K, van Embden JD.** 1999. Molecular epidemiology of tuberculosis in The Netherlands: a nationwide study from 1993 through 1997. J Infect Dis **180:**726–736. http://dx.doi.org/10.1086/314930.
14. **Weniger T, Krawczyk J, Supply P, Niemann S, Harmsen D.** 2010. MIRU-VNTR*plus*: a web tool for polyphasic genotyping of *Mycobacterium tuberculosis* complex bacteria. Nucleic Acids Res **38:**W326-3W31. http://dx.doi.org/10.1093/nar/gkq351.
15. **Jasmer RM, Hahn JA, Small PM, Daley CL, Behr MA, Moss AR, Creasman JM, Schecter GF, Paz EA, Hopewell PC.** 1999. A molecular epidemiologic analysis of tuberculosis trends in San Francisco, 1991 to 1997. Ann Intern Med **130:**971–978. http://dx.doi.org/10.7326/0003-4819-130-12-199906150-00004.
16. **Mulder C, van Deutekom H, Huisman EM, Meijer-Veldman W, Erkens CG, van Rest J, Borgdorff MW, van Leth F.** 2011. Coverage and yield of tuberculosis contact investigations in The Netherlands. Int J Tuberc Lung Dis **15:**1630–1637. http://dx.doi.org/10.5588/ijtld.11.0027.
17. **Gehre F, Otu J, DeRiemer K, de Sessions PF, Hibberd ML, Mulders W, Corrah T, de Jong BC, Antonio M.** 2014. Deciphering the growth behavior of *Mycobacterium africanum*. PLoS Negl Trop Dis **7:**e2220. http://dx.doi.org/10.1371/journal.pntd.0002220.
18. **Albanna AS, Reed MB, Kotar KV, Fallow A, McIntosh FA, Behr MA, Menzies D.** 2011. Reduced transmissibility of East African Indian strains of *Mycobacterium tuberculosis*. PLoS One **6:**e25075. http://dx.doi.org/10.1371/journal.pone.0025075.
19. **Buu TN, Huyen MN, Lan NT, Quy HT, Hen NV, Zignol M, Borgdorff MW, Cobelens FG, van Soolingen D.** 2009. The Beijing genotype is associated with young age and multidrug-resistant tuberculosis in rural Vietnam. Int J Tuberc Lung Dis **13:**900–906.
20. **Majoor CJ, Magis-Escurra C, van Ingen J, Boeree MJ, van Soolingen D.** 2011. Epidemiology of *Mycobacterium bovis* disease in humans, The Netherlands, 1993 to 2007. Emerg Infect Dis **17:**457–453. http://dx.doi.org/10.3201/eid1703.101111.
21. **Sunder S, Lanotte P, Godreuil S, Martin C, Boschiroli ML, Besnier JM.** 2009. Human-to-human transmission of tuberculosis caused by *Mycobacterium bovis* in immunocompetent patients. J Clin Microbiol **47:**1249–1251. http://dx.doi.org/10.1128/JCM.02042-08.
22. **Etchechoury I, Valencia GE, Morcillo N, Sequeira MD, Imperiale B, López M, Caimi K, Zumarraga MJ, Cataldi A, Romano MI.** 2010. Molecular typing of *Mycobacterium bovis* isolates in Argentina: first description of a person to person transmission case. Zoonoses Public Health **57:**375–381. http://dx.doi.org/10.1111/j.1863-2378.2009.01233.x.
23. **Yang C, Luo T, Sun G, Qiao K, Sun G, DeRiemer K, Mei J, Gao Q.** 2012. *Mycobacterium tuberculosis* Beijing strains favor transmission but not drug resistance in China. Clin Infect Dis **55:**1179–1187. http://dx.doi.org/10.1093/cid/cis670.
24. **Toungoussova OS, Mariandyshev A, Bjune G, Sandven P, Caugant DA.** 2003. Molecular epidemiology and drug resistance of *Mycobacterium tuberculosis* isolates in the Archangel prison in Russia: predominance of the W-Beijing clone family. Clin Infect Dis **37:**665–672. http://dx.doi.org/10.1086/377205.
25. **Langlois-Klassen D, Senthilselvan A, Chui L, Kunimoto D, Saunders LD, Menzies D, Long R.** 2013. Transmission of *Mycobacterium tuberculosis* Beijing strains, Alberta, Canada, 1991 to 2007. Emerg Infect Dis **19:**701–711. http://dx.doi.org/10.3201/eid1905.121578.
26. **Gagneux S, Small PM.** 2007. Global phylogeography of *Mycobacterium tuberculosis* and implications for tuberculosis product development. Lancet Infect Dis **7:**328–337. http://dx.doi.org/10.1016/S1473-3099(07)70108-1.
27. **Lohmann EM, Koster BF, le Cessie S, Kamst-van Agterveld MP, van Soolingen D, Arend SM.** 2012. Grading of a positive sputum smear and the risk of *Mycobacterium tuberculosis* transmission. Int J Tuberc Lung Dis **16:**1477–1484. http://dx.doi.org/10.5588/ijtld.12.0129.
28. **Borgdorff MW, Nagelkerke NJD, de Haas PEW, van Soolingen D.** 2001. Transmission of *Mycobacterium tuberculosis* depending on the age and sex of source cases. Am J Epidemiol **154:**934–943. http://dx.doi.org/10.1093/aje/154.10.934.
29. **Marais BJ, Gie RP, Schaaf HS, Hesseling AC, Obihara CC, Starke JJ, Enarson DA, Donald PR, Beyers N.** 2004. The natural history of childhood intrathoracic tuberculosis: a critical review of literature from the prechemotherapy era. Int J Tuberc Lung Dis **8:**392–402.
30. **Wallinga J, Teunis P, Kretzschmar M.** 2006. Using data on social contacts to estimate age-specific transmission parameters for respiratory-spread infectious agents. Am J Epidemiol **164:**936–944. http://dx.doi.org/10.1093/aje/kwj317.
31. **Lönnroth K, Williams BG, Stadlin S, Jaramillo E, Dye C.** 2008. Alcohol use as a risk factor for tuberculosis: a systematic review. BMC Public Health **8:**http://dx.doi.org/10.1186/1471-2458-8-289.
32. **Szabo G.** 1997. Alcohol's contribution to compromised immunity. 1997. Alcohol Health Res World **21:**30–38.
33. **Carballo-Diéguez A, Sahs J, Goetz R, el Sadr W, Sorell S, Gorman J.** 1994. The effect of methadone on immunological parameters among HIV-positive and HIV-negative drug users. Am J Drug Alcohol Abuse **20:**317–329. http://dx.doi.org/10.3109/00952999409106017.
34. **Alonzo NC, Bayer BM.** 2002. Opioids, immunology, and host defenses of intravenous drug abusers. Infect Dis Clin North Am **16:**553–569. http://dx.doi.org/10.1016/S0891-5520(02)00018-1.
35. **Portevin D, Gagneux S, Comas I, Young D.** 2011. Human Macrophage Responses to Clinical Isolates from the *Mycobacterium tuberculosis* Complex Discriminate between Ancient and Modern Lineages. PLoS Pathog **7:**e1001307. http://dx.doi.org/10.1371/journal.ppat.1001307.

Chapter IV

To be or not to be a pseudogene- a molecular epidemiological approach to the mclx genes and its impact in *Mycobacterium tuberculosis*

PLOS | ONE

RESEARCH ARTICLE

To Be or Not to Be a Pseudogene: A Molecular Epidemiological Approach to the *mclx* Genes and Its Impact in Tuberculosis

Catarina Lopes Santos[1,2☯]*, Hanna Nebenzahl-Guimaraes[1,2,3☯], Marta Vaz Mendes[4], Dick van Soolingen[3,5], Margarida Correia-Neves[1,2]

1 Life and Health Sciences Research Institute (ICVS), School of Health Sciences, University of Minho, Braga, Portugal, **2** ICVS/3B's, PT Government Associate Laboratory, Braga/Guimarães, Portugal, **3** National Institute for Public Health and the Environment (RIVM), Bilthoven, Netherlands, **4** IBMC—Instituto de Biologia Molecular e Celular, Universidade do Porto, Porto, Portugal, **5** Department of Medical Microbiology, Radboud University Nijmegen Medical Centre, Nijmegen, Netherlands

☯ These authors contributed equally to this work.
* catarinasantos@ecsaude.uminho.pt

OPEN ACCESS

Citation: Santos CL, Nebenzahl-Guimaraes H, Mendes MV, van Soolingen D, Correia-Neves M (2015) To Be or Not to Be a Pseudogene: A Molecular Epidemiological Approach to the *mclx* Genes and Its Impact in Tuberculosis. PLoS ONE 10(6): e0128983. doi:10.1371/journal.pone.0128983

Academic Editor: Igor Mokrousov, St. Petersburg Pasteur Institute, RUSSIAN FEDERATION

Received: January 23, 2015

Accepted: May 4, 2015

Published: June 2, 2015

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by Fundação para a Ciência e Tecnologia (FCT), Portugal, and cofunded by Programa Operacional Regional do Norte (ON.2—O Novo Norte), Quadro de Referência Estratégico Nacional (QREN), through the Fundo Europeu de Desenvolvimento Regional (FEDER), and from Projeto Estratégico – LA 26 – 2013–2014 (PEst-C/SAU/LA0026/2013). H.N.-G. received a personal FCT Grant (SFRH/BD/33902/2209). The funders had no role in study design, data collection

Abstract

Tuberculosis presents a myriad of symptoms, progression routes and propagation patterns not yet fully understood. Whereas for a long time research has focused solely on the patient immunity and overall susceptibility, it is nowadays widely accepted that the genetic diversity of its causative agent, *Mycobacterium tuberculosis*, plays a key role in this dynamic. This study focuses on a particular family of genes, the *mclx*s (*Mycobacterium* cyclase/LuxR-like genes), which codify for a particular and nearly mycobacterial-exclusive combination of protein domains. *mclx*s genes were found to be pseudogenized by frameshift-causing insertion (s)/deletion(s) in a considerable number of *M. tuberculosis* complex strains and clinical isolates. To discern the functional implications of the pseudogenization, we have analysed the pattern of frameshift-causing mutations in a group of *M. tuberculosis* isolates while taking into account their microbial-, patient- and disease-related traits. Our logistic regression-based analyses have revealed disparate effects associated with the transcriptional inactivation of two *mclx* genes. In fact, *mclx2* (Rv1358) pseudogenization appears to be primarily driven by the microbial phylogenetic background, being mainly related to the Euro-American (EAm) lineage; on the other hand, *mclx3* (Rv2488c) presents a higher tendency for pseudogenization among isolates from patients born on the Western Pacific area, and from isolates causing extra-pulmonary infections. These results contribute to the overall knowledge on the biology of *M. tuberculosis* infection, whereas at the same time launch the necessary basis for the functional assessment of these so far overlooked genes.

and analysis, decision to publish, or preparation of the manuscript.

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

Introduction

Tuberculosis (TB) is an air-borne contagious disease that remains responsible for high rates of morbidity and mortality worldwide: it is estimated that in 2013, nine million people fell ill with TB and 1.5 million died from it [1]. TB's rate of incidence is declining slowly (1.5% per year in average between 2000 and 2013) [1], which is somehow counter-intuitive given the financial effort put into research and prevention frameworks aiming towards its eradication. The reasons for this halting TB twilight, besides those related with health policies, are the still missing links in the understanding of the disease establishment on its latent or active form and TB transmission. TB has many different facets—from a life-lasting silent infection to an active and potentially deathly disease—and classically most of this variability has been attributed to the hosts' immune competence. However, and more recently, the role of the etiological agent—*Mycobacterium tuberculosis* (Mtb)—genotype has been gaining more relevance, as several studies came up demonstrating that small genetic variations in clinical isolates or laboratory strains have a significant impact not only in strict microbial characteristics, such as antibiotic resistance (for instance, [2]), but also in factors related with the host-microorganism relationship dynamics, such as disease progression and/or ability to modulate the host's immune response ([3] and references therein).

This work presented here is focused on a family of genes that is almost exclusive of *Mycobacterium* spp. and particularly abundant in members of the *M. tuberculosis* complex. Although they have been seldom addressed from a functional point of view, their structure suggests they might be involved in transcriptional regulation and response to quorum sensing (*sensu lato*) stimuli (i.e., communication within bacterial cells or between bacteria and their hosts) [4]. Its uniqueness relies on codifying for a particular combination of domains: an N-terminal CHD (cyclase homology domain) and a C-terminal LuxR HTH (helix turn helix) domain; depending on the domain identification algorithm, an AAA (ATPases associated with several cellular activities)/NB-ARC domain may also be identified between the other two [4–6]. Whereas the AAA/NB-ARC domain has the general function of binding and hydrolysing ATP and/or participating in the protein oligomerization ([5] and references therein), the CHD is the catalytic domain of the class III nucleotidyl (adenylyl/guanylyl) cyclases [5,6], and the LuxR is a DNA-binding domain mostly (but not exclusively) known to be associated with quorum sensing transcription modulation [4,7,8]. In the genome of the reference strain Mtb H37Rv, one can find three genes that codify for proteins with this particular domain composition: Rv0386, Rv1358 and Rv2488c [4–6]. For practical reasons, these genes will be referred to as *mclx1*, *mclx2* and *mclx3* (from *Mycobacterium* cyclase/LuxR-like genes), respectively, in the remainder of this manuscript.

Taking into account the overall lack of physiological and functional studies on Mclx proteins, they can either be viewed as putative cyclase proteins with additional non-cyclase domain(s) attached [5,6], or as putative transcriptional regulators from the LuxR family that may respond to, or be modulated by, ATP or cAMP [4]. Interestingly, the genome of the reference strain Mtb H37Rv codifies for 16 genes with a CHD domain, some of which are predicted to be transmembranar and others (including those from the Mclx family) predicted to be soluble [5,6]. Both their frequency and their unique diversity in terms of domains composition suggest that the cyclase activity may be a key point in Mtb fitness. In order to complete their function —bind ATP (or, less commonly, GTP) and convert it to the secondary signal cAMP (or cGMP)—nucleotidyl cyclases require a series of conserved residues that have been previously characterized and that are responsible for binding a divalent metal, for stabilizing the transition state species and for selecting and/or attaching the substrate (either ATP or GTP) [5,6]. Interestingly, only the Mclx1 has all the necessary residues for cyclase activity, as Mclx2 lacks one of

the residues necessary for the metal binding and a transition-stabilizing asparagine, Mclx3 lacks a transition-stabilizing asparagine. Moreover, both Mclx2 and Mclx3 seem to lack the substrate selectivity residues [5]. Additionally, the Mclx1 is the only family member that has been functionally characterized: not only was this protein found to have a significant (20%) guanylyl cyclase side activity, besides its adenylyl activity [9], but was also found to have a role in virulence [10]. In fact, Mclx1 was found to be required for a cAMP burst in macrophages upon infection that destabilizes the macrophage immune response: loss of Mclx1 resulted in a reduction in the production of tumor necrosis factor (TNF), and a decrease in the bacterial survival and in the immunopathology in the animal tissues [10].

The general association between the LuxR domain and quorum-sensing mechanisms, together with the lack of some canonical residues for the cyclase activity in the Mclx2 and Mclx3, suggest that these proteins may indeed be quorum-sensing-like transcriptional regulators that respond to or bind to ATP and/or cAMP in an allosteric modulatory fashion. A possible relation between cyclic nucleotides and quorum-sensing (*sensu lato*) is not new: a dynamic relationship between cyclic nucleotides as signals and quorum-sensing regulatory mechanisms has been observed before, either directly or through CRP (cAMP-binding proteins), in organisms such as *Vibrio vulnificus* [11], *Vibrio fischeri* [12], and *Vibrio cholerae* [13], among others.

The study presented here describes an integrative analysis combining genomics and epidemiology and is focused on the clinical consequences of *mclx* variation. Interestingly, we have found a scattered pattern of pseudogenization among *mlcx2* and *mclx3* genes, with implications at the level of patients' demographic characteristics and TB clinical manifestations. As so, this report establishes a link between the functionality of the proteins encoded by these two genes and the virulence-related fitness and host adaptation ability of the Mtb.

Materials and Methods

Screening and alignment of the *mclx* genes from public available genomes

The presence and transcriptional integrity of the *mclx* genes was analysed in a panel of Mtb complex strains and clinical isolates which genome had been completely sequenced. These organisms were selected from the Genome database of the National Center for Biotechnology Information (NCBI), limiting the search by organism—*Mycobacterium tuberculosis* complex (taxid: 77643)—and including only genomes with the status "Complete" or "Scaffolds or contigs". This search was performed on the 20th of March 2014 and yielded 187 organisms. To retrieve the *mclx* gene sequences from these organisms, each genome or set of scaffolds was used as a reference against which the *mclx* sequences from the reference strain Mtb H37Rv (Rv0386, Rv1358 and Rv2488c) were mapped, using the assembling tools from Geneious R7.1.4 (Biomatters) [14]. Instances when the putative *mclx* orthologue spanned more than one scaffold/contig lead to the elimination of the respective organism from the analyses, as to avoid sequencing misreads potentially attributed to scaffolds/contigs junctions. One hundred and fifty strains/clinical isolates were retained for further analyses: two of *Mycobacterium africanum*, 12 of *Mycobacterium bovis*, eight of *Mycobacterium cannetti*, and 128 of Mtb (Table 1). The *mclx* genes were identified and individually aligned against their reference orthologue from Mtb H37Rv to identify nucleotide substitutions, insertions and deletions, using the ClustalW algorithm [15] available in the Geneious software [14] (S1A Fig).

For most Mtb strains/clinical isolates, the information on the lineage could be retrieved from the literature and/or information on the genome. For the few cases in which this information could not be found, each genome or set of scaffolds was used as a reference against which the regions of difference (RD) from the reference strain Mtb H37Rv, described by Gagneux

Table 1. Pairwise identity and indels in the *mclx* genes in a panel of 150 Mtb complex strains/clinical isolates[1].

Organism	Lineage	*mclx* 1	*mclx* 2	*mclx* 3
Maf GM041182	WA-2	**99.97%—del -1406 (T)**	99.90%	99.90%
Maf K85	WA-2	**99.90%—del -1406 (T); del—2481 C)**	99.90%	99.97%
Mbv 04–303	-	99.97%	99.90%	100.00%
Mbv AF2122/97	-	99.90%	99.90%	100.00%
Mbv AN5	-	100.00%	99.90%	100.00%
Mbv BCG ATCC 35733	-	100.00%	99.90%	99.97%
Mbv BCG ATCC 35740	-	99.90%	99.90%	99.97%
Mbv BCG China	-	100.00%	99.90%	**99.80%—ins—2198–2202 (GGCGG)**
Mbv BCG Frappier	-	100.00%	99.90%	99.97%
Mbv BCG Korea 1168P	-	100.00%	99.90%	99.97%
Mbv BCG Mexico	-	100.00%	99.90%	99.97%
Mbv BCG Moreau	-	100.00%	99.90%	99.97%
Mbv BCG Pasteur 1173P2	-	100.00%	99.90%	99.90%
Mbv BCG Tokyo 172	-	100.00%	99.90%	99.97%
Mcn CIPT 140010059	-	99.90%	99.90%	93.40%—ins—2519–2521 (CCA)
Mcn CIPT 140060008	-	99.80%	99.60%	93.50%—ins—2521–2523 (ACC)
Mcn CIPT 140070002	-	99.90%	99.90%	98.90%
Mcn CIPT 140070005	-	99.70%	99.60%	98.60%
Mcn CIPT 140070007	-	99.30%	99.50%	99.00%
Mcn CIPT 140070008	-	99.70%	99.90%	99.10%
Mcn CIPT 140070013	-	99.30%	98.50%	98.80%
Mcn CIPT 140070017	-	99.00%	99.20%	97.00%—ins—2531–2533 (GCC)
Mtb '98-R604 INH-RIF-EM'	EAm	100.00%	**99.90%—ins—840 (T)**	99.97%—del—2716 (A)
Mtb 02_1987	EAs	99.90%	99.97%	100.00%
Mtb 1034	EAs	99.97%	99.97%	100.00%
Mtb 210	EAs	99.97%	99.97%	99.97%
Mtb 43–16836	IO	100.00%	99.97%	**99.90%—ins—1393–1394 (TA)**
Mtb 7199–99	EAm	100.00%	99.90%	100.00%
Mtb BT1	EAs	99.97%	99.97%	99.97%
Mtb BT2	EAs	99.97%	99.97%	99.97%
Mtb BTB05-552	EAm	100.00%	99.97%	100.00%
Mtb BTB05-559	EAm	100.00%	99.97%	100.00%
Mtb C	EAm	99.50%—ins—3195 (T); ins—3211 (G); ins—3218 (G)	99.90%	99.90%—ins—2919 (G); ins—2992 (G)
Mtb CAS/NITR204	EAI	**98.90%- 15 del/ 1 ins**	99.20%- 13 del	**98.70%- 19 del/ 2 ins**
Mtb CCDC5079	EAs	99.97%	99.90%—del—2523–2525 (CGA)	99.97%
Mtb CCDC5180	EAs	99.97%	99.97%	99.97%
Mtb CDC1551	EAm	100.00%	99.96%	100.00%
Mtb CDC1551A		100.00%	99.97%	100.00%
Mtb CTRI-2	EAm	100.00%	**99.90%—ins—840 (T)**	100.00%
Mtb EAI5		100.00%	99.90%	**99.90%—del—1388–1392 (TTGCG)**
Mtb EAI5/NITR206	IO	99.90%	99.97%	100.00%
Mtb EAS054	IO	99.97%	99.97%	**99.80%—del—1388–1392 (TTGCG)**
Mtb F11	EAm	99.97%	**99.90%—ins—840 (T)**	100.00%
Mtb FJ05194	EAs	100.00%	99.90%	100.00%

(*Continued*)

Table 1. (*Continued*)

Organism	Lineage	*mclx* 1	*mclx* 2	*mclx* 3
Mtb GuangZ0019	EAm	100.00%	99.97%	100.00%
Mtb H37Ra	EAm	100.00%	100.00%	100.00%
Mtb H37RvCO	EAm	100.00%	100.00%	100.00%
Mtb HKBS1	EAs	99.97%	99.97%	99.97%
Mtb HM	EAm*	99.97%	**99.90%—ins—840 (T)**	99.00%—ins—1921–1953 (TG...TG)
Mtb HN878	EAs	99.97%	99.97%	99.97%
Mtb INS_MDR	EAm	100.00%	99.97%	100.00%
Mtb INS_SEN	EAm	100.00%	99.97%	100.00%
Mtb INS_XDR	EAm	100.00%	99.97%	100.00%
Mtb KZN 1435	EAm	100.00%	**99.90%—ins—840 (T)**	99.97%
Mtb KZN 4207	EAm	100.00%	**99.90%—ins—840 (T)**	99.97%
Mtb KZN 605	EAm	100.00%	**99.90%—ins—840 (T)**	99.97%
Mtb KZN R506	EAm	100.00%	**99.90%—ins—840 (T)**	99.97%
Mtb KZN V2475	EAm	100.00%	**99.90%—ins—840 (T)**	99.97%
Mtb MTB-489	?	99.97%	100.00%	99.97%
Mtb NA-A0008	?	**99.97%—del—871 (C)**	99.90%	**99.80%—del—1388–1392 (TTGCG); del—2044 (C)**
Mtb NA-A0009	?	**99.90%—del—871 (C); ins—2308–2310 (AG); del—2901 (C)**	99.97%	**99.80%—del—1388–1392 (TTGCG); del—2104 (G)**
Mtb NCGM2209	EAs	99.97%	99.97%	100.00%
Mtb OSDD071	EAI	100.00%	99.97%	99.90%
Mtb OSDD105	EAm	100.00%	99.90%	**99.97%—del—900 (C)**
Mtb OSDD493	EAm	**99.97%—del—352 (C)**	99.97%	100.00%
Mtb PanR0201	EAm	100.00%	**99.90%—ins—840 (T)**	100.00%
Mtb PanR0202	EAm	100.00%	99.97%	100.00%
Mtb PanR0205	EAm	100.00%	**99.90%—ins—840 (T)**	99.97%
Mtb PanR0206	EAm	100.00%	99.90%	100.00%
Mtb PanR0207	EAm	100.00%	**99.90%—ins—840 (T)**	99.97%
Mtb PanR0208	EAm	100.00%	99.97%	100.00%
Mtb PanR0209	EAm	100.00%	**99.90%—ins—840 (T)**	99.97%
Mtb PanR0304	EAm	100.00%	**99.90%—ins—840 (T)**	100.00%
Mtb PanR0305	EAm	100.00%	99.97%	100.00%
Mtb PanR0306	EAm	100.00%	**99.90%—ins—840 (T)**	99.97%
Mtb PanR0307	EAm	100.00%	**99.90%—ins—840 (T)**	99.97%
Mtb PanR0308	EAm	100.00%	**99.90%—ins—840 (T)**	100.00%
Mtb PanR0309	EAm	100.00%	99.97%	100.00%
Mtb PanR0313	EAm	100.00%	**99.90%—ins—840 (T)**	100.00%
Mtb PanR0314	EAm	100.00%	**99.90%—ins—840 (T)**	99.97%
Mtb PanR0315	EAm	100.00%	99.90%	100.00%
Mtb PanR0316	EAm	100.00%	99.97%	99.97%
Mtb PanR0317	EAm	100.00%	**99.90%—ins—840 (T)**	99.97%
Mtb PanR0401	EAm	100.00%	**99.90%—ins—840 (T)**	99.97%
Mtb PanR0402	EAm	100.00%	**99.10%—ins—840 (T); del—3466 (G)**	99.97%
Mtb PanR0403	EAm	100.00%	**99.90%—ins—840 (T)**	100.00%
Mtb PanR0404	EAm	100.00%	**99.90%—ins—840 (T)**	100.00%
Mtb PanR0405	EAm	100.00%	**99.90%—ins—840 (T)**	99.97%

(*Continued*)

Table 1. (*Continued*)

Organism	Lineage	*mclx* 1	*mclx* 2	*mclx* 3
Mtb PanR0409	EAm	100.00%	**99.90%—ins—840 (T)**	100.00%
Mtb PanR0410	EAm	100.00%	**99.90%—ins—840 (T)**	99.97%
Mtb PanR0411	EAm	100.00%	**99.90%—ins—840 (T)**	99.97%
Mtb PanR0412	EAm	100.00%	**99.90%—ins—840 (T)**	99.97%
Mtb PanR0501	EAm	100.00%	**99.90%—ins—840 (T)**	99.97%
Mtb PanR0503	EAm	100.00%	**99.90%—ins—840 (T)**	100.00%
Mtb PanR0505	EAm	100.00%	**99.90%—ins—840 (T)**	99.97%
Mtb PanR0602	EAm	100.00%	**96.50%- 5 del/ 4 ins**	100.00%
Mtb PanR0603	EAm	100.00%	**99.90%-ins—840 (T)**	99.97%
Mtb PanR0604	EAm	100.00%	**99.90%-ins—840 (T)**	99.97%
Mtb PanR0605	EAs	99.97%	99.97%	99.97%
Mtb PanR0606	EAs	99.97%	99.97%	99.97%
Mtb PanR0607	EAm	100.00%	**99.90%—ins—840 (T)**	99.97%
Mtb PanR0609	EAm	100.00%	**99.90%—ins—840 (T)**	99.97%
Mtb PanR0610	EAm	100.00%	**99.90%—ins—840 (T)**	100.00%
Mtb PanR0611	EAm	99.97%	**99.90%—ins—840 (T)**	99.97%
Mtb PanR0702	EAm	100.00%	**99.90%—ins—840 (T)**	99.97%
Mtb PanR0703	EAm	99.97%	**99.90%—ins—840 (T)**	99.97%
Mtb PanR0704	EAm	99.97%	**99.90%—ins—840 (T)**	99.97%
Mtb PanR0707	EAm	100.00%	**99.90%—ins—840 (T)**	99.97%
Mtb PanR0708	EAm	99.97%	**99.90%—ins—840 (T)**	99.97%
Mtb PanR0801	EAm	100.00%	99.90%	100.00%
Mtb PanR0802	EAm	100.00%	**99.90%—ins—840 (T)**	99.97%
Mtb PanR0803	EAm	100.00%	**99.90%—ins—840 (T)**	99.97%
Mtb PanR0804	EAm	100.00%	**99.90%—ins—840 (T)**	99.97%
Mtb PanR0805	EAm	99.97%	**99.90%—ins—840 (T)**	99.97%
Mtb PanR0902	EAm	100.00%	99.90%	100.00%
Mtb PanR0903	EAm	100.00%	**99.90%—ins—840 (T)**	100.00%
Mtb PanR0904	EAm	100.00%	**99.90%—ins—840 (T)**	99.97%
Mtb PanR0906	EAm	99.97%	**99.90%—ins—840 (T)**	99.97%
Mtb PanR0907	EAm	100.00%	99.90%	100.00%
Mtb PanR0909	EAm	100.00%	**99.90%—ins—840 (T)**	99.97%
Mtb PanR1005	EAm	100.00%	**99.90%—ins—840 (T)**	99.97%
Mtb PanR1006	EAm	100.00%	**99.90%—ins—840 (T)**	99.97%
Mtb PanR1007	EAm	99.97%	**99.90%—ins—840 (T)**	99.97%
Mtb PanR1101	EAm	100.00%	**99.90%—ins—840 (T)**	100.00%
Mtb PR05	?	100.00%	99.97%	**99.80%—del—1388–1392 (TTGCG)**
Mtb R1207	EAs	99.90%	99.97%	100.00%
Mtb RGTB327	EAm*	**99.90%—ins—454 (G); ins—684 (A)**	**99.90%—ins—840 (T); ins—1331 (C)**	**99.90%—ins—766 (G); ins—2844–2845 (GG)**
Mtb RGTB423	IO*	99.97%—ins—2821 (C)	**99.90%—ins—390 (A)**	**99.70%—ins—1121 (A); ins—2160–2162 (GCC); ins—3017 (G)**
Mtb S96-129	EAm	100.00%	99.97%	100.00%
Mtb Beijing/NITR203	EAs	99.90%	99.90%	99.97%
Mtb Erdman = ATCC 35801	EAm	100.00%	99.90%	100.00%
Mtb Haarlem	EAm	99.10%- 7 ins	99.90%	**99.90%—del—2264 (G);del—2362 (G)**

(*Continued*)

Table 1. (*Continued*)

Organism	Lineage	*mclx* 1	*mclx* 2	*mclx* 3
Mtb OSDD515	EAm	100.00%	99.97%	**99.97%—del—938 (T)**
Mtb SUMu001	?	**99.97%—del—952 (G)**	100.00%	100.00%
Mtb SUMu002	?	**99.97%—del—952 (G)**	99.97%	99.97%—ins—2487 (A)
Mtb SUMu003	?	100.00%	99.97%	100.00%
Mtb SUMu004	?	100.00%	99.97%	100.00%
Mtb SUMu005	?	100.00%	99.97%	100.00%
Mtb SUMu006	?	100.00%	99.97%	100.00%
Mtb SUMu008	?	**99.97%—del—952 (G)**	99.97%	100.00%
Mtb SUMu010	?	100.00%	100.00%	100.00%
Mtb SUMu011	?	100.00%	100.00%	100.00%
Mtb SUMu012	?	**99.90%—del—2254 (G); del—2962 (G)**	100.00%	100.00%
Mtb UM 1072388579	?	100.00%	99.90%	100.00%
Mtb UT205	EAm	100.00%	**99.90%—ins—840 (T)**	100.00%
Mtb W-148	EAs	99.97%	99.90%—del—2523–2525 (CGA)	99.97%
Mtb WX3	EAs	99.97%	99.97%	99.97%
Mtb X122	EAs	99.97%	99.90%—del—2523–2525 (CGA)	99.97%
Mtb XDR1219	EAs	99.97%	99.97%	99.97%
Mtb XDR1221	EAs	99.90%	99.97%	100.00%

[1]The pairwise identity refers to the % of conserved residues of each gene after aligning it with its orthologue from the reference strain *M. tuberculosis* H37Rv; the putative pseudogenes are highlighted in bold; indel ocurrences are described by their type (ins, insertion; del, deletion) and by their location (considering an alignment with the reference mclx genes from the Mtb H37Rv strain), except when the total number of isolated occurences exceeds 3, in which case only their frequency is indicated; Maf, *Mycobacterium africanum*; Mvb, *Mycobacterium bovis*; Mcn, *Mycobacterium cannetti*; Mtb, *Mycobacterium tuberculosis*; WA-2, West-African 2; EAm, Euro-American lineage; EAs, East-Asian lineage; IO, Indo-Oceanic; EAI, East-African Indian lineage

*, indicates that the information on the lineage was obtained by aligning the H37Rv RD with the genome/scaffolds of the respective organism.

doi:10.1371/journal.pone.0128983.t001

et al. [16] in the definition of the six phylogeographical lineages, were mapped, and the occurrence of the described long sequence polymorphisms was used for the definition of the lineage (marked with an "*" in Table 1). In a few cases the RDs fell in scaffolds/contigs junctions, and therefore the presence of polymorphisms could not be accurately determined, precluding the determination of the lineage (marked with a "?" in Table 1).

Screening and alignment of the *mclx* genes from clinical isolates of an epidemiologically characterized cohort

To get an insight into the epidemiological/clinical features that may be associated with the occurrence of the *mclx* pseudogenization, the *mclx* transcriptional integrity was analysed in a diversified panel of Mtb clinical isolates. This panel is composed of 140 organisms collected and isolated in the Netherlands from 1993 to 2011, which are fully characterized from an epidemiological and clinical point of view (S1 Table) [17]. Demographic and clinical information, provided by the Registration Committee of the Netherlands Tuberculosis Register (NTR) that approved this retrospective study, were linked to the isolates on the basis of gender, date of birth, year of diagnosis and postal code. For 100 of these clinical isolates, whole-genome sequencing was performed ([17] and Nebenzahl-Guimaraes *et al.*, unpublished data) and the single nucleotide polymorphisms and INDELs (insertions or deletions) were called against the

reference strain H37Rv using Breseq software (version 0.23) with a minimum depth of 15x [18] (S1B Fig). SNPs with low-quality evidence (i.e. possible mixed read alignment) or within 5 bp of an INDEL were discarded. The presence of INDELs within the *mclx* coding regions and their potential to disrupt the open reading frame was evaluated. For the remaining 40 clinical isolates, the *mclx* genes were PCR-amplified in 3 overlapping fragments, each of 1200 to 1400 base pairs (see S2 Table for information on the primers and PCR conditions), and the purified PCR products were sequenced (by GATC Biotech). Each gene fragment was amplified twice and all fragments were sequenced in both directions. The final sequences were mapped against and aligned with their orthologues from the Mtb H37Rv reference strain, using the Geneious R7.1.4 (Biomatters) [14] software (S1C Fig). All alignments were visually inspected and a conservative approach was applied: whenever the sequencing results failed to converge to an obvious consensus, the gene status (functional/pseudogene) was considered to be unknown. For that reason, a few clinical isolates were not considered in the final analyses (final $n = 127$).

Statistical analyses

To envisage the relationship between the clinical and epidemiological features of the TB infection and the status of the *mclx* genes, a separate logistic regression-based analysis was carried out for *mclx2* and *mclx3*. Firstly, simple binary logistic regressions were performed to identify the significant predictors of each *mclx* genes status. Afterwards, two multiple binary logistic regressions were performed considering only those considered to be significant predictor variables ($p<0.2$ in the univariate analysis). Particularly in the case of *mclx3*, a sequential binary regression model was performed considering three sets of variables (patient-related, microorganism-related and disease-related). All statistical analyses were performed using the IBM SPSS Statistics, version 22 (IBM).

Phylogenetic tree construction

The FASTA files of publicly available NCBI strains were downloaded and pair aligned with the H37Rv reference sequence (NC_000962.gbk) using MAUVE v2.4.0. Multiple sequence alignments (MSA) using 62 robust SNP markers that have been shown to construct high resolution and reproducible phylogenies [19] were then made for 100 of the clinical isolates and 100 of the publicly available NCBI strains. The MSAs were subsequently used to generate a parsimony-based tree using the DNA parsimony algorithm version 3.69 from the Phylip package.

Results

A number of *mclx* genes among the *M. tuberculosis* complex species are likely pseudogenized

The *mclx* genes were previously identified as a family of genes nearly exclusive to the *Mycobacterium* genus and particularly abundant among the members of Mtb complex [4]. As a distinctive feature, these genes encode proteins with a particular domain architecture, including a CHD (cyclase homology domain) and a LuxR HTH domain. In order to acquire a better understanding of the distribution of this particular group of genes among the Mtb complex, an alignment-based screening was performed against a diversified panel of organisms with their genome fully-sequenced, including elements from the species *M. africanum*, *M. bovis*, *M. cannetti* and Mtb. Interestingly, this strategy revealed that even though *mclx* genes have in general an overall high degree of sequence conservation, usually with more than 99% nucleotides identical to their orthologues from the reference strain Mtb H37Rv, in a number of them a few INDELs were present, which caused a disruption in the open reading frame leading to the

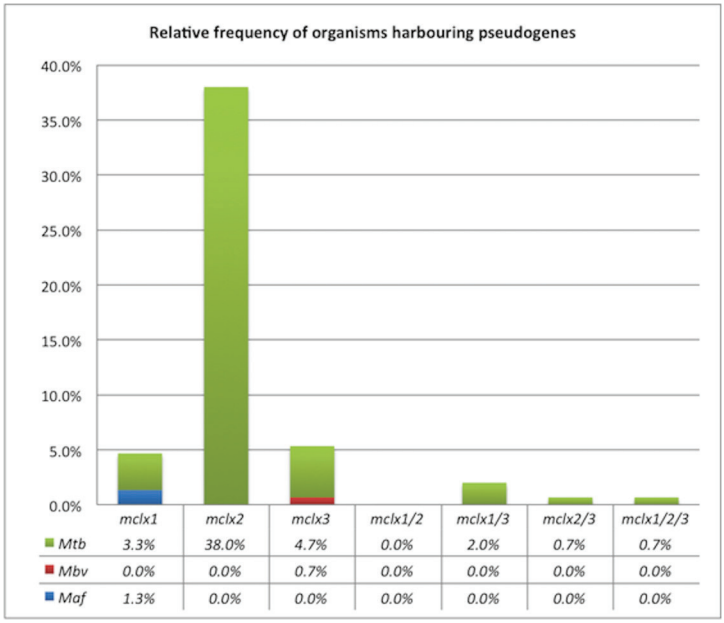


Fig 1. Relative frequency of the organisms harbouring pseudogenes in each of the *mclx* genes. The percentages are relative to the total number of strains/organisms listed in Table 1, and are stratified according to the three main species: Mtb (*Mycobacterium tuberculosis*), Mbv (*Mycobacterium bovis*) and Maf (*Mycobacterium africanum*).

doi:10.1371/journal.pone.0128983.g001

accumulation of stop codons and to the truncation of the respective sequence (Table 1). To avoid misinterpretations, and keeping in mind that gene size can vary to a certain extent without necessarily implying loss of function, the following criterion was established: any given *mclx* gene was considered to be a pseudogene whenever its sequence was truncated in more than 25% of the size of the corresponding Mtb H37Rv reference orthologue. While this criterion would require further functional validation, we consider it appropriate for this initial screening, as it likely minimizes the chance of false positives (i.e., genes that could be considered pseudogenes given a small size variation but that in fact retain their full functionality).

Overall, our analysis revealed that the pseudogenization is a rather common phenomenon among the *mclx* genes: 51.3% of the organisms in the analysed panel have at least one of their *mclx* genes truncated (Table 1 and Fig 1). As shown in Fig 1, the occurrence of the pseudogenes is not evenly distributed among the three different genes: the number of species with a *mclx2* pseudogene is much higher than that of species that suffered pseudogenization in the *mclx1* and/or the *mclx3*. In fact, whereas 7.3% and 8.7% of the organisms in the analysed panel have a pseudogenized *mclx1* and *mclx3*, respectively, 39.3% do so for the *mcx2*. Accordingly, the percentage of organisms with more than 1 *mclx* pseudogenized is relatively low (3.3%). Notwithstanding, it should be highlighted that in three (out of the four) organisms with two pseudogenes, the pseudogenization events occur in the *mclx1* and *mclx*3 (Table 1 and Fig 1), the two genes with lower pseudogenization occurrence, suggesting that the simultaneous loss of *mclx2* and any other *mclx* may be somehow harmful to the microorganisms.

Epidemiological assessment of the *mclx* pseudogenization

In order to evaluate whether there was a relationship between the inactivation of these genes in certain bacterial isolates and the epidemiological and clinical characteristics of the disease caused by those isolates, the presence and functionality of the *mclx* genes was analysed in a

panel of Mtb strains isolated in the Netherlands from 1993 to 2011 and fully characterized regarding their epidemiological and clinical features (S1 and S2 Tables). Even though all organisms in this panel were isolated in the Netherlands, only a minority of them (22.0%) were from native Dutch individuals. In fact, the patients' birth region is quite diversified: whereas 29.9% of the patients were born in Europe, 18.1% were born in Africa, 18.1% in the Eastern Mediterranean area, 13.4% in the South East Asia, 10.2% in the Western Pacific and 9.4% in the Americas. Accordingly, and given the strong phylogeographical nature of the Mtb lineages, this panel of isolates holds representatives of the EAm lineage (51.2%), Indo-Oceanic (IO) lineage (24.4%), East-African Indian (EAI) lineage (14.2%), and East-Asian (EAs) lineage (7.9%). From the classical risk factors commonly associated with active TB, the one that stands out in this panel is the origin from an endemic region. As for the other frequently-mentioned risk factors, their occurrence is rather low: only 3.1% of the strains were known to be isolated from homeless patients, 8.7% from known drug and/or alcohol users, and 14.2% from patients with co-morbidities (10.2% were HIV-positive, 3.1% had diabetes mellitus, 0.8% reported a malignancy and none were diagnosed with renal insufficiency or had been through organ transplantation). Regarding the TB localization, 64.6% of the isolates were retrieved from patients diagnosed with pulmonary TB, 18.9% from patients with extra-pulmonary TB, and 15.7% from patients reported to have both pulmonary and extra-pulmonary TB. Microbial transmissibility was defined following the work of Nebenzahl-Guimaraes *et al.* [17], and 61.4% of the clinical isolates analysed were considered "transmissible". In what concerns the transcriptional integrity of the *mclx* genes, no pseudogenes were found among *mclx1* following the criteria described in the material and methods, whereas 18.9% and 24.4% of the *mclx2* and *mclx3*, respectively, had suffered pseudogenization (S1 Table).

To identify the significant predictors of each *mclx2* and *mclx3* gene status, simple binary logistic regressions were performed (Table 2). Interestingly, the results were dissimilar for both genes, i.e., the independent variables for which the different categories presented odd ratios (ORs) for pseudogenization statistically different from the reference were not the same for *mclx2* and *mclx3*. The only variable that had a statistically significant association with the pseudogenization for both genes was “transmissibility” (Table 2). However, the effect of this variable had a different direction in each of the genes, i.e., non-transmissible isolates were around four times more likely to carry a pseudogenized copy of *mclx2*, but around two times more likely to carry a non-pseudogenized copy of *mclx3*.

The other variable revealing an association to *mclx2* was ethnicity—being a native Dutch represented a 4.5-fold increased risk of carrying a pseudogene (Table 2). The calculation of the ORs for the different Mtb lineages did not yield significant results, as most of the isolates with the *mclx2* pseudogenized belong to the EAm lineage (with a single exception—S1 Table), and for a number of lineages the number of pseudogenizations is null. However, isolates belonging to the EAI lineage did have a decreased risk for pseudogenization when compared to isolates belonging to the EAm lineage. Among the completely sequenced and publicly available Mtb complex genomes, there was also only one strain outside the EAm lineage that had a pseudogenized *mclx2* (Table 1).

Concerning the *mclx3* four other factors besides transmissibility were shown to have a significant association with its pseudogenization: gender, birth region, house setting and localization of the TB infection (Table 2). Being a female represented a 2.2-increased risk for having a pseudogenized form of *mclx*3, whereas living in an urban area represented a decrease of this risk to 0.345. On the other hand, isolates from strictly extra-pulmonary infections had an increased OR (more than five-fold higher) for *mclx3* pseudogenization when compared to strictly pulmonary strains. Finally, birth region was strongly associated with the *mclx3* gene status, with no pseudogenes identified in isolates from patients born in Africa, a decreased OR (0.886

Table 2. Univariate ORs for *mclx*2 and *mclx*3 pseudogenization (*p* values <0.2 are highlighted).

independent variables		*n*	univariate ORs (95% CI)	
			***mclx*2 pseudogenization**	***mclx*3 pseudogenization**
patient-related				
age		126	**0.973 (0.942–1.006); *p* = 0.103**	1.008 (0.983–1.033); *p* = 0.553
gender			*p* = 0.624	***p* = 0.061**
	female	47	1.255 (0.507–3.106)	**2.202 (0.965–5.024)**
	male	79	1 (ref)	**1 (ref)**
birth region			*p* = 0.589	***p* = 0.005**
	Africa	23	0.679 (0.216–2.137); *p* = 0.508	**0.000 (0.000-.); *p* = 0.998**
	The Americas	12	0.385 (0.073–2.022); *p* = 0.259	**0.886 (0.158–4.974); *p* = 0.890**
	Eastern Mediterranean	23	0.288 (0.072–1.154); *p* = 0.079	**0.664 (0.154–2.874); *p* = 0.584**
	Europe	38	1 (ref)	**1 (ref)**
	South East Asia	17	0.000 (0.000-.); *p* = 0.998	**3.100 (0.873–11.007); *p* = 0.080**
	Western Pacific	13	0.000 (0.000-.); *p* = 0.999	**53.143 (5.896–478.992); *p*<0.001**
ethnicity			***p* = 0.002**	*p* = 0.391
	native dutch	28	**4.583 (1.738–12.088)**	0.626 (0.215–1.823)
	foreign-born	97	**1 (ref)**	1 (ref)
house setting			*p* = 0.787	***p* = 0.033**
	rural	81	1 (ref)	**1 (ref)**
	urban	45	0.878 (0.343–2.248)	**0.345 (0.129–0.918)**
BCG vaccination			*p* = 0.668	*p* = 0.751
	no	27	1.306 (0.385–4.431)	0.842 (0.291–2.433)
	yes	39	1 (ref)	1 (ref)
co-morbidities			*p* = 0.371	*p* = 0.414
	no or unknown	109	1 (ref)	1 (ref)
	yes	18	0.494 (0.106–2.312)	0.579 (0.156–2.148)
alcohol or drug use			*p* = 0.949	*p* = 0.243
	no or unknown	116	1 (ref)	1 (ref)
	yes	11	0.949 (0.192–4.707)	0.287 (0.035–2.334)
homelessness			*p* = 0.753	*p* = 0.978
	no or unknown	123	1 (ref)	1 (ref)
	yes	4	1.499 (0.144–14.573)	1.033 (0.104–10.310)
microorganism-related				
lineage			*p* = 0.220	-
	EAI	18	0.107 (0.013–0.860); *p* = 0.036	-
	EAm	65	1 (ref)	-
	EAs	10	0.000 (0.000-.); *p* = 0.999	-
	IO	31	0.000 (0.000-.); *p* = 0.998	1
antibiotic resistance			*p* = 0.659	*p* = 0.979
	none or unknown	118	1 (ref)	1 (ref)
	resistant	8	1.455 (0.275–7.696)	1.023 (0.196–5.349)
transmissibility			***p* = 0.003**	***p* = 0.097**
	no	49	**4.242 (1.650–10.907)**	**0.467 (0.190–1.148)**
	yes	78	**1 (ref)**	**1 (ref)**
disease-related				
disease localization			*p* = 0.573	***p* = 0.002**
	pulmonary TB	82	1 (ref)	**1 (ref)**
	extra-pulmonary TB	24	0.589 (0.156–2.222); *p* = 0.435	**5.279 (1.983–14.049); *p* = 0.001**
	pulmonary and extra-pulmonary TB	20	1.375 (0.435–4.343); *p* = 0.587	**0.788 (0.205–3.038); *p* = 0.730**

doi:10.1371/journal.pone.0128983.t002

and 0.664) for pseudogenization in patients born in the Americas and Eastern Mediterranean, and an increased OR (3.100 and 53.143) for those born in South East Asia and Western Pacific, when compared to Europe (Table 2).

To detect confounding and/or mediation factors in the relation between the different variables with a significant association, multivariate binary logistic regression analyses were performed for the pseudogenization of *mclx2* and *mclx3* (Tables 3 and 4). These analyses included, for each case, all variables that presented a *p* below 0.200 in the univariate binary logistic regression (Table 2). These variables were organized into three different blocks (patient-related, microorganism-related and disease-related), which were sequentially added to each multivariate model.

For *mclx2*, a single model was built including the microorganism-related variable transmissibility, and the patient-related variables age and ethnicity (Table 3). Transmissibility was no longer significant upon correcting for age and ethnicity. In fact, in the multivariate model only ethnicity and age present significant associations: microorganisms isolated from native Dutch have an increased tendency to be carriers of *mclx2* pseudogenes, as do microorganisms isolated from younger people. However, it should be noticed that the *p* value for the microorganism lineage is close to the cut-off (0.200), and one of its categories actually has a *p* value of 0.036. Adding this variable to the multivariate model, both age and ethnicity lose its significance (S3 Table).

For *mclx3*, three different models were built: the first one using only the disease-related variable TB localization, the second one including the microorganism-related variable transmissibility, and the third one incorporating the patient-related variables gender, house setting and birth region (Table 4). As for transmissibility, gender and house setting, their associations to *mclx3* status were no longer significant after correcting for the other variables in the model. However, birth region remained as a significant variable, with patients born in the Western Pacific having an increased (77-fold) probability of carrying a pseudogenized form of this gene when compared to patients from Europe (Table 4). Since no pseudogenes were found among strains isolated from African patients, it was not possible to perform the mathematical computation of the OR and confidence interval for this category (and for the-Log likelihood of the

Table 3. Multivariate logistic regression model for *mclx*2 pseudogenization (*p* values < 0.05 are highlighted).

		multivariate ORs (95% CI)
***mclx*2**		**Model 1**
		patient-related
age		**0.952 (0.909–0.997); *p* = 0.038**; *B* = -0.049; *S.E.* = 0.024; *Wald* = 4.307
ethnicity	native dutch	**6.628 (1.708–25.714); *p* = 0.006**; *B* = 1.891; *S.E.* = 0.692; *Wald* = 7.475
	foreign-born	**1 (ref)**
		microorganism-related
transmissibility	no	1.918 (0.594–6.193); *p* = 0.276; *B* = 0.651; *S.E.* = 0.598; *Wald* = 1.186
	yes	1 (ref)
Omnibus Test (chi-square/*p*)		20.039/ *p*<0.001
Cox & Snell R^2		0.149
Nagelkerke R^2		0.242
Hosmer and Lemeshow (chi-square/*p*)		5.568/ *p* = 0.695
n		124

doi:10.1371/journal.pone.0128983.t003

Table 4. Multivariate logistic regression models for *mclx*3 pseudogenization (*p* values < 0.05 are highlighted).

		multivariate ORs (95% CI)		
***mclx*3**		**Model 1**	**Model 2**	**Model 3**
		patient-related		
gender	female	-	-	0.532(0.149–1.893); *p* = 0.330; *B* = -0.631; *S.E.* = 0.648; *Wald* = 0.950
	male	-	-	1 (ref)
house setting	rural	-	-	1 (ref)
	urban	-	-	0.433 (0.118–1.593); *p* = 0.208; *B* = -0.837; *S.E.* = 0.665; *Wald* = 1.586
birth region		-	-	***p* = 0.009**; *Wald* = 15.307
	Africa	-	-	**0.000 (0.000-.); *p* = 0.998**; *B* = -20.087; *S.E.* = 8006.116; *Wald* = 0.000
	The Americas	-	-	**0.929 (0.148–5.849); *p* = 0.938**; *B* = -0.074; *S.E.* = 0.939; *Wald* = 0.006
	Eastern Mediterranean	-	-	**0.433 (0.077–2.446); *p* = 0.344**; *B* = -0.837; *S.E.* = 0.883; *Wald* = 0.897
	Europe	-	-	**1 (ref)**
	South East Asia	-	-	**3.479 (0.657–18.431); *p* = 0.143**; *B* = 1.247; *S.E.* = 0.851; *Wald* = 2.149
	Western Pacific	-	-	**77.372 (6.357–941.774); *p* = 0.001**; *B* = 4.349; *S.E.* = 1.275; *Wald* = 11.631
		microorganism-related		
transmissibility	no	-	0.761 (0.280–2.070); *p* = 0.593; *B* = -0.273; *S.E.* = 0.510; *Wald* = 0.285	2.953 (0.648–13.458); *p* = 0.162; *B* = 1.083; *S.E.* = 0.774; *Wald* = 1.957
	yes	-	1 (ref)	1 (ref)
		disease-related		
disease localization		***p* = 0.002**; *Wald* = 12.466	***p* = 0.009**; *Wald* = 9.447	***p* = 0.025**; *Wald* = 7.342
	pulmonary TB	**1 (ref)**	**1 (ref)**	**1 (ref)**
	extra-pulmonary TB	**5.279 (1.983–14.049); *p* = 0.001**; *B* = 1.664; *S.E.* = 0.499; *Wald* = 11.097	**4.702 (1.630–13.560); *p* = 0.004;** *B* = 1.548; *S.E.* = 0.540; *Wald* = 8.206	**9.894 (1.825–53.654); *p* = 0.008**; *B* = 2.292; *S.E.* = 0.863; *Wald* = 7.060
	pulmonary and extra-pulmonary TB	**0.788 (0.205–3.038); *p* = 0.730**; *B* = -0.238; *S.E.* = 0.688; *Wald* = 0.120	**0.769 (0.199–2.978); *p* = 0.704**; *B* = -0.262; *S.E.* = 0.691; *Wald* = 0.144	**3.609 (0.637–20.434); *p* = 0.147**; *B* = 1.283; *S.E.* = 0.885; *Wald* = 2.105
Omnibus Test (chi-square/*p*)		12.555/ *p* = 0.002	12.843/ *p* = 0.005	56.253/ *p*<0.001
Cox & Snell R^2		0.095	0.097	0.360
Nagelkerke R^2		0.141	0.144	0.536
Hosmer and Lemeshow (chi-square/*p*)		0.000/ *p* = 1.000	0.081/ *p* = 0.994	6.225/ *p* = 0.514
n		126		

doi:10.1371/journal.pone.0128983.t004

model no final solution could be found). However, excluding African individuals from the sample had a negligible effect on the calculation of the parameters for the other categories/variables and on the overall significance of this model (data not shown). Finally, disease localization remains as a significant variable even after correcting for the microorganism- and

patient-related variables. Clinical isolates from strictly extra-pulmonary infections have a nearly 10-fold increased probability of carrying a pseudogenized form of *mclx3* when compared to isolates from strictly pulmonary forms of the disease (Model 3, Table 4). This increased propensity for *mclx3* pseudogenized forms is maintained for isolates from disseminated (pulmonary and extra-pulmonary) infections, although in a non-significant way (Model 3, Table 4).

To avoid phylogenetic redundancy, i.e., to ensure that the observed results were due to actual relations observed in the sample and not to the relative abundance of certain genotypes, the analyses were repeated using a single representative for each VNTR and RFLP type (the excluded isolates are annotated in the S1 Table). The results were similar to the previous ones in both the univariate (S4 Table) and multivariate (S5 and S6 Tables) analyses, supporting the initial deductions.

To gain some phylogenetic insight into the distribution of these pseudogenization events, a phylogenetic tree encompassing a number of genomes analysed in this article was constructed and the position of putative pseudogene-causing INDELS annotated (Fig 2). The frameshift-causing INDEL events in the different *mclx* genes are not unique in each organism, but are often found repeated across different strains/clinical isolates. On the other hand, although most pseudogenes in a given gene are concentrated in the same part of the tree, a few others appear scattered throughout the organisms, preventing a clear phylogenetic signal. Particularly, although most strains with pseudogenization events among *mclx* genes belong to Mtb, two *M. africanum* representatives have pseudogenized forms of the *mclx1* and there is one *M. bovis* BCG with a pseudogenized *mclx3* (Table 1, Figs 1 and 2). This somewhat dispersed distribution of pseudogenization events, together with the fact that different INDELs occur in different strains/clinical isolates but result in the pseudogenization of the same gene, suggest that the pseudogenization of each *mclx* may have occurred more than once in their phylogenetic history. This is consistent with a scenario where strong selective pressures are at the basis of these inactivation events, leading to the same overall result—the pseudogenization of a given *mclx*, although sometimes following different pathways (different INDELs), as opposed to a scenario of random evolution, where the pseudogenization of a given gene would have likely occurred once and dispersed throughout the lineage.

Discussion

The analysis of the genotypic variability of the *mclx* genes revealed a scattered pattern of pseudogenization among the Mtb complex strains and clinical Mtb isolates. The occurrence of INDELs in the different *mclx* genes was not homogenous between the two panels explored, the most striking difference being that *mclx1* was pseudogenized in 7.3% of the publicly available strains compared to none in the dataset of clinical isolates. The absence of West-African 2 representatives in the latter may have contributed to this, as the pseudogenization of *mclx1* is particularly common in this lineage [20] and quite rare amongst others. Taking into account the role played by its codified protein in the macrophages' initial immune response [10], one can argue that its inactivation may be one significant aspect in *M. africanum* (West-African 2 members) virulence attenuation. The percentages of the pseudogenized *mclx2* and *mclx3* differed between both panels, the former being more common amongst the publicly available strains and the latter more so in the clinical isolate dataset. These differences are likely related to the degree of clustering in both panels. Whereas the clustering is limited among the studied clinical isolates, and could actually be controlled for without an impact in the main results, one cannot access such information regarding the genome-sequenced and publicly available strains. The fact that a few of them are laboratory strains, and a number of others have been likely isolated from the same given TB outbreak (such as the SUMu or the PanR collections,

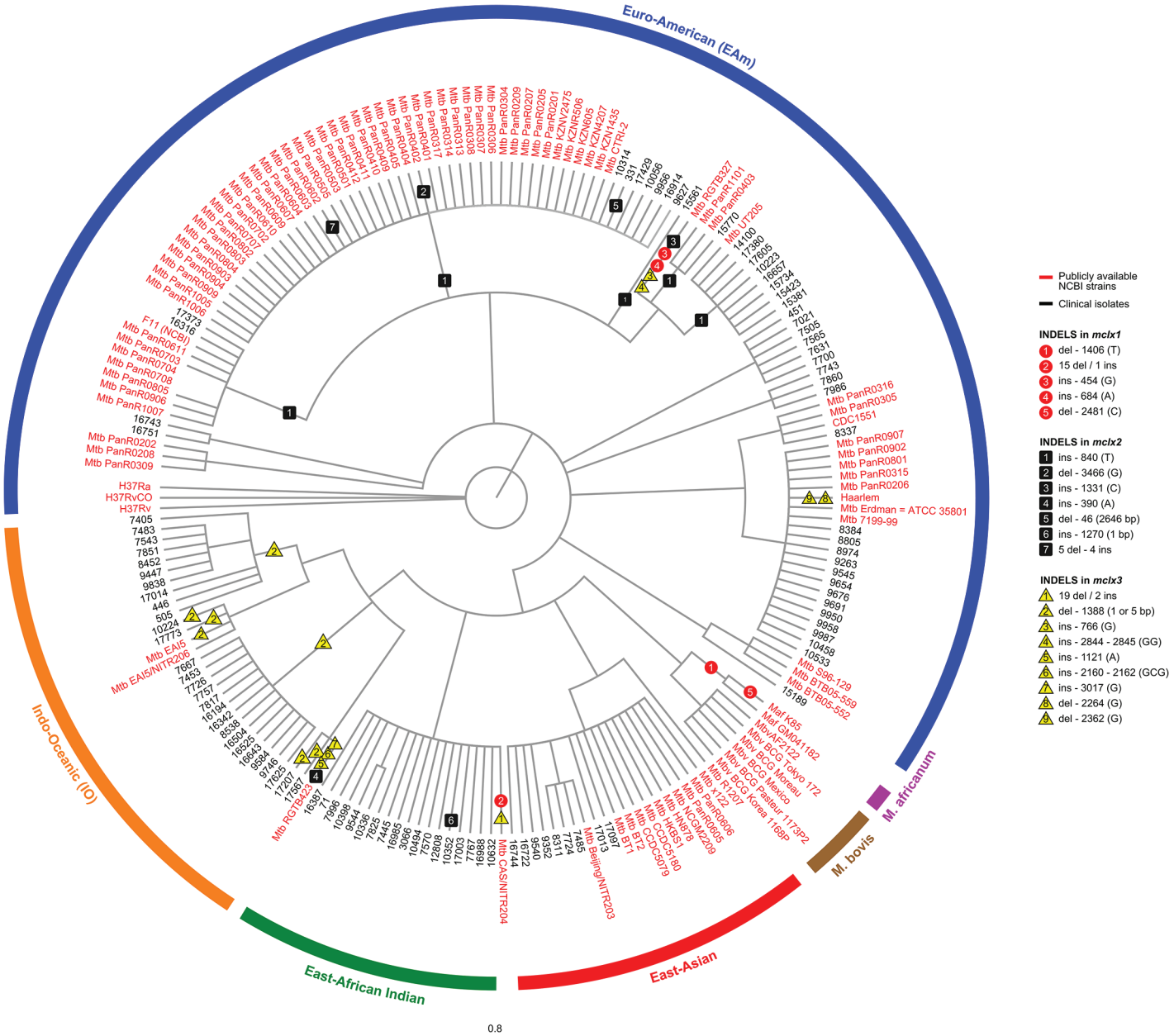


Fig 2. A parsimony-based phylogenetic tree depicting the distribution of pseudogenization events across the 3 *mclx* genes in both epidemiologically characterized clinical isolates (n = 100; denoted in black) and publicly available NCBI strains (n = 100; denoted in red). LSP-defined MTBC lineages/sublineages are colour-coded and indicated in the outer arc (Purple – *Mycobacterium africanum*; Brown—*Mycobacterium bovis*; blue—Euro-American lineage; red—East-Asian lineage; orange—Indo-Oceanic; green—East-African Indian lineage).

doi:10.1371/journal.pone.0128983.g002

representing 6.7% and 39.3% of this panel, respectively), hints at the existence of higher genetic relatedness. In this context, certain genetic features may appear more common solely because they occur in an overrepresented genotype.

The existence of three *mclx* genes in each genome raises the hypothesis of functional redundancy among them. However, that would likely have as consequence a random pattern of inactivation, resulting in similar pseudogenization ratios for each gene, which is not supported by the data. Moreover, the disparate results in the logistic regression analysis in terms of

significant variables buttresses the hypothesis that selective pressures and/or the clinical consequences of inactivating *mclx2* or *mclx3* are dissimilar. This is in agreement with the previously published functional analysis of the *mclx1*—Agarwall *et al*. mutated other adenylyl cyclases besides *mclx1*, namely *mclx3*, and for none of them significant effects were noticed in survival upon competing at mouse lungs, nor differences in macrophage cAMP levels compared to the wild-type, as it was for the *mclx*1, strongly suggesting that they have different functions [10].

The univariate and multivariate binary logistic regression analyses uncovered a number of statistically significant relationships that highlight the potential impact of the *mclx* genes functionality in aspects related with the microorganism biology and fitness, TB development and patients' demography. In *mclx2*, the main factor associated with pseudogenization appears to be the microorganism lineage, with all but one organism carrying a pseudogenized gene belonging to the EAm lineage among the panel of strains used for this analysis. Accordingly, in the publicly available genomes there is only one pseudogenization of *mclx2* outside the EAm lineage. Age and ethnicity might be significant factors as well, with the multivariate model showing that native Dutch patients present an approximately 6.6-fold higher risk of carrying a pseudogenized *mclx2*, and younger patients having a decreased OR for these forms. Although this could suggest some degree of adaptation it is also true that 78.6% of the native Dutch isolates actually belong to the EAm lineage, and both age and ethnicity lose significance in the multivariate model controling for lineage. As such, even if age and ethnicity do play a role in the pseudogenization of *mclx*2, this should be a rather limited one.

For *mclx3*, the results were quite dissimilar from those of *mclx2*. Significant variables in the multivariate model for *mclx3* include patient birth region and disease localization. Given the strong phylogeographical structure of Mtb lineages, the relation between the patients' birth region and *mclx3* pseudogenization can either be interpreted as a reflection of a phylogenetic signal (the effect of the microorganisms' lineage by itself cannot be evaluated for this gene, as all pseudogenes are found among strains belonging to the IO lineage in the analysed panel) or due to differences in the individuals from different regions (either genetic or socially/culturally- implemented). The relationship with disease localization remains consistently significant after correcting for all the other variables with a statistically significant signal in the univariate analysis. This suggests that *mclx3* plays a key role in the establishment of a pulmonary infection—and therefore its absence causes an adjustment of the infection to the extra-pulmonary space—or, conversely, that Mclx3 function prevents the infection from spreading.

Several risk factors for extra-pulmonary forms of TB have been addressed and characterized previously, both in what comes to the microbial influence [21–23], and also regarding host factors [24]. Concerning genetic microbial features, large INDEL polymorphisms in a phospholipase C gene, *plcD*, have been significantly associated with extra-pulmonary forms of TB when compared to strains without a *plcD* interruption [22]. On the other hand, the study of the same kind of mutations occurring in other genes from the same family, *plcA*, *plcB* and *plcC*, failed to show such a correlation [23]. This present study parallels this: *mclx3* is strongly and significantly associated with extra-pulmonary TB, but such is not the case for its paralogue *mclx2*. In what concerns host features, gender, ethnicity and HIV-status have all been found to be significant risk factors for extra-pulmonary forms of TB [24]. Whereas in this study we could not access the patients' ethnicity (concerning race/skin colour), both gender and birth region (a possible proxy) were accounted for in the multivariate model. As a precaution, the data was re-analysed integrating the *mclx3* pseudogenization status as a putative risk factor for extra-pulmonary infections (as opposed to strictly pulmonary and/or disseminated infections) and correcting for all host factors previously associated with this form of the disease: age, gender, HIV serological status, birth region and ethnicity. In accordance the pseudogenization of *mclx3* appears as an independent and highly significant risk factor for extra-pulmonary TB (S7 Table). Finally,

another microbial factor that should be taken into consideration is the microbial lineage. A previous report has demonstrated that, compared to the EAs lineage, the EAm, IO and the EAI lineages are significantly associated with extra-pulmonary forms, even after correcting for relevant host factors [21]. This is particularly important in the context of this study, as almost all *mclx3* pseudogenes are found among members of the IO lineage (and actually all of them in the panel used for the regression analysis) and therefore could suggest that the *mclx3* pseudogenization is a mere phylogenetic signal. Since the occurrence of *mclx3* pseudogenes in the analysed sample is restricted to IO strains, it is not possible to correct for the lineage in the multivariate models. However, previous reports support that the EAI lineage represents a fairly similar risk for extra-pulmonary infections as the IO one [21]. Notwithstanding, in the analysed sample and among the EAI lineage there is only one case (5.6%) of strictly extra-pulmonary infection, a value that deviates significantly from the 13 cases (41.9%) observed for the IO lineage (S8 Table). Although it is not possible to completely disregard the phylogenetic hypothesis, our results suggest that the *mclx*3 pseudogenization is one of the factors that favor extra-pulmonary forms of TB, and its prevalence among the IO lineage justifies that same tendency in this lineage. Conversely, the lack of *mclx3* pseudogenes among the studied EAI could help to justify the missing tendency for extra-pulmonary infections in this particular sample.

In the context of this manuscript, pseudogene is referred to as any gene whose coding sequence has been abruptly terminated by a large or small INDEL event, leading to the accumulation of stop codons and the precocious ending of the putatively codified peptide. It does not, by any means, reflect a status of overall non-functionality. Pseudogenes can have a number of different functions in the cell. In this context, it is important to highlight the *Mycobacterium leprae*. Although *M*. *leprae* holds a large collection of pseudogenes in its genome (approximately 50%), an interestingly high number of them are actually expressed (43%), and some even vary their expression patterns upon infection or in different leprosy patients, suggesting that they can play a role in the virulence of this microorganism [25–28]. More often than not, this expression occurs from pseudogenes that have stop codons in their reading frames, as is the case of the *mclx* addressed in this study. Therefore, the *mclx* pseudogenes should be regarded as potentially functional genes, although codifying smaller proteins/peptides than their orthologues or displaying non-codifying functions. Supporting this hypothesis, one study has previously referred an over-expression of the *mclx2* after the induction of an alternative sigma factor (*sigF*) [29]. Interestingly, the SigF binding site was located within the *mclx2* coding region, resulting in a 250 residues-shorter protein. This suggests that shorter versions of at least this *mclx* may hold an important role under defined conditions.

Gene pseudogenization has been often associated with the absence of purifying selective pressures, which allow the accumulation of nucleotide substitutions and INDELs. However, in this case, the high degree of sequence conservation at a nucleotide level suggests otherwise. Frameshift-causing INDELs are sometimes the only difference in the sequences when compared with their orthologues from the reference genome Mtb H37Rv. Conversely, the *mclx* genes from the closely related *M*. *cannetti* hold a much higher degree of sequence divergence but are not pseudogenized. This suggests that the pseudogenization of the *mclx* genes is either recent and/or the result of defined selective pressures, as opposed to a longer process of genome erosion in the absence of selection. This work, by describing a family of genes selectively pseudogenized in certain isolates, reinforces the recent trend to complement immunological data with the study of bacterial evolution in order to fully understand—and control—TB.

Supporting Information

S1 Fig. A. Snapshot of a ClustalW alignment in the Geneious software calling a one base-pair insertion in codon 279 of the *mclx2* gene in a publicly available NCBI strain. B. Snapshot of the breseq software calling the same insertion in an epidemiologically characterized clinical isolate whose genome has been fully sequenced. Displayed are color-coded Illumina sequencing reads mapping to the H37Rv reference sequence (singled out at the top). **C.** Snapshot of a ClustalW alignment in the Geneious software calling the same insertion in an epidemiologically characterized clinical isolate in which only the *mclx* genes have been sequenced.
(PDF)

S1 Table. Clinical and epidemiological characteristics of the 127 Mtb isolates analysed by binary logistic regression.
(PDF)

S2 Table. PCR primers and conditions for *mclx* amplification.
(PDF)

S3 Table. Multivariate ORs for *mclx*2 including the "microorganism lineage" as a variable.
(PDF)

S4 Table. Univariate ORs for *mclx*2 and *mclx*3 pseudogenization (clinical isolates with unique RFLP and VNTR).
(PDF)

S5 Table. Multivariate logistic regression models for *mclx*2 pseudogenization (clinical isolates with unique RFLP and VNTR).
(PDF)

S6 Table. Multivariate logistic regression models for *mclx*3 pseudogenization (clinical isolates with unique RFLP and VNTR).
(PDF)

S7 Table. Multivariate logistic regression models for extra-pulmonary TB infections.
(PDF)

S8 Table. Crosstabulation of Mtb lineage *vs*. local of infection.
(PDF)

Acknowledgments

The authors are grateful to the staff of the RIVM mycobacteriological laboratory for their work on the RFLP and VNTR typing of *M. tuberculosis* isolates.

Author Contributions

Conceived and designed the experiments: CLS HN-G MVM MC-N. Performed the experiments: CLS HN-G. Analyzed the data: CLS HN-G DvS MC-N. Contributed reagents/materials/analysis tools: DvS MVM MC-N. Wrote the paper: CLS HN-G.

References

1. Global Tuberculosis Report 2014: World Health Organization.
2. Zhang H, Li D, Zhao L, Fleming J, Lin N, Wang T, et al. (2013) Genome sequencing of 161 *Mycobacterium tuberculosis* isolates from China identifies genes and intergenic regions associated with drug resistance. Nat Genet 45: 1255–1260. doi: 10.1038/ng.2735 PMID: 23995137

3. Coscolla M, Gagneux S (2010) Does *M. tuberculosis* genomic diversity explain disease diversity? Drug Discov Today Dis Mech 7: e43–e59. PMID: 21076640
4. Santos CL, Correia-Neves M, Moradas-Ferreira P, Mendes MV (2012) A walk into the LuxR regulators of Actinobacteria: phylogenomic distribution and functional diversity. PLoS One 7: e46758. doi: 10.1371/journal.pone.0046758 PMID: 23056438
5. Shenoy AR, Sivakumar K, Krupa A, Srinivasan N, Visweswariah SS (2004) A survey of nucleotide cyclases in actinobacteria: unique domain organization and expansion of the class III cyclase family in *Mycobacterium tuberculosis*. Comp Funct Genomics 5: 17–38. doi: 10.1002/cfg.349 PMID: 18629044
6. Shenoy AR, Visweswariah SS (2006) Mycobacterial adenylyl cyclases: Biochemical diversity and structural plasticity. FEBS letters 580: 3344–3352. PMID: 16730005
7. Fuqua C, Winans SC, Greenberg EP (1996) Census and consensus in bacterial ecosystems: the LuxR-LuxI family of quorum-sensing transcriptional regulators. Annu Rev Microbiol 50: 727–751. PMID: 8905097
8. Patankar AV, Gonzalez JE (2009) Orphan LuxR regulators of quorum sensing. FEMS Microbiol Rev 33: 739–756. doi: 10.1111/j.1574-6976.2009.00163.x PMID: 19222586
9. Castro LI, Hermsen C, Schultz JE, Linder JU (2005) Adenylyl cyclase Rv0386 from *Mycobacterium tuberculosis* H37Rv uses a novel mode for substrate selection. FEBS Journal 272: 3085–3092. PMID: 15955067
10. Agarwal N, Lamichhane G, Gupta R, Nolan S, Bishai WR (2009) Cyclic AMP intoxication of macrophages by a *Mycobacterium tuberculosis* adenylate cyclase. Nature 460: 98–102. doi: 10.1038/nature08123 PMID: 19516256
11. Kim SP, Kim CM, Shin SH (2012) Cyclic AMP and cyclic AMP-receptor protein modulate the autoinducer-2-mediated quorum sensing system in *Vibrio vulnificus*. Curr Microbiol 65: 701–710. doi: 10.1007/s00284-012-0218-0 PMID: 22961036
12. Lyell NL, Colton DM, Bose JL, Tumen-Velasquez MP, Kimbrough JH, Stabb EV (2013) Cyclic AMP receptor protein regulates pheromone-mediated bioluminescence at multiple levels in *Vibrio fischeri* ES114. J Bacteriol 195: 5051–5063. doi: 10.1128/JB.00751-13 PMID: 23995643
13. Krasteva PV, Fong JC, Shikuma NJ, Beyhan S, Navarro MV, Yildiz FH, et al. (2010) *Vibrio cholerae* VpsT regulates matrix production and motility by directly sensing cyclic di-GMP. Science 327: 866–868. doi: 10.1126/science.1181185 PMID: 20150502
14. Drummond AJ, Ashton B, Buxton S, Cheung M, Cooper A, Duran C, et al. (2011) Geneious v5.4, Available from http://www.geneious.com/.
15. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, et al. (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23: 2947–2948. PMID: 17846036
16. Gagneux S, DeRiemer K, Van T, Kato-Maeda M, de Jong B, Narayanan S, et al. (2006) Variable host-pathogen compatibility in *Mycobacterium tuberculosis*. Proc Natl Acad Sci USA 103: 2869–2873. PMID: 16477032
17. Nebenzahl-Guimaraes H, Borgdorff MW, Murray MB, van Soolingen D (2014) A novel approach—the propensity to propagate (PTP) method for controlling for host factors in studying the transmission of *Mycobacterium tuberculosis*. PLoS One 9: e97816. doi: 10.1371/journal.pone.0097816 PMID: 24849817
18. Barrick JE, Yu DS, Yoon SH, Jeong H, Oh TK, Schneider D, et al. (2009) Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. Nature 461: 1243–1247. doi: 10.1038/nature08480 PMID: 19838166
19. Coll F, McNerney R, Guerra-Assuncao JA, Glynn JR, Perdigao J, Viveiros M, et al. (2014) A robust SNP barcode for typing *Mycobacterium tuberculosis* complex strains. Nature Communcations 5, Article number: 4812.
20. Bentley SD, Comas I, Bryant JM, Walker D, Smith NH, Harris SR, et al. (2012) The genome of *Mycobacterium africanum* West African 2 reveals a lineage-specific locus and genome erosion common to the *M. tuberculosis* complex. PLoS Negl Trop Dis 6: e1552. doi: 10.1371/journal.pntd.0001552 PMID: 22389744
21. Click ES, Moonan PK, Winston CA, Cowan LS, Oeltmann JE (2012) Relationship between *Mycobacterium tuberculosis* phylogenetic lineage and clinical site of tuberculosis. Clin Infect Dis 54: 211–219. doi: 10.1093/cid/cir788 PMID: 22198989
22. Yang Z, Yang D, Kong Y, Zhang L, Marrs CF, Foxman B, et al. (2005) Clinical relevance of *Mycobacterium tuberculosis plcD* gene mutations. Am J Respir Crit Care Med 171: 1436–1442. PMID: 15805187
23. Kong Y, Cave MD, Yang D, Zhang L, Marrs CF, Foxman B, et al. (2005) Distribution of insertion- and deletion-associated genetic polymorphisms among four *Mycobacterium tuberculosis* phospholipase C genes and associations with extrathoracic tuberculosis: a population-based study. J Clin Microbiol 43: 6048–6053. PMID: 16333097

24. Yang Z, Kong Y, Wilson F, Foxman B, Fowler AH, Marrs CF, et al. (2004) Identification of risk factors for extrapulmonary tuberculosis. Clin Infect Dis 38: 199–205. PMID: 14699451

25. Nakamura K, Akama T, Bang PD, Sekimura S, Tanigawa K, Wu H, et al. (2009) Detection of RNA expression from pseudogenes and non-coding genomic regions of *Mycobacterium leprae*. Microb Pathog 47: 183–187. doi: 10.1016/j.micpath.2009.06.006 PMID: 19555754

26. Singh P, Cole ST (2011) *Mycobacterium leprae*: genes, pseudogenes and genetic diversity. Future Microbiol 6: 57–71. doi: 10.2217/fmb.10.153 PMID: 21162636

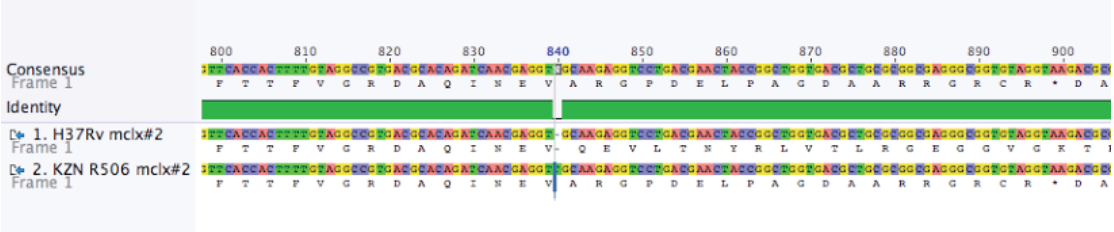
27. Suzuki K, Nakata N, Bang PD, Ishii N, Makino M (2006) High-level expression of pseudogenes in *Mycobacterium leprae*. FEMS Microbiol Lett 259: 208–214. PMID: 16734781

28. Williams DL, Slayden RA, Amin A, Martinez AN, Pittman TL, Mira A, et al. (2009) Implications of high level pseudogene transcription in *Mycobacterium leprae*. BMC Genomics 10: 397. doi: 10.1186/1471-2164-10-397 PMID: 19706172

29. Hartkoorn RC, Sala C, Uplekar S, Busso P, Rougemont J, Cole ST (2012) Genome-wide Definition of the SigF Regulon in *Mycobacterium tuberculosis*. J Bacteriol. 194(8):2001–9. doi: 10.1128/JB.06692-11 PMID: 22307756

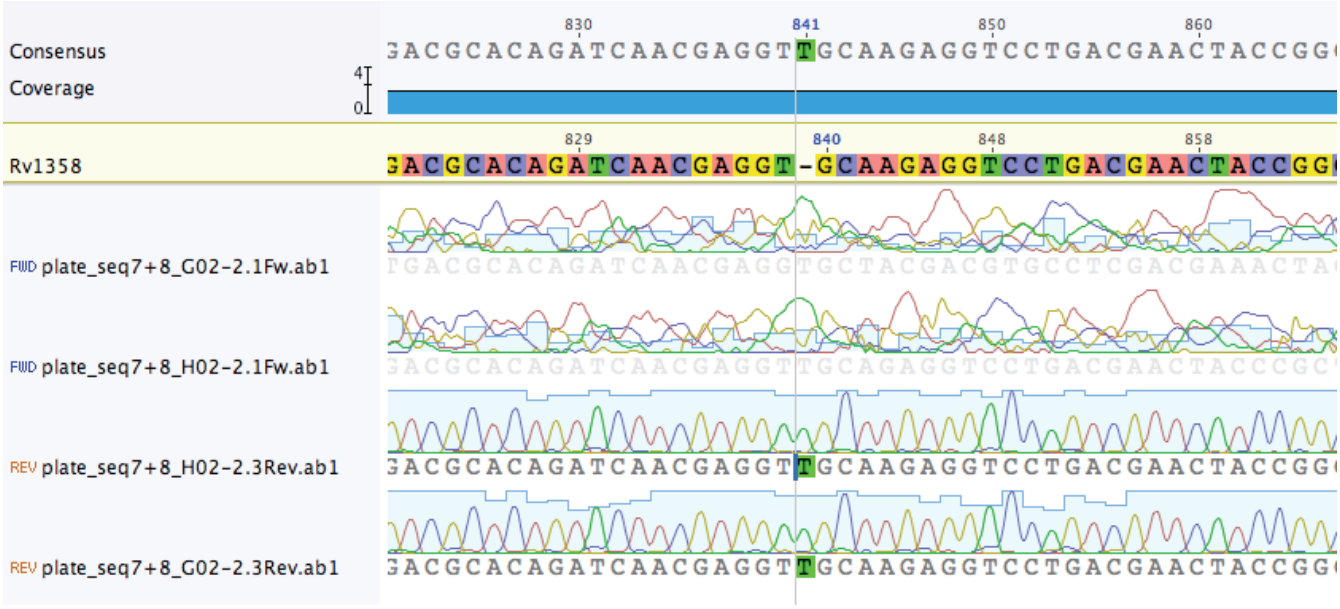
Supporting Figure 1

A



B

C



Supporting Table 1

iso	transm	year	birth_c	birth_reg	gen	age	TB_local	lineage	LSP_lineage	rx	ethn	alc_drug	homeless	h_set	BGC	mclx1	mclx2	mclx3	RFLP	VNTR	HIV	DM	malig	r_insuf	o_transp	med	co_morb
71	1	2008	Nigeria	af		1	58	3 CAS	EAI		0	1	0	0	2	1	0	0	0	0	9001110	0	0	0	0	0	0
331	0	2009	Senegal	af		1	24	1 LAM	EAm		0	1	0	0	2	9	0	del-1526658-2646	0	0	9003256	0	0	0	0	0	0
446	1	2009	Myanmar	sea		0	57	2 EAI	IO		0	1	0	0	1	9	0	0	del-2799489-1	0	9001132	0	0	0	0	0	0
451	0	2009	Ecuador	am		0	38	1 H	EAm		0	1	0	0	1	1	0	0	0	0	9003263	0	0	0	0	0	0
505	0	2009	Netherlands	eu		1	47	1 EAI	IO		0	0	0	0	1	0	0	0	del-2799489-5	0	9003246	0	0	0	0	0	0
1160	1	2010	Cameroon	af		0	36	2 LAM	EAm		0	1	0	0	2	9	9	0	0	0	9001822	1	0	0	0	0	1
3066	0	2002	Morocco	em		0	25	1 CAS	EAI		0	1	0	0	1	1	0	0	0	1181	0	0	0	0	0	0	0
7021	1	2001	Cameroon	af		1	30	3 HAARLEM	EAm		0	1	0	0	2	9	0	0	0	860	0	1	0	0	0	0	1
7405	1	2001	Somalia	em		1	18	3 EAI	IO		0	1	0	0	1	0	0	0	del-2799489-5	295123	0	0	0	0	0	0	0
7445	1	2002	India	sea		1	30	3 CAS	EAI		0	1	0	0	2	9	0	0	0	986	0	0	0	0	0	0	0
7453	1	2002	Indonesia	sea		0	25	2 EAI	IO		0	1	0	0	1	0	0	0	del-2799489-5	858	0	0	0	0	0	0	0
7483	1	2002	Vietnam	wp		0	24	2 EAI	IO		0	1	0	0	1	9	0	0	del-2799489-1	295017	0	0	0	0	0	0	0
7485	1	2002	Algeria	af		1	19	1 BEIJING	EAs		0	1	0	0	1	1	0	0	0	1151	0	0	0	0	0	0	0
7505	1	2002	Turkey	eu		1	24	1 H	EAm		0	1	0	0	1	1	0	0	0	653	0	0	0	0	0	0	0
7543	1	2002	Vietnam	wp		1	40	1 EAI	IO		1	1	0	0	1	0	0	0	del-2799489-5	1128134	0	0	0	0	0	0	0
7565	1	2002	Afganistan	em		1	22	1 H	EAm		0	1	0	0	1	1	0	0	0	859	0	0	0	0	0	0	0
7570	0	2002	Somalia	em		0	36	3 CAS	EAI		0	1	0	0	1	9	0	0	0	1134	0	0	0	0	0	0	0
7587	1	2002	Burma	sea		1	45	1 EAI	IO		1	1	0	0	1	1	9	0	del-2799489-5	295138	0	1	0	0	0	0	1
7631	1	2002	Cape Verde	af		1	19	3 LAM	EAm		0	1	0	0	2	9	0	0	0	744	0	0	0	0	0	0	0
7667	0	2002	Netherlands	eu		1	47	1 EAI	IO		0	0	0	0	1	0	0	0	del-2799489-5	5865	0	0	0	0	0	0	0
7688	1	2002	Eritrea	af		1	39	2 T	EAm		0	1	0	0	1	9	9	0	0	154	0	0	0	0	0	0	0
7696	1	2002	Thailand	sea		0	31	3 BEIJING	EAs		0	1	0	0	2	9	0	0	0	5891	0	0	0	0	0	0	0
7700	1	2002	Armenia	eu		1	39	1 H	EAm		0	1	0	0	1	0	0	0	0	1170	0	0	0	0	0	0	0
7724	1	2002	Afganistan	em		1	24	1 BEIJING	EAs		0	1	0	0	2	9	0	0	0	1151	0	0	0	0	0	0	0
7726	1	2002	Philippines	wp		0	22	1 EAI	IO		0	1	0	0	1	1	0	0	del-2799489-5	465	0	0	0	0	0	0	0
7743	1	9	9	9	9	999	9 LAM	EAm			9	0	0	0		9	0	0	0	744	0	0	0	0	0	0	0
7757	1	2002	Netherlands	eu		0	83	2 EAI	IO		0	0	0	0	1	9	0	0	del-2799489-5	549	0	0	0	1	0	0	1
7767	1	2002	Somalia	em		0	31	1 CAS	EAI		0	1	0	0	1	1	0	0	0	1178	0	0	0	0	0	0	0
7817	1	2002	Philippines	wp		0	32	2 EAI	IO		0	1	0	0	2	9	0	0	del-2799489-5	949	0	0	0	0	0	0	0
7825	1	2002	Pakistan	em		1	52	1 CAS	EAI		0	1	0	0	2	9	0	0	0	556	0	0	0	0	0	0	0
7851	1	2002	Vietnam	wp		0	26	2 EAI	IO		0	1	0	0	1	1	0	0	del-2799489-5	295017	0	0	0	0	0	0	0
7860	1	2002	Suriname	am		1	54	1 H	EAm		0	1	0	0	2	9	0	0	0	994	0	0	0	0	0	0	0
7986	1	2003	Turkey	eu		1	40	1 T	EAm		0	1	0	0	1	9	0	0	0	272	0	0	0	0	0	0	0
7996	0	2003	Eritrea	af		1	17	1 CAS	EAI		0	1	0	0	1	9	0	0	0	1449	0	0	0	0	0	0	0
8222	1	2003	Italy	eu		1	58	1 U		9	1	1	0	0	1	0	0	0	0	73	0	0	0	0	0	0	0
8311	0	2003	Guinea	af		1	21	1 BEIJING	EAs		0	1	0	0	2	9	0	0	0	6446	0	0	0	0	0	0	0
8337	1	2003	Etiopie	af		1	44	1 LAM	EAm		0	1	0	0	2	9	0	0	0	286005	0	0	0	0	0	0	0
8384	1	2003	Morocco	em		1	65	2 T	EAm		0	1	0	0	2	1	0	0	0	97	0	0	1	0	0	0	1
8452	1	2003	Vietnam	wp		1	30	1 EAI	IO		0	1	0	0	1	1	0	0	del-2799489-5	1128134	9000031	0	0	0	0	0	0
8538	1	2004	Indonesia	sea		1	65	2 EAI	IO		0	1	0	0	1	1	0	0	del-2799489-5	1212	9000085	0	0	0	0	0	0
8805	1	2004	Indonesia	sea		1	62	1 H	EAm		0	1	0	0	1	1	0	0	0	227	9000157	0	0	0	0	0	0
8974	1	2004	Turkey	eu		1	28	1 T	EAm		0	1	0	0	2	1	0	0	0	154	9000791	0	0	0	0	0	0
8989	1	2005	Kenya	af		0	31	3 LAM	EAm		0	1	0	0	2	9	0	ins-1527449-1	0	7004	9000802	1	0	0	0	0	1
9263	1	2005	Irak	em		1	37	2 H	EAm		0	1	0	0	1	1	0	0	0	528	9001150	0	0	0	0	0	0
9352	1	2005	Netherlands	eu		1	45	1 BEIJING	EAs		0	0	1	0	1	0	0	0	0	1504	9000467	0	0	0	0	0	0
9447	1	2006	Vietnam	wp		0	29	1 EAI	IO		0	1	0	0	1	1	0	0	del-2799489-5	1128134	9001362	0	0	0	0	0	0
9540	1	2006	Indonesia	sea		0	48	1 BEIJING	EAs		0	1	0	0	1	9	0	0	0	1189	9000726	0	0	0	0	0	0
9544	1	2006	Irak	em		1	31	1 CAS	EAI		0	1	0	0	1	9	0	0	0	870	9000289	0	0	0	0	0	0
9545	1	2006	Etiopie	af		1	27	1 T	EAm		1	1	0	0	1	9	0	0	0	154	9001899	1	0	0	0	0	1
9584	1	2006	Philippines	wp		0	50	2 EAI	IO		0	1	0	0	1	1	0	0	del-2799489-5	465	9000270	0	0	0	0	0	0
9627	1	2006	Morocco	em		0	28	3 LAM	EAm		0	1	0	0	1	1	0	ins-1527449-1	0	1472	9001596	0	0	0	0	0	0
9654	1	2006	Swaziland	af		0	35	3 H	EAm		0	1	0	0	1	1	0	0	0	227	9001257	0	0	0	0	0	0
9676	1	2006	Afganistan	em		0	71	2 H	EAm		0	1	0	0	1	9	0	0	0	859	9001653	0	0	0	0	0	0

Supporting Table 2

gene	primers' sequence		localization in the genome (nt)
mclx1 (463411-466668)	1.1Fw	CGGTGGCGTCGCTTCGACAT	463328-463347
	1.3Rev	CACCGAGGCCCACAGCGTC	464700-464682
	1.2 Fw	GGCCGGACTTTCGCCTCACC	464486-464505
	1.2Rev	CTGGTAGGCGAGCGCGAAGG	465780-465761
	1.3Fw	CCCAAGAGGCACGCGAGCTG	465569-465588
	1.1Rev	CCGTCCCCGAACGCCAATCA	466695-466676
mclx2 (1526612-1530091)	2.1Fw	CCAGCGGTTTCCTACGGGCG	1526542-1526561
	2.3Rev	CGCCGGCAGATCTCGCTCAC	1527960-1527941
	2.2Fw	TGGGTGCCTGCCCGGAGTTA	1527762-1527781
	2.2Rev	TCTGCGCCAGGCAGGCAAAC	1529081-1529062
	2.3Fw	CCGAGGCGATCGAGCTGGC	1528881-1528899
	2.1Rev	GCGACAACGCGCAGAAGAGC	1530170-1530151
mclx3 (2797467-2800880)	3.1Fw	ACCTTTGGTCGCTGGCTGGC	2797439-2797458
	3.3Rev	GGATCTGGCGCGACCGTGG	2798733-2798715
	3.2Fw	CCCTGCCAGAGATTCGCCGC	2798497-2798516
	3.2Rev	GCGGCTCTGATCGTCGCGTT	2799830-2799811
	3.3Fw	CAGGGCGAGGTTGTCGGCAG	2799633-2799652
	3.1Rev	CACGGGCACTGTAGGTCCGC	2800950-2800931

Supporting Table 3

***mclx*#2**			**multivariate ORs (95% CI)**
			Model 2
patient-related	**age**		0.958 (0.912-1.007) *p*=0.089 *B*=-0.043; *S.E.*=0.025 *Wald*=2.897
	ethnicity	native dutch	3.702 (0.811-16.898) *p*=0.091 *B*=1.309; *S.E.*=0.775 *Wald*=2.854
		foreign-born	1 (ref)
microorganism-related	**transmissibility**	no	1.754 (0.439-7.000) *p*=0.426 *B*=0.562; *S.E.*=0.706 *Wald*=0.632
		yes	1 (ref)
	lineage		*p*=0.357 *Wald*= 3.237
		EAI	0.136 (0.015-1.196) *p*=0.072 *B*=-1.998; *S.E.*=1.111 *Wald*=3.237
		EAm	1 (ref)
		EAs	*p*=0.999 0.000 (0.000-.) B=-20.230; S.E.=12299.315 Wald=0.000
		IO	*p*=0.998 0.000 (0.000-.) *B*=-20.251; *S.E.*=7120.660 *Wald*=0.000
Omnibus Test (chi-square/*p*)			40.620/ *p*<0.001
Cox & Snell R^2			0.285
Nagelkerke R^2			0.459
Hosmer and Lemeshow (chi-square/*p*)			0.792/ *p*=0.999
n			121

Supporting Table 4

independent variables			*n*	univariate ORs (95% CI)	
				mclx#2 pseudogenization	*mclx*#3 pseudogenization
patient-related	**age**		105	.970 (0.937-1.004) *p*=0.080	1.016 (0.988-1.045) *p*=0.264
	gender			*p*=0.463	*p*=0.178
		female	36	1.436 (0.546-3.773)	1.895 (0.747-4.809)
		male	69	1 (ref)	1 (ref)
	birth region			*p*=0.558	*p*=0.020
		Africa	19	0.885 (0.268-2.916) *p*=0.840	.000 (.000-.) *p*=0.998
		The Americas	10	0.479 (0.088-2.621) *p*=0.396	0.444 (0.048-4.116) *p*=0.475
		Eastern Mediterranean	20	0.213 (0.042-1.075) *p*=0.061	0.706 (0.161-3.103) *p*=0.645
		Europe	35	1 (ref)	1 (ref)
		South East Asia	13	0.000 (0.000-.) *p*=0.999	3.429 (0.872-13.483) *p*=0.078
		Western Pacific	8	0.000 (0.000-.) *p*=0.999	28.000 (2.942-266.467) *p*=0.004
	ethnicity			*p*=0.003	*p*=0.602
		native dutch	27	4.606 (1.669-12.714)	0.745 (0.247-2.250)
		foreign-born	77	1 (ref)	1 (ref)
	house setting			*p*=0.632	*p*=0.030
		rural	67	1 (ref)	1 (ref)
		urban	38	0.783 (0.288-2.131)	0.276 (0.087-0.883)
	BCG vaccination			*p*=0.721	*p*=0.873
		no	25	1.263 (0.351-4.551)	1.098 (0.349-3.458)
		yes	30	1 (ref)	1 (ref)
	co-morbidities			*p*=0.328	*p*=0.255
		no or unknown	89	1 (ref)	1 (ref)
		yes	17	0.460 (0.097-2.183)	0.406 (0.086-1.917)
	alcohol or drug use			*p*=0.824	*p*=0.280
		no or unknown	95	1 (ref)	1 (ref)
		yes	11	0.833 (0.167-4.167)	0.313 (0.038-2.578)
	homelessness			*p*=0.831	*p*=0.909
		no or unknown	102	1 (ref)	1 (ref)
		yes	4	1.286 (0.127-12.998)	1.145 (0.114-11.538)
microbe-related	**lineage**			*p*=0.267	-
		EAI	15	0.119 (0.015-0.972) *p*=0.047	-
		EAm	56	1 (ref)	-
		EAs	8	0.000 (0.000-.) *p*=0.999	-
		IO	24	0.000 (0.000-.) *p*=0.998	1
	antibiotic resistance			*p*=0.770	*p*=0.881
		none or unknown	97	1 (ref)	1 (ref)
		resistant	8	1.283 (0.241-6.846)	1.136 (0.214-6.033)
	transmissibility			*p*=0.006	*p*=0.076
		no	48	4.333 (1.537-12.217)	0.412 (0.154-1.098)
		yes	58	1 (ref)	1 (ref)
disease-related	**local of infection**			*p*=0.559	*p*=0.001
		pulmonary TB	71	1 (ref)	1 (ref)
		extra-pulmonary TB	16	0.533 (0.109-2.609) *p*=0.438	9.091 (2.741-30.153) *p*<0.001
		pulmonary and extra-pulmonary TB	18	1.436 (0.442-4.665) *p*=0.547	1.091 (0.270-4.408) *p*=0.903

Supporting Table 5

***mclx*#2**			**multivariate ORs (95% CI)**
			Model 1
patient-related	**age**		0.950 (0.903-0.999) *p*=0.047 *B*=-0.051; *S.E.*=0.026 *Wald*=3.940
	ethnicity	native dutch	6.845 (1.704-27.502) *p*=0.007 *B*=1.924; *S.E.*=0.710 *Wald*=7.349
		foreign-born	1 (ref)
microbe-related	**transmissibility**	no	1.790 (0.485-6.606) *p*=0.382 *B*=0.582; *S.E.*=0.666 *Wald*=0.764
		yes	1 (ref)
Omnibus Test (chi-square/*p*)			18.640/ p<0.001
Cox & Snell R^2			0.166
Nagelkerke R^2			0.260
Hosmer and Lemeshow (chi-square/*p*)			5.027/ *p*=0.755
n			103

Supporting Table 6

mclx#3			multivariate ORs (95% CI)		
			Model 1	Model 2	Model 3
patient-related	**gender**	female	-	-	0.632 (0.162-2.471) *p*=0.510 *B*=-0.458; *S.E.*=0.695 *Wald*=0.434
		male	-	-	1 (ref)
	house setting	rural	-	-	1 (ref)
		urban	-	-	0.287(0.061-1.352) *p*=0.114 *B*=-1.248; *S.E.*=0.791 *Wald*=2.492
	birth region		-	-	*p*=0.062 *Wald*=10.505
		Africa	-	-	0.000 (0.000-.) *p*=0.998 *B*=-20.135; *S.E.*=8708.272 *Wald*=0.000
		The Americas	-	-	0.547 (0.053-5.636) *p*=0.612 *B*=-0.604; *S.E.*=1.190 *Wald*=0.257
		Eastern Mediterranean	-	-	0.320 (0.049-2.087) *p*=0.233 *B*=-1.141; *S.E.*=0.957 *Wald*=1.420
		Europe	-	-	1 (ref)
		South East Asia	-	-	3.025 (0.470-19.483) *p*=0.244 *B*=1.107; *S.E.*=0.950 *Wald*=1.357
		Western Pacific	-	-	24.851 (1.933-319.528) *p*=0.014 *B*=3.213; *S.E.*=1.303 *Wald*=6.080
microbe-related	**transmissibility**	no	-	0.728 (0.241-2.199) *p*=0.573 *B*=-0.318; *S.E.*=0.564 *Wald*=0.317	2.079 (0.421-10.259) *p*=0.369 *B*=0.732; *S.E.*=0.815 *Wald*=0.807
		yes	-	1 (ref)	1 (ref)
disease-related	**local of infection**		*p*=0.001 *Wald*= 13.647	*p*=0.006 *Wald*=10.383	*p*=0.025 *Wald*=7.345
		pulmonary TB	1 (ref)	1 (ref)	1 (ref)
		extra-pulmonary TB	9.091 (2.741-30.153) *p*<0.001 *B*=2.207; *S.E.*=0.612 *Wald*=13.018	7.860 (2.163-28.569) *p*=0.002 *B*=2.062; *S.E.*=0.658 *Wald*=9.806	13.464(1.951-92.939) *p*=0.008 *B*=2.600; *S.E.*=0.986 *Wald*=6.958
		pulmonary and extra-pulmonary TB	1.091 (0.270-4.408) *p*=0.903 *B*=0.087; *S.E.*=0.712 *Wald*=0.015	1.056 (0.259-4.294) *p*=0.940 *B*=0.054; *S.E.*= 0.716 *Wald*=0.006	4.571 (0.691-30.246) *p*= 0.115 *B*=1.520; *S.E.*=0.964 *Wald*=2.484
Omnibus Test (chi-square/*p*)			14,269/ *p*=0.001	14.587/ *p*=0.002	42.240/ *p*<0.001
Cox & Snell R^2			0.127	0.130	0.331
Nagelkerke R^2			0.193	0.197	0.503
Hosmer and Lemeshow (chi-square/*p*)			0.000/ *p*=1.000	0.301/ *p*=0.960	6.527/ *p*=0.480
n			105		

Supporting Table 7

independent variables		*n*	univariate ORs (95% CI)	
			exclusively extrapulmonary (vs. pulmonary and disseminated)	**extrapulmonary or both** (vs. exclusivetly pulmonary)
age		124	1.059 (1.019-1.102) *p*=0.004 *B*=0.058; *S.E.*= 0.020 *Wald*= 8.420	1.020 (0.990-1.051) *p*=0.186 *B*=0.020; *S.E.*=0.015 *Wald*=1.749
gender	female	46	6.575 (1.896-22.802) *p*=0.003 *B*=1.883; *S.E.*=0.634 *Wald*= 8.810	5.708 (2.216-14.701) *p*<0.001 *B*=1.742; *S.E.*=0.483 *Wald*=13.021
	male	78	1 (ref)	1 (ref)
birth region			*p*=0.900 *Wald*=1.608	*p*=0.179 *Wald*=7.610
	Africa	23	6.883×10^{8} (0.000-.) *p*=0.999 *B*=20.350; *S.E.*=11556.217 *Wald*=0.000	11.220 (0.710-177.347) *p*=0.086 *B*=2.418; *S.E.*=1.408 *Wald*= 2.947
	The Americas	12	1.307×10^{8} (0.000-.) *p*=0.999 *B*=18.689; *S.E.*=11556.217 *Wald*=0.000	1.874 (0.104-33.760) *p*=0.670 *B*=0.628; *S.E.*=1.475 *Wald*=0.181
	Eastern Mediterranean	23	5.254×10^{8} (0.000-.) *p*=0.999 *B*=20.080; *S.E.*= 11556.217 *Wald*=0.000	6.247 (0.429-90.932) *p*=0.180 *B*=1.832; *S.E.*=1.366 *Wald*= 1.798
	Europe	36	1 (ref)	1 (ref)
	South East Asia	17	4.683×10^{8} (0.000-.) *p*=0.999 *B*=19.965; *S.E.*= 11556.217 *Wald*=0.000	3.314 (0.213-51.496) *p*=0.392 *B*=1.198; *S.E.*=1.400 *Wald*=0.733
	Western Pacific	13	3.313×10^{8} (0.000-.) *p*=0.999 *B*=19.619; *S.E.*=11556.217 *Wald*=0.000	1.077 (0.058-19.930) *p*=0.960 *B*=0.075; *S.E.*=1.489 *Wald*=0.003
ethnicity	native dutch	27	9.348×10^{7} (0.000-.) *p*=0.999 *B*=18.353; *S.E.*=11556.217 *Wald*=0.000	0.995 (0.064-15.571) *p*=0.997 *B*=-0.005; *S.E.*=1.403 *Wald*=0.000
	foreign-born	97	1 (ref)	1 (ref)
HIV	negative	111	1 (ref)	1 (ref)
	positive	13	0.497 (0.044-5.627) *p*=0.572 B=-0.700; S.E.= 1.239 Wald=0.319	1.195 (0.279-5.114) *p*=0.810 B=0.178; S.E.=0.742 Wald=0.058
***mclx*3 status**	pseudogene	30	8.259 (1.674-40.761) *p*=0.010 *B*=2.111; *S.E.*=0.814 *Wald*=6.720	4.994 (1.430-17.436) *p*=0.012 *B*=1.608; *S.E.*=0.638 *Wald*=6.354
	functional	94	1 (ref)	1 (ref)
Omnibus Test (chi-square/*p*)			41.614/ *p*<0.001	33.355/ *p*<0.001
Cox & Snell R^2			0.285	0.236
Nagelkerke R^2			0.456	0.324
Hosmer and Lemeshow (chi-square/*p*)			5.289/ *p*=0.726	5.420/ *p*=0.712
n			124	

Supporting Table 8

			Lineage				Total
			EAI	EAm	EAs	IO	
local of infection	pumonary TB	Count	12_a	43_a	9_a	15_a	79
		% within Lineage	66,7%	67,2%	90,0%	48,4%	64,2%
	extra-pulmonary TB	Count	1_a	10_a	$0_{a, b}$	13_b	24
		% within Lineage	5,6%	15,6%	0,0%	41,9%	19,5%
	pulmonary+extra-pulmonary TB	Count	5_a	11_a	1_a	3_a	20
		% within Lineage	27,8%	17,2%	10,0%	9,7%	16,3%
Total		Count	18	64	10	31	123
		% within Lineage	100,0%	100,0%	100,0%	100,0%	100,0%

Each subscript letter denotes a subset of LSP_lineage categories whose column proportions do not differ significantly from each other at the 0.05 level (collumn proportions compared by the z-test with p-values adjusted by the Bonferroni method).

Chapter V

Convergent genetic markers in *Mycobacterium tuberculosis* are associated with transmissibility and altered immune responses

Convergent genetic markers in *Mycobacterium tuberculosis* are associated with transmissibility and altered immune responses

Hanna Nebenzahl-Guimaraes[1,2,3 ¶], Arjan van Laarhoven[4 ¶], Maha R. Farhat[5 ¶], Valerie A.C.M. Koeken[4], Jornt J. Mandemakers[6], Mihai G. Netea[4], Megan Murray[7,8 †,*], Reinout van Crevel[4 †,*], Dick van Soolingen[1,9 †]

AFFILIATIONS:

National Institute for Public Health and the Environment (RIVM), Bilthoven, The Netherlands.

Life and Health Sciences Research Institute (ICVS), School of Health Sciences, University of Minho, Braga, Portugal.

ICVS/3B's, PT Government Associate Laboratory, Braga/Guimarães, Portugal.

Department of Internal Medicine and Radboud Center for Infectious Diseases, Radboud University Medical Center, Nijmegen, the Netherlands

Pulmonary and Critical Care Division, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02114 Massachusetts, USA.

University of Wageningen, 6706 KN, Wageningen, the Netherlands

Department of Global Health and Social Medicine, Harvard Medical School, Boston, MA 02115, Massachusetts, USA.

Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts, USA

Department of Medical Microbiology, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands

¶ Joint first authors

† Joint last authors

Abstract

Successful transmission of tuberculosis (TB) depends on human behavior, host immune responses and *Mycobacterium tuberculosis* virulence factors. We sought to identify mycobacterial genetic markers associated with increased transmissibility, and examined whether these markers lead to altered *in vitro* immune responses. Using a comprehensive TB registry and strain collection in the Netherlands, we identified *M. tuberculosis* strains either least or most likely to be transmitted after controlling for host associated behavioral factors. Through whole genome sequencing of 100 strains, we identified the loci *espE, PE-PGRS33, PE-PGRS56*, Rv0197, Rv2813-2814c and Rv2815-2816c as targets of convergent evolution among transmissible strains. We validated four of these regions in an independent set of strains, and demonstrated that mutations in these targets affected *in vitro* monocyte and T-cell cytokine production, reactive oxygen species release and neutrophil apoptosis. These findings suggest that *M. tuberculosis* shows convergent evolution associated with enhanced transmissibility *in vivo* and altered immune responses *in vitro*.

Author summary

Tuberculosis is a highly contagious airborne disease that is transmitted from one person to the next through coughing, sneezing, or talking. There are several factors that determine the likelihood of transmission and these include the severity of infection and other patient related factors as well as factors related to the bacteria itself. To study bacterial genetic factors associated with tuberculosis transmission, we used the Dutch National TB registry that contains data on all TB patients, their infecting bacteria and the size of the associated circle of transmission. After adjustment for patient factors, we selected 100 bacterial samples that were either highly or poorly transmissible and studied them with whole genome sequencing. We identified six bacterial DNA regions to be associated with TB transmission. In the laboratory, we validated these regions by studying the response of human white blood cells to extracts from a subset of the tuberculosis bacteria that carried or did not carry mutations in these DNA regions. We show that there are differences in the immune response that associate with these genetic changes. In conclusion, we identified novel genetic regions that appear to be important for transmission of *M. tuberculosis* and maybe relevant targets for disease surveillance.

Introduction

When patients with active pulmonary TB cough, they generate small droplet nuclei containing the pathogen *M. tuberculosis*, which can then be transmitted to others through the respiratory route. Successful transmission requires that viable bacteria enter the lungs, evade killing by the innate immune system, and replicate intracellularly. If a series of transmission events occurs over a relatively short time, one can identify a group of patients with *M. tuberculosis* strains that are genotypically highly similar. Epidemiologists often use molecular fingerprinting to characterize the genetic similarity among a group of strains; strains that share a molecular fingerprint are described as "clustered"[1-3] and are inferred to be the result of recent transmission rather than the reactivation of a previous infection.

Host factors affect tuberculosis (TB) transmission and disease progression[4,5], but recent molecular epidemiologic studies have shown that *M. tuberculosis* strains also differ in their ability to cause pulmonary disease[6-8], their proclivity to infect contacts[9,10] or cause secondary cases[11-13]. This variability may reflect the strains' ability to subvert innate[14-17] and/or adaptive[1-3,6-8] immunity, or their ability to exploit the host immune system by inducing a detrimental inflammatory response[4,5,9,10] leading to tissue damage[1-3,6-8] and the formation of cavities that enable disease spread[4,5,9,10]. Cytokines play a pivotal role in these events; insufficient production of pro-inflammatory cytokines may lead to uncontrolled mycobacterial growth, while overproduction may lead to tissue damage[11-13].

Phylogenetic differences in cytokine response[14-17] suggest that specific microbial genetic determinants may underlie transmission related phenotypes. Several studies have used *M. tuberculosis* mutants *in vitro* and experimental models to identify the role of a few individual genes on transmission-associated phenotypes[18,19]. However, further elucidation of the full spectrum of genes affecting transmission could improve our understanding of the host-pathogen relationship in TB.

To control for host factors and isolate mycobacterial factors of transmissibility, we used the Netherlands' country-wide TB registry that stores patient data and *M. tuberculosis* isolates for all new culture positive cases of TB since 1993. We performed whole genome sequencing of 100 strains that were either more or less clustered than would be expected from their patient risk factors like sputum bacterial load, history of drug use or homelessness etcetera. We identified loci under positive selection for clustering by analyzing whole *M. tuberculosis* genomes from clustered and unclustered isolates for evidence of convergence. Following the hypothesis that clustered strains have consistent genetic differences compared to unclustered ones, and that the genes or intergenic regions implicated in these differences affect the host immune response, we performed a functional validation of the newly identified targets of independent mutation (TIMs) by measuring *in vitro* cytokine production and neutrophil responses.

Results

STRAIN SELECTION FOR SEQUENCING

We aimed to compare strains that caused clusters of tuberculosis *in the absence* of obvious patient-related risk factors for clustering with unique (non-clustered) strains isolated from patients with a high likelihood of being part of a cluster (e.g. a homeless man with grade 3 sputum smear-positivity). The National Institute for Public Health and the Environment (RIVM) in the Netherlands stores all *M. tuberculosis* complex strains (>13,000) isolated in the Netherlands and their DNA fingerprints since 1993, and also has accompanying information on risk factors for clustering. Using this data, we calculated the Cluster Propensity to Propagate (CPP) as a summary measure of risk for transmission

of the patients belonging to that particular tuberculosis cluster (**Table S1**) [20]. This CPP was calculated for 10,389 patient isolates. Following current practice, DNA fingerprinting by RFLP and MIRU-VNTR was used to define molecular clustering as a proxy for the relative transmissibility of a *M. tuberculosis* strain[21,22]. For whole genome sequencing, we selected 100 strains aiming for maximum contrast of the transmissibility phenotype: 66 clustered strains with low CPPs and 34 unclustered strains with a high CPP (**Fig. S1**). Strains for the clustered phenotype were picked at random from 56 unique cluster fingerprints (5 pairs of strains came from within the same cluster). To increase our power we matched the number and type of strain lineages in the clustered and unclustered group[23,24]. The 100 selected strains were all drug sensitive, belonged to patients originating from 44 different countries, and were representative of the four major *M. tuberculosis* lineages.

TARGETS OF INDEPENDENT MUTATIONS (TIMS)

To identify genetic markers of clustering, we performed next generation whole genome sequencing. We constructed a Bayesian phylogenetic tree of the 100 sequenced strains, based on a Multiple Sequence Alignment of single nucleotide substitutions (SNPs) called against reference strain H37Rv (**Fig. 1**). We conducted two parallel phylogenetic evolutionary convergence tests (PhyC) to identify either individual nucleotide positions, or genes and intergenic regions where cluster-associated mutations occur frequently and along disparate locations in the phylogenetic tree. Region-level PhyC detected four genes and two intergenic regions as significant targets of independent mutation (TIMs) ($p < 0.05$) (**Table 1**). A total of 12 SNPs, 2 insertions and 31 deletions were found in these TIMs, including 1 SNP and 2 deletions that were also significant by the site-level phyC test (**Table S2**). TIMs in the two *PE-PGRS* genes occurred solely in clustered branches, while those in *espE,* Rv0197, Rv2813-2814c and Rv2815-2816c were also found in unclustered branches, but at a lower rate than in clustered branches (depicted for *espE* in **Fig. S2**).

We validated these results in an independent public dataset of whole genome sequences of clustered (n=96) and unclustered (n=47) *M. tuberculosis* strains[23,24]. These strains were collected from patients of different geographical backgrounds and were predominantly drug resistant (**Table S3**). PhyC confirmed four out of six genes or intergenic regions (**Table 1**) including Rv0197, in which it detected the same nonsynonymous coding site (234,477TG). The TIMs occurring in the two PE-PGRS genes could not be validated, as their occurrence in the original dataset was restricted to lineage 1, which only made up 3.4% of the validation dataset.

DELETERIOUS EFFECT OF TIMS ON PROTEINS

We used two protein prediction algorithms, I-Mutant v2.0 and PolyPhen-2[25], to predict the functional impact of the significant SNPs on the structure and function of their respective proteins. All 12 SNPs in genes Rv0197 and *espE* are predicted to adversely affect the respective proteins (**Table S4**). Two TIMs in Rv0197 (234,265GT and 234,477TG) result in a STOP codon and truncation of the protein, whilst two TIMs in *PE-PGRS33* and *PE-PGRS56* are frameshift mutations likely to have functional consequences.

ASSOCIATION BETWEEN TIMS AND INDUCTION OF CYTOKINE RESPONSES

Genetic variation associated with transmissibility is likely to influence the initial host response. Next, we therefore examined *in vitro* cytokine responses in strains with and without convergent changes. Since mycobacterial lineages are known to induce differential cytokine responses[26], we only used strains of two lineages (1 and 4), both of which had strains with mutations in at least four out of six genes or intergenic regions (**Table S5**). Nineteen clinical strains were selected. H37Rv was added for quality control of the experiments, but not included in the analysis. Peripheral blood mononuclear cells (PBMCs) from 12 healthy donors were stimulated with 3 µg/mL of heat-killed, bead-disrupted *M. tuberculosis* lysate from all 20 strains for 4 hours (for TNF-α) and 24 hours (TNF-α and other monocyte-derived cytokines) and 7 days (T cell-derived cytokines) using a previously established method[16,27]. We used a mixed effects regression model that accounts for between-donor variation and lineage effects. This model exploits within-donor variation and covariance between cytokine concentrations to maximize statistical power. We first tested whether the six respective genes or intergenic regions were associated with an immunological phenotype for three sets of assays (monocyte cytokines, T-cell cytokines, and PMN responses). In secondary analyses, we compared levels of individual cytokines and PMN assays. In both analyses, significance was determined at $\alpha = 0.05/6$, Bonferroni corrected for the six genes or intergenic regions tested.

Mutations in three of the targets we identified, *espE*, *PE-PGRS33* and Rv2813-2814, were associated with alterations in monocyte cytokine production ($p < 10^{-4}$, **Table 2, Fig. 2**). In the secondary analysis (see **Table S6**), mutations in *espE* were associated with decreased production of IL-10 (- 26%, $p = 1.7 \times 10^{-8}$, **Fig. 3A**) and to a lesser extent early TNF-α (- 18% $p = 8.0 \times 10^{-3}$); mutations in *PE-PGRS33* were associated with decreased IL-10 (- 18%, $p = 1.9 \times 10^{-3}$); and mutations in Rv2813-2814c were associated with increased production of early (+ 29%, $p = 6.2 \times 10^{-3}$) and late (+ 33%, $p = 2.5 \times 10^{-3}$) TNF-α, IL-1β (+ 30%, $p = 7.7 \times 10^{-3}$) and IL-10 (+ 19%,

Fig. 1: *Consensus Bayesian phylogenetic tree.*
Legend: Clustered strains and M. tuberculosis lineages are highlighted.

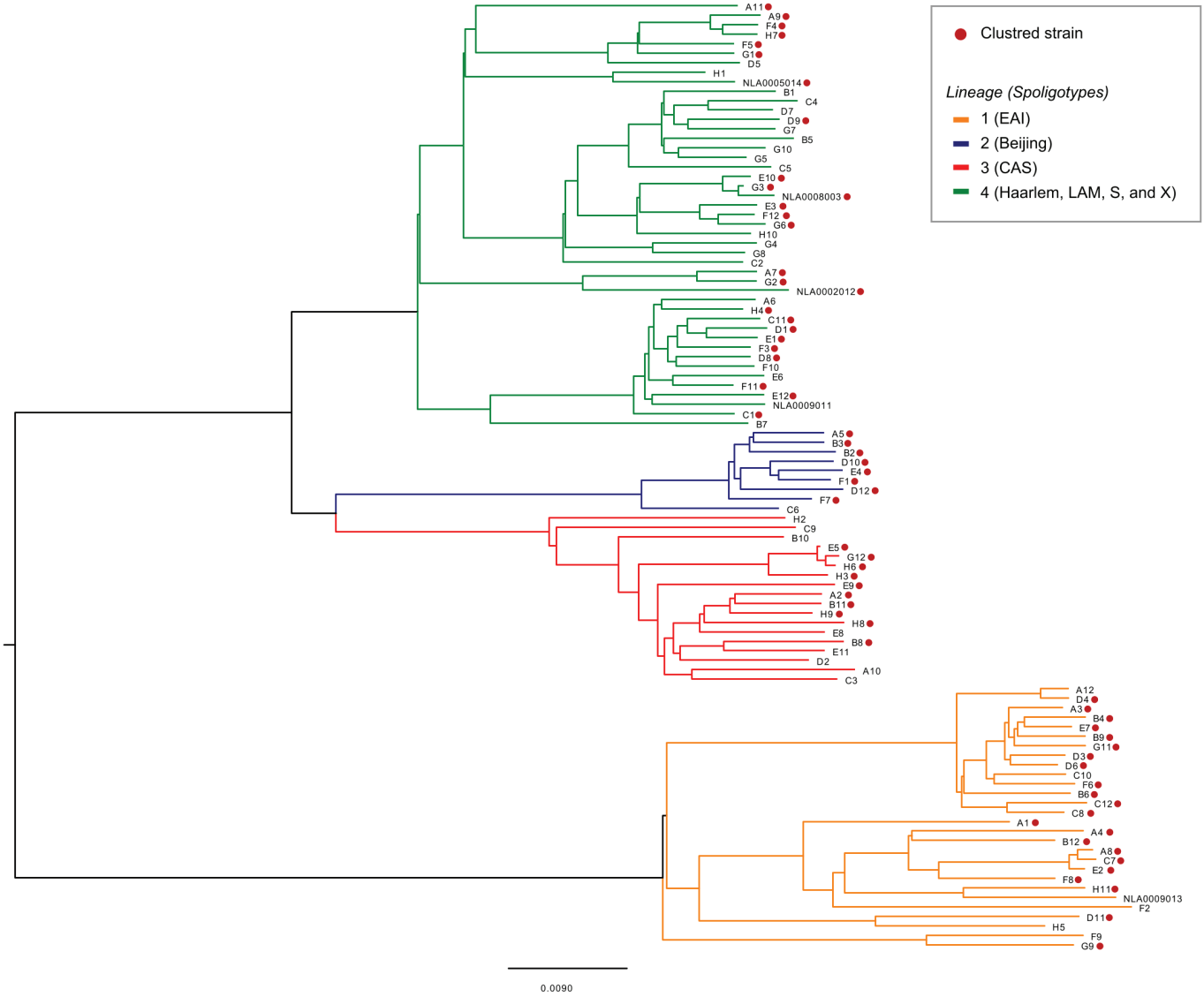


Table 1: *Significant genes or intergenic regions by PhyC.*

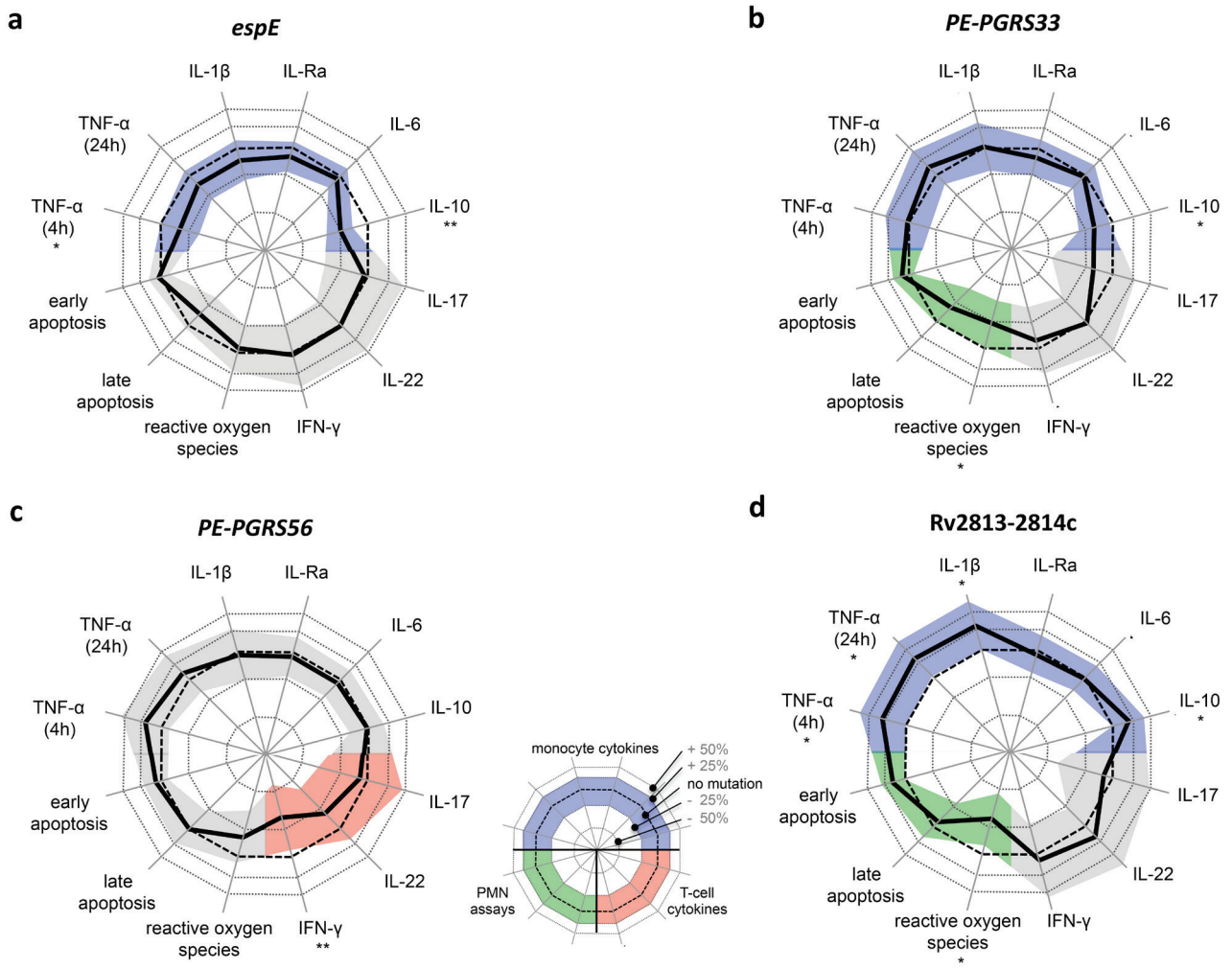
	Original dataset (n = 100)				Validation dataset (n = 143)			
	Strains with mutations, deletions and insertions (N)				Strains with mutations, deletions and insertions (N)			
Gene/region [Rv number]	Clustering	Non-clustering	p-value	Lineages with cases	Clustering	Non-clustering	p-value	Lineages with cases
espE [Rv3864]	10	1	0.0377	1, 3, 4	10	2	0.0232	1, 3, 4
PE-PGRS33 [Rv1818c]	16	0	0.0006	1	8	2	0.0779	1, 4
PE-PGRS56 [Rv3512]	13	0	0.0052	1, 4	1	0	1	4
unnamed [Rv0197]	20	12	0.0214	1, 2, 3, 4	26	12	0.0362	1, 2, 3, 4
unnamed [Rv2813-2814c]	20	6	0.0458	1, 3, 4	22	3	0.0001	1, 3, 4
unnamed [Rv2815-2816c]	18	4	0.0178	1, 4	22	5	0.0105	1, 4

Table 2: *Overall response to M. tuberculosis strain with or without mutations in the six genes or intergenic regions for each of the assay groups.*

Gene or intergenic region	monocyte cytokines (df=6)	T-cell cytokines (df=3)	PMNs (df=3)
espE	1.33×10^{-6}	0.961	0.077
PE-PGRS33	2.83×10^{-5}	0.345	3.99×10^{-5}
PE-PGRS56	0.039	5.35×10^{-3}	0.021
Rv0197	0.017	0.224	0.343
Rv2813-2814c	7.47×10^{-6}	0.309	5.79×10^{-8}
Rv2815-2816c	0.025	0.027	0.151

Legend: Significance (in bold) is determined at α = 0.05/6 = 0.0083, corrected for the six genes or intergenic regions tested.

Fig. 2: *Response to M. tuberculosis strain with or without mutations in the four TIMs that showed an effect in primary analysis. Relative differences for individual assays in the secondary analysis are indicated by the difference between the thick black line (mutation present) and the thin reference line (no mutation) for each of the TIMs (a) espE, (b) PE-PGRS33, (c) PE-PGRS56, or (d) Rv2813-2814c that significantly influenced at least one assay group. Shaded area: 95% confidence interval, corrected for the fact that six genes or intergenic regions were tested for each assay (z = 2.64). Legend: * p < 0.05/6 = 0.0083. ** significant after further correcting for number of assays per group, i.e. 0.05/(6*6) for monocyte cytokines and p < 0.05/(6*3) for T-cell cytokines and PMN assays. Significance in the primary analysis is indicated by a colored confidence interval for monocyte cytokines (blue), T-cell cytokines (red) and PMN assays (green).*

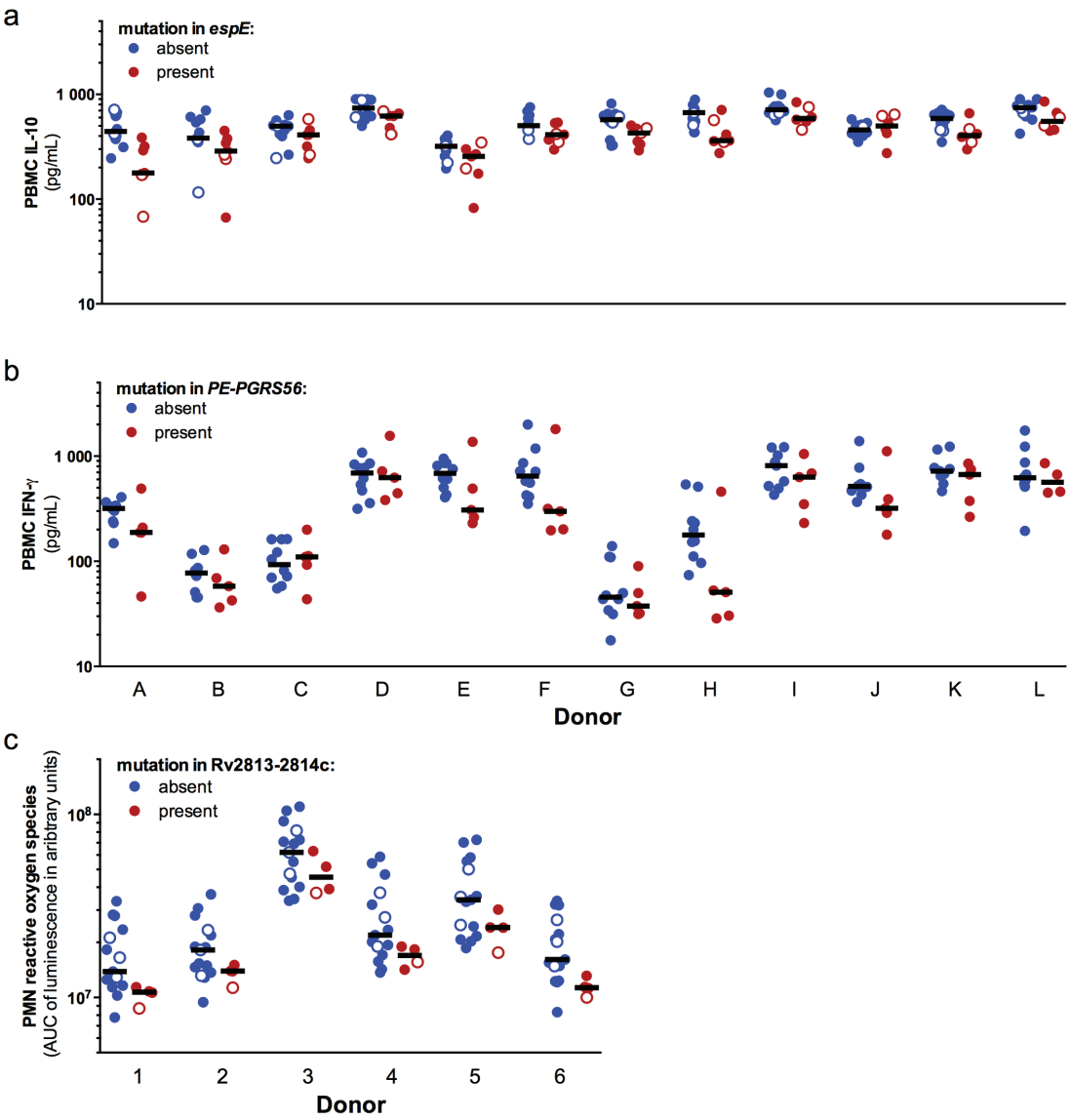


p = 1.9 x 10^{-3}). Of the six genes or intergenic regions, only *PE-PGRS56* affected T-cell cytokine responses (p = 5.4 x 10^{-3}), and in our secondary analysis, this was associated with lower IFN-γ production (- 34%, p = 1.6 x 10^{-3}, **Fig. 3B**).

ASSOCIATION BETWEEN TIMS AND RESPONSE OF NEUTROPHILS

We next examined the effect of TIMs on *in vitro* responses of neutrophils, given their putative role in transmission and clinical manifestation of TB[28]. We stimulated isolated polymorphonuclear cells (PMNs, largely consisting of neutrophils) with the 20 *M. tuberculosis* strains, measuring the induction of reactive oxygen species (ROS) using luminol-enhanced chemiluminescence for one hour (6 donors), and neutrophil apoptosis and cell death with flow cytometry after six hours (8 donors). In the mixed effects model, we found that the TIMs *PE-PGRS33* and Rv2813-2814c affected PMN responses (both p < 10^{-4}). In secondary analysis, ROS production was lower (- 31%, p = 4.8 x 10^{-4}, **Fig. 3C**) and early apoptosis higher (+ 15%, p = 3.6 x 10^{-3}) for Rv2813-2814c, while ROS production was also lower for *PE-PGRS33* (- 24%, p = 4.1 x 10^{-3}).

Fig. 3: *In vitro responses of selected assays TIMs. Stimulation was performed with lysate of M. tuberculosis strains from lineage 1 (filled) and lineage 4 (open) that did not harbor (blue) or harbored (red) a mutation in (A) espE or (B) PE-PGRS56 or (c) Rv2813-2814c. PBMCs of twelve healthy donors (A-L) were stimulated and (A) IL-10 was measured after 24h, (B) IFN-γ after 7 days. Because no mutations occurred in PE-PGRS genes in strains from lineage 4, only lineage 1 strains were used in the analysis for PE-PGRS33 genes and subsequently displayed here. (C) PMNs of six healthy donors (1-6) were stimulated with lysate of strains from lineage 1 (filled) and lineage 4 (open) that did not harbor (blue) or harbored (red) a mutation in Rv2813-2814c. Reactive oxygen species were measured by luminol-enhanced chemiluminescence and plotted in arbitrary units of the area under the curve (AUC) of the measurement over the first hour after stimulation.*



Discussion

We identified six genes and intergenic regions (*espE*, *PE-PGRS33*, *PE-PGRS56*, Rv0197, Rv2813-14c and Rv2815-16c) as targets of independent mutation (TIMs) in clustered *M. tuberculosis* strains. We confirmed four out of six genes and intergenic regions in a second dataset despite differences in lineages and drug resistance profiles between the original and validation datasets. The TIMs we identified are predicted to alter the function of their respective proteins, supporting the hypothesis that they confer a selective advantage for transmission. Finally, four of six identified genes or intergenic regions were associated with altered cytokine production or PMN responses.

Experimental studies on the identified genes or intergenic regions support their potential roles in increasing the transmissibility of *M. tuberculosis* (**Table S7**). In addition, other genomic epidemiological studies may support the role of the genes we identified in mycobacterial transmission. Non-synonymous SNPs (albeit different from the ones identified in this study) and a frameshift mutation in *espE* were found to be more common in *M. africanum* strains relative to H37Rv [29], and to be implied in their reduced ability to induce a CD4-cell ESAT-6 induced IFN-γ host response[30,31]. Similarly, a previous study identified a Large Sequence Polymorphism (LSP) associated with clustering in a gene (MT1801) encoding Molybdopterin oxidoreductase, which is also encoded by Rv0197[32]. Another study reported that a *M. tuberculosis* strain responsible for a large outbreak in the UK harbored an insertion in position 3,121,877 of intergenic region Rv2815-16c[33], adjacent to the 2bp deletion in 3,121,879 observed in our own study. Finally, clinical strains with large insertions or deletions (INDELs) and frameshift mutations in the *PE-PGRS33* protein have been linked to both the clustered phenotype and absence of lung cavitation[34].

Four out of six genes or intergenic regions with TIMs associated with clustering of TB showed a clear and statistically significant effect on monocyte or T-cell cytokine production or PMN responses at the group level. *M. tuberculosis* strains with TIMs in *PE-PGRS56* induced lower production of IFN-γ, which is unequivocally seen as a key factor in protection against TB[11,35]. Some of the monocyte cytokines, however, can act as a double-edged sword, and may have different roles in different parts of the TB life cycle that all contribute to clustering. A lower IL-10 for example, found in association with TIMs in *espE* and *PE-PGRS33* in our study, may prevent its inhibiting effect on tissue damage, while a higher amount, as associated with TIMs in Rv2813-2814c could also decrease intracellular killing of *M. tuberculosis*[11]. In zebrafish models, a high TNF-α as we observed in Rv2813-2814c, has been shown to lead to 'necroptosis' which favors bacterial outgrowth[36], while a low TNF-α as we observed in PE-PGRS33 might favor breakthrough to disease as in patients treated with TNF-α blocking therapy[37]. In line with a previous comparison of *in vitro* cytokine responses to different *M. tuberculosis* lineage strains [14], TNF-α and IL-6 induction in our study was higher in lineage 4 (ancient) compared to lineage 1 (modern) strains. These lineage effects may depend on strain selection, as shown by another study by Reiling et al.[38] that found opposite results. The aim of our study was not to discern lineage effects, but we corrected for these effects in our statistical model.

With regard to neutrophils, strains harboring cluster-associated mutations in Rv2813-2814c induced lower ROS production and early apoptosis, and strains with mutations in *PE-PGRS33* lower ROS production. Neutrophils are considered protective during early infection, when they are recruited to the site of infection, phagocytose mycobacteria[28] or mycobacteria infected macrophages[39], and resist mycobacterial growth using reactive oxygen species (ROS)[39]. Children with chronic granulomatous disease (CGD) have a reduced oxidative burst and are more susceptible to TB[40]. Neutrophil ROS also correlates with apoptosis[15], which is thought to contain mycobacterial growth and facilitate antigen presentation but may also contribute to mycobacterial spread[41].

This study was limited by several factors. The inclusion of additional key host factors that may influence disease transmissibility, such as exposure time (i.e. via prospective household contact data) and pulmonary cavitation, could improve our ability to isolate bacterial factors influencing transmissibility in the future. The difference in drug resistance profiles, and possibly other related parameters, such as treatment efficacy between the original and validation cohort of strains for the phyC test, could have introduced bias in measurement of the transmissibility phenotype (**Table S3**). However, validation of genetic markers associated with transmission in this separate dataset reduces the risk of false positive findings.

Of note, we performed *in vitro* cellular stimulations aiming to find biological support for the epidemiological associations identified through convergent evolutionary analysis, and not to identify specific effects of individual TIMs on *in vitro* cellular responses. Such effects cannot be identified in this study, as multiple TIMs were present in single strains in this dataset (**Table S4**). For this, additional studies using mutagenesis or recombineering to isolate the mutational effects should be performed. It is no surprise that no single pattern of cytokine production or PMN response was found for the six genes

or intergenic region, as *M. tuberculosis* has different strategies to subvert or resist the host immune system or use it to its advantage.

In summary, we present evidence from an evolutionary convergence analysis that six *M. tuberculosis* genes or intergenic regions confer a selective advantage promoting the transmission of *M. tuberculosis* and/or TB disease progression, and that these genetic elements influence the response of the host to the mycobacteria. These findings serve as an important step forward in the quest for an improved understanding of the microbial genetic determinants of TB transmission.

Materials and Methods

STUDY DESIGN

Strains were selected taking into account their cluster size and cluster propensity to propagate (CPP), a summary measure of the contribution of the hosts' risk factors towards clustering. In the overall RIVM dataset of 10,389 strains, we found CPP to be significantly higher in clustered versus unclustered strains, although the CPP rapidly plateaus with increasing cluster size (**Fig. S1**) [38]. As to our knowledge there are no power calculators to guide the design of studies associating genomic variants with the transmissibility phenotype [23,42]. We arbitrarily chose 100 strains for whole genome sequencing: 66 unclustered strains with an average CPP of 1.02 (sd=0.3) that was higher than the overall average of 0.84 (sd=0.12). We also chose 34 clustered strains with an average CPP of 0.75 (sd=0.006) that was lower than the overall average above (**Fig. S1**). After variant calling and phylogeny reconstruction, PhyC[43] was used to identify genetic loci that displayed significantly higher variation in the clustered group. Second, these loci were validated by repeating the PhyC analysis in an independent dataset that comprised of 96 clustered strains and 47 unclustered strains. Third, out of the first dataset, 19 strains were recultured, and heat-killed and bead-beated to perform functional experiments. After stimulation of PBMCs, cytokine responses were measured (6 experiments of two healthy donors each), and after stimulation of PMNs reactive oxygen species (3 x 2 donors) and apoptosis (4 x 2 donors) was measured. Multivariate mixed models were applied to exploit covariance between assays and to control for inter-donor variability.

ACCESSION CODES

All sequences have been rendered publically available through NCBI. The complete genome sequence for reference strain H37Rv was accessed from GenBank accession NC_000962.3. Raw sequences for the 200 strains from Bryant et al. are available at the European Nucleotide Archive (ENA) under accession ERP000111.

RIVM DATASET OF STRAINS

The National Institute for Public Health and the Environment (RIVM) in Bilthoven, The Netherlands, serves as a reference laboratory for the secondary laboratory diagnosis of all TB cases in The Netherlands, offering identification, drug susceptibility testing, and molecular typing. Strains recovered from patients between 1993 and 2009 underwent IS6110 and polymorphic GC-rich sequence (PGRS) restriction fragment length polymorphism (RFLP) typing and those from 2004 onwards to variable number of tandem repeat (VNTR) typing. Clusters were defined as groups of patients who shared TB strains with identical RFLP or VNTR patterns or, if strains had fewer than five IS6110 copies, identical PGRS RFLP patterns[44]. DNA fingerprints of all nationwide *M. tuberculosis* complex strains and their cluster status have been stored in a database since 1993. Demographic and clinical information, provided by the Registration Committee of the Netherlands Tuberculosis Register (NTR), were linked to the strains on the basis of gender, date of birth, year of diagnosis and postal code. Phylogenetic lineages were ascertained based on a combination of spoligotyping, MIRU-typing and Restriction Fragment Length Polymorphisms (RFLP)-pattern similarity, as previously described[20].

SEQUENCING, ALIGNMENT AND VARIANT CALLING

DNA was extracted from all strains using standard methods and was sequenced on an Illumina HiSeq 2500 instrument using reads of 50bp in length in the paired-end modus. The average genome coverage was approximately 100x. The FASTQ sequence reads were generated using the Illumina Casava pipeline version 1.8.3. Initial quality assessment was based on data passing the Illumina Chastity filtering. Subsequently, reads containing adapters and/or PhiX control signal were removed using an in-house filtering protocol. The second quality assessment was based on the remaining reads using the FASTQC quality control tool version 0.10.0. The quality of the FASTQ sequences was enhanced by trimming off low-quality bases using the "Trim sequences" option of the CLC Genomics Workbench version 6.5. The quality-filtered sequence reads were then puzzled into a number of contig sequences using the previously mentioned software. SNPs were called against reference strain H37Rv using Breseq software (version 0.23) using a minimum threshold of 15x coverage [45]. Mutations with low-quality evidence (i.e. possible mixed read alignment) or within 5 bp of an INDEL (insertion or deletion) were

discarded. Due to the higher likelihood of false-positive calls in PE-PGRS genes, the two site-specific deletions in *PE-PGRS33* and *PE-PGRS56* significantly associated to transmissibility were manually checked to confirm that they did not fall within repetitive regions (**Fig. S3**).

PHYLOGENY CONSTRUCTION

The phylogeny was constructed on the basis of multiple-sequence alignment of the *M. tuberculosis* whole-genome sequences. Single nucleotide polymorphisms (SNPs) occurring in repetitive elements, including PE/PPE and PGRS genes, were excluded to avoid any concern about inaccuracies in read alignment in those portions of the genome. The final concatenate of SNPs was used to construct phylogenetic trees using three different methods: parsimony (PHYLIP dnapars algorithm v3.68), Bayesian Markov chain Monte Carlo (MCMC) (MrBayes v3.2) and maximum-likelihood (PhyML v3.0) using the GTR model with eight categories for the gamma model. One hundred bootstrap re-samplings were performed for each tree, except for the Bayesian tree, where posterior probabilities on the branches were used as a measure of confidence. The three trees were fully consistent between each other, and we used the Bayesian tree for all subsequent analyses.

PHYLOGENETIC CONVERGENCE TEST FOR SELECTION (PHYC)

PhyC is a test for positive natural selection based on homoplasy or parallel evolution, and is well suited for the study of clonal pathogens such as *Mycobacterium tuberculosis* . It has been shown in prior work on drug resistance to have a higher sensitivity (and likely also specificity) than the dN/dS method [43]. The PhyC test was conducted here as previously described[43] with two modifications: (1) We used Carmin-sokal parsimony for reconstruction of the phenotypic states as we thought this better mirrors our assumption that transmissibility evolves unidirectionally (*i.e.* from less to more transmissible); (2) We also performed ancestral reconstruction of INDELs using FASTML and maximum-likelihood criteria[46]. For each nucleotide position in the genome, we counted the number of convergent SNPs and INDELs in clustered and unclustered branches. We controlled for the occurrence of SNPs or INDELs in strains belonging to the same cluster (as defined by MIRU- or RFLP-typing) by counting only one strain per cluster. Given that some background convergence is expected owing to neutral mutation and sequence error, even without positive selection, we assessed the significance of each convergent SNP or INDEL compared to the empirical background distribution using a permutation test, as previously described[23]. As this was a permutation test based on the observed frequency distribution for all variants across the genome, a 0.05 P-value threshold was used. In parallel we ran the convergence test grouping SNPs and INDELs by the gene or intergenic region in which they occurred. We used the same empirical resampling strategy a list of significant regions.

PHYC VALIDATION DATASET

Strains in the validation dataset were assigned two phenotypes: clustered (belonging to a cluster of minimum size of 3, n=96) and unclustered (having a unique fingerprint and no epidemiologic links reported from contact investigation, n=47). All clustered strains belong to clusters of different fingerprints, in order to eliminate redundant results caused by highly similar (or effectively clonal) strains. Since epidemiological data (host risk factors) was not available for these strains, we could not take the strain's CPP into account, and hence the phenotype was defined solely using clustering status. For 19 clustered strains, single end 36bp read sequencing was previously performed which made calling INDELs unreliable (**Table S5**).

STRAIN SELECTION FOR IMMUNOLOGICAL EXPERIMENTS

Mycobacterial lineage is known to influence host immune response[26]. We therefore only selected strains from the WGS dataset belonging to lineages 1 and 4, both of which had four to six TIMs represented, and could therefore include lineage as a factor in the statistical model. Nineteen out of twenty-one strains could be re-cultured, fifteen of lineage 1 and four of lineage 4. Fifteen of the strains were of the clustered phenotype, four unclustered. The unclustered strains had a maximum of one mutation in the genes or intergenic regions associated to increased transmissibility, whilst all clustered strains had at least one TIM in the genes of interest (range 1–6), with the exception of strain F6 which had zero (**Fig. S4**, **Table S3)**. H37Rv, the most well characterized strain of *M. tuberculosis*, was included as a reference strain.

MYCOBACTERIAL CULTURE AND STANDARDIZATION

Strains were grown on a shaking platform to determine the growth curve, and then regrown to harvest mid-log (OD_{600} 0.6-0.8 for all strains). Strains were heat-killed, washed in PBS, lysed mechanically by bead-beating and divided in two aliquots. The first was used for stimulation experiments and to measure protein concentration by bicinchoninic acid (BCA) protein, and the second was freeze-dried to determine dry weight and after resuspension used for the ROS experiments. Protein-to-dryweight ratio did not differ substantially for one of the isolates **(Fig. S5**). Prior to this study, standardization experiments were performed to determine the optimal

moment of harvesting (mid-log phase) and processing method (bead-beating), and confirm the reproducibility of cellular responses. PBMCs of six donors were stimulated with three batches of H37Rv. Mean standard deviation over six donors was 0.33 for TNF-α (coefficient of variation [CV] 5.5%) and 0.45 for IL-1β (CV 6.1%).

PBMC CYTOKINE STIMULATION EXPERIMENTS

PBMCs from buffy coats obtained from 12 healthy volunteers (Sanquin Bloodbank, Nijmegen, the Netherlands) over a density gradient using Ficoll-paque were stimulated in duplicate in 96-well round-bottom plates with 3 ug/mL of the different strains in a total volume of 200 µl and incubated at 37 °C in a 5% CO_2 environment (in the presence of 10% human pooled serum for 7 day stimulation). Cytokines were measured batch-wise using ELISA after 4h (TNF-α, using R&D Systems, Minneapolis, Minnesota, USA), 24h (TNF-α, IL-1β and IL-1Ra, R&D; IL-6 and IL-10 using Sanquin, Amsterdam, the Netherlands) or 7 days (IL-17 and IL-22, R&D; IFN-γ, Sanquin) stimulation.

PMN REACTIVE OXYGEN SPECIES AND APOPTOSIS EXPERIMENTS

Polymorphonuclear cells (PMNs) were isolated from EDTA blood from 8 other healthy volunteers using ficoll-paque, cleared from erythrocytes by hypotonic lysis (two times) buffer and washed two times in cold PBS. Reactive oxygen species were measured in 6 volunteers in white 96-well flat-bottom plates using luminol-enhanced chemiluminescence (5-amino-2,3,dihydro-1,4-phtalazinedione, Sigma-Aldrich, St. Louis, Missouri, USA). PMNs were stimulated in 240 µl at $1{\cdot}10^6$ cells/mL in 0.5% BSA HBSS with culture medium alone, zymosan (final concentration: 833 µl/mL) or the 20 different *M. tuberculosis strains* (10 µg/mL), and chemiluminescence was measured at 37 °C for the next 60 min, after which the area under the curve for luminescence was calculated. For apoptosis, PMNs of 8 volunteers were stimulated for 6 hours with the different strains (10 µg/mL), IL-1β as anti-apoptotic control and cyclohexamide as positive control, after which Annexin V-FITC conjugate (Av, BioVision, Milpitas, California, USA) and propium iodide (PI) were added. Annexin V stains phosphatidylserine translocating from the inner to the outer leaflet of the membrane, marking early apoptosis. PI stains nuclei from cells that are permeable, reflecting cell death, either from advanced apoptosis or necrosis[47]. Flow cytometric analysis using Cytomics FC50 was used to distinguish Av^-/PI^- (alive) Av^+/PI^- (early apoptotic) and Av^+/PI^+ (advanced apoptotic / necrotic) populations. Different concentrations and time-points were tested first for ROS and apoptosis assays (**Fig. S5**).

DATA ANALYSIS AND STATISTICS

To control for inter-donor variability and exploit covariance between outcome measures, results were analysed using a multivariate mixed model for each assay group (monocyte cytokines, T-cell cytokines and PMNs). Cytokine concentrations and ROS area under the curve (AUCs) were Ln-transformed. PBMC and PMN experiments were performed on different donors and therefore tested in different models that included a set of fixed effects for each combination of donor, assay, and lineage group (lineage 1 and 4) to account for variability between donors, assays, and between lineages and a random effect for strain to model the dependency structure in the data. For each strain, the absence/presence of TIMs in a gene or intergenic region was tested using assay-specific dummy indicators. First, we tested whether the presence of TIMs was associated with differential response by comparing multivariate models with and without TIMs indicators using Likelihood Ratio tests (**Table 2**). Bonferroni correction was applied for 6 tests. Secondly, we examined the predicted differences and further corrected for the number of assays in each group. All analyses were performed using Stata MP, version 12.1. Cytokine radar graphs show the percentage change for each of the assays, plotted on a logarithmic axis.

ETHICS STATEMENT

The Registration Committee of the Netherlands Tuberculosis Register (NTR) approved the retrospective access to strains and provided demographic and clinical information for patients. Because the data are de-identified by name, DNA fingerprinting results from the RIVM were linked on the basis of sex, date of birth, year of diagnosis and postal code. Peripheral blood mononuclear cells (PBMCs) were isolated from volunteers with written informed consent and approval from the Ethics Committee of Radboud University Medical Center, Nijmegen, the Netherlands.

ACKNOWLEDGMENTS:

The authors would like to thank Aldert Zomer and Sacha van Hijum for their bioinformatics assistance with SNP calling; Jessica de Beer and Arnout Mulder for culturing of mycobacterial strains; Jeroen de Keijzer for processing them; Ekta Lachmandas, Bas Blok, Mark Gresnigt and Cor Jacobs for assisting in immunological experiments; and Professor Jelle Goeman for statistical advice.

References and Notes

1. Rakotosamimanana N, Raharimanga V, Andriamandimby SF, Soares JL, Doherty TM, Ratsitorahina M, et al. Variation in Gamma Interferon Responses to Different Infecting Strains of *Mycobacterium tuberculosis* in Acid-Fast Bacillus Smear-Positive Patients and Household Contacts in Antananarivo, Madagascar. Clinical and Vaccine Immunology. 2010;17: 1094–1103. doi:10.1128/CVI.00049-10

2. WHO. World Health Organization (2012) Global tuberculosis control: WHO Report 2012. Geneva; 2012;: 1–100.

3. Manca C, Tsenova L, Bergtold A, Freeman S, Tovey M, Musser JM, et al. 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. Proc Natl Acad Sci USA. 2001;98: 5752–5757. doi:10.1073/pnas.091096998

4. López B, AGUILAR D, Orozco H, BURGER M, Espitia C, Ritacco V, et al. A marked difference in pathogenesis and immune response induced by different *Mycobacterium tuberculosis* genotypes. Clin Exp Immunol. 2003;133: 30–37.

5. Kik SV, Verver S, van Soolingen D, de Haas PEW, Cobelens FG, Kremer K, et al. Tuberculosis Outbreaks Predicted by Characteristics of First Patients in a DNA Fingerprint Cluster. Am J Respir Crit Care Med. 2008;178: 96–104. doi:10.1164/rccm.200708-1256OC

6. Aguilar León D, Hanekom M, Mata D, van Pittius NCG, van Helden PD, Warren RM, et al. *Mycobacterium tuberculosis* strains with the Beijing genotype demonstrate variability in virulence associated with transmission. Tuberculosis. Elsevier Ltd; 2010;90: 319–325. doi:10.1016/j.tube.2010.08.004

7. Kong Y, Cave MD, Zhang L, Foxman B, Marrs CF, Bates JH, et al. Population-based study of deletions in five different genomic regions of *Mycobacterium tuberculosis* and possible clinical relevance of the deletions. Journal of Clinical Microbiology. 2006;44: 3940–3946. doi:10.1128/JCM.01146-06

8. Kato-Maeda M, Shanley CA, Ackart D, Jarlsberg LG, Shang S, Obregon-Henao A, et al. Beijing sublineages of *Mycobacterium tuberculosis* differ in pathogenicity in the guinea pig. Clin Vaccine Immunol. 2012. doi:10.1128/CVI.00250-12

9. Jones-López EC, Kim S, Fregona G, Marques-Rodrigues P, Hadad DJ, Molina LPD, et al. Importance of cough and M. tuberculosis strain type as risks for increased transmission within households. PLoS ONE. 2014;9: e100984. doi:10.1371/journal.pone.0100984

10. Albanna AS, Reed MB, Kotar KV, Fallow A, McIntosh FA, Behr MA, et al. Reduced Transmissibility of East African Indian Strains of *Mycobacterium tuberculosis*. Neyrolles O, editor. PLoS ONE. 2011;6: e25075. doi:10.1371/journal.pone.0025075.t003

11. O'Garra A, Redford PS, McNab FW, Bloom CI, Wilkinson RJ, Berry MPR. The Immune Response in Tuberculosis. Annu Rev Immunol. 2013;31: 475–527. doi:10.1146/annurev-immunol-032712-095939

12. Kato-Maeda M, Kim EY, Flores L, Jarlsberg LG, Osmond D, Hopewell PC. Differences among sublineages of the East-Asian lineage of *Mycobacterium tuberculosis* in genotypic clustering. int j tuberc lung dis. 2010;14: 538–544.

13. de Jong BC, Hill PC, Aiken A, Awine T, Antonio M, Adetifa IM, et al. Progression to Active Tuberculosis, but Not Transmission, Varies by *Mycobacterium tuberculosis*Lineage in The Gambia. J INFECT DIS. 2008;198: 1037–1043. doi:10.1086/591504

14. Portevin D, Gagneux S, Comas I, Young D. Human Macrophage Responses to Clinical Isolates from the *Mycobacterium tuberculosis* Complex Discriminate between Ancient and Modern Lineages. Bessen DE, editor. PLoS Pathogens. 2011;7: e1001307. doi:10.1371/journal.ppat.1001307.g009

15. Romero MM, Balboa L, Basile JI, López B, Ritacco V, la Barrera de SS, et al. Clinical Isolates of *Mycobacterium tuberculosis* Differ in Their Ability to Induce Respiratory Burst and Apoptosis in Neutrophils as a Possible Mechanism of Immune Escape. Clinical and Developmental Immunology. 2012;2012: 1–11. doi:10.1155/2012/152546

16. van Laarhoven A, Mandemakers JJ, Kleinnijenhuis J, Enaimi M, Lachmandas E, Joosten LAB, et al. Low Induction of Proinflammatory Cytokines Parallels Evolutionary Success of Modern Strains within the *Mycobacterium tuberculosis* Beijing Genotype. Infection and Immunity. 2013;81: 3750–3756. doi:10.1128/IAI.00282-13

17. Wang C, Peyron P, Mestre O, Kaplan G, van Soolingen D, Gao Q, et al. Innate Immune Response to *Mycobacterium tuberculosis* Beijing and Other Genotypes. Ahmed N, editor. PLoS ONE. 2010;5: e13594. doi:10.1371/journal.pone.0013594.t001

18. Cadieux N, Parra M, Cohen H, Maric D, Morris SL, Brennan MJ. Induction of cell death after localization to the host cell mitochondria by the *Mycobacterium tuberculosis* PE_PGRS33 protein. Microbiology (Reading, Engl). 2011;157: 793–804. doi:10.1099/mic.0.041996-0

19. Shin D-M, Jeon BY, Lee H-M, Jin HS, Yuk J-M, Song C-H, et al. *Mycobacterium tuberculosis* eis regulates autophagy, inflammation, and cell death through redox-dependent signaling. PLoS Pathogens. 2010;6: e1001230. doi:10.1371/journal.ppat.1001230

20. Nebenzahl-Guimaraes H, Borgdorff MW, Murray MB, van Soolingen D. A Novel Approach - The Propensity to Propagate (PTP) Method for Controlling for Host Factors in Studying the Transmission of *Mycobacterium Tuberculosis*. Neyrolles O, editor. PLoS ONE. 2014;9: e97816. doi:10.1371/journal.pone.0097816.s001

21. Small PM, Hopewell PC, Singh SP, Paz A, Parsonnet J, Ruston DC, et al. The epidemiology of tuberculosis in San Francisco. A population-based study using conventional and molecular methods. N Engl J Med. 1994;330: 1703–1709. doi:10.1056/NEJM199406163302402

22. Alland D, Kalkut GE, Moss AR, McAdam RA, Hahn JA, Bosworth W, et al. Transmission of tuberculosis in New York City. An analysis by DNA fingerprinting and conventional epidemiologic methods. N Engl J Med. 1994;330: 1710–1716. doi:10.1056/NEJM199406163302403

23. Farhat MR, Shapiro BJ, Sheppard SK, Colijn C, Murray M. A phylogeny-based sampling strategy and power calculator informs genome-wide associations study design for microbial pathogens. Genome Medicine. 2014;6: 1–14. doi:10.1186/s13073-014-0101-7

24. Bryant JM, rch ACS, van Deutekom H, Harris SR, de Beer JL, de Jager V, et al. Inferring patient to patient transmission of *Mycobacterium tuberculosis* from whole genomesequencing data. BMC Infectious Diseases. BMC Infectious Diseases; 2013;13: 1–1. doi:10.1186/1471-2334-13-110

25. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. correspondence. Nat Methods. Nature Publishing Group; 2010;7: 248–249. doi:10.1038/nmeth0410-248

26. Coscolla M, Gagneux S. Consequences of genomic diversity in *Mycobacterium tuberculosis*. Semin Immunol. Elsevier Ltd; 2014;26: 431–444. doi:10.1016/j.smim.2014.09.012

27. Chen Y-Y, Chang J-R, Huang W-F, Hsu S-C, Kuo S-C, Sun J-R, et al. The pattern of cytokine production in vitro induced by ancient and modern Beijing *Mycobacterium tuberculosis* strains. PLoS ONE. 2014;9: e94296. doi:10.1371/journal.pone.0094296

28. Lowe DM, Redford PS, Wilkinson RJ, O'Garra A, Martineau AR. Neutrophils in tuberculosis: friend or foe? Trends in Immunology. Elsevier Ltd; 2012;33: 14–25. doi:10.1016/j.it.2011.10.003

29. Gehre F, Otu J, DeRiemer K, de Sessions PF, Hibberd ML, Mulders W, et al. Deciphering the Growth Behaviour of Mycobacterium africanum. Small PLC, editor. PLoS Negl Trop Dis. 2013;7: e2220. doi:10.1371/journal.pntd.0002220.t003

30. de Jong BC, Hill PC, Brookes RH, Gagneux S, Jeffries DJ, Otu JK, et al. Mycobacterium africanum elicits an attenuated T cell response to early secreted antigenic target, 6 kDa, in patients with tuberculosis and their household contacts. J INFECT DIS. 2006;193: 1279–1286. doi:10.1086/502977

31. Tientcheu LD, Sutherland JS, de Jong BC, Kampmann B, Jafali J, Adetifa IM, et al. Differences in T-cell responses between *Mycobacterium tuberculosis* and Mycobacterium africanum-infected patients. Eur J Immunol. 2014;44: 1387–1398. doi:10.1002/eji.201343956

32. Talarico S, Ijaz K, Zhang X, Mukasa LN, Zhang L, Marrs CF, et al. Tuberculosis. Tuberculosis. Elsevier Ltd; 2011;91: 244–249. doi:10.1016/j.tube.2011.01.007

33. Yesilkaya H, Forbes KJ, Shafi J, Smith R, Dale JW, Rajakumar K, et al. The genetic portrait of an outbreak strain. Tuberculosis. 2006;86: 357–362. doi:10.1016/j.tube.2005.08.019

34. Talarico S, Cave MD, Foxman B, Marrs CF, Zhang L, Bates JH, et al. Association of *Mycobacterium tuberculosis* PE_PGRS33 polymorphism with clinical and epidemiological characteristics. Tuberculosis. Elsevier Ltd; 2007;87: 338–346. doi:10.1016/j.tube.2007.03.003

35. van Crevel R, Ottenhoff THM, van der Meer JWM. Innate immunity to *Mycobacterium tuberculosis*. Clinical Microbiology Reviews. 2002;15: 294–309.

36. Roca FJ, Ramakrishnan L. TNF Dually Mediates Resistance and Susceptibility to Mycobacteria via Mitochondrial Reactive Oxygen Species. Cell. Elsevier Inc; 2013;153: 521–534. doi:10.1016/j.cell.2013.03.022

37. Keane J, Gershon S, Wise RP, Mirabile-Levens E, Kasznica J, Schwieterman WD, et al. Tuberculosis associated with infliximab, a tumor necrosis factor alpha-neutralizing agent. N Engl J Med. 2001;345: 1098–1104. doi:10.1056/NEJMoa011110

38. Reiling N, Homolka S, Walter K, Brandenburg J, Niwinski L, Ernst M, et al. Clade-Specific Virulence Patterns of *Mycobacterium tuberculosis* Complex Strains in Human Primary Macrophages and Aerogenically Infected Mice. mBio. 2013;4: e00250–13–e00250–13. doi:10.1128/mBio.00250-13

39. Yang C-T, Cambier CJ, Davis JM, Hall CJ, Crosier PS, Ramakrishnan L. Neutrophils Exert Protection in the EarlyTuberculous Granuloma by Oxidative Killing of Mycobacteria Phagocytosed from Infected Macrophages. Cell Host and Microbe. Elsevier Inc; 2012;12: 301–312. doi:10.1016/j.chom.2012.07.009

40. Lee PPW, Chan K-W, Jiang L, Chen T, Li C, Lee T-L, et al. Susceptibility to mycobacterial infections in children with X-linked chronic granulomatous disease: a review of 17 patients living in a region endemic for tuberculosis. Pediatr Infect Dis J. 2008;27: 224–230. doi:10.1097/INF.0b013e31815b494c

41. Moraco AH, Kornfeld H. Cell death and autophagy in tuberculosis. Semin Immunol. Elsevier Ltd; 2014;26: 497–511. doi:10.1016/j.smim.2014.10.001

42. Read TD, Massey RC. Characterizing the genetic basis of bacterial phenotypes using genome-wide association studies: a new direction for bacteriology. 2014;: 1–11. doi:10.1186/s13073-014-0109-z

43. Farhat MR, Shapiro BJ, Kieser KJ, Sultana R, Jacobson KR, Victor TC, et al. Genomic analysis identifies targets of convergent positive selection in drug-resistant *Mycobacterium tuberculosis*. Nat Genet. Nature Publishing Group; 2013;: 1–9. doi:10.1038/ng.2747

44. van Soolingen D, de Haas PE, Hermans PW, Groenen PM, van Embden JD. Comparison of various repetitive DNA elements as genetic markers for strain differentiation and epidemiology of *Mycobacterium tuberculosis*. Journal of Clinical Microbiology. 1993;31: 1987–1995.

45. Barrick JE, Yu DS, Yoon SH, Jeong H, Oh TK, Schneider D, et al. Genome evolution and adaptation in a long-term experiment with Escherichia coli. Nature. 2009;461: 1243–1247. doi:10.1038/nature08480

46. Ashkenazy H, Penn O, Doron-Faigenboim A, Cohen O, Cannarozzi G, Zomer O, et al. FastML: a web server for probabilistic reconstruction of ancestral sequences. Nucleic Acids Research. 2012;40: W580–W584. doi:10.1093/nar/gks498

47. Galluzzi L, Aaronson SA, Abrams J, Alnemri ES, Andrews DW, Baehrecke EH, et al. Guidelines for the use and interpretation of assays for monitoring cell death in higher eukaryotes. Nature Publishing Group; 2009;16: 1093–1107. doi:10.1038/cdd.2009.44

Supplementary Materials

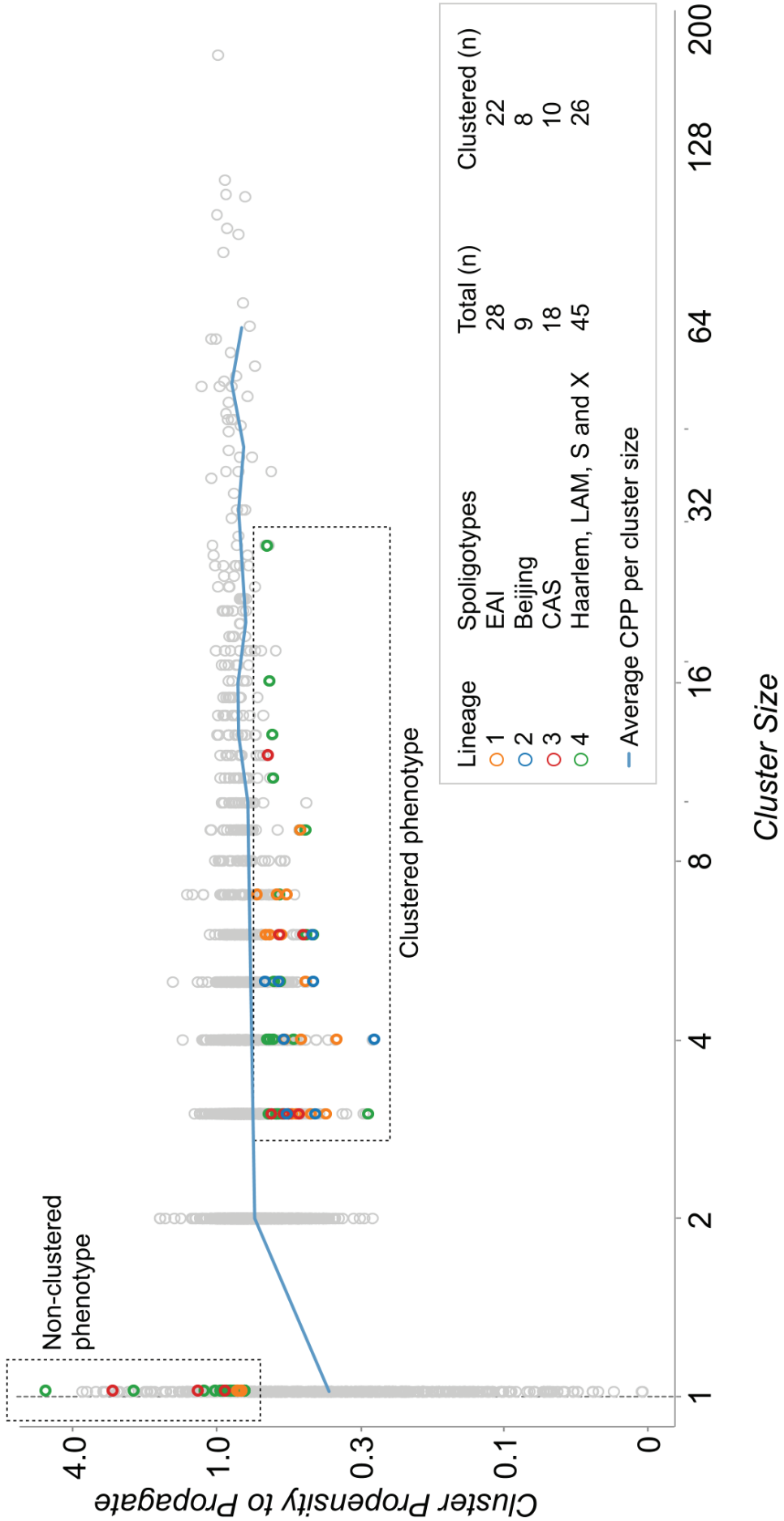


Fig. S1: *Selection of strains from clustered and unclustered phenotypes. Grey circles represent strains from the overall RIVM dataset, whilst colored circles denote the 100 strains selected for WGS and the phylogenetic lineage they belong to.*

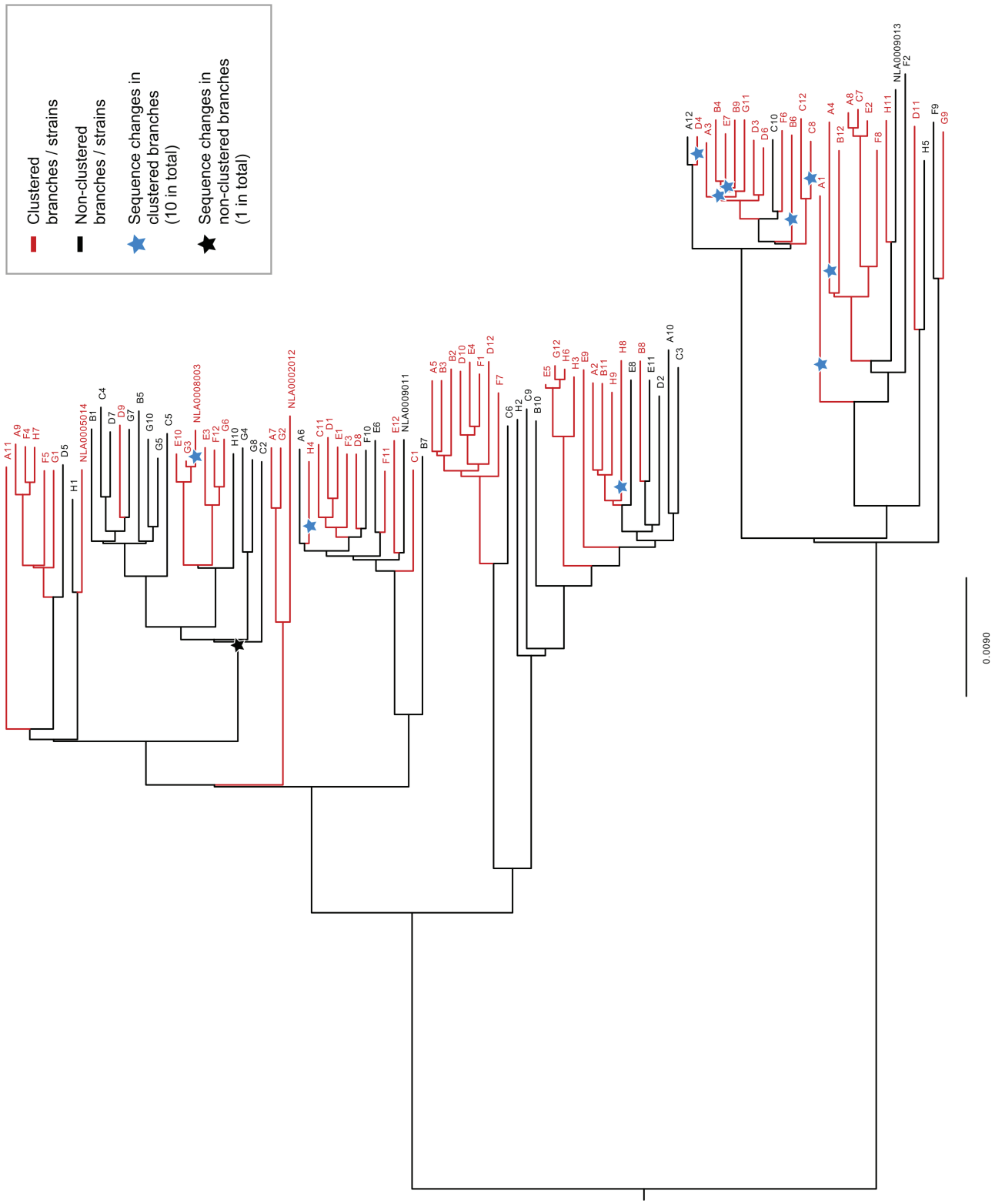


Fig. S2: *Pairwise convergence for SNPs in gene espE (Rv3864). Branches in red represent clustered isolates, stars denote locations in the tree where TIMs associated with increased clustering occurred.*

Fig. S3: *Diagram demonstrating breseq calling a 1 base-pair deletion in the PE_PGRS33 gene. Displayed are 40 color-coded Illumina sequencing reads mapping to the H37Rv reference sequence (singled out at the top and bottom). Visual inspection of the deleted site confirms that it does not occur in a region containing uniformly lower base quality scores.*

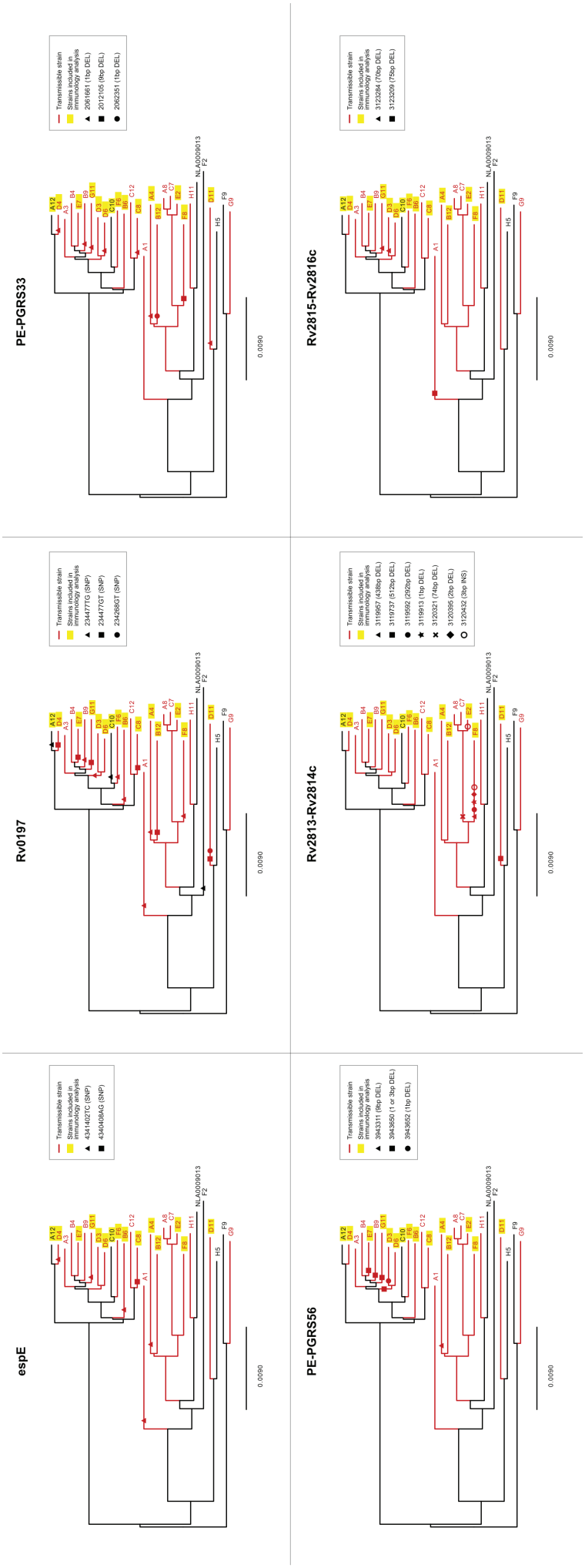


Fig. S4: *Distribution of TIMs in the six genes or intergenic regions associated to clustering across the 15 lineage 1 strains selected for functional validation studies. Phylogenetic relationship between strains taken from the lineage 1 portion of the original Bayesian tree in Figure 1.*

Fig. S5: *(a) M. tuberculosis H37Rv protein concentration versus dry weight for the nineteen strains (with sequence ID) used in the immunological experiments. (b) PMN reactive oxygen production with different concentrations of H37Rv M. tuberculosis lysate (3.2, 10 or 32 µg/mL). (c) PMN apoptosis stimulation time and dose response optimisation after stimulation with IL-1β, cycloheximide or H37Rv M. tuberculosis lysate (0.1, 1 or 10 µg/mL) at 3, 6, 18 or 48 hours. Combined results of two experiments with in total six donors. * = not measured.*

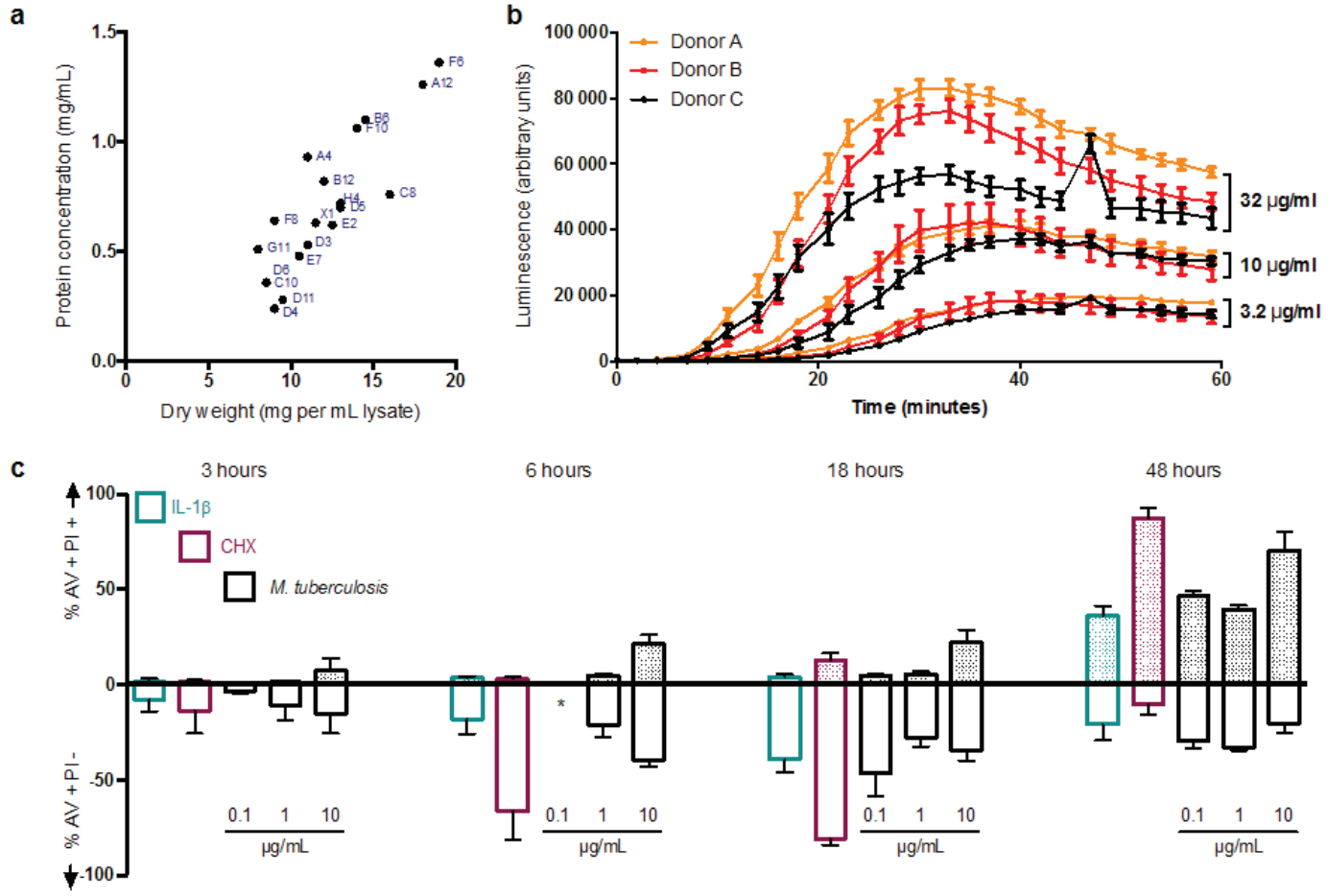


Table S1: *Summary of risk factors combined to calculate the Patient Propensity to Propagate (PPP)*.*

Category	Odds ratio	Case group
Sex	1	males
	0.87	females
Age at diagnosis	1.05	0-15
	1	16-30
	0.86	31-45
	0.77	46-60
	0.49	61-75
	0.19	76-90
	0.12	>90 years
Disease classification	1	pulmonary
	0.76	extrapulmonary
	0.90	pulmonary + extrapulmonary
Smear-positivity	1	no
	1.17	yes
Alcohol consumption	1	no
	1.29	yes
Drug-use	1	no
	2.75	yes
Homelessness	1	no
	1.58	yes
Traveler to endemic areas	1	no
	0.58	yes
Origin	1	native Dutch
	0.28	foreign-born (Asia)
	0.76	foreign-born (Africa)
	1.06	foreign-born (America)
	0.43	foreign-born (Europe)

** The geometric mean of the PPPs across the cluster was used to calculate the Cluster Propensity to Propagate (CPP).*

Table S2: *Genomic positions of SNPs and indels associated with the clustering phenotype.*

	Mutations, deletions and insertions (N)		
Genomic position: polymorphism	**In clustering strains**	**In non-clustering strains**	**p-value**
espE			
4341369: T=>G	1	0	1
4341402: C=>T	1	0	1
4341402: T=>C	5	0	0.8
4340408: A=>G	1	0	1
4340330: G=>T	1	0	1
4341224: G=>C	1	0	1
PE-PGRS33			
2061661: Δ1bp	15	0	**0.0002**
2062105: Δ9bp	1	0	1
2062351: Δ1bp	1	0	1
PE-PGRS56			
3943650: Δ3bp, Δ1bp	12	0	**0.0027**
3944270: Δ9 bp	1	0	1
3943311: Δ9 bp	1	0	1
3941910: Δ9 bp	1	0	1
Rv0197			
234082: G=>A	1	0	1
234242: C=>T	1	0	1
234265: G=>T	1	0	1
234477: T=>G	1	1	1
234477: G=>T	16	8	**0.0179**
Rv2813-2814c			
3119737: Δ512 bp	1	0	1
3120432: +AGC, +AGCA	4	0	0.9998
3120031: Δ438 bp, Δ218 bp	12	3	0.2542
3120395: Δ2 bp	1	0	1
3119592: Δ292 bp	1	0	1
3120321: Δ74 bp	5	1	1
3119913: Δ1 bp	2	1	1
3119663: Δ221 bp, Δ74 bp	2	1	1
3119957: Δ438 bp	1	0	1
3120469: Δ1,725 bp	1	0	1
Rv2815-2816c			
3122774: Δ144bp	1	0	1
3122549: Δ72bp	2	1	1
3122847: Δ144bp	3	0	1
3123209: Δ75bp	1	1	1
3122122: Δ72bp, Δ350bp	4	0	0.9998
3121879: Δ2bp	1	1	1
3123284: Δ70bp, Δ142bp	8	1	0.4468

Table S3: *Description of validation set of strains including host risk factors.*

	Original dataset		Validation dataset	
	Clustered strains (n=66)	Unclustered strains (n=34)	Clustered strains (n=96)	Unclustered strains (n=47)
Publication source				
Farhat et al.[1]	-	-	50 (52%)	47 (100%)
Bryant et al.[2]	-	-	46 (48%)	-
Mutations called				
SNPs & indels	66 (100%)	34 (100%)	77 (80%)	47 (100%)
SNPs only	-	-	19 (20%)	-
Lineage				
1 (EAI)	22 (34%)	6 (18%)	4 (4%)	1 (2%)
2 (Beijing)	8 (12%)	1 (3%)	21 (22%)	12 (26%)
3 (CAS)	10 (15%)	8 (23%)	3 (3%)	11 (23%)
4 (EAM)	26 (39%)	19 (56%)	62 (65%)	18 (38%)
M. bovis	-	-	1 (1%)	-
Unclassified/T	-	-	5 (5%)	5 (11%)
Patient origin				
Europe	10 (15%)	16 (47%)	12 (13%)	12 (26%)
Africa	13 (20%)	8 (23.5%)	22 (23%)	12 (26%)
Asia	40 (61%)	2 (6%)	4 (4%)	18 (38%)
The Americas	3 (4%)	8 (23.5%)	9 (9%)	1 (2%)
Unknown	-		49 (51%)	4 (8%)
Drug resistance profile				
Susceptible	66 (100%)	34 (100%)	35 (36%)	-
Mono-resistant	-	-	5 (5%)	-
MDR	-	-	36 (38%)	37 (79%)
XDR	-	-	4 (4%)	-
Unknown	-	-	16 (17%)	10 (21%)
Gender				
Male	35 (53%)	23 (68%)	NA	
Female	31 (47%)	11 (32%)		
Age at diagnosis				
0-15	-	1 (3%)	NA	
16-30	25 (38%)	13 (38%)		
31-45	12 (18%)	15 (44%)		
46-60	15 (23%)	5 (15%)		
61-75	10 (15%)	-		
76-90	4 (6%)	-		
>90 years --	-	-		
Disease classification				
Pulmonary	36 (54.5%)	27 (79%)	NA	
Extrapulmonary	20 (30%)	1 (3%)		
Pulmonary + Extrapulmonary	10 (15.5%)	6 (18%)		
Smear positivity				
No	38 (58%)	6 (18%)	NA	
Yes	28 (42%)	28 (82%)		
Alcohol consumption				
No	66 (100%)	32 (94%)	NA	
Yes	-	2 (6%)		
Drug-use				
No	65 (98%)	29 (85%)	NA	
Yes	1 (2%)	5 (15%)		
Homelessness				
No	66 (100%)	33 (97%)	NA	
Yes		1 (3%)		
Traveler to endemic areas				
No	63 (95%)	34 (100%)	NA	
Yes	3 (5%)	-		

Table S4: *Protein prediction for TIMs associated to an increased clustering phenotype.*

			Protein Prediction	
Gene position	**Nucleotide change**	**Amino acid change**	**I-Mutant [#]**	**PolyPhen [&]**
espE				
61	**T=>G**	**L21V**	Large Decrease of Stability	NA
139	A=>G	M47V	Large Decrease of Stability	NA
955	G=>C	V319L	Large Decrease of Stability	NA
1100	T=>G	L367R	Large Decrease of Stability	NA
1133	**T=>C**	**V378A**	Large Decrease of Stability	NA
Rv0197				
344	G=>T	G115V	No effect	Probably damaging
1334	G=>T	R445L	No effect	Probably damaging
1852	G=>A	V618M	Large Decrease of Stability	Probably damaging
2012	C=>T	A671V	No effect	Probably damaging
2035	G=>T	E679 (STOP)	*Prediction not possible*	*Prediction not possible*
2038	G=>T	V680F	Large Decrease of Stability	No effect
2247	**T=>G**	**Y749 (STOP)**	*Prediction not possible*	*Prediction not possible*

NA: No homologs of espE were found therefore protein prediction was not possible.
Entries in bold denote that backwards mutations of these polymorphisms also occurred.
[#] I-mutant predicts free energy changes of protein stability upon a point mutation under different conditions
[&] PolyPhen predicts the possible impact of an amino acid substitution on the structure and function of a human protein using straightforward physical and comparative considerations.

Table S5: *Count of TIMs in the six genes or intergenic regions associated with clustering across the 19 strains selected for functional validation studies.*

		TIMs in individual genes or intergenic regions of interest (n)					
Sequencing ID	**Lineage**	**espE**	**PE-PGRS33**	**PE-PGRS56**	**Rv0197**	**Rv2813-2814c**	**Rv2815-2816c**
Clustered							
A4	1	1	1	1	0	0	0
B12	1	0	1	0	1	0	0
B6	1	1	0	0	0	0	0
C8	1	1	1	0	1	0	0
D11	1	0	1	0	2	1	0
D3	1	0	0	2	0	0	1
D4	1	1	1	0	1	0	0
D6	1	0	1	1	0	0	0
E2	1	0	0	0	0	2	0
E7	1	0	0	1	1	0	0
F6	1	0	0	0	0	0	0
F8	1	0	1	0	0	5	0
G11	1	1	1	1	1	0	1
X1	4	1	0	0	0	0	0
H4	4	1	0	0	1	0	0
Non-clustered							
A12	1	0	0	0	0	0	0
C10	1	0	0	0	0	0	0
D5	4	0	0	0	0	1	0
F10	4	0	0	0	0	0	1

Table S6: *Secondary analysis of the results of the immunological experiments. Summary statistics of each assay (left panel), detailed statistics of the mixed model (middle panel) and predicted margins (right panel).*

		Assay characteristics			Mixed model results of presence or absence of TIM on assay					Mixed model predicted margins	
Assay ~ Gene or intergenic region		**mean**	**SD**	**unit**	**standardized beta (se)**		**p[&]**	**Δ%**	**(95 % CI)[#]**	**no TIM**	**with TIM**
Monocyte cytokines											
IL-10 (24 h)	H37Rv	5.99	0.30	Ln(pg/mL)							
	clinical isolates	6.14	0.44	Ln(pg/mL)							
espE					-0.67	(0.12)	**1.72×10^{-8}**	- 26	(-35 – -15)	6.16	5.87
PE-PGRS33					-0.45	(0.14)	**0.0019**	- 18	(-31 – -3)	6.16	5.96
PE-PGRS56					-0.04	(0.15)	0.7938	- 2	(-18 – 17)	6.15	6.13
Rv0197					-0.26	(0.13)	0.0489	- 11	(-24 – 4)	6.15	6.04
Rv2813-14c					0.39	(0.15)	**0.0082**	19	(0 – 41)	6.14	6.31
Rv2815-16c					-0.02	(0.18)	0.8905	- 1	(-19 – 22)	6.15	6.13
IL-1Ra (24 h)	H37Rv	9.72	0.65	Ln(pg/mL)							
	clinical isolates	9.83	0.49	Ln(pg/mL)							
espE					-0.19	(0.12)	0.1193	- 9	(-22 – 7)	9.83	9.74
PE-PGRS33					-0.21	(0.14)	0.1417	- 10	(-25 – 9)	9.83	9.73
PE-PGRS56					-0.10	(0.15)	0.5336	- 5	(-22 – 17)	9.83	9.78
Rv0197					-0.14	(0.13)	0.3000	- 7	(-21 – 11)	9.83	9.76
Rv2813-14c					-0.08	(0.15)	0.5875	- 4	(-21 – 16)	9.83	9.79
Rv2815-16c					0.10	(0.18)	0.5659	5	(-17 – 32)	9.82	9.87
IL-1β (24 h)	H37Rv	7.75	0.78	Ln(pg/mL)							
	clinical isolates	8.57	0.67	Ln(pg/mL)							
espE					-0.20	(0.12)	0.0951	- 13	(-29 – 8)	8.58	8.45
PE-PGRS33					0.02	(0.14)	0.8863	1	(-22 – 31)	8.57	8.59
PE-PGRS56					-0.04	(0.15)	0.8067	- 3	(-26 – 28)	8.57	8.55
Rv0197					0.07	(0.13)	0.5860	5	(-17 – 33)	8.57	8.62
Rv2813-14c					0.39	(0.15)	**0.0077**	30	(0 – 69)	8.56	8.83
Rv2815-16c					-0.10	(0.18)	0.5746	- 6	(-32 – 28)	8.57	8.51
IL-6 (24 h)	H37Rv	9.86	0.48	Ln(pg/mL)							
	clinical isolates	9.91	0.45	Ln(pg/mL)							
espE					-0.09	(0.12)	0.4703	- 4	(-16 – 11)	9.92	9.88
PE-PGRS33					-0.06	(0.14)	0.6855	- 3	(-18 – 16)	9.91	9.89
PE-PGRS56					-0.10	(0.15)	0.4986	- 5	(-20 – 15)	9.91	9.87
Rv0197					-0.08	(0.13)	0.5717	- 3	(-17 – 13)	9.91	9.88
Rv2813-14c					-0.02	(0.15)	0.9046	- 1	(-17 – 18)	9.91	9.90
Rv2815-16c					0.07	(0.18)	0.6820	3	(-16 – 27)	9.91	9.94

Assay ~ Gene or intergenic region		Assay characteristics mean	SD	unit	Mixed model results of presence or absence of TIM on assay standardized beta (se)		p[&]	Δ%	(95 % CI)[#]	Mixed model predicted margins no TIM	with TIM
TNF-α (24 h)	H37Rv	5.81	0.67	Ln(pg/mL)							
	clinical isolates	6.50	0.64	Ln(pg/mL)							
espE					-0.21	(0.12)	0.0794	- 13	(-28 – 7)	6.51	6.38
PE-PGRS33					0.19	(0.14)	0.1998	13	(-12 – 44)	6.49	6.61
PE-PGRS56					0.14	(0.15)	0.3505	10	(-15 – 42)	6.50	6.59
Rv0197					0.14	(0.13)	0.3030	9	(-13 – 36)	6.49	6.58
Rv2813-14c					0.44	(0.15)	**0.0025**	33	(4 – 70)	6.49	6.77
Rv2815-16c					-0.05	(0.18)	0.7604	- 3	(-28 – 30)	6.50	6.47
TNF-α (4 h)	H37Rv	5.86	0.76	Ln(pg/mL)							
	clinical isolates	6.30	0.63	Ln(pg/mL)							
espE					-0.32	(0.12)	**0.0080**	- 18	(-33 – 0)	6.31	6.11
PE-PGRS33					0.03	(0.14)	0.8377	2	(-20 – 30)	6.30	6.31
PE-PGRS56					0.28	(0.15)	0.0663	20	(-8 – 55)	6.29	6.47
Rv0197					-0.01	(0.13)	0.9170	- 1	(-21 – 24)	6.30	6.29
Rv2813-14c					0.40	(0.15)	**0.0062**	29	(1 – 65)	6.29	6.54
Rv2815-16c					-0.41	(0.18)	0.0199	- 23	(-43 – 4)	6.30	6.04
T-cell cytokines											
IFN-γ (7 d)	H37Rv	5.22	0.85	Ln(pg/mL)							
	clinical isolates	5.57	1.09	Ln(pg/mL)							
espE					0.02	(0.12)	0.8346	3	(-26 – 43)	5.57	5.60
PE-PGRS33					-0.08	(0.12)	0.5262	- 8	(-36 – 31)	5.59	5.50
PE-PGRS56					-0.38	(0.12)	**0.0016**	- 34	(-54 – -7)	5.62	5.20
Rv0197					-0.01	(0.11)	0.9602	- 1	(-28 – 38)	5.58	5.57
Rv2813-14c					0.06	(0.13)	0.6308	7	(-27 – 58)	5.57	5.64
Rv2815-16c					-0.33	(0.15)	0.0254	- 30	(-54 – 7)	5.60	5.24
IL-17 (7 d)	H37Rv	5.78	1.43	Ln(pg/mL)							
	clinical isolates	5.20	1.34	Ln(pg/mL)							
espE					-0.03	(0.12)	0.8123	- 4	(-37 – 47)	5.21	5.17
PE-PGRS33					-0.16	(0.13)	0.2229	- 19	(-48 – 28)	5.23	5.02
PE-PGRS56					-0.05	(0.13)	0.6655	- 7	(-41 – 45)	5.21	5.13
Rv0197					-0.06	(0.12)	0.5930	- 8	(-39 – 39)	5.21	5.13
Rv2813-14c					-0.08	(0.14)	0.5868	- 10	(-45 – 48)	5.21	5.11
Rv2815-16c					0.03	(0.15)	0.8517	4	(-39 – 78)	5.20	5.24

Assay ~ Gene or intergenic region		Assay characteristics mean	SD	unit	Mixed model results of presence or absence of TIM on assay standardized beta (se)		p[&]	Δ%	(95 % CI)[#]	Mixed model predicted margins no TIM	with TIM
IL-22 (7 d)	H37Rv	5.72	0.98	Ln(pg/mL)							
	clinical isolates	6.12	1.15	Ln(pg/mL)							
espE					0.01	(0.12)	0.9238	1	(-29 – 44)	6.12	6.14
PE-PGRS33					0.03	(0.12)	0.8284	3	(-29 – 50)	6.12	6.15
PE-PGRS56					-0.19	(0.12)	0.1229	- 19	(-44 – 16)	6.14	5.93
Rv0197					0.13	(0.11)	0.2473	16	(-18 – 64)	6.10	6.25
Rv2813-14c					0.14	(0.13)	0.3130	17	(-22 – 75)	6.11	6.27
Rv2815-16c					-0.13	(0.15)	0.3622	- 14	(-45 – 34)	6.13	5.98
PMN assays											
Early apoptosis (6 h)	H37Rv	33.05	11.27	%							
	clinical isolates	39.4	11.6	%							
espE					0.10	(0.15)	0.5032	3	(-8 – 14)	39.1	40.3
PE-PGRS33					0.26	(0.16)	0.0968	8	(-5 – 20)	39.1	42.1
PE-PGRS56					0.23	(0.17)	0.1883	7	(-7 – 20)	39.0	41.6
Rv0197					-0.05	(0.14)	0.7380	- 1	(-13 – 10)	39.4	38.8
Rv2813-14c					0.51	(0.18)	**0.0036**	15	(1 – 29)	38.9	44.9
Rv2815-16c					-0.12	(0.18)	0.5137	- 4	(-18 – 11)	39.4	38.0
Late apoptosis (6 h)	H37Rv	10.03	5.09	%							
	clinical isolates	28.7	14.3	%							
espE					-0.28	(0.15)	0.0551	- 14	(-33 – 5)	29.0	25.0
PE-PGRS33					-0.37	(0.16)	0.0187	- 18	(-38 – 2)	29.2	23.9
PE-PGRS56					0.02	(0.17)	0.9191	1	(-22 – 24)	28.4	28.7
Rv0197					-0.25	(0.14)	0.0781	- 13	(-32 – 6)	29.0	25.3
Rv2813-14c					-0.12	(0.18)	0.4994	- 6	(-29 – 17)	28.6	26.9
Rv2815-16c					0.28	(0.18)	0.1305	14	(-11 – 39)	28.2	32.2
ROS (1 h)	H37Rv	16.99	0.44	Ln(AUC)							
	clinical isolates	16.97	0.62	Ln(AUC)							
espE					-0.08	(0.15)	0.6043	- 5	(-25 – 21)	16.97	16.92
PE-PGRS33					-0.44	(0.15)	**0.0041**	- 24	(-41 – -2)	17.01	16.74
PE-PGRS56					-0.34	(0.17)	0.0463	- 19	(-38 – 7)	16.98	16.77
Rv0197					-0.10	(0.14)	0.4875	- 6	(-26 – 19)	16.97	16.91
Rv2813-14c					-0.60	(0.17)	**0.0005**	- 31	(-48 – -9)	16.99	16.62
Rv2815-16c					0.16	(0.19)	0.4144	10	(-20 – 52)	16.96	17.06

Table S7: *Summary of relevant experimental findings on the functions of the six TIMs.*

Gene	Experimental findings
Rv0197 codes for a possible oxidoreductase.	Contains the binding motif for molybdenum cofactor, a key component in TB pathogenesis.[3] Upregulated during higB expression in M. tuberculosis H37Rv (important for bacterial survival under stress conditions encountered during infection).[4]
espE codes for an ESX-1 secretion-asociated protein.	Essential for survival of *M. tuberculosis* in C57BL/6J mouse macrophage.[5] Rv3616c, a homologue of espE is essential for in vivo survival of M. tuberculosis in C57BL/6J mice.[6] Homologue of Rv3864 (espA) is an Esx-1 substrate required for virulence of *M. tuberculosis* in C57BL/6J and BALB/C-SCID mice.[7]
PE-PGRS33 encodes a suface exposed protein.	Enhances survival of *M. smegmatis* and induces induces necrosis in macrophages of C57BL/6 mice.[8] PE-PGRS33 protein co-localises to the host mitochondria of T-Rex cell lines and induces apoptosis and necrosis.[9] Variations in the polymorphic repeats of the PGRS domain of *M. smegmatis* attenuate the gene's TNF-α inducing ability.[10]
PE-PGRS56 no data on function	One of the 10 most dominant *M. tuberculosis* H37Rv proteins found within both 30- and 90-day infected guinea pig lung samples.[11]
Rv2813-14c not coding, intergenic region.	No published data.
Rv2815-2816c contains promotor to both Rv2814c and Rv2815c	Rv2815 is implied in induction of PI-measured cell death.[12]

REFERENCES TO SUPPLEMENTARY MATERIAL

1. Farhat, M. R., Shapiro, B. J., Sheppard, S. K., Colijn, C. & Murray, M. A phylogeny-based sampling strategy and power calculator informs genome-wide associations study design for microbial pathogens. 1–14 (2014). doi:10.1186/s13073-014-0101-7

2. Bryant, J. M. *et al.* Inferring patient to patient transmission of *Mycobacterium tuberculosis* from whole genomesequencing data. *BMC Infectious Diseases* **13,** 1–1 (2013).

3. Williams, M., Mizrahi, V. & Kana, B. D. Molybdenum cofactor: A key component of *Mycobacterium tuberculosis*pathogenesis? *Critical Reviews in Microbiology* **40,** 18–29 (2014).

4. Schuessler, D. L. *et al.* Induced ectopic expression of HigB toxin in *Mycobacterium tuberculosis*results in growth inhibition, reduced abundance of a subset of mRNAs and cleavage of tmRNA. *Mol. Microbiol.* n/a–n/a (2013). doi:10.1111/mmi.12358

5. Rengarajan, J., Bloom, B. R. & Rubin, E. J. Genome-wide requirements for *Mycobacterium tuberculosis* adaptation and survival in macrophages. *Proc. Natl. Acad. Sci. U.S.A.* **102,** 8327–8332 (2005).

6. Sassetti, C. M., Boyd, D. H. & Rubin, E. J. Genes required for mycobacterial growth defined by high density mutagenesis. *Mol. Microbiol.* **48,** 77–84 (2003).

7. Fortune, S. M. *et al.* Mutually dependent secretion of proteins required for mycobacterial virulence. *Proc. Natl. Acad. Sci. U.S.A.* **102,** 10676–10681 (2005).

8. Dheenadhayalan, V., Delogu, G. & Brennan, M. J. Expression of the PE_PGRS 33 protein in Mycobacterium smegmatis triggers necrosis in macrophages and enhanced mycobacterial survival. *Microbes and Infection* **8,** 262–272 (2006).

9. Cadieux, N. *et al.* Induction of cell death after localization to the host cell mitochondria by the *Mycobacterium tuberculosis* PE_PGRS33 protein. *Microbiology (Reading, Engl.)* **157,** 793–804 (2011).

10. Basu, S. *et al.* Execution of macrophage apoptosis by PE_PGRS33 of *Mycobacterium tuberculosis* is mediated by Toll-like receptor 2-dependent release of tumor necrosis factor-alpha. *J. Biol. Chem.* **282,** 1039–1050 (2007).

11. Kruh, N. A., Troudt, J., Izzo, A., Prenni, J. & Dobos, K. M. Portrait of a Pathogen: The *Mycobacterium tuberculosis* Proteome In Vivo. *PLoS ONE* **5,** e13938 (2010).

12. Zhang, H. *et al. Mycobacterium tuberculosis* Rv2185c contributes to nuclear factor-κB activation. *Mol. Immunol.* **66,** 147–153 (2015).

Chapter VI

General Discussion

6.1 SYNTHESIS OF STUDIES

Many of the studies to date on *M. tuberculosis* (Mtb) transmission have focused on host risk factors. The study of different Mtb lineages have revealed measurable differences in virulence and immune responses and suggested differences in patient-to-patient transmissibility independent of host factors (Reiling et al. 2013; Sarkar et al. 2012; Krishnan et al. 2011). However, to date there has not been a systematic study of the genetic determinants of differences in transmissibility.

This thesis begins with the description of a novel method - the Propensity to Propagate (PPP) - to adjust for host risk factors when quantifying transmissibility, as described in Chapter 2. An overall scoring of host risk factors behind a cluster of strains (CPP) was found to not be equal across four phylogenetic lineages in the Netherlands. This finding emphasizes the importance of controlling for host-related factors in order to attain comparability in measuring the ability of difference strains/lineages of Mtb to propagate. Many studies on investigating the relationship between phylogenetic lineages and transmissibility actually do not control for host risk factors (Table 1).

Applying the PPP method to the large database of molecular-typed strains in the Netherlands revealed no significant differences in average cluster size of four different phylogenetic lineages, which goes against our hypothesis that a bacterial factor, such as phylogenetic lineage, also accounts for differences in transmissibility. However, we did find evidence of phylogenetic lineages influencing more specific indices of transmissibility, namely, a decreased ability to infect and a lower secondary case rate in ancient phylogenetic lineages (*M. africanum* and EAI) compared to their modern counterparts (Euro-American, Beijing, and CAS). An indication that phylogenetic lineages may be more transmissible via different mechanisms (increased ability to enter the lungs of hosts versus an ability to more rapidly progress to disease) is relevant in that each calls for different control strategies. A more physical method such as quarantining, for example, might be more effective for strains known to be more infective, while a greater focus on starting adequate therapy as early as possible is crucial for strains known to progress faster to disease.

A survey of the literature shows no real consensus to date regarding potential differences in ability to transmit by phylogenetic lineage. As Table 1 demonstrates, studies greatly vary by time span of sample collection, sample size, lineages compared (including which lineage was used as the reference) and which host risk factors were adjusted for. Even within the same country (Canada), three studies produce conflicting findings regarding the transmissibility of lineage 2 (Beijing), although they do agree on the lower transmissibility of lineage 1. In general, the geographically widespread lineage 2 (whose emergence is hypothetically linked to enhanced pathogenicity, leading to increased transmissibility and rapid progression from infection to active disease) (Buu et al. 2009; M. Hanekom et al. 2007) shows increased clustering in higher prevalence settings rather than in lower prevalence ones. The association between lineage 2 strains and drug resistance could account for this, as compensatory mutations have been associated with transmission of MDR Beijing genotype strains in China (Li et al. 2016). The lower clustering observed for lineage 1, as was also found in our study in Chapter 3, seems to span across both settings.

A first step towards discovering more specific genetic regions behind a particular phenotype involves checking for the absence/presence of mutations in the genes of interest between the two phenotypes. In Chapter 4, we looked at the frequency of frameshift-causing indels in Mycobacterium cyclase/LuxR-like genes (*mclx*s) across different phylogenetic lineages and, using a regression-based model, their association with patient-, disease- and microorganism-related factors, including transmissibility. While this approach is justified when functional studies strongly support the role of particular regions/genes behind a phenotype of interest, its main shortcoming is that it does not control for phylogeny (that is, distinguishing between having the same phenotype due to the repeated and independent emergence of recent mutations versus due to a more deep-rooted sharing of a common ancestor).

To disentangle adaptive loci from other mutations fixed in the clonal background, one can look for loci that have an excess of functional changes, such as the ratio of non-synonymous to synonymous mutations (dN:dS), or other metrics that explicitly measure deviations from the expected pattern of amino acid substitution (Shapiro et al. 2009). Evolutionary convergence analysis is a solution that

***Table 1**: Overview of studies investigating association between phylogenetic lineages and transmission.*

Transmission indicators	Time span of sample collection	Sample size	Lineages (L) compared	Host risk factors adjusted for	Setting	Prevalence	Findings	Reference
RFLP/spoligotyping clustering rate	16 years	1379	L2 and Others	Gender, age, ethnicity, sputum smear status, bacillary load, results of chest radiography, drug resistance	Canada	Low	Non-L2 strains were significantly more associated with transmission clusters than L2 strains (difference disappeared after adjusting for host risk factors).	(Klassen et al. 2013)
Secondary case-rate ratios	11 years	604	L2, L3 and L4	None	USA	Low	L4 strains were three times more likely to generate a secondary compared to non-L4 strains. The Indo-Oceanic lineage had a significantly lower secondary case-rate ratio and L2 the lowest.	(Gagneux et al. 2006)
RFLP/spoligotyping clustering rate TST conversion	11 years	678	L1 and Others (Ref.)	Age and probability of previous latent TB	Canada	Low	L1 associated with lower rates of transmission; L2 strains were found not to be associated with enhanced transmissibility.	(Albanna et al. 2011)
Spoligotyping and MIRU24 clustering	4 years	2016	L1, L2, L3, L4, L5, L6 and M. bovis BCG	Gender, age >65 years, FB, homeless, Aboriginal	Canada	Low	L2 and M. bovis/BCG were significantly associated with genotypic clustering; L1 and L4 with less.	(Tuite et al. 2013)
RFLP clustering	14 years	535 (all DR)	L2 and Others	None	Sweden	Low	No significant difference in clustering between L2 and non-L2 strains	(Ghebremichael et al. 2010)
VNTR clustering rate	2 years	274	Ancient L2, modern L2 and Others	None	Japan	Low	Clustering was significantly higher in modern L2 than ancient L2 and Others.	(Wada et al. 2009)
RFLP clustering rate	11 years	325	L2 sub-lineages 1 to 7	None	South Africa (urban setting)	High	Clustering was strongly associated with the L2 sublineages.	(Hanekom et al. 2007)
RFLP clustering rate	1 year	114	L2 and Others	None	Russia (prison)	High	A significantly higher proportion of L2 strains were clustered compared to non-L2.	(Toungoussova et al. 2003)
Clustering rate based on combination of RFLP, VNTR and spolityping	3 years	2207	L1 (Ref.), L2 and Others	Age, gender, residence, year of inclusion, TB treatment history, Resistance to streptomycin, ethambutol, and MDR	Vietnam	High	Clustering was associated with L1 compared to L2.	(Buu et al. 2012)
Genetic linkage (up to 10 SNPs difference)	3 years	1346	L1, L2, L3 and L4 (Ref.)	Age, gender, smear positivity, HIV status, previous TB, INH resistance, place of residence and birth place	Malawi	High	L2 and L3 strains were more likely to be clustered and in larger clusters and L1 strains were less likely to be clustered and were in smaller clusters.	(Guerra-Assuncao et al. 2015)

has been shown to be well suited for the study of clonal pathogens such as Mtb (Read & Massey 2014; Guerra-Assuncao et al. 2015). In a perfectly clonal population, genomes are related by a single phylogenetic tree, rather than a more complicated network structure that represents recombination. Alleles that arise independently multiple times in different branches (and are, thus, incongruous with the tree) stand out as candidate examples of convergent evolution. Thus after adjustment for patient factors using the PPP method, we selected 100 bacterial samples that were either highly or poorly transmissible and subjected them to whole genome sequencing and evolutionary convergence analysis. We identified 6 bacterial DNA regions - *espE*, *PE-PGRS33*, *PE-PGRS56*, Rv0197, Rv2813-14c and Rv2815-16c - to be associated with Mtb transmission and validated these regions by studying the response of human white blood cells to extracts from a subset of the Mtb that carried or did not carry mutations in these DNA regions.

It is interesting to note that the mutations associated to increased transmissibility described in Chapter 5, with the exception of those in Rv0197, were not observed in the successful lineage 2 strains. Since differences in dissemination and virulence of ancient versus modern Beijing lineages have been documented (Ribeiro et al. 2014), it is imperative that future phylogenetic studies looking into transmission distinguish between the two sub-lineages. A larger study allowing for a stratification of the analysis by phylogenetic lineages would be needed in order to assess targets of independent mutation (TIMs) that are universal across lineages and those that are specific. Since it is postulated that phylogenetic lineages have resulted from adaptation to hosts of the particular geographic location where they circulate (Gagneux et al. 2006), it is conceivable that mutations conferring increased transmissibility could also be lineage-specific. Quite recently, the presence of non-synonymous SNPs in putative genes coding for DNA repair enzymes *mutT2, mutT4, ogt* (postulated to confer a mutator phenotype to facilitate spreading of the pathogen) were detected with variable percentages in all of modern lineage 2 sublineages, but absent in ancient ones (Chang et al. 2011).

6.2 LIMITATIONS

The described PPP method in Chapter 2 for adjustment for host risk factors and designation of a CPP scoring per cluster only does so approximately. The future inclusion of additional epidemiological and clinical factors, such as degree of exposure to index case (Bailey et al. 2002; Marks et al. 2000), or presence of pulmonary cavitation in the host (Jones-López et al. 2014), can further improve the calculation of CPP and hence more accurately pinpoint those strains that are more transmissible due to bacterial factors. It is however theoretically impossible to fully control for a host's propensity to propagate, since genetic factors and biomarkers for susceptibility and progression to disease are still actively being discovered and tend to be controversial (Minchella et al. 2015; Elliott et al. 2015; Jiang et al. 2015; Salem & Gros 2013). In addition, the highly skewed distribution of genotypic cluster sizes found in the Netherlands has been suggested as a sign of superspreading (Ypma et al. 2013). This heterogeneity in the number of secondary cases caused per infectious individual is currently not accounted for in the calculation of CPPs using the geometric mean of PPPs. As such, the existence of one superspreader with a particularly high PPP can offset the overall CPP scoring of a cluster composed of hosts with otherwise relatively low PPPs.

The challenge of identifying bacterial factors responsible for increased transmissibility also stems from the ambiguity in the definition and quantification of the "transmissibility" phenotype itself. Unlike with drug resistance, where the endpoint to be measured is clearly defined (i.e. minimum inhibitory concentrations), the widely used clustering rates from molecular typing can be calculated at different resolutions (i.e. in Wada et al., a VNTR cluster is defined as two or more isolates sharing 19 identical VNTR alleles, while in our study we used a stricter definition of all 24 identical VNTR alleles) (Wada et al. 2009; Nebenzahl-Guimaraes et al. 2015). The rough quantification of "transmissibility" can also be broken down into further specific components, such as the ability of the bacteria to spread (infectivity) followed by the likelihood of breaking down to disease in the host (pathogenicity, or virulence), as described in Chapter 3. How exactly the latter is related to transmissibility is yet unclear. More virulent strains could lead to large clusters if virulence was associated with increased transmission rates or increased rates of disease after host infection (Valway et al. 1998). However, more virulent strains

could have less opportunity to transmit if the severity of symptoms leads to early treatment or death, thus reducing the duration of the infectious period.

There is also no “gold standard” “non-transmissible” strain to compare more transmissible strains to because it is currently unknown how transmissible the reference strain selected for the study in chapter 5 (H37Rv) really is. For example, in a rabbit inhalation model evaluated by Lurie’s Pulmonary Tubercle Count Method, H37Rv tubercles were found to be larger and contain more bacilli than a CDC1551 strain, a clinical isolate reported to be hypervirulent and to grow faster than other isolates (Bishai et al. 1999).

It should be noted that the methods to quantify tuberculin-skin test (TST) conversion and number of secondary cases following the index case as applied in Chapter 3 are influenced by programmatic factors, such as the thoroughness of contact investigations, and the underlying proportion of the population that is BCG vaccinated (Menzies 2000). Whilst stratifying by ethnicity can certainly help reduce the latter bias, even in the Netherlands where a solid TB control program is in place, there is still room for improvement for contact investigation in particular risk factor groups, such as immigrant patients (Mulder et al. 2011; Mulder et al. 2012).

6.3 IMPLICATIONS FOR FURTHER RESEARCH

The study findings support the following research priorities:

Conduct a prospective validation study on cohort transmission phenotypes

A prospective household contact study including TST conversion data, which would be far more accurate at ascertaining instances of transmission, is warranted to confirm our findings on differences in transmissibility across phylogenetic lineages. In terms of improving the transmissibility phenotype by adjusting for host risk factors, more data not only on index but also secondary cases would allow us to adjust for patient risk factors across the transmission chain i.e. rates of latent TB treatment and existing medical risk factors in secondary cases (Kumari & Meena 2014; Ayele et al. 2015), which could influence the likelihood of progression to disease or susceptibility to infection of the host, respectively.

Expand evolutionary convergence analysis

Our study applying evolutionary convergence analysis on a limited set of 100 WGS strains can be built upon in various ways. To begin with, we could easily increase the statistical leverage by improving the selection of strains to compare (for example, by pairing closely related strains) (Farhat et al. 2014). Targeted sequencing, using molecular inversion probe technology, of the identified candidate TIMs instead of WGS of entire strains would also allow for a larger sample size.

Given increased statistical power, the analysis could also be repeated looking at more specific phenotypic outcomes, such as either infectivity or pathogenicity i.e. comparing strains from patients with a high number of positive TST contacts to those with none, regardless of the final cluster size to which they belong to. In this way we would be able to identify SNPs/genes specifically related to the bacteria's ability to spread (infectivity).

Initiate validation experiments using mutant strains and animal models

The *in vitro* studies on cytokine and neutrophil responses in Chapter 5 were performed merely to provide some independent biological support for molecular-epidemiological associations. Due to the co-occurrence of TIMs in the strains used for the immunological validation experiments, we could not definitely discern what genetic variation was responsible for which effect. Demonstrating a causal association between the genes and transmissibility would involve directly manipulating the bacteria i.e. creating knockouts of one (or more) of the putative TIMs in a fixed reference strain and measuring the outcome in phenotype, much in the same way as has been done in allelic exchange experiments of mutations conferring drug resistance (Appendix 1).

Another avenue through which to validate the candidate TIMs is via experiments in animal models. Several reports have demonstrated that the severity and clinical manifestations of TB depend on differences in the immunogenicity and pathogenicity of the infecting strains of Mtb (Manca et al. 2001; Dormans et al. 2004; López et al. 2003; Malik & Godfrey-Faussett 2005). Progressive pulmonary TB by the intratracheal route or aerosol inhalation in different mouse strains have been used to examine

the course of infection in terms of strain virulence (mouse survival, lung bacillary load, histopathology) and immune responses (cytokine expression determined by real-time PCR). A newer mouse model of transmissibility consisting of prolonged cohousing (up to 60 days) of infected and naıve animals has been used to assess the ability of strains to be transmitted, measuring lung bacillus loads of the naıve animal and cutaneous delayed type hypersensitivity (DTH) against mycobacterial antigens as markers of disease and transmission (Marquina-Castillo et al. 2009). In this study they reassuringly found that rapid death, higher bacterial loads, more tissue damage, immunological responses consistent with a Th2 response and transmission of infection to contact animals correlated with indicators of transmission in the community, such as size of cluster and TST reactivity or rapid progression to disease among household contacts. Guinea pigs have also often been used in studies on infectivity because they are highly susceptible to infection by human Mtb, more so than mice and rabbits and perhaps as much as AIDS patients; in addition, their immunological response seems to be more similar to that of humans than mice (Young 2009). The big limitation around these models is that none of these animals transmit TB efficiently, unlike bovine strains and other animal adapted strains (such as *Mycobacterium microti* in voles) that are transmitted in the wild (Dharmadhikari & Nardell 2008). Thus, a small-animal model that can be used to study TB transmission has yet to be developed.

6.4 REFERENCES

Albanna, A.S. et al., 2011. Reduced transmissibility of East African Indian strains of *Mycobacterium tuberculosis*. *PloS ONE*, 6(9), p.e25075.

Ayele, H.T. et al., 2015. Isoniazid prophylactic therapy for the prevention of tuberculosis in HIV infected adults: A systematic review and meta-analysis of randomized trials. *PLoS ONE*, 10(11).

Bailey, W.C. et al., 2002. Predictive model to identify positive tuberculosis skin test results during contact investigations. *Jounal of the American Medical Association*, 287(8), p.996–1002.

Bishai, W.R. et al., 1999. Virulence of *Mycobacterium tuberculosis* CDC1551 and H37Rv in rabbits evaluated by Lurie's pulmonary tubercle count method. *Infection and Immunity*, 67(9), p.4931–4934.

Buu, T.N. et al., 2012. Increased transmission of *Mycobacterium tuberculosis* Beijing genotype strains associated with resistance to streptomycin: a population-based study. *PloS one*, 7(8), p.e42323.

Buu, T.N. et al., 2009. The Beijing genotype is associated with young age and multidrug-resistant tuberculosis in rural Vietnam. *International Journal of Tuberculosis and Lung Disease*, 13(7), p.900–906.

Chang, J.R. et al., 2011. Genotypic analysis of genes associated with transmission and drug resistance in the Beijing lineage of *Mycobacterium tuberculosis*. *Clinical Microbiology and Infection*, 17(9), p.1391–1396.

Dharmadhikari, A.S. & Nardell, E.A., 2008. What animal models teach humans about tuberculosis. *American Journal of Respiratory Cell and Molecular Biology*, 39(5), p.503–508.

Dormans, J. et al., 2004. Correlation of virulence, lung pathology, bacterial load and delayed type hypersensitivity responses after infection with different *Mycobacterium tuberculosis* genotypes in a BALB/c mouse model. *Clinical and Experimental Immunology*, 137(3), p.460–468.

Elliott, T.O.J.P. et al., 2015. Dysregulation of apoptosis is a risk factor for tuberculosis disease progression. *Journal of Infectious Diseases*, 212(9), p.1469–1479.

Farhat, M.R. et al., 2014. A phylogeny-based sampling strategy and power calculator informs genome-wide associations study design for microbial pathogens. *Genome Medicine*, 6(11), p.101.

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), p.2869–2873.

Ghebremichael, S. et al., 2010. Drug resistant *Mycobacterium tuberculosis* of the Beijing genotype does not spread in Sweden. *PLoS One*, 5(5), p.e10893.

Guerra-Assuncao, J.A. et al., 2015. Recurrence due to relapse or reinfection with *Mycobacterium tuberculosis*: A whole-genome sequencing approach in a large, population-based cohort with a high HIV infection prevalence and active follow-up. *Journal of Infectious Diseases*, 211(7), p.1154–1163.

Guerra-Assuncao, J. et al., 2015. Large-scale whole genome sequencing of *M. tuberculosis* provides insights into transmission in a high prevalence area. *Elife*, 4.

Hanekom, M. et al., 2007. A recently evolved sublineage of the *Mycobacterium tuberculosis* Beijing strain family is associated with an increased ability to spread and cause disease. *Journal of Clinical Microbiology*, 45(5), p.1483–1490.

Hanekom, M. et al., 2007. Evidence that the Spread of *Mycobacterium tuberculosis* Strains with the Beijing Genotype Is Human Population Dependent. *Journal of Clinical Microbiology*, 45(7), p.2263–2266.

Jiang, D. et al., 2015. The variations of IL-23R are associated with susceptibility and severe clinical forms of pulmonary tuberculosis in Chinese Uygurs. *BMC infectious diseases*, 15(1), p.550.

Jones-López, E.C. et al., 2014. Importance of Cough and M. tuberculosis Strain Type as Risks for Increased Transmission within Households. *PloS one*, 9(7), p.e100984.

Klassen, D.L.- et al., 2013. Transmission of *Mycobacterium tuberculosis* Beijing strains, Alberta, Canada, 1991-2006. *Centers for Disease Control and Prevention*, 19(5), p.1–16.

Krishnan, N. et al., 2011. *Mycobacterium tuberculosis* lineage influences innate immune response and virulence and is associated with distinct cell envelope lipid profiles. *PLoS ONE*, 6(9).

Kumari, P. & Meena, L.S., 2014. Factors Affecting Susceptibility to *Mycobacterium tuberculosis*: a Close View of Immunological Defence Mechanism. *Applied Biochemistry and Biotechnology*, 174(8), p.2663–2673.

Li, Q. et al., 2016. Compensatory Mutations of Rifampicin Resistance Are Associated with transmission of multidrug resistant *Mycobacterium tuberculosis* Beijing genotype strains in China. *Antimicrobial Agents and Chemothery*, pii:AAC.02358-15.

López, B. et al., 2003. A marked difference in pathogenesis and immune response induced by different *Mycobacterium tuberculosis* genotypes. *Clinical and Experimental Immunology*, 133(1), p.30–37.

Malik, A.N.J. & Godfrey-Faussett, P., 2005. Effects of genetic variability of *Mycobacterium tuberculosis* strains on the presentation of disease. *The Lancet infectious diseases*, 5(3), p.174–183.

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), p.5752–5757.

Marks, S.M. et al., 2000. Outcomes of contact investigations of infectious tuberculosis patients. *American Journal of Respiratory and Critical Care Medicine*, 162(6), p.2033–2038.

Marquina-Castillo, B. et al., 2009. Virulence, immunopathology and transmissibility of selected strains of *Mycobacterium tuberculosis* in a murine model. *Immunology*, 128(1), p.123–133.

Menzies, D., 2000. What does tuberculin reactivity after bacille Calmette-Guérin vaccination tell us? *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, 31 Suppl 3, p.S71–S74.

Minchella, P.A. et al., 2015. Iron homeostasis and progression to pulmonary tuberculosis disease among household contacts. *Tuberculosis (Edinburgh, Scotland)*, 95(3), p.288–93.

Mulder, C. et al., 2012. Adherence by Dutch Public Health Nurses to the National Guidelines for Tuberculosis Contact Investigation. *PLoS ONE*, 7(11).

Mulder, C. et al., 2011. Coverage and yield of tuberculosis contact investigations in the Netherlands. *International Journal of Tuberculosis and Lung Disease*, 15(12), p.1630–1636.

Nebenzahl-Guimaraes, H. et al., 2015. Transmission and progression to disease of *mycobacterium tuberculosis* phylogenetic lineages in the Netherlands. *Journal of Clinical Microbiology*, 53(10), p.3264–3271.

Read, T.D. & Massey, R.C., 2014. Characterizing the genetic basis of bacterial phenotypes using genome-wide association studies: a new direction for bacteriology. *Genome Medicine*, 6(11), p.109.

Reiling, N. et al., 2013. Clade-specific virulence patterns of *Mycobacterium tuberculosis* complex strains in human primary macrophages and aerogenically infected mice. *mBio*, 4(4).

Ribeiro, S.C.M. et al., 2014. *Mycobacterium tuberculosis* strains of the modern sublineage of the beijing family are more likely to display increased virulence than strains of the ancient sublineage. *Journal of Clinical Microbiology*, 52(7), p.2615–2624.

Gagneux, S., 2006. Variable host – pathogen compatibility in *Mycobacterium tuberculosis*. *Proceedings of the National Academy of Sciences of the United States of America*, 103(8), p.2869–2873.

Salem, S. & Gros, P., 2013. Genetic determinants of susceptibility to mycobacterial infections: IRF8, a new kid on the block. *Advances in Experimental Medicine and Biology*, 783, p.45–80.

Sarkar, R. et al., 2012. Modern lineages of *Mycobacterium tuberculosis* exhibit lineage-specific patterns of growth and cytokine induction in human monocyte-derived macrophages. *PLoS ONE*, 7(8).

Shapiro, B.J. et al., 2009. Looking for Darwin's footprints in the microbial world. *Trends in microbiology*, 17(5), p.196–204.

Toungoussova, O.S. et al., 2003. Molecular epidemiology and drug resistance of *Mycobacterium tuberculosis* isolates in the Archangel prison in Russia: predominance of the W-Beijing clone family. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, 37(5), p.665–672.

Tuite, A.R. et al., 2013. Epidemiological evaluation of spatiotemporal and genotypic clustering of *mycobacterium tuberculosis* in Ontario, Canada. *International Journal of Tuberculosis and Lung Disease*, 17(10), p.1322–1327.

Valway, S.E. et al., 1998. An outbreak involving extensive transmission of a virulent strain of *Mycobacterium tuberculosis*. *New England Journal of Medicine*, 338(10), p.633–639.

Wada, T. et al., 2009. High transmissibility of the modern Beijing *Mycobacterium tuberculosis* in homeless patients of Japan. *Tuberculosis*, 89(4), p.252–255.

Young, D., 2009. Animal models of tuberculosis. *European Journal of Immunology.*, 39(8), p.2011–2014.

Ypma, R.F. et al., 2013. A sign of superspreading in tuberculosis: highly skewed distribution of genotypic cluster sizes. Epidemiology, 24(3), p.395–400.

Appendix

Journal of Antimicrobial Chemotherapy Advance Access published September 20, 2013

J Antimicrob Chemother
doi:10.1093/jac/dkt358

Journal of Antimicrobial Chemotherapy

Systematic review of allelic exchange experiments aimed at identifying mutations that confer drug resistance in *Mycobacterium tuberculosis*

Hanna Nebenzahl-Guimaraes[1–4]*†, Karen R. Jacobson[5,6]†, Maha R. Farhat[7] and Megan B. Murray[1,8]

[1]Department of Epidemiology, Harvard School of Public Health, Boston, MA 02115, USA; [2]National Institute for Public Health and the Environment (RIVM), Bilthoven, The Netherlands; [3]Life and Health Sciences Research Institute (ICVS), School of Health Sciences, University of Minho, Braga, Portugal; [4]ICVS/3B's, PT Government Associate Laboratory, Braga/Guimarães, Portugal; [5]Section of Infectious Diseases, Boston University School of Medicine and Boston Medical Center, Boston, MA 02118, USA; [6]DST/NRF Centre of Excellence for Biomedical TB Research/MRC Centre for Molecular Cellular Biology, Division of Molecular Biology and Human Genetics, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa; [7]Pulmonary and Critical Care Division, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02114, USA; [8]Department of Global Health and Social Medicine, Harvard Medical School, Boston, MA 02115, USA

*Corresponding author. National Institute for Public Health and the Environment (RIVM), PO Box 1, 3720 BA, Bilthoven, The Netherlands.
Tel: +31618333976; E-mail: hanna.guimaraes@gmail.com
†These authors contributed equally.

Received 4 June 2013; returned 18 July 2013; revised 1 August 2013; accepted 13 August 2013

Background: Improving our understanding of the relationship between the genotype and the drug resistance phenotype of *Mycobacterium tuberculosis* will aid the development of more accurate molecular diagnostics for drug-resistant tuberculosis. Studies that use direct genetic manipulation to identify the mutations that cause *M. tuberculosis* drug resistance are superior to associational studies in elucidating an individual mutation's contribution to the drug resistance phenotype.

Methods: We systematically reviewed the literature for publications reporting allelic exchange experiments in any of the resistance-associated *M. tuberculosis* genes. We included studies that introduced single point mutations using specialized linkage transduction or site-directed/*in vitro* mutagenesis and documented a change in the resistance phenotype.

Results: We summarize evidence supporting the causal relationship of 54 different mutations in eight genes (*katG*, *inhA*, *kasA*, *embB*, *embC*, *rpoB*, *gyrA* and *gyrB*) and one intergenic region (*furA-katG*) with resistance to isoniazid, the rifamycins, ethambutol and fluoroquinolones. We observed a significant role for the strain genomic background in modulating the resistance phenotype of 21 of these mutations and found examples of where the same drug resistance mutations caused varying levels of resistance to different members of the same drug class.

Conclusions: This systematic review highlights those mutations that have been shown to causally change phenotypic resistance in *M. tuberculosis* and brings attention to a notable lack of allelic exchange data for several of the genes known to be associated with drug resistance.

Keywords: *M. tuberculosis*, microbial susceptibility tests, genetics, SNPs, *in vitro* resistance

Introduction

The 2012 WHO report on global tuberculosis (TB) surveillance suggests that only one in five patients with drug-resistant TB are diagnosed and appropriately treated.[1] Patients with undiagnosed drug resistance have higher morbidity and mortality than patients with drug-susceptible disease, and may continue to spread drug-resistant TB in their communities.[2] The WHO has stated that a major challenge for drug-resistant TB control is the lack of laboratory capacity to diagnose resistance.[1] Newer, molecular-based diagnostics detect mutations conferring drug resistance and offer advantages for the identification of resistance in *Mycobacterium tuberculosis* (Mtb) over traditional culture-based techniques, including a more rapid turnaround time and a lower level of skill required to run the tests.[3–5] A thorough understanding of which mutations encode drug resistance in Mtb will be helpful in focusing research aimed at elucidating the underlying mechanisms of resistance and in supporting the development of more accurate molecular diagnostic tests for patient care.

Epidemiological studies of Mtb drug resistance have largely focused on the association of specific mutations with the drug resistance phenotype, primarily through the comparison of

mutations in specific genes in resistant clinical strains with drug-susceptible counterparts.[6–8] This approach, however, cannot definitively establish causality between the mutation and the resistance phenotype. Studies using direct bacterial genetic manipulation to identify the mutations that cause Mtb drug resistance can better elucidate the individual mutation's contribution to the drug resistance phenotype and uncover whether additional factors, like synergy, strain background or interactions between mutations, modulate this relationship. The purpose of this systematic review is to clarify mutation–phenotype relationships in Mtb by identifying which mutations have been causally linked to Mtb drug resistance and in what context these causal observations have been made.

Methods

Definitions

Two types of mutation were included in this study: (i) non-synonymous nucleotide substitutions, denoted by x#y, where x represents the wild-type amino acid, # the codon number and y the variant amino acid; and (ii) non-coding (ribosomal RNA, promoter, intergenic regions) nucleotide substitutions, denoted by #xy, where # refers to the position relative to the start of the non-coding region, x is the wild-type nucleotide base and y is the variant nucleotide base. For phenotype measurements, we defined MIC as the lowest concentration of drug that inhibits bacterial growth and IC_{50} as the concentration of drug required to inhibit supercoiling activity by 50%. We describe a mutation leading to any increase in MIC as *causative* of resistance; mutations that increase the MIC above the accepted critical concentration for medical diagnostic testing is said to be causing *clinical* levels of resistance.[9]

Literature search

Using the search strategy described in Table 1, we identified peer-reviewed primary research studies that reported the effect of creating specific mutations in resistance-associated genes on the drug resistance phenotypes of Mtb strains. We searched the PubMed and EMBASE databases from January 1980 to June 2012, using combinations of the keywords listed in Table 1. Bibliographies of articles selected for further review were hand-searched and additional references not previously identified were added as appropriate. We performed full-text mining of keywords in search theme 4 'Introduction of mutation' on articles retrieved by PubMed and EMBASE using search themes 1–3. This additional step was undertaken to capture articles that did not have these keywords in the title or abstract.

Study selection criteria

Methods to investigate phenotype causation have included (i) gene knockouts and complementation of the resulting null mutants; (ii) increasing transcription of the gene, leading to its overexpression; and (iii) *in vitro* selection of drug-resistant clones by plating susceptible strains on serial dilutions of a drug.[10–12] While the former methods shed light on whether the entire gene is essential for resistance, spontaneous mutants with a resistance phenotype may include compensatory mutations. Allelic exchange techniques, which introduce specific point mutations into a gene of interest, do not have these limitations and directly define the causative role for mutations in drug resistance, making it our method of choice for this review.

We included studies if they met the following criteria: (i) single point mutations within a putative resistance gene were introduced into Mtb strains using specialized linkage transduction or site-directed/*in vitro* mutagenesis; and (ii) a change in the resistance phenotype was documented. The resistance phenotypes were reported as MIC measurements or IC_{50} results performed before and after the introduction of a mutation. Researchers have demonstrated a quinolone structure–activity relationship for the *gyrA/B* protein complex, in which inhibition of supercoiling activity by 50% (IC_{50}) correlates well (better than DNA cleavage) with inhibition of Mtb growth by the fluoroquinolones (FQs).[13] We included studies that used liquid- or solid-based media for drug susceptibility testing.

We excluded manuscripts that (i) studied mycobacterial species other than Mtb; (ii) created knockout or overexpression of a gene instead of a single point mutation; (iii) did not specify the host strain used when measuring the MIC effect; (iv) did not state how the unique transfer of the intended point mutation was confirmed; or (v) did not have a phenotypic result (MIC or IC_{50}). We excluded *in vitro* selected mutations in order to remove the potential effects of compensatory mutations.

Data extraction

For every study that met our eligibility criteria, two of three authors (H. N.-G., K. R. J. and M. R. F.) independently reviewed the data and one additional author (M. B. M.) adjudicated differences between the authors. From each publication, the following information was extracted by two authors (H. N.-G. and K. R. J.): authors; publication year; gene; amino acid and nucleotide coordinates of the mutation; host strain and method used to introduce the mutation; method used to confirm introduction of the mutation; resistance genotypic and phenotypic susceptibility methods; and phenotypic results. Additional details and clarifications were obtained via personal correspondence by one of the authors (H. N.-G.).

Results

Of the 489 publications that we identified, 444 were excluded after abstract review. We performed full-text reviews of the remaining 45 papers and excluded a further 25. We identified 433 more papers through an additional text-mining step. Seventeen of these were selected through title and abstract review, but 16 were excluded upon full-text review. In total, 21 articles were selected for inclusion and final data extraction (Figure 1).

Isoniazid

We identified studies examining 11 different putative isoniazid resistance mutations in four Mtb genes: *katG*, the *furA-katG* intergenic region, *inhA* and *kasA*. Of these 11 mutations, 7 were shown to confer resistance to isoniazid (Table 2). No two studies looked at the same point mutation.

Mutations that caused isoniazid resistance

Pym *et al.*[14] investigated two point mutations in *katG* using host strain INH34, a clinical isolate with inherent up-regulation of *ahpC*.[15] The use of this strain ensured that any phenotypic differences detected among the INH34 transformants could not be due to the emergence of compensatory mutations in the promoter region of *ahpC*. Both *katG* S315T and T275P caused isoniazid resistance. Vilcheze *et al.*[16] introduced mutation S94A into the *inhA* gene of an H37Rv Mtb reference strain and found that it conferred a >5-fold increase in resistance to both isoniazid and ethambutol. Richardson *et al.*[17] demonstrated that a *katG* W300G H37Rv transformant caused a 1280-fold increase in isoniazid MIC, while Ando *et al.*[18] found that complementing mutations (−7GA, −10AC and −12GA) into a clinical isolate with a deleted *furA-katG* gene conferred low-level isoniazid resistance (0.1–1 mg/L).

Table 1. Search strategy to identify studies of mutations documented to confer resistance by evidence of genetic experiment

	Search theme			
	1. organism	2. drug resistance	3. mutation	4. method of introducing mutation
PubMed database				
Medical Subject Headings (MeSH) terms	1. 'mycobacterium tuberculosis'	1. 'drug resistance', OR 2. 'microbial sensitivity tests'	1. 'mutation', OR 2. 'amino acid substitution', OR 3. 'mutagenesis, site-directed', OR 4. 'codon'	NA
text terms	1. 'mycobacterium tuberculosis', OR 2. 'm tuberculosis', OR 3. 'mtb'	1. 'resistance', OR 2. 'mic', OR 3. 'inhibitory concentration', OR 4. 'drug susceptibility'	1. 'mutation*', OR 2. 'mutagenesis', OR 3. 'mutant*', OR 4. 'nonsense', OR 5. 'missense', OR 6. 'frameshift', OR 7. 'codon*', OR 8. 'transduction'	1. 'isogenic', OR 2. 'engineered', OR 3. 'mutagenesis', OR 4. 'recombinant', OR 5. 'site-directed', OR 6. 'allelic', OR 7. 'transduction', OR 8. 'wild-type', OR 9. 'induced', OR 10. 'introduced'
EMBASE				
Emtree tool	1. 'mycobacterium tuberculosis'	1. 'drug resistance'	1. 'mutation', OR 2. 'site-directed mutagenesis', OR 3. 'amino acid substitution', OR 4. 'codon'	NA
text terms	1. 'mycobacterium tuberculosis', OR 2. 'm tuberculosis', OR 3. 'mtb'	1. 'resistance', OR 2. 'mic', OR 3. 'mics', OR 4. 'inhibitory concentration', OR 5. 'drug susceptibility' 6. 'dst'	1. 'mutations*' OR 2. 'mutagenesis', OR 3. 'mutant*', OR 4. 'nonsense', OR 5. 'missense', OR 6. 'frameshift', OR 7. 'codon*', OR 8. 'transduction'	1. 'isogenic', OR 2. 'engineered', OR 3. 'mutagenesis', OR 4. 'recombinant', OR 5. 'site-directed', OR 6. 'allelic', OR 7. 'linkage transduction', OR 8. 'wild-type', OR 9. 'induced', OR 10. 'introduced'

NA, not available.

Mutations that had no effect on isoniazid resistance

Pym *et al.*[14] complemented a resistant *furA-katG* deletion Mtb mutant with a *katG* gene carrying the A139V mutation and found that it restored isoniazid susceptibility. This demonstrates that A139V does not confer resistance in this strain. Vilcheze *et al.*[16] found that *kasA* G312S and F413L mutations in H37Rv caused no detectable changes in isoniazid MIC. Ando *et al.*[18] found that mutation C41T in *furA* resulted in no appreciable change in MIC relative to the Mtb strain harbouring the wild-type *furA* gene (Table 3).

Ethambutol

We identified six studies examining nine different putative ethambutol resistance mutations in the gene *embB*. Four of these

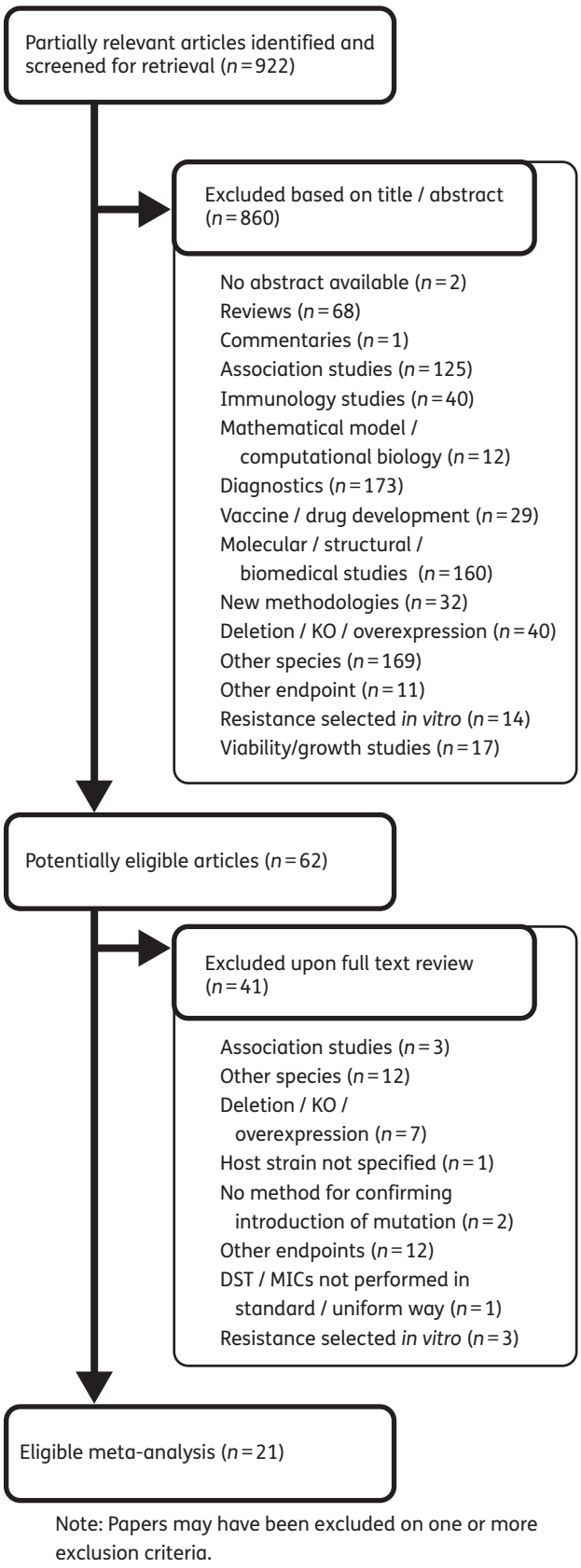


Figure 1. Study selection process and reasons for exclusion of studies. KO, knock-out.

investigated the same three mutations [Met306Ile (ATA), Met306Ile (ATC) and Met306Val]. We identified one additional study that examined five putative ethambutol resistance mutations in *embC*. All nine *embB* mutations and one *embC* mutation were shown to confer resistance to ethambutol.

Mutations that caused ethambutol resistance

Two different clinical strains were chosen by Safi *et al.*[19] as host backgrounds for introducing *embB* gene mutations into codon 306: drug-susceptible 210 belonging to the W-Beijing family, the *embB* sequence of which is identical to that of laboratory strain H37Rv; and the ethambutol-resistant clinical isolate 5310 that had been reverted back to wild-type *embB* sequence. Mutations M306V, M306L, M306I (ATA) and M306I (ATC) all caused ethambutol resistance (MIC > 4 mg/L) when incorporated into wild-type strain 210 and strain 5310.[19]

Starks *et al.*[20] introduced the *embB* M306V allele into H37Rv and Beijing F2, resulting in a 4-fold ethambutol MIC increase, while M306I resulted in a 2-fold increase in both host strains. Plinke *et al.*[21] found a 4-fold increase in ethambutol MIC for mutations M306V and M306I (ATA) and a 2-fold increase for mutation M306I (ATC) when introduced into H37Rv. These, however, remained below the critical clinical cut-off value. Goude *et al.*[22] showed that mutation M70I introduced into H37Rv presented a small increase in ethambutol MIC (4 mg/L), insufficient to render it clinically resistant.

Safi *et al.*[23] also looked at the role of common mutations found in clinical strains with high-level ethambutol resistance at the *embB* 406 and 497 codons. They substituted the wild-type clinical Mtb 210 strain *embB* G406 codon with G406A, G406D, G406C or G406S, all of which led to ethambutol resistance. Replacing the wild-type *embB* Q497 codon in strain 210 with the Q497R codon also increased the ethambutol MIC (Table 2).

Mutations that potentiated susceptibility or had no effect on ethambutol resistance

Goude *et al.*[22] introduced a point mutation at the conserved aspartate D294G that had previously been shown to affect the activity of *embC* in *Mycobacterium smegmatis* to determine whether a similar effect would be seen in Mtb. They found that D294G and a further two mutations introduced into codon 300 of the *embC* arabinosyltransferase (M300L and M300V) increased susceptibility to ethambutol. Mutation M300I had no resistance effect (Table 3).

Rifamycins

We identified four studies that examined five putative single rifamycin resistance mutations and three double mutations in Mtb. The most common rifamycin amino acid substitutions in clinical strains (*rpoB* codons 531, 526 or 516) and two additional mutations were shown to individually confer resistance to rifamycins.[23]

Mutations that caused resistance to rifamycins

Williams *et al.*[24] investigated the causal relationship between specific amino acid changes and three rifamycins (rifampicin, rifapentine and rifabutin) by incorporating mutations D516V, H526Y and S531L into the *rpoB* gene of Mtb H37Rv. Mutant alleles S531L and H526Y conferred high-level resistance to the three rifamycins

Table 2. Mutations shown to confer resistance to isoniazid, ethambutol or RIF

Drug	Gene	Host strain	Substitution	MIC (mg/L)	Reference
Isoniazid	*katG*	INH34[a]	WT[b]	0.1	Pym *et al.*, 2002[14]
			S315T	5	
			T275P	>10	
		H37Rv	WT	0.1	Richardson *et al.*, 2009[17]
			W300G	128	
	furA-katG intergenic region	NCGM2836[c]	WT[b]	0.1	Ando *et al.*, 2011[18]
			G7A	0.4	
			A10C	0.4	
			G12A	0.15	
	inhA	H37Rv	WT	0.1	Vilcheze *et al.*, 2006[16]
			S94A	0.5	
Ethambutol	*embB*	210[d]	WT	2	Safi *et al.*, 2010[23]
			G406S	6	
			G406A	7	
			G406D	7	
			G406C	7	
			Q497R	12	
		210[d]	WT	2	Safi *et al.*, 2008[19]
			M306I (ATA)	7	
			M306I (ATC)	7	
			M306L	8.5	
			M306V	14	
		5310[e]	WT	3	
			M306I (ATA)	16	
			M306I (ATC)	16	
			M306L	20	
			M306V	28	
		H37Rv	WT	1	Plinke *et al.*, 2011[21]
			M306I (ATC)	2	
			M306I (ATA)	4	
			M306V	4	
		H37Rv	WT	5	Starks *et al.*, 2009[20]
			M306V	20	
			M306I (ATA)	10	
		Beijing F2	WT	5	
			M306V	20	
			M306I (ATA)	10	
	embC	H37Rv	WT	3	Goude *et al.*, 2009[22]
			T270I	4	
Rifampicin	*rpoB*	H37Rv	WT	0.25	Williams *et al.*, 1998[24]
			D516V	32	
			H526Y	64	
			S531L	>64	
			WT	≤0.015[f]	
			H526Y	16[f]	
			S531L	16[f]	
			WT	0.03[g]	
			D516V	16[g]	
			H526Y	16[g]	
			S531L	>64[g]	
		H37Ra	WT	1.5	Zaczek *et al.*, 2009[26]
			D516V	50	
			H526Y	25	

Continued

Systematic review

Table 2. *Continued*

Drug	Gene	Host strain	Substitution	MIC (mg/L)	Reference
			S531L	50	
			S512I+D516G	6.2	
			Q513L	6.2	
			M515I+D516Y	6.2	
			D516Y	3.1	
		KL1936[h]	WT	1.5	
			H526D	50	
			D516V	25	
			Q513L	12.5	
			S531L	50	
			Q510H+D516Y	6.2	
			S512I+D516G	6.2	
			M515I+D516Y	6.2	
			D516Y	6.2	
		KL463[i]	WT	1.5	
			H526D	50	
			D516V	25	
			Q513L	50	
			S531L	50	
			Q510H+D516Y	6.2	
			S512I+D516G	6.2	
			M515I+D516Y	6.2	
			D516Y	3.1	
		H37Ra[j]	WT	1.5	
			H526D	50	
			D516V	25	
			S531L	50	
			Q510H+D516Y	6.2	
			S512I+D516G	6.2	
			Q513L	6.2	
			M515I+D516Y	6.2	
			D516Y	6.2	
		H37Ra	WT	<0.1	Siu *et al.*, 2011[27]
			S531L	64	
			V146F	64	
			I572F	8–16	

WT, wild-type.
[a]Δ*furA*-Δ*katG* clinical isolate resistant to isoniazid and with inherent up-regulation of *ahpC*.
[b]Complemented with the wild-type *katG* gene.
[c]Δ*furA*-Δ*katG* clinical isolate resistant to isoniazid.
[d]Drug-susceptible clinical strain, member of the W-Beijing family.
[e]Clinical isolate resistant to ethambutol.
[f]Rifabutin.
[g]Rifapentine.
[h]Rifampicin-susceptible clinical strain containing PrpoB natural promoter.
[i]Rifampicin-susceptible clinical strain.
[j]Containing a modified heat shock promoter (Phsp65).

(Table 2). Clones containing mutation D516V showed resistance to rifampicin and rifapentine but susceptibility to rifabutin. Gill and Garcia[25] also found that these three mutant alleles led to elevation of IC_{50} values for rifampicin, rifabutin and rifaximin. They found that the rifabutin IC_{50} was elevated less by mutations S531L and D516V than by H526Y. Zaczek *et al.*[26] explored whether the background Mtb strain affected the change in the rifampicin MIC. All strain backgrounds (H37Ra, KL1936 and KL463) containing *rpoB* genes with mutations H526D, D516V or S531L had high-level rifampicin resistance.

Noting that 5% of clinical strains with rifampicin resistance do not have mutations in the 81 bp region of *rpoB*, Siu *et al.*[27] aimed to identify mutations located outside this rifampicin resistance-determining region. They found that H37Ra transformants

Table 3. Mutations shown not to confer resistance to isoniazid, ethambutol or rifampicin

Drug	Gene	Host strain	Substitution	MIC (mg/L)	Reference
Isoniazid	*katG*	INH34[a]	A139V	0.1	Pym *et al.*, 2002[14]
	furA-katG intergenic region	NCGM2836[a]	C41T	0.1	Ando *et al.*, 2011[18]
	kasA	H37Rv	G312S	0.1	Vilcheze *et al.*, 2006[16]
			F413L	0.1	
Ethambutol	*embC*	H37Rv	M300L	0.5[b]	Goude *et al.*, 2009[22]
			M300I	3	
			M300V	0.5[b]	
			D294G	0.5[b]	
Rifampicin	*rpoB*	H37Rv	D516V	≤0.015[c]	Williams *et al.*, 1998[24]
		H37Ra	Q510H+D516Y	1.5	Zaczek *et al.*, 2009[26]

[a]Δ*furA*-Δ*katG* clinical isolate resistant to isoniazid and with inherent up-regulation of *ahpC*.
[b]Mutations shown to increase susceptibility to isoniazid, ethambutol or rifampicin.
[c]Rifabutin.

containing mutations S531L or V146F were resistant to rifampicin. Transformants containing mutation I572L had a rifampicin MIC raised to 8–16 mg/L. Although V146F and I572L conferred resistance, the authors noted that they are rarely seen in clinical strains.

Zaczek *et al.*[26] found that D516Y conferred low-level resistance in strains KL453, KL1936 and H37Ra. The double mutations Q510H+D516Y, S512I+D516G and M515I+D516Y all conferred low-level resistance in these strains. Hence, the substitutions in position 516 (D/Y; D/G), even when supported with Q510H, M515I or S512I, did not result in a large increase in the rifampicin MIC. The authors therefore concluded that a mutation D/Y or D/G at 516 is not sufficient to confer clinical rifampicin resistance in Mtb, in contrast to D/V, which does confer clinical resistance.

Mutation(s) with an effect on resistance to rifamycins that varied by host strain

Zaczek *et al.*[26] found that mutation Q513L led to high-level rifampicin resistance in strain KL463, lower-level resistance in KL1936 and no significant increase in MIC in H37Ra (Table 2).

FQs

We identified nine articles that studied the causal relationship between mutations in *gyrA/B* in Mtb and resistance to FQs. All except one of these studies measured IC_{50} rather than MIC as the resistance outcome. Not all FQs have the same effect on Mtb. Ofloxacin and ciprofloxacin have bacteriostatic antimycobacterial activity, whereas moxifloxacin shows high bactericidal activity.[25]

gyrA

Mutations that caused FQ resistance Onodera *et al.*[28] found that *gyrA* mutations A90V and A90V+D94V greatly increased the IC_{50} of levofloxacin and ciprofloxacin compared with the wild-type (Table 4). Aubry *et al.*[29] reported that *gyrA* mutations A90V, D94G and D94H led to increased IC_{50}s of four FQs; in addition, mutation A90V+D94G had an additive effect as a double mutant. Matrat *et al.*[30] found that transformants bearing *gyrA* G88A and G88C were more resistant than wild-type gyrase to inhibition by FQs. The increases in IC_{50} for G88C were higher than for G88A with respect to gatifloxacin, levofloxacin and moxifloxacin and similar for ofloxacin. Malik *et al.*[31] reported that the A74S mutation increased the MIC 2-fold to 4-fold for each FQ tested, which is slightly above the critical concentration. While the single D94G mutation conferred resistance, the addition of A74S to D94G had a synergistic effect, further increasing the MICs of all FQs tested by 2-fold to 8-fold over those for the single D94G mutation. Kim *et al.*[32] found that IC_{50} values of levofloxacin, ciprofloxacin and gatifloxacin against DNA gyrase containing S95+D94G were 2-fold greater than those against DNA gyrase containing S95 with A74S+D94G, which was higher than the wild-type.

Mutations that increased susceptibility or had no effect on FQ resistance Aubry *et al.*[29] reported that *gyrA* mutations T80A and T80A+A90G led to a reduced IC_{50}; A90G alone did not affect the FQ IC_{50} (Table S1, available as Supplementary data at *JAC* Online). Malik *et al.*[31] also found that the *gyrA* double mutation T80A+A90G had no significant effect on MICs and actually decreased the MIC for ofloxacin. Transformants with G247S and A384V, located outside the *gyrA* quinolone resistance-determining region (QRDR), had similar FQ MICs compared with negative controls.

Matrat *et al.*[33] looked to identify the minimum number of mutations needed to increase FQ susceptibility in Mtb to levels similar to those in *Escherichia coli*. An A83S mutation in *gyrA* was sufficient to decrease moxifloxacin IC_{50} to a susceptible range for *E. coli*. To decrease the ofloxacin IC_{50} to a susceptible range similar to *E. coli*, the A83S mutation had to be coupled with a second substitution, either M74I in *gyrA* or R447K in *gyrB*. Modification of the vicinity of A83 (residues 84 and 85) did not have any effect on FQ susceptibility.

Kim *et al.*[32] explored whether lineage-specific amino acid residues affect FQ resistance. They conducted *in vitro* IC_{50} studies using recombinant DNA gyrase bearing an S95 residue in *gyrA*. The wild-type (*gyrA* containing S95) and *gyrA* containing A74S with the S95 demonstrated similar levels of *in vitro* FQ susceptibility. The authors believed the reason that this mutation did not show the higher FQ resistance described in previous reports was because those earlier strains from China were

Table 4. Mutations shown to cause resistance to at least one FQ

Gene	Host strain[a]	Substitution	Ofloxacin IC_{50} or MIC[b] (mg/L)	Ciprofloxacin IC_{50} or MIC[c] (mg/L)	Levofloxacin IC_{50} or MIC[d] (mg/L)	Gatifloxacin IC_{50} or MIC (mg/L)	Moxifloxacin IC_{50} or MIC[e] (mg/L)	Reference
gyrA	—	WT	—	12.2	13.9	—	—	Onodera *et al.*, 2001[28]
		A90V	—	>400	>400	—	—	
		A90V+D94V	—	>400	>400	—	—	
	—	WT	10	—	12	2.5	2	Aubry *et al.*, 2006[29]
		A90V	100	—	55	20	35	
		D94G	350	—	170	70	50	
		D94H	800	—	320	150	90	
		A90V+D94G	>1600	—	>1600	>320	>160	
	—	WT	10	—	5	4	4	Matrat *et al.*, 2006[30]
		G88A	40	—	30	7	10	
		G88C	50	—	100	>128	35	
	—	WT (S95)	—	18	34	9	—	Kim *et al.*, 2012[32]
		D94G+S95	—	196	310	76	—	
		A74S+D94G+S95	—	107	171	48	—	
	H37Rv or Erdman	WT	0.5	<0.25–0.5	<0.25	—	<0.25	Malik *et al.*, 2012[31]
	H37Rv	A74S+D94G	16–32	16	16	—	4–16	
		A90V	2–4	2–4	0.5–2	—	0.5–1	
	Erdman	A74S	1–2	1	1	—	0.5–1	
		A90V	2–8	4	0.5–4	—	0.5–1	
	CDC1551	D94G	8	8	8	—	2	
gyrB	—	N510D	120	—	500	45	35	Aubry *et al.*, 2006[29]
	—	WT	—	7	22	9	16	Kim *et al.*, 2011[36]
		E540V	—	251	82	37	61	
	—	WT	10	—	8	3	2.5	Pantel *et al.*, 2012[37]
		D500A	22	—	25	8	6	
		N538T	28	—	24	14	12	
		T539P	30	—	17	13	12	
		E540V	80	—	64	>20	>20	
	H37Rv or Erdman	WT	0.5	<0.25—0.5	<0.25	—	<0.25—0.5	Malik *et al.*, 2012[31]
		N538D	4	4	2	—	1	
		T539P	0.5–1	1	0.5–1	—	0.5–1	
		N538K	2	2	1	—	1–2	
	H37Rv	E540V	4	2	1–2	—	0.5–1	
		D500H	4–8	1–2	2–4	—	<0.25–0.5	
		D500N	4	1	2	—	<0.25–0.5	
		N538D+T546M	2	4	2	—	1	
		N538T+T546M	0.5	2	0.5	—	0.5–1	
		A543V	2	1	1	—	0.5–1	
		E540D	0.5	0.5	0.5	—	2–4	
		R485C+T539N	4–8	2	2–4	—	2	

Erdman	E540V	4	2–4	2	—	1
	D500H	4–8	1	2–4	—	0.5
	D500N	4	2	2	—	0.5
	N538D+T546M	4	8	2	—	<0.25–1
	N538T+T546M	0.5	2	0.5	—	<0.25–1
	T539N	2	2	1	—	1
	E540D	0.5–1	0.5–1	0.5	—	2
	R485C+T539N	8	4	4	—	4

[a]IC_{50}s were determined directly on recombinant *gyrA* and *gyrB* subunits produced in *E. coli* plasmids. All references except for Malik *et al.*,[31] used IC_{50}s.
[b]Critical concentration: MIC > 2 mg/L.
[c]Critical concentration: MIC > 2 mg/L.
[d]Critical concentration: MIC > 1 mg/L.
[e]MIC > 0.5 mg/L (low-level resistance) and >2 mg/L (high-level resistance).

Beijing, which contains threonine at position 95, which may already enhance resistance by altering interactions between α4 and α3 helices.[34,35]

gyrB

Mutations that caused FQ resistance Aubry *et al.*[29] found that the N510D mutation in *gyrB* led to an IC_{50} elevation (Table 4). Kim *et al.*[36] found that a gyrase bearing the E540V amino acid substitution in *gyrB*, mimicking a clinical strain from Bangladesh, was highly resistant to inhibition by four FQs. Pantel *et al.*[37] reported that D500A and N538T (located in the QRDR) and T539P (located outside the QRDR) conferred low-level resistance, in contrast to E540V (also outside the QRDR), which conferred higher-level resistance. In contrast to the findings of Kim *et al.*[36] and Pantel *et al.*,[37] Malik *et al.*[31] found that the resistance pattern of the E540V mutation was dependent on the genetic background of the mutated strain. In H37Rv, E540V conferred consistent susceptibility to ciprofloxacin but conferred resistance to levofloxacin and ofloxacin and low-level resistance to moxifloxacin. In the Erdman background, E540V exhibited cross-resistance to all four FQs tested during one round of testing but was susceptible to moxifloxacin on repeat testing. In addition, Malik *et al.*[31] found that transformants harbouring D500A had increased MICs for levofloxacin and ofloxacin (at least 4-fold), which were still considered in the susceptible range; the MICs for ciprofloxacin and moxifloxacin were unaffected.

Malik *et al.*[31] report that transformants harbouring *gyrB* D500H or D500N were resistant to levofloxacin and ofloxacin but susceptible to ciprofloxacin and moxifloxacin. The N538D-containing transformant exhibited resistance to all four FQs. The N538D+T546M double mutation conferred resistance to all of the FQs tested when introduced into Erdman but did not significantly increase the MIC to a greater extent than N538D alone. The N538D+T546M double mutation resulted in slightly different results in the H37Rv genetic background, where it was resistant to ciprofloxacin, levofloxacin and moxifloxacin but susceptible to ofloxacin. Transformants carrying another variant at codon 538, N538T, plus T546M were susceptible to all FQs tested. These data suggest that T546M does not play a synergistic role in FQ resistance and N538T does not confer resistance. The R485C and T539N *gyrB* mutations each independently increased the MIC, but not to clinical resistance levels. The T539N mutation did confer low-level resistance to moxifloxacin in the Erdman strain. When introduced together into H37Rv, *gyrB* R485C+T539N conferred resistance to ofloxacin, levofloxacin and moxifloxacin; the same double mutation in Erdman conferred resistance to all four FQs tested. Based on these results, R485C and T539N individually increase the FQ MIC slightly but in combination they act synergistically to increase the MIC above the critical concentration to confer clinical resistance.

Malik *et al.*[31] found that the T539P mutation alone increased the levofloxacin MIC, but not above the critical concentration; this mutation did not substantially affect the MIC for any other FQ. Both A543T and A543V increased (2-fold to 4-fold) the MICs for levofloxacin, ciprofloxacin and ofloxacin but had no effect on the moxifloxacin MIC. These were still below the accepted critical concentration for clinical resistance. The N538K mutation exhibited low-level resistance to moxifloxacin and increased the MICs (4-fold) of ciprofloxacin, ofloxacin and levofloxacin, although these increases were not sufficient to be considered resistant.

Mutations that increased susceptibility or had no effect on FQ resistance Pantel *et al.*[38] studied eight substitutions in *gyrB* (D473N, P478A, R485H, S486F, A506G, A547V, G551R and G559A) and found that none of them was implicated in FQ resistance (Table S1, available as Supplementary data at *JAC* Online). Malik *et al.*[31] found that *gyrB* M330I, V340L and T546M did not confer resistance to any FQ tested. Transformants with D533A were also susceptible to all four FQs. T546M did not confer FQ resistance. Matrat *et al.*[33] found that the R447K substitution conferred increased susceptibility.

Discussion

In this systematic review we identified papers that introduced drug resistance-conferring mutations into eight genes (*katG*, *inhA*, *kasA*, *embB*, *embC*, *rpoB*, *gyrA* and *gyrB*) and one intergenic region (*furA-katG*). Within these genomic regions, 25 individual mutations plus 3 double mutations caused clinical resistance to first-line drugs, and 8 resulted in no change in inhibitory concentration. A further 18 individual mutations and 7 double mutations caused clinical resistance to one or more FQs, with 26 individual mutations and 4 double mutations conferring no change in FQ inhibitory concentrations (Tables 2–4 and Table S1, available as Supplementary data at *JAC* Online).

Several studies found that mutations can have a different effect on the drug MIC, depending on the background strain into which it is introduced. For example, the *rpoB* mutation Q513L led to high-level resistance to rifampicin in strain KL463, lower-level resistance in strain KL1936 and no significant increase in MIC in H37Ra. In *embB*, mutation M306I (ATC) caused a moderately higher MIC in strain 5310 compared with strain 210 and H37Rv. Similarly, *embB* mutation M306I (ATA) resulted in varied levels of MIC: 7 mg/L in strain 210; 16 mg/L in strain 5310; 10 mg/L in a Beijing F2 strain; and both 4 mg/L and 10 mg/L in two H37Rv-derived strains in two different studies. Depending on what value is chosen as the critical concentration cut-off, the latter H37Rv could be considered 'susceptible' and the former 'resistant'.[9]

Although all MICs consistently increased, such discrepancies underline the limitations of the currently accepted critical concentration cut-offs in determining clinical 'resistance', and suggest that epistasis between the introduced mutations and other genetic variation elsewhere in the genome plays an important role in influencing the resistance phenotype. Mutation–mutation interactions have been previously noted to influence the drug resistance phenotype of other pathogens such as HIV.[39] The observation of epistasis influencing the drug resistance phenotype in Mtb challenges the reductionist view that one 'correct' mutation is sufficient to result in resistance to a particular drug, and supports the more comprehensive study of additional genes in the Mtb genome that can modulate or contribute to the resistance phenotype in an alternative 'multi-hit' model.

This systematic review also demonstrates that the same drug resistance mutations can cause varying levels of resistance to different members of the same drug class, For example, the D500H mutation in *gyrB* led to resistance to earlier-generation FQs (ofloxacin, levofloxacin) but not moxifloxacin. Likewise, clones containing the D516V mutation in *rpoB* showed resistance to rifampicin and rifapentine but maintained susceptibility to rifabutin. This finding is consistent with similar observations made in clinical strains that exhibited rifampicin resistance with rifabutin susceptibility by current cut-offs.[40]

It is possible that these observations may be overemphasized by the current, arguably arbitrary, drug concentration cut-offs for clinical resistance. However, the observation that in isogenic backgrounds the same mutation leads to smaller increments in MIC for some members of the same drug class, coupled with the known higher pharmacological potency of some of these agents seems likely to have treatment implications. To date, there are no direct clinical or pharmacological data to support the clinical efficacy of treating Mtb resistant to one member of the FQ or rifamycin drug class with another member, but observations of improved treatment outcomes for patients with extensively drug-resistant TB (by definition resistant to a member of the FQ drug class) who were treated with later-generation FQs (levofloxacin or moxifloxacin) provide some indirect support for this notion.[41–44] Evidence from this review thus emphasizes the importance of further studying FQs and alternative rifamycins to assess their clinical value in the treatment of Mtb resistant to other members of the same drug class.

This systematic review highlights some notable lack of allelic exchange data for several of the genes known to be associated with drug resistance. Notably, we found no studies that met our inclusion criteria which studied *pncA*, *rrs*, *inhA* promoter region, or *ethA* encoding resistance to the drugs pyrazinamide, streptomycin, the aminoglycosides (amikacin, capreomycin, kanamycin) and ethionamide. Even within the genes studied, only a subset of the common mutations was studied in most cases. For example, we found no report of allelic exchange experiments performed at codon 91 of *gyrA*, or codons 446, 447, 461, 494, 501 and 504 of *gyrB*, codons that have previously been associated with FQ resistance in clinical strains.[45,46]

Rapid molecular assays for detecting drug resistance are currently limited, with GeneXpert (Cepheid) only testing for rifampicin resistance, the sensitivity of the GenoType MTBDR test (Hain Lifescience) for the detection of isoniazid resistance reported to be in the 80%–90% range[47–49] and the GenoType MTBDRsl assay showing a low level of performance for FQs, aminoglycosides and ethambutol (reported sensitivities of 87%–89%, 21%–100% and 39%–57%, respectively).[50–53] Their accuracy is largely dependent on the strength of the association between a specific mutation and the resistance phenotype. These and further allelic exchange studies may point towards recommendations for improving the diagnostic accuracy of molecular-based resistance assays, depending on their correlation with the frequency of these mutations found in clinical strains. For example, including *embB* mutations in codon 406, shown to increase the ethambutol MIC to a clinically significant level in this review and also observed in clinical isolates in India, Russia and the USA,[54–56] could improve the sensitivity for detecting resistance to ethambutol in those particular geographic settings. An updatable database on mutations associated with resistance worldwide, such as TBDReaMDB, may serve as a cross-check for the clinical relevance of including newly identified mutations from allelic exchange studies into diagnostic tests.[46] Finally, the reviewed allelic exchange experiments suggest that mutation Q513L in *rpoB*, currently assayed in the GeneXpert pipeline, does not result in a consistent increase in rifampicin MIC, depending on the strain background genome. This may have an impact on GeneXpert's specificity.

It is critical to note that drug susceptibility testing, although the gold standard, is not 100% accurate. A lack of concordance with

resistance screening may therefore not necessarily imply that resistance has been missed. It has been shown that *in vitro* data do not necessarily correlate with *in vivo* data and vice versa. For example, mutations leading to only slightly raised *in vitro* rifampicin resistance may indeed have clinical significance,[57] while mutations with dramatic *in vitro* effects may be unfit *in vivo* and hence very rare in patient isolates.[12] Whole genome sequencing and convergence analysis may be particularly useful in identifying potential mutations of interest requiring confirmation.[58,59]

This systematic review highlights the current understanding of the causal relationships of different mutations on phenotypic resistance in Mtb as studied via allelic exchange. Given increasing reports of Mtb strains with higher levels of drug resistance worldwide, this review provides new suggestions for drug resistance diagnostics development and highlights some gaps in our knowledge of genotype–phenotype relationships that are worth further study.

Funding

This work was supported by the Portuguese Foundation for Science and Technology (FCT) (SFRH/BD/33902/2009 to H. N.-G.), the National Institutes of Health/Fogarty International Center (1K01 TW009213 to K. R. J.), departmental funds of the pulmonary division of Massachusetts General Hospital to M. R. F. and the National Institutes of Health/NIAID (U19 A1076217 to M. B. M.).

Transparency declarations

None to declare.

Supplementary data

Table S1 is available as Supplementary data at *JAC* Online (http://jac.oxfordjournals.org/).

References

1 WHO. *Global Tuberculosis Report 2012*. Geneva: WHO. http://apps.who.int/iris/bitstream/10665/75938/1/9789241564502_eng.pdf (2 May 2013, date last accessed).

2 Menzies D, Benedetti A, Paydar A *et al*. Standardized treatment of active tuberculosis in patients with previous treatment and/or with mono-resistance to isoniazid: a systematic review and meta-analysis. *PLoS Med* 2009; **6**: e1000150.

3 Evans CA. GeneXpert—a game-changer for tuberculosis control? *PLoS Med* 2011; **8**: e1001064.

4 Kim SY, Kim H, Kim SY *et al*. The Xpert® MTB/RIF assay evaluation in South Korea, a country with an intermediate tuberculosis burden. *Int J Tuberc Lung Dis* 2012; **16**: 1471–6.

5 Jacobson KR, Theron D, Kendall EA *et al*. Implementation of GenoType MTBDRplus reduces time to multidrug-resistant tuberculosis therapy initiation in South Africa. *Clin Infect Dis* 2012; **56**: 503–8.

6 Siddiqi N, Shamim M, Hussain S *et al*. Molecular characterization of multi-drug resistant isolates of *Mycobacterium tuberculosis* from patients in North India. *Antimicrob Agents Chemother* 2002; **46**: 443–50.

7 Tang K, Sun H, Zhao Y *et al*. Characterization of rifampin-resistant isolates of *Mycobacterium tuberculosis* from Sichuan in China. *Tuberculosis* 2013; **93**: 89–95.

8 Tessema B, Beer J, Emmrich F *et al*. Analysis of gene mutations associated with isoniazid, rifampicin and ethambutol resistance among *Mycobacterium tuberculosis* isolates from Ethiopia. *BMC Infect Dis* 2012; **10**: 12–37.

9 Bottger EC. The ins and outs of *Mycobacterium tuberculosis* drug susceptibility testing. *Clin Microbiol Infect* 2011; **17**: 1128–34.

10 Boshoff HI, Mizrahi V. Purification, gene cloning, targeted knockout, overexpression, and biochemical characterization of the major pyrazinamidase from *Mycobacterium smegmatis*. *J Bacteriol* 1998; **180**: 5809–14.

11 Heym B, Stavropoulos E, Honore N *et al*. Effects of overexpression of the alkyl hydroperoxide reductase AhpC on the virulence and isoniazid resistance of *Mycobacterium tuberculosis*. *Infect Immun* 1997; **65**: 1395–401.

12 Bergval IL, Schuitema AR, Klatser PR *et al*. Resistant mutants of *Mycobacterium tuberculosis* selected *in vitro* do not reflect the *in vivo* mechanism of isoniazid resistance. *J Antimicrob Chemother* 2009; **64**: 515–23.

13 Aubry A, Pan X, Fisher LM *et al*. *Mycobacterium tuberculosis* DNA gyrase: interaction with quinolones and correlation with antimycobacterial drug activity. *Antimicrob Agents Chemother* 2004; **48**: 1281–8.

14 Pym AS, Saint-Joanis B, Cole ST. Effect of *katG* mutations on the virulence of *Mycobacterium tuberculosis* and the implication for transmission in humans. *Infect Immun* 2002; **70**: 4955–60.

15 Pym AS, Domenech P, Honoré N *et al*. Regulation of catalase–peroxidase (KatG) expression, isoniazid sensitivity and virulence by *furA* of *Mycobacterium tuberculosis*. *Mol Microbiol* 2001; **40**: 879–89.

16 Vilcheze C, Wang F, Arai M *et al*. Transfer of a point mutation in *Mycobacterium tuberculosis inhA* resolves the target of isoniazid. *Nat Med* 2006; **12**: 1027–9.

17 Richardon E, Lin S, Pinsky B *et al*. First documentation of isoniazid reversion in *Mycobacterium tuberculosis*. *Int J Tuberc Lung Dis* 2009; **13**: 1347–54.

18 Ando H, Kitao T, Miyoshi-Akiyama T. Downregulation of *katG* expression is associated with isoniazid resistance in *Mycobacterium tuberculosis*. *Mol Microbiol* 2011; **79**: 1615–28.

19 Safi H, Sayers B, Hazbon MH *et al*. Transfer of *embB* codon 306 mutations into clinical *Mycobacterium tuberculosis* strains alters susceptibility to ethambutol, isoniazid, and rifampin. *Antimicrob Agents Chemother* 2008; **52**: 2027–34.

20 Starks AM, Gumusboga A, Plikaytis BB *et al*. Mutations at *embB* codon 306 are an important molecular indicator of ethambutol resistance in *Mycobacterium tuberculosis*. *Antimicrob Agents Chemother* 2009; **53**: 1061–6.

21 Plinke C, Walter K, Alv S *et al*. *Mycobacterium tuberculosis embB* codon 306 mutations confer moderately increased resistance to ethambutol. *Antimicrob Agents Chemother* 2011; **55**: 2891–6.

22 Goude R, Amin A, Chatterjee D *et al*. The arabinosyltransferase EmbC is inhibited by ethambutol in *Mycobacterium tuberculosis*. *Antimicrob Agents Chemother* 2009; **53**: 4138–46.

23 Safi H, Fleischmann RD, Peterson SN *et al*. Allelic exchange and mutant selection demonstrate that common clinical *embCAB* gene mutations only modestly increase resistance to ethambutol in *Mycobacterium tuberculosis*. *Antimicrob Agents Chemother* 2010; **54**: 103–8.

24 Williams DL, Spring L, Collins L *et al*. Contribution of *rpoB* mutations to development of rifamycin cross-resistance in *Mycobacterium tuberculosis*. *Antimicrob Agents Chemother* 1998; **42**: 1853–7.

25 Gill SK, Garcia GA. Rifamycin inhibition of WT and Rif-resistant *Mycobacterium tuberculosis* and *Escherichia coli* RNA polymerase *in vitro*. *Tuberculosis* 2011; **91**: 361–9.

26 Zaczek A, Brzostek A, Augustynowicz-Kopec E *et al*. Genetic evaluation of relationship between mutations in *rpoB* and resistance of *Mycobacterium tuberculosis* to rifampin. *BMC Microbiol* 2009; **9**: 10.

27 Siu GKH, Zhang Y, Lau TCK *et al*. Mutations outside the rifampicin resistance-determining region associated with rifampicin resistance in *Mycobacterium tuberculosis*. *J Antimicrob Chemother* 2011; **66**: 730–3.

28 Onodera Y, Tanaka M, Sato K. Inhibitory activity of quinolones against DNA gyrase of *Mycobacterium tuberculosis*. *J Antimicrob Chemother* 2001; **47**: 447–50.

29 Aubry A, Veziris N, Cambau E *et al*. Novel gyrase mutations in quinolone-resistant and -hypersusceptible clinical isolates of *Mycobacterium tuberculosis*: functional analysis of mutant enzymes. *Antimicrob Agents Chemother* 2006; **50**: 104–12.

30 Matrat S, Veziris N, Mayer C *et al*. Functional analysis of DNA gyrase mutant enzymes carrying mutations at position 88 in the A subunit found in clinical strains of *Mycobacterium tuberculosis* resistant to fluoroquinolones. *Antimicrob Agents Chemother* 2006; **50**: 4170–3.

31 Malik S, Willby M, Sikes D *et al*. New insights into fluoroquinolone resistance in *Mycobacterium tuberculosis*: functional genetic analysis of *gyrA* and *gyrB* mutations. *PLoS One* 2012; **7**: e39754.

32 Kim H, Nakajima C, Kim YU *et al*. Influence of lineage-specific amino acid dimorphisms in GyrA on *Mycobacterium tuberculosis* resistance to fluoroquinolones. *Jpn J Infect Dis* 2012; **65**: 72–4.

33 Matrat S, Aubry A, Mayer C *et al*. Mutagenesis in the $\alpha3\alpha4$ GyrA helix and in the toprim domain of GyrB refines the contribution of *Mycobacterium tuberculosis* DNA gyrase to intrinsic resistance to quinolones. *Antimicrob Agents Chemother* 2008; **52**: 2909–14.

34 Sun Z, Zhang J, Zhang X *et al*. Comparison of *gyrA* gene mutations between laboratory-selected ofloxacin-resistant *Mycobacterium tuberculosis* strains and clinical isolates. *Int J Antimicrob Agents* 2008; **31**: 115–21.

35 Shi R, Zhang J, Li C *et al*. Emergence of ofloxacin resistance in *Mycobacterium tuberculosis* clinical isolates from China as determined by *gyrA* mutation analysis using denaturing high-pressure liquid chromatography and DNA sequencing. *J Clin Microbiol* 2006; **44**: 4566–8.

36 Kim H, Nakajima C, Yokoyama K *et al*. Impact of the E540V amino acid substitution in GyrB of *Mycobacterium tuberculosis* on quinolone resistance. *Antimicrob Agents Chemother* 2011; **55**: 3661–7.

37 Pantel A, Petrella S, Veziris N *et al*. Extending the definition of the GyrB quinolone resistance-determining region in *Mycobacterium tuberculosis* DNA gyrase for assessing fluoroquinolone resistance in *M. tuberculosis*. *Antimicrob Agents Chemother* 2012; **56**: 1990–6.

38 Pantel A, Petrella S, Matrat S *et al*. DNA gyrase inhibition assays are necessary to demonstrate fluoroquinolone resistance secondary to *gyrB* mutations in *Mycobacterium tuberculosis*. *Antimicrob Agents Chemother* 2011; **55**: 4524–9.

39 Hirsch MS, Günthard HF, Schapiro JM *et al*. Antiretroviral drug resistance testing in adult HIV-1 infection: 2008 recommendations of an International AIDS Society-USA panel. *Top HIV Med* 2008; **16**: 266–85.

40 Schön T, Juréen P, Chryssanthou E *et al*. Rifampicin-resistant and rifabutin-susceptible *Mycobacterium tuberculosis* strains: a breakpoint artefact? *J Antimicrob Chemother* 2013; **68**: 2074–7.

41 Jacobson KR, Tierney DB, Jeon CY *et al*. Treatment outcomes among patients with extensively drug-resistant tuberculosis: systematic review and meta-analysis. *Clin Infect Dis* 2010; **51**: 6–14.

42 Rustomjee R, Lienhardt C, Kanyok T *et al*. A phase II study of the sterilizing activities of ofloxacin, gatifloxacin and moxifloxacin in pulmonary tuberculosis. *Int J Tuberc Lung Dis* 2008; **12**: 128–38.

43 Wang JY, Wang JT, Tsai TH *et al*. Adding moxifloxacin is associated with a shorter time to culture conversion in pulmonary tuberculosis. *Int J Tuberc Lung Dis* 2010; **14**: 65–71.

44 Takiff H, Guerrero E. Current prospects for the fluoroquinolones as first-line tuberculosis therapy. *Antimicrob Agents Chemother* 2011; **55**: 5421–29.

45 Maruri F, Sterling TR, Kaiga AW. A systematic review of gyrase mutations associated with fluoroquinolone-resistant *Mycobacterium tuberculosis* and a proposed gyrase numbering system. *J Antimicrob Chemother* 2012; **67**: 819–31.

46 Sandgren A, Strong M, Muthukrishnan P *et al*. Tuberculosis drug resistance mutation database. *PLoS Med* 2009; **6**: e2.

47 Cavusoglu C, Turhan A, Akinci P *et al*. Evaluation of the Genotype MTBDR assay for rapid detection of rifampin and isoniazid resistance in *Mycobacterium tuberculosis* isolates. *J Clin Microbiol* 2006; **44**: 2338–42.

48 Somoskovi A, Dormandy J, Mitsani D *et al*. Use of smear-positive samples to assess the PCR-based genotype MTBDR assay for rapid, direct detection of the *Mycobacterium tuberculosis* complex as well as its resistance to isoniazid and rifampin. *J Clin Microbiol* 2006; **44**: 4459–63.

49 Aslan G, Tezcan S, Emekdas G. Evaluation of the genotype MTBDR assay for rapid detection of rifampin and isoniazid resistance in clinical *Mycobacterium tuberculosis* complex clinical isolates. *Mikrobiyol Bul* 2009; **43**: 217–26.

50 Hillemann D, Rusch-Gerdes S, Richter E. Feasibility of the GenoType MTBDRsl assay for fluoroquinolone, amikacin, capreomycin, and ethambutol resistance testing of *Mycobacterium tuberculosis* strains and clinical specimens. *J Clin Microbiol* 2009; **47**: 1767–72.

51 Huang WL, Chi TL, Wu MH *et al*. Performance assessment of the GenoType MTBDRsl test and DNA sequencing for detection of second-line and ethambutol drug resistance among patients infected with multidrug-resistant *Mycobacterium tuberculosis*. *J Clin Microbiol* 2011; **49**: 2502–8.

52 Brossier F, Veziris N, Aubry A *et al*. Detection of GenoType MTBDRsl test of complex mechanisms of resistance to second-line drugs and ethambutol in multidrug-resistant *Mycobacterium tuberculosis* complex isolates. *J Clin Microbiol* 2010; **48**: 1683–9.

53 Said HM, Kock MM, Ismail NA *et al*. Evaluation of the GenoType MTBDRsl assay for susceptibility testing of second-line anti-tuberculosis drugs. *Int J Tuberc Lung Dis* 2012; **16**: 104–9.

54 Srivastava S, Garg A, Ayyagari A *et al*. Nucleotide polymorphism associated with ethambutol resistance in clinical isolates of *Mycobacterium tuberculosis*. *Curr Microbiol* 2006; **53**: 401–5.

55 Ramaswamy SV, Amin AG, Goksel S *et al*. Molecular genetic analysis of nucleotide polymorphisms associated with ethambutol resistance in human isolates of *Mycobacterium tuberculosis*. *Antimicrob Agents Chemother* 2000; **44**: 326–36.

56 Srivastava S, Ayyagari A, Dhole TN *et al*. *embB* nucleotide polymorphisms and the role of *embB*306 mutations in *Mycobacterium tuberculosis* resistance to ethambutol. *Int J Med Microbiol* 2009; **299**: 269–80.

57 Van Deun A, Barrera L, Bastian I *et al*. *Mycobacterium tuberculosis* strains with highly discordant rifampin susceptibility test results. *J Clin Microbiol* 2009; **47**: 3501–6.

58 Hazbon MH, Motiwala AS, Cavatore M *et al*. Convergent evolutionary analysis identifies significant mutations in drug resistance targets of *Mycobacterium tuberculosis*. *Antimicrob Agents Chemother* 2008; **52**: 3369–76.

59 Farhat MR, Shapiro BJ, Kieser KJ *et al*. Convergent evolution reveals targets of positive selection in drug resistant *Mycobacterium tuberculosis* strains. *Nat Genet* 2013; in press.