The Evolution of Drug Resistant Mycobacterium Tuberculosis

Professor Sarah M. Fortune Christopher B. Ford

The evolution of drug resistant *Mycobacterium tuberculosis*

Abstract

Mycobacterium tuberculosis (Mtb) poses a global health catastrophe that has been compounded by the emergence of highly drug resistant Mtb strains. We used whole genome sequencing (WGS) to directly compare the accumulation of mutations in Mtb isolated from cynomolgus macaques with active, latent and early reactivation disease. Based on the distribution of single nucleotide polymorphisms (SNPs) observed, we calculated the mutation rates for these disease states. Our data suggest that during latency, Mtb acquires a similar number of chromosomal mutations as would be expected to emerge in a logarithmically growing culture over the same period of time despite reduced bacterial replication during latent infection. The pattern of polymorphisms suggests that the mutational burden *in vivo* is due to oxidative DNA damage.

We next sought to determine why some strains of Mtb are preferentially associated with high-level drug resistance. We demonstrate that Mtb strains from the East Asian lineage acquire drug resistances *in vitro* more quickly than Mtb strains from the Euro-American lineage. Their higher drug resistance rate *in vitro* reflects a higher basal mutation. Moreover, the *in vitro* mutation rate correlates well with the bacterial mutation rate in humans as determined by whole genome sequencing of clinical isolates. Finally, using an agent-based model, we show that the observed differences in mutation rate predict a significantly higher probability of multi-drug resistance in patients infected with East Asian lineage strains of Mtb.

Lastly, we sought to determine the mechanisms Mtb uses to proofread nascently polymerized DNA. Through fluctuation analysis of deletion mutants of two potential polIIIε

homologs, we demonstrate that neither is responsible for the maintenance of DNA replication fidelity. To explore the possibility that one of these homologs, Rv3711c, participates in an unknown redundant pathway, we used transposon capture and sequence (TraCS) to identify genes conditionally essential in an Rv3711c deletion mutant. Our analysis suggests that while Rv3711c does not participate in proofreading, it may act in an alternative novel DNA repair pathway. Taken together, our fluctuation analysis and TraCS data suggest that mycobacteria do not use canonical methods of proofreading to maintain genomic fidelity.

Table of Contents

List of Figures and Tables *(in order of reference in text)*

Chapter 1

Chapter 2

Chapter 3

Chapter 3 (continued)

Chapter 4

Chapter 1 – Introduction

1.1 *Mycobacterium tuberculosis* infection and treatment

Mycobacterium tuberculosis (Mtb), a pathogen of both ancient and modern man[1-3], continues to cause a global health catastrophe that has in recent times been compounded by the emergence of highly drug resistant strains of Mtb[4-7]. Infection follows inhalation of aerosolized bacteria, and leads to either active disease or, in the majority of patients, latent infection[8] (**Figure 1.1)**. From the time of infection, these latently infected individuals have a decreasing likelihood of reactivating and developing active disease. Immunocompromise, through HIV infection, senescence or as a result of medication, greatly increases the chance of reactivation. In active tuberculosis, patients harbor large a large bacterial burden at diagnosis; in contrast latent infection is thought to be characterized by smaller population sizes with potentially reduced capacity for growth and mutation.

The emergence of drug resistant strains has greatly compromised treatment, leading to increased mortality, increased costs, and a desperate need for novel antibiotics. Since the advent of antibiotic therapy for the treatment of tuberculosis, drug resistance has been an ever-present complication to successful treatment[9]. This is in large part due to the mutational capacity of the bacterial population – a product of mutation rate and bacterial population size. Multidrug therapy was advanced as a solution to the rapid emergence of drug resistance during monotherapy[10], and is in large part successful[11]. Though the initial combination of PAS and streptomycin has since been supplanted by newer, more effective antibiotics[12], the fundamental observation remains unchanged. Successful treatment of tuberculosis requires prolonged simultaneous therapy with multiple antibiotics[11].

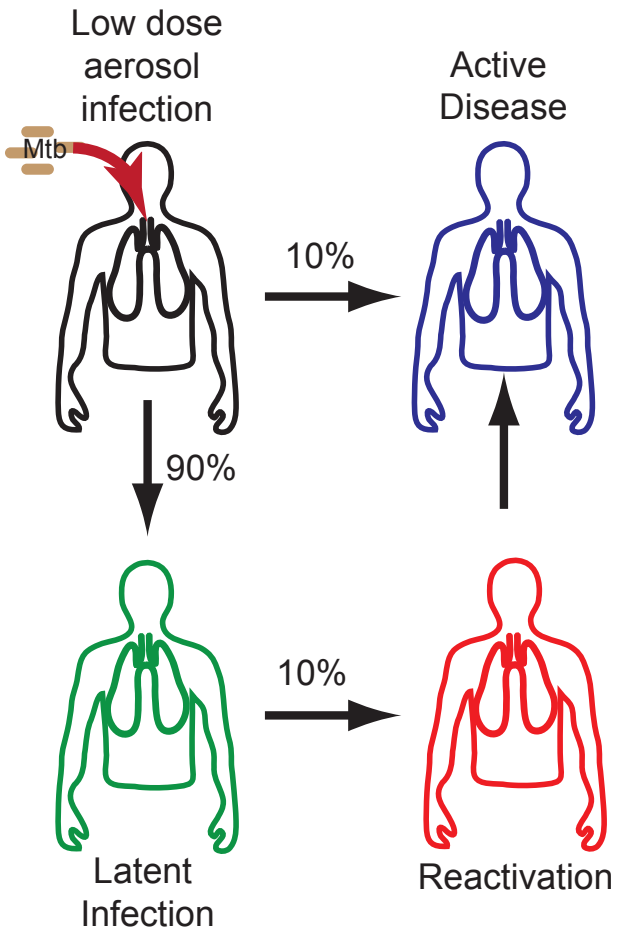


Figure 1.1 Aerosol spread of Mtb leads either to active disease or latent infection. Inhalation of aerosolized droplets containing Mtb leads to phagocytosis of Mtb by resident aveolar macrophages. In patients developing active disease, bacterial replication proceeds unchecked within the macrophage phagosome, leading to the development of primary and disseminated granulomas as disease progresses. In latently infected individuals, bacterial dissemination is controlled, and the infection fails to spread or cause symptoms. Latently infected individuals are at risk for later reactivation, leading to active disease.

Despite the application of multiple antibiotics, strains of Mtb resistant to many or all antibiotics have arisen[5,6,13,14]. As mycobacteria have little capacity for horizontal gene transfer and reside in a restricted and isolated niche, genetic diversity is principally driven by chromosomal mutation. Indeed, in Mtb, all drug resistances are the result of chromosomal mutation and depend on the bacterium's capacity for mutation during the course of infection. In the absence of horizontal gene transfer, the most basic model of mutation predicts that the probability of drug resistance is the product of each rate of resistance for the components of multidrug therapy and bacterial population size. While estimates of *in vivo* bacterial burden at the time of diagnosis and treatment are difficult to obtain, over the range of possible values, multidrug resistance should emerge rarely if at all[15].

How does Mtb generate sufficient genetic diversity to develop multidrug resistance? To properly address this question it is important to first understand the natural genetic diversity found in populations of Mtb. With the rapid development of whole genome sequencing (WGS) technologies, we have unprecedented capacity to detect genetic diversity. WGS is particularly powerful when applied to Mtb, which is characterized by a relatively low amount of genetic diversity that requires high resolution to be fully captured. WGS data has allowed us to reconstruct the phylogeny of Mtb, and in the process learn a great deal about it's geographic distribution[16,17]. More recently, studies have investigated the dynamics of evolution, transmission and treatment across shorter time scales[18-22]. By sequencing strains from small outbreaks and single infections, groups have sought to understand the unique evolutionary dynamics inherent in the spread of tuberculosis[20]. Using WGS, we can address a broad range of topics - from questions on the transmission and fitness of clinical strains to how Mtb evolves over long and short time scales. Here we have reviewed the insights gained from the use of

WGS and discuss those areas still to be addressed, moving from global (phylogeography), to local (transmission chains and circulating strain diversity), to the single patient (clonal heterogeneity), to the bacterium itself (evolutionary studies), and finally discussing the platform of WGS, its strengths and current limitations.

1.2 Genetic Diversity in *M. tuberculosis.*

1.2.1 Global diversity

WGS has been proposed as a sort of "gold standard" for strain typing in Mtb. As such, it clarifies previous strain typing approaches used for phylogenetic and epidemiologic studies. The standard genotyping methods are based on repetitive elements that provide limited functional information and are highly prone to convergent evolution, limiting their application to phylogenic reconstructions. The discriminative power of these methods varies, the results of different methods do not always agree, and the diversity of the markers can complicate the analysis[16,23]. For example, Niemann et al sequenced two isolates of the rapidly spreading Mtb Beijing genotype clone from a high incidence region (Karakalpakstan, Uzbekistan)[24]. The isolates possessed the same genotype by IS6110 restriction fragment length polymorphism (RFLP) and mycobacterial interspersed repetitive unit – variable number tandem repeat (MIRU-VNTR) patterns; however, WGS demonstrated that they differed at 131 separate sites, including one large deletion. Thus, typing methods can miss substantial amounts of genetic diversity, and where the overall diversity of circulating clones is limited, standard typing measures are insufficient to discriminate between strains.

As the cost of WGS drops, it is more feasible to consider WGS as a primary tool for the typing of Mtb strains that is not subject to the limitations of standard methodologies. However, use of WGS sequence data is subject to its own set of errors. To demonstrate the power and the

pitfalls of this approach, we undertook a phylogenetic analysis of Mtb strains using all of the publically available Mtb genomes. We utilized 55 whole (or nearly whole) genomes downloaded and assembled from GenBank, the Broad Institute, or obtained through personal communication from the authors of published papers by January 15 of 2011 (**Table 1.1**). While each genome included more than 4 million bases, these genomes were not fully assembled, nor were they annotated in the same manner. To enable an inclusive comparison, we blasted these genomes against annotated genes of the reference strain F11 to form multiple strain alignments for each gene, including all available strains. Variable sequence positions or SNPs, where at least 1 of 55 sequences differed from other sequences, were then extracted and concatenated, in the order they appear in the reference F11 genome, to create an abbreviated multiple alignment of SNP positions from all 55 sequences. A phylogenetic tree was built from these sequences by molecular parsimony using PAUP[25] (**Figure 1.2a**). Genes or regions of the genes that were misaligned, where one or several sequences had a high density of SNPs in close proximity, indicating possible either sequencing or alignment problems, were excluded from the analysis using a computational clean-up algorithm; the resulting trees are provided in Figure 1. The 17740 aligned positions were in 3324 genes and included a mutation in at least one of 55 strains that differed from the other sequences.

The relatively high number of SNPs we identified is in part the result of natural variation, as we have included all genes from 55 nearly complete globally diverse strains, and it is in part a consequence of the technical error present in some sequences. In particular, several strains out of the 55 available were highly enriched in regional clusters of SNPs, suggesting potentially problematic base calls (**Table 1.1**). The number of clustered SNPs was unusually high in several strains (8 strains have greater than 1000 regional clusters of SNPS), which indicates these

sequences tended to be noisier than the others. While such clustered SNP regions were excluded from our phylogenetic analysis, the not-clustered, stand-alone SNPs in those same strains were included in our collection of SNPs for phylogenetic analysis. It is difficult to algorithmically resolve whether any given mutation is of biological origin or is a sequencing artifact. Not surprisingly, the strains that have the highest number of clustered SNPs (**Table 1.1**) also have unusually long terminal branches (**Figure 1.2a**), a further indication of a higher frequency of sequencing error in these strains. As our intent was to review the whole genome data currently publically available, however, we included all 55 sequences here. While the long terminal branch lengths are likely to be contributed to by sequencing error, the approximate location of these sequences in the tree is of interest, as the phylogenetically informative SNPs that determine the branching order are likely to be valid. Also, all 55 genomes may offer useful sequence information in particular genes of interest. However, potential problems in these strains should be considered for any subsequent analysis. In particular, polymorphisms relating to drug resistance and any further studies based on these whole genomes should include manual inspection of the alignments. The reconstructed phylogeny of 55 whole genomes (**Figure 1.1a**) was compared to a phylogeny based on a minimal set 45 SNP positions determined by Filliol et al[16] to be sufficient to resolve and differentiate Mtb and *M. bovis* isolates (**Figure 1.1b**). In general, all main lineages were preserved and there is a high level of correspondence between the two trees. Two of the three Beijing strain sub-lineages present in the whole genome tree converged into 1 sub-lineage in the 45-SNP tree. Additionally, F11, the KZN strains, and the SUMu strains from Canada were related but clearly distinguishable on the whole genome tree, but they were collapsed on the 45-SNP trees. In contrast, one sub-lineage of the Beijing

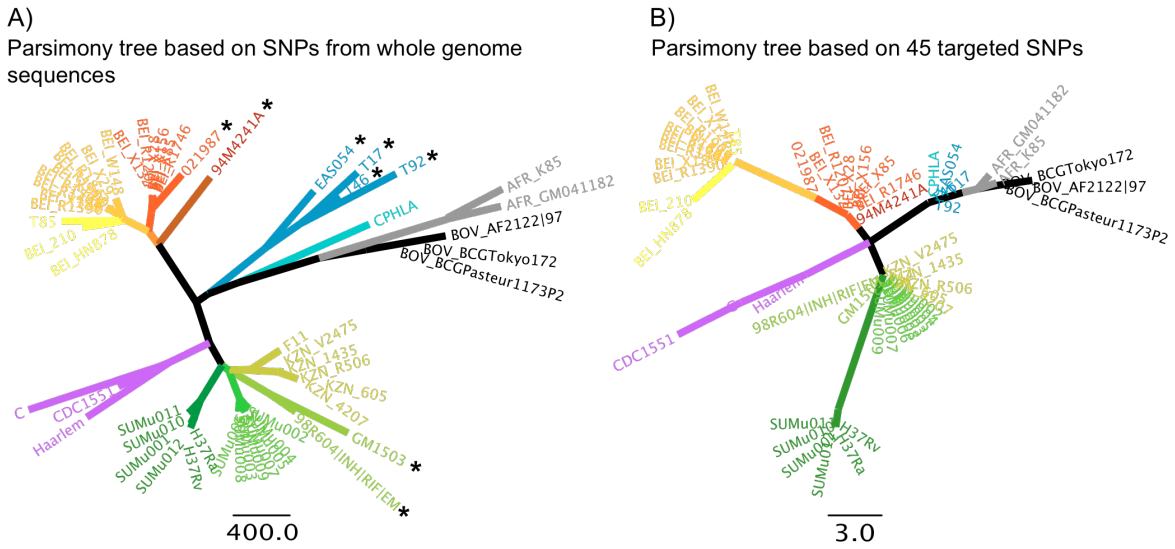


Figure 1.2 Comparison of phylogenies based on whole genome and targeted SNPs from the available WGS data (Table 1.1). The phylogenies were reconstructed using molecular parsimony in PAUP software[25]. Different lineages are colored in distinct colors, with sub-lineages colored in different shades of the same color. **(a)** Parsimony tree based on 17740 SNP positions taken from 3324 genes, where at least 1 of 55 sequences differed from other sequences. SNPs were extracted from multiple sequence alignments created for each gene using the following algorithm. Each annotated gene, from a total of 3941 genes for the reference strain F11, was used to perform pair-wise BLASTN search against the other 54 whole genomes. If the resulting partial sequence was shorter than the gene in F11, it was augmented by adjacent chromosome sequence to the length of the gene of F11. The homologous genes from 55 strains were then assembled into files so that each gene had a corresponding file. After assembling homologous sequences, the sequences were aligned using MUSCLE (version 3.8.31) to generate nucleotide multiple sequence alignments for homologous gene files (total 3596) that has at least one non-identical sequence. Resulting gene alignments were used to generate artificial SNP sequences, which contained a concatenated version of all SNPs identified (where at least 1 of 55 sequences differed from other sequences). To minimize inclusion of sequencing errors in these artificial SNP sequences, we excluded SNPs that appeared to be clustered in a local region, specifically where three SNPs were found in a 10 base pair window when compared to genes in F11. Furthermore we manually checked the files that have are enriched for clusters of mutations and excluded 70 problematic files. Sequences that are highly enriched for clusters of potentially problematic bases (**Table 1.1**) are indicated with the stars. The strains that have the highest number of clustered SNPS (**Table 1.1**) also have unusually long terminal branches, a further indication of sequencing artifacts in these strains. **(b)** Parsimony tree based on 45 SNP positions from Filliol et al[16] taken from the same set of 55 genomes.

genotype formed its own lineage on the 45 SNP trees (orange cluster, **Figure 1.1a**, **1.1b**). Similarly, H37 and close SUMu strains from Canada appeared further from F11, KZN and other SUMu strains from Canada on the 45-SNP tree than they appeared on a whole genome tree. Taken together, this confirms that the 45 positions determined by Filliol et al can successfully resolve the same main lineages that appear in the whole genome analysis and thus can be used for initial characterization of isolates, but WGS provides added resolution, which may be of value to applications that require finer resolution data.

Whenever available for the near full-length genomes, we recorded an isolate's sample history: the year and geographic location of isolate collection, and patient history including place of birth and drug resistance status (**Table 1.1**). This allowed us to relate the isolate's phylogeny with their geographic location and the date of sampling (**Figure 1.3**). In agreement with Filliol et al., the 55 whole genomes fall into 7 major lineages. We found 4 distinctive sub-lineages of the Beijing strain: 2 sub-lineages formed from strains isolated in USA in 1990-1998 and 2 sub-lineages formed from strains isolated in Western Cape, South Africa. Perhaps not surprisingly, one of the South African Beijing lineages also includes a strain isolated in San Francisco in 2002 (**Table 1.1, Figure 1.3**).

In agreement with previous studies[16,17,24], the tree revealed large genetic distances between isolates that were designated related to the Beijing strain. The distances between different strains within the Beijing sub-lineages appear to be roughly comparable to the distances between different F11 and KwaZulu Natal (KZN) lineages, hence the latter are marked by the same light green background color on **Figure 1.3**. The distances between the F11 / KZN sub-cluster and Canadian SUMu sequences and H37Rv and H37Ra (colored in distinct shades of green on Figure 2) are comparable to the distances between Beijing sub-clusters. Thus the

genetic distances in the hierarchy of relationships in Mtb lineages are not always consistently represented by commonly used nomenclature conventions and reference strains.

1.2.2 Local Diversity

WGS derived phylogenies build a framework upon which questions about evolution, transmission, and drug resistance can be asked. The last of these is of particular interest as drug resistant strains have stymied the treatment of tuberculosis, leading to the need for novel therapeutics and carefully designed treatment regimens. When different levels of drug resistance are observed among highly related strains, such as the KZN strains and the Beijing strains from the Western Cape of South Africa, it provides an opportunity for improved clarity and resolution in mapping the acquisition and spread of drug resistance.

Using such comparisons, it has become clear that highly drug resistant strains are emerging independently in the same geographic locale to a greater degree than previously appreciated. For example, in the Western Cape region of South Africa, the Beijing XDR strains are closely intermingled with MDR strains (**Figure 1.2**). Standard genotyping methods suggested that the Beijing XDR strains emerged once but were undergoing clonal expansion and transmission in the region. However, WGS-based phylogenetic analysis revealed the independent appearance of distinct XDR resistance mutations within different MDR sub-lineages of the Beijing genotype[19]. Similarly, in KwaZulu Natal, South Africa, the XDR KZN strains, isolated in 2005 and 2006, have a slightly longer phylogenetic distance from their most recent common ancestor of the KZN lineage than the MDR KZN strains, isolated in 1994, suggesting at first glance that the XDR KZN strains evolved stepwise from the MDR strains. However, upon further inspection of the WGS data, the XDR strains and MDR strains have different rifampicin

Table 1.1. Summary of 55 strains used in the analysis

Strain	Genotype	GB ID	Source	Place of isolation	Year	DR Status	Clustered SNPs	Comment
F11	F11	148719718	Broad CDB	Western Cape SA		DS	0	
V2475	KZN	297718568	GB, loerger2009	Durban, KZN, SA	1994	MDR	3	Resistant to RIF INH STR
1435	KZN	253318418	Broad CDB	Durban, KZN, SA	1994	MDR	28	Resistant to RIF INH STR
R506	KZN	297718569	GB, loerger2009	Durban, KZN, SA	2006	XDR	3	Resistant to INH RIF STR OFL KAN
605	KZN	289552250	Broad CDB	Tugela Ferry, KZN, SA	2005	XDR	28	First full XDR genome, resistant to INH RIF STR OFL KAN ETH
4207	KZN	295687135	Broad	Durban, KZN, SA	1995	DS	87	
GM1503		193506183	Broad DDB	Gambia	2002		**1501** *	
98R604 **		220031339	Broad CDB			MDR	**1403** *	genomesonline.org notes the patient from KZN, SA
SUMu001..012		multiple	Broad NMR	Canada		DS	28..281	12 endemic Canadian strains
H37Ra	H37	148503909	GB		1905	DS	38	Avirulent strain derived from virulent H37
H37Rv	H37	57116681	GB		1905	DS	43	Virulent strain derived from original virulent H37
Haarlem	Haarlem	115299839	Broad CDB	Netherlands		MDR	590	Has an accelerated transmission rate in crowded conditions
CDC1551		50952454	GB	KY/ TN, USA	~1994	DS	387	1994-1996 outbreak in a rural community; highly contagious
C		81248475	Broad CDB	New York, USA	1997	DS	569	Caused a large proportion of new tuberculosis cases in New York
HN878	Beijing	315064806	GB	Houston, TX, USA	1990s	DS	27	Hyper-virulent; part of TB outbreak in the 1990s
210	Beijing	261746034	GB	TX, US	~1993	DS	382	1993-1995 outbreak. One of the most virulent M.tb strains;
T85		189490718	Broad DDB	San Francisco, CA	1998		617	Patient born in China
R1390	Beijing	Not in GB	loerger2010	Western Cape, SA		Mono	27	Resistant to INH
X189	Beijing	Not in GB	loerger2010	Western Cape, SA		XDR	27	Resistant to INH RIF AMI CAP OFL KAN
R1441	Beijing	Not in GB	loerger2010	Western Cape, SA		Mono	27	Resistant to INH
R1505	Beijing	Not in GB	loerger2010	Western Cape, SA		MDR	27	Resistant to INH RIF
X122	Beijing	312987796	GB, loerger2010	Western Cape, SA		preXDR	27	Resistant to INH RIF OFL
R1909	Beijing	Not in GB	loerger2010	Western Cape, SA		MDR	27	Resistant to INH RIF EMB
R1842	Beijing	Not in GB	loerger2010	Western Cape, SA		Mono	27	Resistant to INH
X29	Beijing	Not in GB	loerger2010	Western Cape, SA		preXDR	27	Resistant to INH RIF AMI STR KAN
X132	Beijing	Not in GB	loerger2010	Western Cape, SA		preXDR	9	Resistant to INH RIF AMI CAP STR KAN
R1207	Beijing	312984523	GB, loerger2010	Western Cape, SA		MDR	9	Resistant to INH RIF
X28	Beijing	Not in GB	loerger2010	Western Cape, SA		XDR	9	Resistant to INH RIF AMI CAP OFL STR KAN
X156	Beijing	Not in GB	loerger2010	Western Cape, SA		preXDR	9	Resistant to INH RIF AMI CAP STR KAN
X85	Beijing	Not in GB	loerger2010	Western Cape, SA		XDR	9	Resistant to INH RIF AMI OFL KAN ETH
R1746	Beijing	Not in GB	loerger2010	Western Cape, SA		MDR	9	Resistant to INH RIF
W148	Beijing	326590567	Broad CDB			MDR	324	Highly multi-drug resistant
021987		189215797	Broad DDB	San Francisco, CA	2002		**1876** *	Patient born in South Korea
94M4241A		189212844	Broad DDB	San Francisco, CA	1994		**1665** *	Patient born in China
EAS054		189213750	Broad DDB	San Francisco, CA	1993		**1043** *	Patient born in India
T17		192338437	Broad DDB	San Francisco, CA	1995		**1808** *	Patient born in The Philippines
T46		225935560	Broad DDB	San Francisco, CA	1996		**1514** *	Patient born in The Philippines
T92		189213624	Broad DDB	San Francisco, CA	1999		**3200** *	Patient born in The Philippines
CPHLA		225935638	Broad DDB	California			638	Patient born in South Africa
K85	M.africanum	225935793	Broad DDB	Netherlands			896	
GM041182	M.africanum	Not in GB	TBDB	West Africa			820	
AF2122|97	M.bovis	31742509	GB	United Kingdom	1997		869	Isolated from a cow with necrotic lesions in lung
BCGTokyo172	M.bovis	224771496	GB				603	BCG vaccine substrain
BCGPasteur **	M.bovis	121491530	GB				731	Was used to produce BCG vaccine

Abbreviations: GB – GenBank; Broad – Broad Institute M.tb databases; CDB – Comparative Database; DDB – Diversity Database; NMR – Natural Mutation Rate Project;
TBDB – genome.tbdb.org; loerger2010 – loerger *et al*, 2010 (PMID: 21110864); loerger2009 – loerger *et al*, 2009 (PMID: 19890396)
* Strains that are highly enriched for clusters of potentially problematic bases; ** 98R604 and BCGPasteur strains have longer names: 98R604|INH|RIF|EM and BCGPasteur1173P2

Table 1.1 (Continued)
We utilized 55 whole (or nearly whole) genomes downloaded and assembled from GenBank, the Broad Institute, or obtained through personal communication from the authors of published papers by January of 2011. This table provides the number of clustered SNPs for each sequence. For several strains high numbers of clustered SNPs were unusual (8 strains have greater than 1000 regional clusters of SNPS), and indicated these sequences tended to be more problematic than the others. While such clustered SNP regions were excluded from our phylogenetic analysis, the not-clustered, stand-alone SNPs in those same strains were included in our collection of SNPs for phylogenetic analysis. Whenever available for the near full-length genomes, we recorded isolate's sample history: the year and geographic location of isolate collection, and patient history including place of birth and drug resistance status. The total number of bases where at least one of 55 strains differed from other sequences was 17740, in 3324 genes.

Figure 1.3

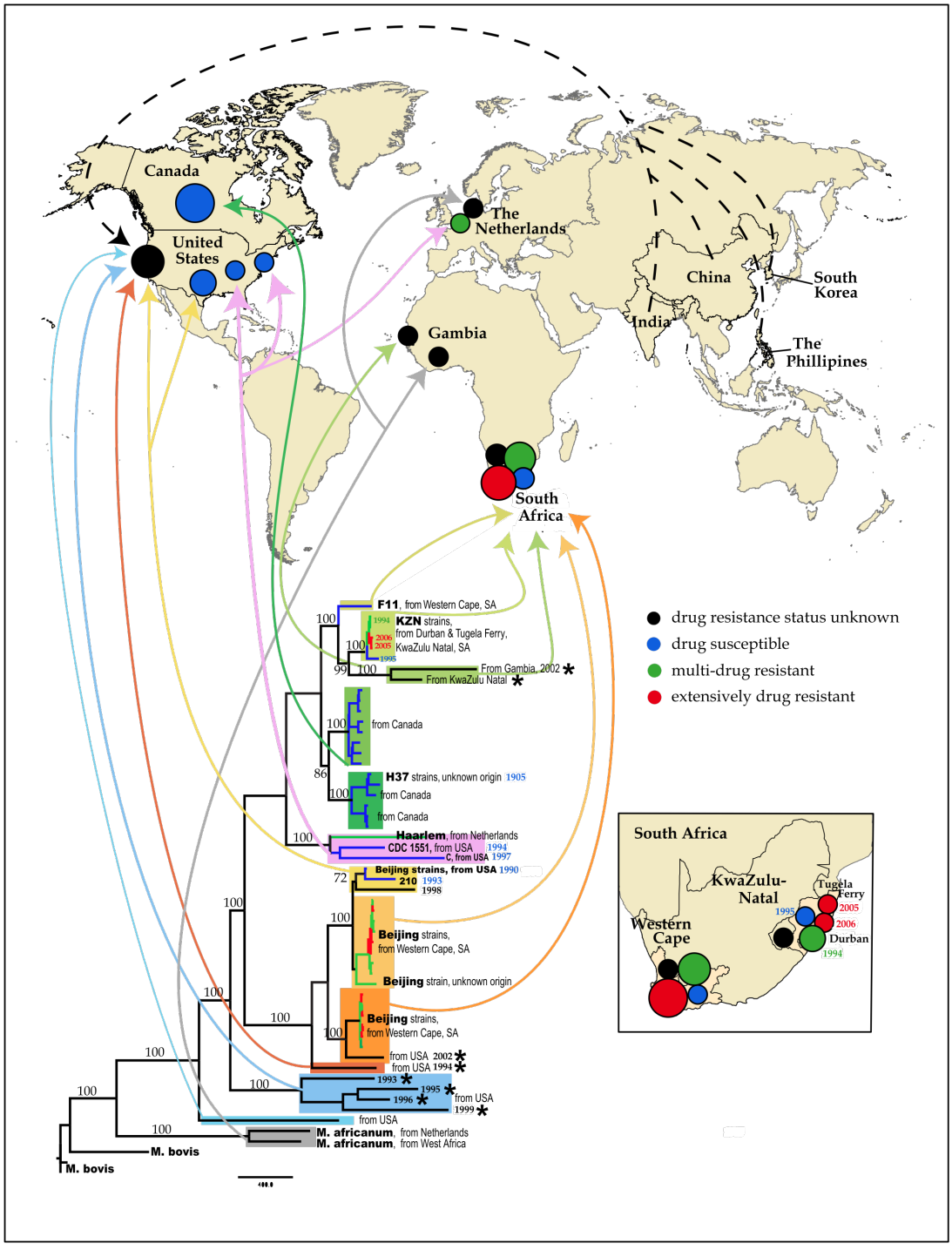

Canada
United States
The Netherlands
China
South Korea
India
The Phillipines
Gambia
South Africa
drug resistance status unknown
drug susceptible
multi-drug resistant
extensively drug resistant
F11, from Western Cape, SA
KZN strains, from Durban & Tugela Ferry, KwaZulu Natal, SA
From Gambia, 2002
From KwaZulu Natal
from Canada
H37 strains, unknown origin 1905
from Canada
from Canada
Haarlem, from Netherlands
CDC 1551, from USA 1994
C, from USA 1997
Beijing strains, from USA 1990
210 1993
1998
Beijing strains, from Western Cape, SA
Beijing strain, unknown origin
Beijing strains, from Western Cape, SA
from USA 2002
from USA 1994
1993
1995
1996
from USA 1999
from USA
M. africanum, from Netherlands
M. africanum, from West Africa
M. bovis
M. bovis
South Africa
KwaZulu-Natal
Western Cape
Tugela Ferry
2005
2006
Durban
1994
1995

Figure 1.3 (Continued) Phylogenetic lineages and geographic mapping. Whenever available for the near full-length genomes, we recorded isolate's sample history: the year and geographic location of isolate collection, and patient history including place of birth and drug resistance status (**Table 1.1**). This allowed us to relate the isolates phylogeny with their geographic location and the date of sampling. Lower part of the figure: same phylogenetic tree that as shown in **Figure 1.2a**. The numbers on the brunches represent bootstrap values obtained using 200 bootstrap replicates. Sequence names are removed except several reference strains, but the most relevant information (genotypes, isolation place, year of sampling) is noted on the tree. As in Figure 1A, sequences that are highly enriched for clusters of potentially problematic bases (Table 1) are indicated with the stars. Each lineage is shown with a distinct background box color, which corresponds to the colors of **Figure 1.2**. The sub-lineages are shown with the shades of the same background box color. For example, the Beijing strain-related lineage is shown in shades of yellow and orange. Additionally, drug resistance status is indicated by branch color. 4 colors for branches are used: black – drug resistant status of the strain is unknown; blue -- drug susceptible (DS); green – mono-resistant to INH or multi-drug resistant (MDR); red – extensively drug resistant (XDR), or pre-XDR (resistant to either fluoroquinolones or aminoglycosides). Arrows that are the same color as the background boxes from which they originate show the geographic places of the sequence isolation on the world map at the upper portion of the figure. The black, blue, green and red circles on the map correspond to the drug status of the isolates, the same coloring scheme as in the tree. The relative size of the circles corresponds to the number of whole genomes isolated in the particular location. The dashed black arrows pointing to the black circle in California represent isolates that were sequenced in San Francisco from patients that were born in India, China, South Korea and the Philippines. The more detail geographic distribution of whole genome isolates from South Africa is shown on insert at the right lower corner of the figure.

and pyrazinamide resistance mutations, indicating that these strains emerged independently from monoresistant isolates[18].

The high resolution of WGS based phylogenetic analysis has been informative in other public health settings, most notably in the context of outbreak tracing. A recent outbreak of Mtb was detected in Vancouver, British Columbia, and by standard typing methods, it was defined as a single clonal outbreak [20]. The authors initially applied MIRU/VNTR to determine a source case and transmission chain; however, the MIRU-VNTR pattern was identical across all isolates and the addition of contact tracing did not reveal a source case. The authors sequenced the genomes of 36 isolates, 32 from the outbreak and 4 historical isolates from the region with an identical MIRU-VNTR pattern. Concatenation of polymorphic loci and subsequent phylogenetic analysis revealed a dendrogram with two primary branches, indicating two distinct transmission chains. Overlaying the phylogenetic analysis with a social network analysis, Gardy et al identified the transmission chain through which the infection spread. This study serves as the prototype for the use of WGS in outbreak tracing for Mtb, demonstrating its effectiveness particularly in low-endemic countries where even high resolution (24-loci) MIRU-VNTR is likely to be uninformative.

1.2.3 Individual diversity

The resolution provided by WGS also allows researchers to investigate long-standing assumptions about the nature of tuberculosis. While Mtb infection is typically thought to be clonally homogeneous, recent studies have challenged this idea suggesting there is more heterogeneity within the infecting bacterial population than expected[26,27]. There are several potential mechanisms by which bacterial population heterogeneity might exist within a host –

(**1**) an individual may be simultaneously infected by multiple strains, (**2**) an individual may be super-infected, or reinfected by a new strain, or (**3**) genetic diversity may arise in the bacterial population spontaneously during the course of infection (**Figure 1.4**). While the level of heterogeneity created by each of these mechanisms varies, WGS provides both the depth of coverage and sensitivity necessary to accurately assess population heterogeneity and begin to probe its functional consequences.

Recurrent Mtb infection is a common clinical problem. Many studies have used a variety of genotyping techniques to identify differences between the original strain and the recurrent one. Where these strains are discordant, patients are typically assumed to have been infected with a second strain[28-33]. Indeed, mixed infections have been reported to occur in up to 54% of patients sampled[34]. Clinically, infection with multiple strains can lead to apparent differences in drug susceptibility[35-37], which may only be one aspect of the differences between strains[31] making diagnosis and treatment significantly more complicated. It is likely that with additional investigation and better sampling methods, mixed infections will be increasingly recognized, and the application of WGS will provide greater resolution in these studies, indentifying genetically distinct though highly related strains.

Within-host evolution of strains is another important source of genetic heterogeneity, and the principle source of *de novo* drug resistances. Saunders et al use deep resequencing of serial sputum isolates to characterize genetic diversity in a single patient infected with drug susceptible Mtb and in whom drug resistance evolved[22]. Serial isolates were obtained over a 12-month period from presentation with an initial drug susceptible infection through stepwise development of multiple drug resistance. Interestingly, only mutations conferring drug resistance were identified, with no additional mutations identified in non-repetitive regions. These results

suggest that neither the host immune system nor exposure to antibiotics generates a hypermutable state in Mtb. However, limits in sampling and tracking *in vivo* populations of Mtb make it difficult to fully understand the evolutionary dynamics of patient isolates, and such isolates can only be obtained from patients with active disease making it difficult to assess the evolution of the bacterium throughout all stages of infection.

Indeed, until recently, it was assumed that Mtb has little capacity to acquire new mutations in the host because the bacterium is presumed to be quiescent through much of that time. However, by applying WGS to Mtb isolates from the cynomolgus macaque model, we have recently shown that the mutation rate (mutations/bp/day) in active and latent disease is roughly equivalent, suggesting that bacteria continue to acquire genetic diversity, even during latency[21] (**Chapter 2**). Additionally, our results suggest that lesions are both genetically independent and genetically distinct, such that bacteria sampled from one lesion may not represent the true diversity present within the patient (**Figure 1.4b)**. Thus, if extrapulmonary dissemination occurs from any one lesion, the extrapulmonary sites would be genetically related to that lesion but distinct from others. These results have particular relevance when taken in the light of our primary sample source – sputa. Only bacteria present in cavitary lesions open to the airway will be present in sputa samples, and the cavitary lesions producing sputa may change dynamically during the course of infection as old lesions resolve and new lesions form. Future studies investigating in-host bacterial diversity may provide additional insight into these issues.

The overall picture of bacterial diversity is clearly a complex one – with diversity existing between patients, within a patient, and within a lesion. Mixed Mtb infection (whether by mixed inoculum, super-infection, or in-host evolution) raises a set of important questions.

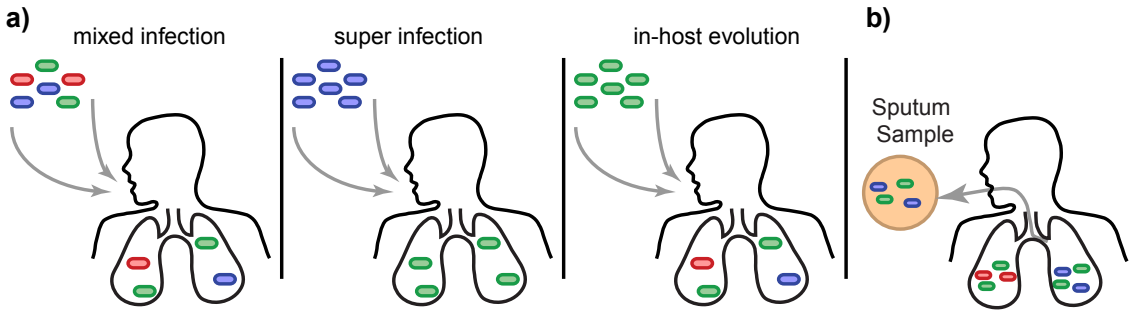


Figure 1.4 Sources of within individual genetic diversity a) In-host heterogeneity complicates the study of the genetic diversity and makes treatment and diagnosis of tuberculosis more difficult. There are several potential mechanisms by which heterogeneity might exist within the host – (**1**) a host may be simultaneously infected by multiple strains, (**2**) a host may be super-infected, or reinfected by a new strain, or (**3**) heterogeneity may arise spontaneously during the course of infection. **b)** Experimental results suggest that lesions are both genetically independent and genetically distinct, such that bacteria sampled from one lesion may not represent the true diversity present within the patient. Thus sputa samples are likely to under-estimate the amount of diversity present within a single patient, and serial sputa samples may not originate from the same lesion within the host.

Could apparent phenotypic classifications (such as drug resistance) based on culture sometimes be incorrect due to mixed populations, leading to misdiagnosis in clinical settings? Can we look at key SNPs from pooled sweeps of colonies to estimate their frequencies in mixed cultures and rapidly detect low frequencies of drug resistance mutations in an individual? Finally, is Mtb superinfection enhanced in HIV high prevalent populations? For HIV-infected individuals, rapid HIV-1 depletion of Mtb-specific CD4 T cells can aggravate Mtb infections[38], raising the possibility that immune dysfunction in HIV-infected people may make them more susceptible to serial Mtb infection, including superinfection with drug resistant strains. WGS will allow us to implement carefully crafted studies to address these important questions.

1.2.4 Bacterial diversity

The accumulation of WGS data allows us to assess the genetic diversity across the genome, seeking signatures of selective pressure. Selection can be quantified by relating the ratio of nonsynonymous genetic changes to synonymous changes (dN/dS), where a dN/dS greater than one is thought to reflect positive selection for increased diversity. Recently, Comas and colleagues used WGS to assess selection in a panel of 21 clinical strains[39]. Not surprisingly, essential genes had a lower dN/dS (0.53) than non-essential genes (0.66), indicating that while the genome overall is under purifying selection, essential genes are under greater purifying selective pressure. The authors then examined an experimentally defined set of T-cell epitopes[40], hypothesizing that these might represent regions of increased functional diversity as a mechanism of immune evasion. Intriguingly, the T-cell epitopes analyzed appear to show the greatest amount of sequence conservation, with a dN/dS less than that of essential genes (0.25 for the epitope-coding region of the ORF). This may suggest that Mtb growth and transmission requires T-cell recognition, and therefore that the bacterium actually benefits from the host T-cell

response. While further work is needed to clarify the dynamics of host-pathogen co-evolution in Mtb, these early results suggest that Mtb may depart from classic paradigms.

1.2.5 Challenges in WGS

While the capacity to perform low cost, high quality whole genome sequencing has transformed phylogenetic and population analyses in Mtb, there are some limitations with the current sequencing methodologies that can significantly skew our interpretation of the data. Most of the analyses described above hinge on the power of WGS to identify SNPs. Deletion or insertion of entire genes or large regions can be detected relatively easily (by absence of expected reads, or presence of novel contigs relative to a reference genome), and gene loss in particular has been frequently noted as a source of variability among mycobacteria[41-43]. However, polymorphisms in repetitive regions, gene duplications, chromosomal rearrangements, and copy-number changes of tandem repeats, are more challenging to detect by next-generation sequencing methods, such as Illumina, and can have significant biologic consequences. Paired-end read technology is quickly becoming the standard in generating WGS data, and offers some solutions to the problem of resolving these otherwise inaccessible regions.

Repetitive Regions: The limitations in sequencing repetitive regions apply to several genes and repeat elements scattered throughout the Mtb genome. These include some lipid biosynthetic genes as well as insertion elements, including IS6110, MIRUs and the clustered regularly interspaced short palindromic repeats (CRISPR) elements, which have been exploited extensively for strain typing. One example of the pitfalls associated with the sequencing of these regions has emerged from the debate over whether there is indeed purifying selection on T cell epitopes as suggested by Comas et al[39]. Uplekar et al. recognized that the genes encoding many of the ESX proteins, a family of secreted proteins that are represent strong CD4 and CD8 T cell

antigens, were extremely homologous and thus poorly assessed by Illumina technology. Thus, the authors used Sanger sequencing to resequence these genes from a panel of clinical isolates and found, contrary to the previously mentioned report, that some of these genes are highly polymorphic, in part because of high levels of recombination[44]. When this diversity is taken into account, these antigens appear to be under diversifying selection. Similarly other important antigens including the PE and PPE genes are simply not accessible to short read sequencing technology and thus may obscure significant antigenic variation.

Genomic Duplications: Large-scale genomic duplications have been observed among mycobacterial strains, and in some cases, are postulated to have an influence on phenotype. For example, some members of the Beijing strain family have recently been found to have a large-scale duplication of ~350kb in the region of Rv3128c to Rv3427c[45], including DosR, the transcriptional regulator of the hypoxic response[46]. This type of polymorphism is difficult to detect with short reads because there are multiple alternative ways to build contigs, producing ambiguity in assembly. Current advances in data analysis, including de Bruijn graph methods[47] and methods of statistical analysis[48] are designed to detect signatures of large duplications, often using variations in depth of coverage to detect when large regions have been copied.

Genomic rearrangements: Genome rearrangements are difficult to detect with short reads because of the localized nature of the lesion. Genomic sequence on either side of the cross-over point might be well-covered by reads, but detection of the rearrangement point itself requires longer reads (e.g. from Roche 454) that span the discontinuity, or paired-end/mate-pair data as evidence of the connectivity. However, genome rearrangements have not been reported among *M. tuberculosis* strains (although there is a chromosomal inversion between *M. tuberculosis* and *M. leprae*[49]. The wild-type (drug-sensitive) KZN 4207 strain isolate from the

KwaZulu-Natal region of South Africa was initially reported to have an inversion (http://www.broadinstitute.org/annotation/genome/mycobacterium_tuberculosis_spp/), although sequencing of the same strain in another lab did not find evidence for this inversion[18]. In practice, the large-scale genomic stability of Mtb justifies the use of a comparative assembly approach[50] in which sequencing data for new *M. tuberculosis* strains are aligned against H37Rv, F11, or the genome sequence of another representative strain for comparative analysis.

1.2.6 Future perspectives

Because of the paucity of genetic diversity in Mtb, WGS is a uniquely powerful tool, providing both the sensitivity to detect rare genetic events, and the broad applicability to detect multiple forms of genetic change. Already, we have learned a great deal about the bacterium and the nature of the disease. Early reports show surprising amounts of heterogeneity in bacterial populations, even between strains with identical MIRU/VNTR, RFLP patterns, or spoligotypes. Many questions remain, however. Perhaps chief amongst these is the true nature of in-host diversity: its clinical consequences and the mechanisms behind it. WGS has the capacity to address these and other questions with minimal bias and unprecedented sensitivity. However, it will be important to remember that current WGS methodologies do not cover all regions of the genome and there is likely to be important biology hidden in these uncharted areas.

Through the use of new tools, such as WGS, and traditional methodologies such as Luria Delbrück fluctuation analysis and molecular genetics, we can begin to quantify the determinants of Mtb mutation in the setting of host infection. Ultimately, these studies will allow us to better understand and predict the emergence of drug resistance and develop treatment protocols geared towards suppressing novel resistances.

1.3 Summary of Aims

The aims of this dissertation probe the dynamics and mechanism of mutation in Mtb, with the ultimate goal of improving our understanding of the evolution of drug resistance in this global pathogen. Prior to this work, all data on drug resistance rates and mutation rates in Mtb were derived from *in vitro* experimentation. In **Chapter 2**, the dynamics of mutation are quantified *in vivo* using the cynomolgus macaque model of disease and WGS. This per base mutation rate is comparable to the per base mutation rate observed *in vitro*. As the estimate of mutation rate was similar in both active disease and latent infection this work suggests that Mtb retains surprising capacity for mutation during latency.

Clinical evidence suggests that strains of the East Asian Beijing sublineage are more commonly associated with drug resistance, relative to their association with drug susceptible disease. To test the hypothesis that certain strains more rapidly acquire drug resistance, in **Chapter 3** we have used a combination of fluctuation analysis and mathematical modeling to demonstrate that the evolution of multidrug resistance is possible before the onset of treatment, and based on the estimated drug resistance parameters of East Asian strains, is more likely in patients infected with strains from this lineage.

The mechanisms responsible for repair, recombination, and replication of DNA in Mtb are unclear. From homology and limited experimental studies, it is becoming clear that these core processes are fundamentally divergent from other organisms. In **Chapter 4**, the role of potential proofreading enzymes is investigated. From both functional studies with exonuclease deletion mutants and transposon capture and sequencing experiments in those same mutants, it is clear that mycobacteria do not employ canonical mechanisms of proofreading to maintain genomic fidelity.

Author Contributions. Christopher B Ford prepared the figures and drafted the manuscript; Karina Yusim, Thomas Ioerger, Shihai Feng, Mary Greene, and Betty Korber conducted phylogenetic analyses; Sarah M Fortune drafted the manuscript.

Chapter 2 – Use of whole genome sequencing to estimate the mutation rate of *Mycobacterium tuberculosis* during latent infection

2.1 Introduction

Mycobacterium tuberculosis (Mtb) has generated a global health catastrophe that has been compounded by the emergence of drug resistant Mtb strains. In active tuberculosis, patients harbor large numbers of replicating organisms and are treated with multiple antibiotics to prevent the emergence of novel drug resistance mutations. In contrast, Mtb from latent infection is thought to have little capacity for mutation and is typically treated with a single antibiotic, isoniazid (INH). Recent epidemiologic studies have found that INH preventive monotherapy (IPT) for latent tuberculosis is associated with INH resistance[51,52]. In Mtb, all drug resistances are the result of chromosomal mutations and depend on the bacterium's capacity for mutation during the course of infection. Therefore, we seek to define the mutational capacity of the bacterium during infection to better predict the rate at which drug resistance can be expected to emerge in active, latent, and reactivated disease.

We used whole genome sequencing to compare the accumulation of mutations in Mtb isolated from cynomolgus macaques with active, latent and reactivated disease. Based on the distribution of SNPs observed, we calculated the mutation rates for these disease states. Our data suggest that Mtb acquires a similar number of chromosomal mutations during latency as occurs during active disease or in a logarithmically growing culture over the same period of time despite reduced bacterial replication during latent infection. The pattern of polymorphisms suggests that the mutational burden *in vivo* is due to oxidative DNA damage. Thus, we demonstrate that Mtb continues to acquire mutations during latency and provide a novel explanation for the

observation that isoniazid monotherapy for latent tuberculosis is a risk factor for the emergence of INH resistance[51,52].

2.2 Results

Conventional approaches to measuring bacterial mutation rates cannot be applied to Mtb *in vivo*. However, high-density whole genome sequencing (WGS) allows us to assess the capacity of Mtb for mutation over the course of infection with minimal bias and maximum sensitivity[53-55]. As the nonhuman primate is the only animal model that mimics the broad range of disease seen in human tuberculosis[56,57], we performed WGS on the infecting strain of Mtb, Erdman, and 33 isolates from nine cynomolgus macaques that represented the three major clinical outcomes of infection (active disease, persistently latent infection and spontaneously reactivated disease after prolonged latency[57]) (**Figure 2.1**). Genome coverage averaged 93% across these isolates, and average read depth was 117x across the genomes (**Supplementary Table 2.1**). Putative polymorphisms were identified using both a scaffolded approach[18,58] and a *de novo* assembly method[59], and polymorphic sites were validated by Sanger sequencing or through independent identification by WGS. Through these analyses, we identified 14 unique single nucleotide polymorphisms (SNPs) (**Figure 2.2**). There was no evidence that these SNPs were present in the inoculum either from repeated deep sequencing and PCR resequencing of the inoculum, or from shared polymorphisms between bacteria from different lesions. While we have used WGS previously to detect insertions and deletions[18], we did not detect either in the 33 genomes analyzed. Within lesions, we identified both shared and independent polymorphisms as would be expected if the SNPs accrued within lesions over the course of the infection.

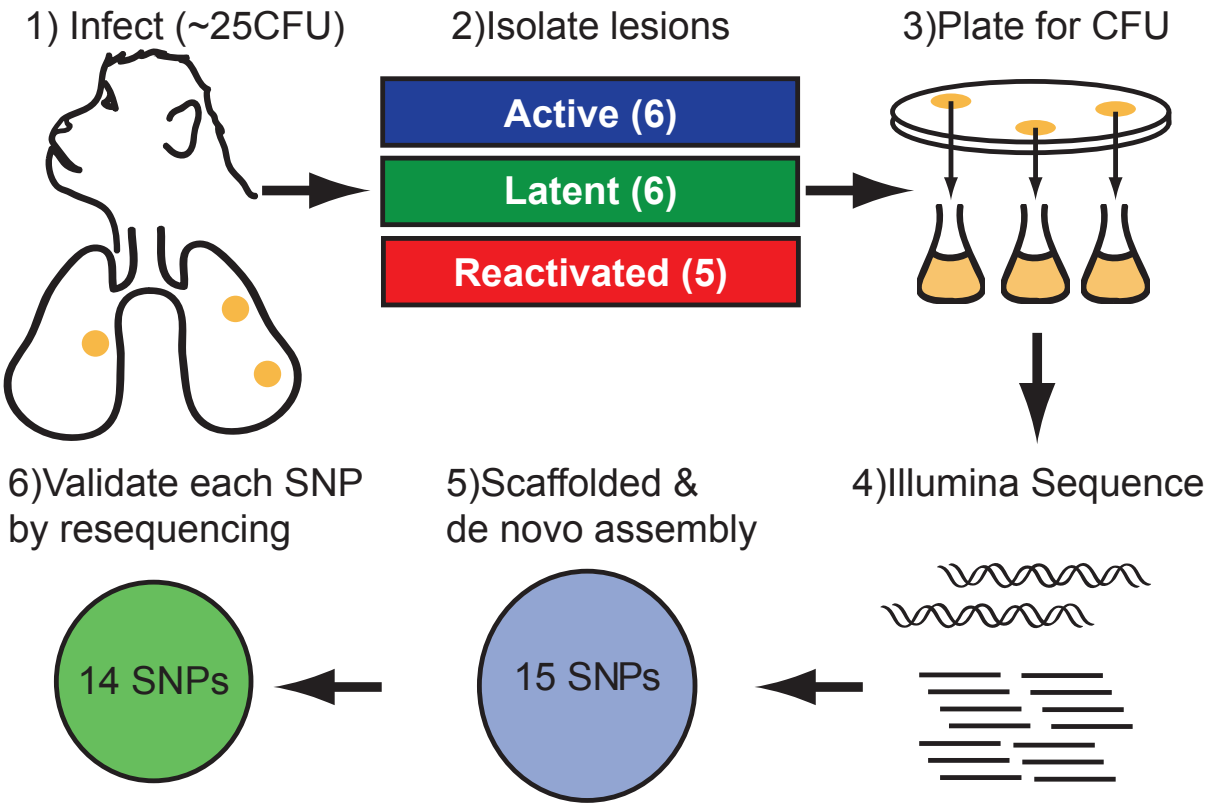


Figure 2.1 Experimental protocol for assessing mutational capacity in different disease states 1) Cynomolgus macaques were infected with ~25CFU of Mtb Erdman via bronchoscopy. 2) Animals were euthanized in the indicated stages of disease for strain isolation. 3) 18 pathologic lesions were plated for bacterial colonies. 33 strains were isolated for WGS. 4) Genomic DNA was isolated from these strains and then analyzed via Illumina sequencing. 5) Reads were assembled using both *de novo* and scaffolded approaches. 15 SNPs were predicted by both methodologies. Insertions and deletions were not detected using either methodology. 6) Sanger sequencing confirmed 14 of the 15 putative SNPs identified by both scaffolded and *de novo* analysis.

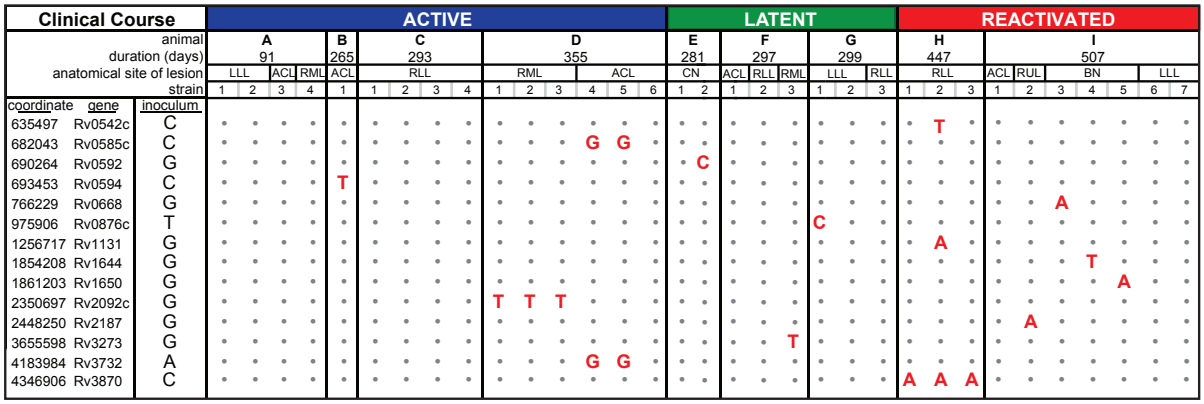


Figure 2.2 WGS identifies SNPs in strains isolated from animals with active, latent, and reactivated latent infection. SNPs were predicted through WGS in 33 Mtb strains isolated from nine cynomolgus macaques at various stages of disease. All SNPs predicted through WGS were confirmed via Sanger sequencing or through independent identification by WGS. Genome coverage and the original notation used to describe each animal are found in **Supplementary Table 1**. The total length of infection in days is listed for each animal below the animal identifier (A-I). Lesion locations are abbreviated as follows: LLL – left lower lobe, RLL – right lower lobe, RML – right middle lobe, RUL – right upper lobe, ACL – accessory lobe, CN – cranial lymph node, BN – bronchial lymph node. Inoculum represents the sequence at the given coordinate of the inoculating strain, Mtb Erdman.

We next sought to quantify the average mutation rate of the bacterium in the different stages of clinical disease. The mutation rate (μ) of a bacterium *in vivo* can be estimated from the number of mutations (m) that have occurred for a genome of known size (N) over a known number of generations (t/g where t is length of infection and g is generation time). However, the generation time of Mtb in humans or non-human primates is unknown. *In vitro*, Mtb has a generation time of approximately 20 hours under nutrient rich conditions[60]. In mice, the bacterial organ burden increases at roughly this rate during the first weeks of infection, but the subsequent onset of the adaptive immune response causes bacterial replication to slow substantially or cease entirely[61,62]. In clinical latency in nonhuman primates and humans, the immune response limits infection to the point that there are no clinical or radiographic signs of overt disease. This is thought to be associated with a dramatic slowing or cessation of bacterial replication, although we have recently shown that lesions from clinically latent cynomolgus macaques display a range of histopathology and bacterial burdens, suggesting a spectrum of bacterial physiologies may occur in latency[63,64].

Because of the inherent uncertainty in the generation time of Mtb *in vivo*, we estimated the mutation rate across a broad range of generation times (18-240 hours), calculating the rate that would be required to generate the number of polymorphisms identified by WGS (**Figure 2.3a-c**). In order to compare the mutation rate of bacteria from each clinical condition, we derived a lower limit for the bacterial mutation rate *in vivo*, which we define as the predicted mutation rate per generation if the *in vivo* generation time were equivalent to the *in vitro* generation time of 20 hours, μ(20hr). While Mtb is likely to have a much longer generation time *in vivo,* especially during prolonged latent infection, we use μ(20hr) as a highly conservative boundary estimate of the *in vivo* mutation rate that allows us to directly compare the mutational capacity of the

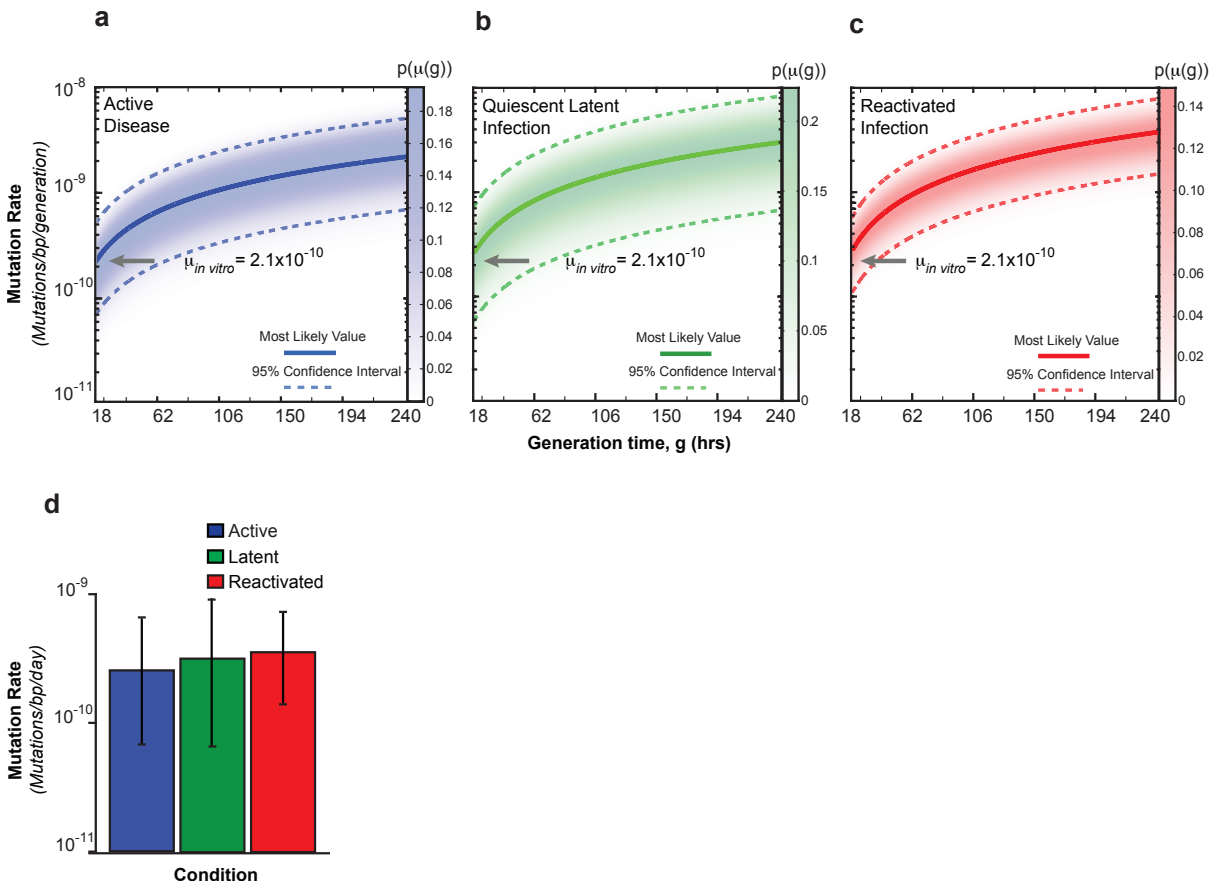


Figure 2.3 The mutational capacity of strains from latency and reactivated disease is similar to that of strains from active disease or *in vitro* growth. (a-c) Mutation rate (μ) was estimated based on the number of unique SNPs (m) observed in each condition (4 active, 3 latent, 7 reactivated). This calculation was performed over a range of generation times (g, 18-240 hours per generation) to allow for the uncertainty in growth rate *in vivo*. The probability of observing μ when g is fixed at any given time was determined to build the probability distribution function around each estimate and to define the 95% confidence intervals. The single base mutation rate of the bacterium during *in vitro* growth ($\mu_{\text{in vitro}}$) was determined by fluctuation analysis (**Supplementary Figs. 1a-c)** and is indicated by an arrow. In each clinical condition, μ_{20} (the predicted mutation rate if the generation time *in vivo* were as rapid as the generation time *in vitro*) is similar to $\mu_{\text{in vitro}}$. Generation time *in vivo* is predicted to be substantially slower than *in vitro*, and thus the mutation rate must be proportionally higher to produce the observed number of SNPs. **(d)** Given the uncertainty in generation time, a mutation rate *per day* can be calculated to determine the rate at which mutations occur regardless of generation time. Mutations occur at a similar rate per day regardless of the disease status of the host. Error bars represent 95% confidence intervals.

bacterium in different *in vivo* conditions. Strikingly, we found that the bacterial population's capacity for mutation, μ(20hr), during latency (**$2.71x10^{-10}$**) and reactivated disease (**$3.03x10^{-10}$**) is equivalent to that of Mtb from animals with active disease (**$2.01\ x10^{-10}$**) (**Table 2.1**). Mutation rate can also be calculated as the number of mutations that occur per day of infection rather than per generation. We therefore calculated the mutation rate per day required for the bacterial populations in each disease state to acquire the number of polymorphisms that we identified by WGS (**Figure 2.3d**). Our data indicate that in macaques with active, latent and reactivated disease, the bacterial populations acquire mutations at the same rate over time, regardless of the number of bacterial replications that have occurred.

We then sought to benchmark these rates against the mutation rate of the bacterium *in vitro*. Luria and Delbrück fluctuation analysis measures the rate at which coding polymorphisms conferring a selectable phenotype arise under stable *in vitro* conditions[65]. While standard and widely applied, this approach is limited in that it only measures the rate of a small set of coding mutations within a single region of the genome and is therefore not as sensitive as WGS. However, the fluctuation analysis derived mutation rate can be converted to a per base mutation rate by defining the number of mutations conferring resistance[66] and then compared to the mutation rate determined with WGS. Using fluctuation analysis and scoring for the acquisition of rifampicin resistance, we found that the rate of resistance is $2.1x10^{-09}$ (**Supplementary Figure 2.1a,b**), consistent with or slightly higher than previously published values for Mtb[67,68]. We then used Sanger sequencing to define the number of coding mutations conferring rifampicin resistance under our growth conditions and found it occurred through ten unique polymorphisms, consistent with previous reports[69] (**Supplementary Figure 2.1c**). Dividing the phenotypic rate

by the target size, we determined that the *in vitro* mutation rate of the inoculating strain (Erdman) is $\mu_{\text{in vitro}}=2.1\text{x}10^{-10}$. Thus μ(20hr), our conservative estimate of the *in vivo* mutation rate from every disease state, is highly similar to the bacterium's mutation rate observed during *in vitro* growth.

Why does the bacterial population in macaques with clinically latent infection acquire mutations at a similar rate to rapidly replicating bacteria *in vitro*? One possibility is that Mtb could be actively dividing during the entire course of prolonged clinical latency, perhaps balanced by robust killing. However, though the generation time of Mtb in animals or humans with latent infection is unknown, several lines of evidence suggest that Mtb replication slows during clinical latency[61-64]. If the generation time slows, the mutation rate would have to be proportionally higher to generate the number of mutations observed. For example, if Mtb from animals with latent infection replicate on average every 135 hours, as in mice after ten weeks of chronic infection[61], the bacterial population must have an average mutation rate per generation of $1.80\text{x}10^{-09}$, nearly an order of magnitude greater than the *in vitro* mutation rate (**Table 2.1**).

An alternative interpretation is that the mutational capacity of Mtb during latent infection is determined primarily by the length of time the organism spends in the host environment rather than the replicative capacity and replicative error of the organism during infection. We noted that eight of the ten polymorphisms that we identified in our isolates from animals with persistent latent or reactivated latent infection are possible products of oxidative damage, either cytosine deamination (GC>AT) or the formation of 8-oxoguanine (GC>TA) (**Figure 2.4a**). This is consistent with the model that Mtb faces an oxidative environment in the macrophage phagolysosome[70,71] and data indicating that genes involved in the repair of oxidative damage are

essential for bacterial survival during murine infection[72]. In addition, we found that the pattern of polymorphisms in Mtb from cynomolgus macaques is similar to the pattern of neutral

Table 2.1 The predicted mutation rate for biologically relevant generation times.

Gen. Time (hrs) (g)	Growth Condition	$\mu(g)_{active}$, (95% CI)[a]	$\mu(g)_{latent}$, (95% CI)[a]	$\mu(g)_{reactivated}$, (95% CI)[a]
20	Rich Media	$2.01x10^{-10}$ ($8.09x10^{-11}$,$4.15x10^{-10}$)	$2.71x10^{-10}$ ($5.57x10^{-11}$,$7.89x10^{-10}$)	$3.03x10^{-10}$ ($1.22x10^{-10}$,$6.24x10^{-10}$)
45	Macrophage	$4.77x10^{-10}$ ($1.30x10^{-10}$,$1.22x10^{-9}$)	$5.99x10^{-10}$ ($1.23x10^{-10}$,$1.75x10^{-9}$)	$6.71x10^{-10}$ ($2.70x10^{-10}$, $1.38x10^{-9}$)
135	Mouse Infection at 10 weeks	$1.43x10^{-9}$ ($3.90x10^{-10}$,$3.66x10^{-9}$)	$1.80x10^{-9}$ ($3.70x10^{-10}$,$5.25x10^{-9}$)	$2.01x10^{-9}$ ($8.09x10^{-10}$,$4.15x10^{-9}$)

[a]Mutation rates were estimated using the equation shown $\mu=m / [N*(t/g)]$, over g = 18 to 240 hours. The generation time (g) was varied from 18 to 240 hours, t represents total time of infection in hours and N is equal to the number of bases sequenced. The values shown represent the predicted μ and 95% confidence intervals of a bacterial population in animals with active, latent or reactivated disease estimated for the indicated, biologically relevant generation times[60,61,73].

polymorphisms that emerged during the evolution of extensively drug resistant Mtb strains in patients from South Africa (**Figure 2.4a**)[18]. Thus, the mutational capacity of Mtb during latent infection as well as the spectrum of those mutations suggests that the dominant source of mutation during latency is oxidative DNA damage rather than replicative error[74] (**Figure 2.4b**). This may occur because the immune response that results in latent infection causes more oxidative damage to the bacterial DNA[63,75] or because a portion of the bacteria may enter a metabolically quiescent state in which DNA repair is diminished[76,77].

2.3 Discussion

Thus, using WGS, we demonstrate that Mtb has greater mutational capacity in latency and early reactivation disease than predicted by *in vitro* measurements of mutation rate and estimates of *in vivo* generation time. These data indicate that Mtb retains the ability to acquire drug resistance mutations during latency. The rate at which clinical drug resistance will emerge after IPT treatment of latently infected individuals harboring an initially drug-sensitive population also depends upon the number of bacteria in a latently infected individual and the rate of reactivation, which is low in immunocompetent individuals. Indeed, there is only a modest increased risk of INH resistance after IPT in immunocompetent populations[51,52], some of which may be attributable to selective killing of susceptible bacterial populations, leaving only resistant populations to reactivate[78]. Our results suggest that in addition to these mechanisms, part of the increased risk of INH resistance after IPT may be due to selection of monoresistant mutants that arise during latency. IPT is now being recommended globally for HIV+ individuals with clinically latent tuberculosis where bacterial burden and the rate of treatment failure may be higher because of immunocompromise[79]. If our data from the macaque model are predictive of the mutational capacity of Mtb in HIV+ individuals, INH monoresistance could arise at a

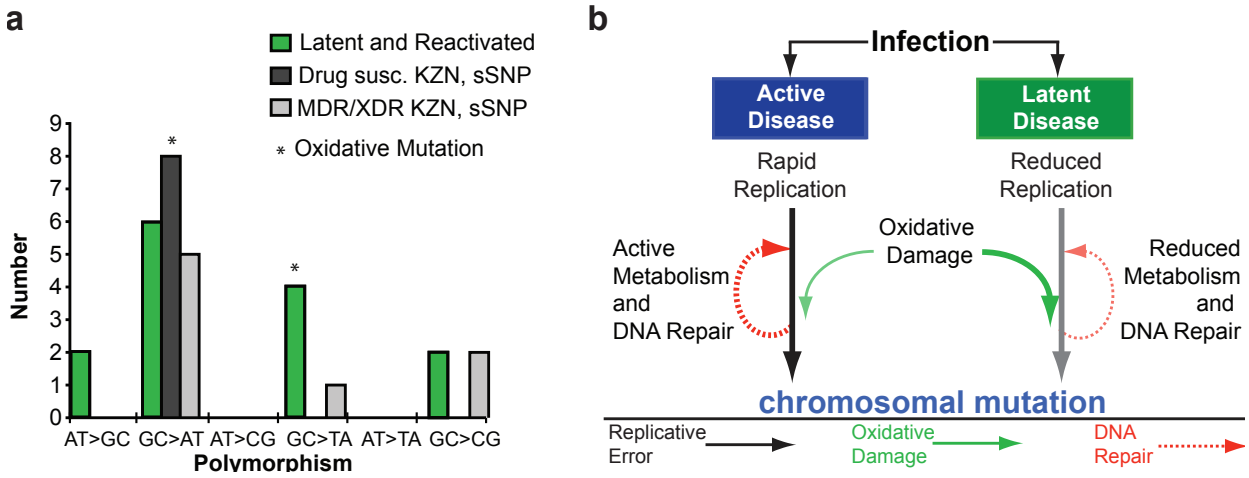


Figure 2.4 Mutations in Mtb isolated from macaques with latent infection and related human isolates are putative products of oxidative damage. (a) Ten of the fourteen mutations observed in this study could be the product of oxidative damage: the deamination of cytosine (GC>AT) or the production of 7,8-dihydro-8-oxoguanine (GC>TA) by the oxidation of guanine. One of each type of mutation observed was seen in active disease (four mutations total). In contrast, eight of ten mutations observed in latent and reactivated disease are potential products of oxidative damage. There is a similar mutational spectra observed in the synonymous SNPs identified by WGS of a set of closely related strains from South Africa[18]. **(b)** These observations lead to a model of mutational pressures on Mtb during active disease and latent infection in which oxidative damage may play a central role in the generation of mutation.

substantial rate. These findings emphasize the importance of drug resistance testing and careful monitoring for treatment failure in these patient populations.

2.5 Materials and Methods

Preparation of isolates

Animals were infected as described previously[57] via bronchoscopy with a small number (~25) of organisms. Like humans, macaques developed either active disease or controlled latent infection. In latency, animals became clinically asymptomatic, without microbiologic or radiographic evidence of disease. Clinically latent animals were followed as described previously[57] for prolonged periods of time in the absence of treatment. Spontaneous reactivation of latent infection occurred in a small number of animals. Animals were euthanized at the indicated times after infection and lesions identified on necropsy were plated for bacterial colonies (**Figure 2**)[57]. Colonies from necropsy were subsequently streaked onto LJ slants and expanded for extraction of genomic DNA. Minimal expansion occurred between isolation of strains and extraction of genomic DNA.

Illumina Sequencing

Two μg of genomic DNA from each isolate were used for sequencing with the Illumina Genome Analyzer (Illumina). Single-read & paired-end read sequencing was performed with read lengths of 36 bases or 75 bases and a target coverage of at least 3 million high quality bases. Libraries were prepared using standard sample preparation techniques recommended by the manufacturer. Libraries were quantified using a Sybr qPCR protocol with specific probes for the ends of the adapters. The qPCR assay measures the quantity of fragments properly adapter-ligated that are appropriate for sequencing. Based on the qPCR quantification, libraries were normalized to 2nM and then denatured using 0.1 N NaOH. Cluster amplification of denatured templates occurred according to manufacturer's protocol using V2 Chemistry and V2 Flowcells (1.4mm channel

width). Sybr Green dye was added to all flowcell lanes to provide a quality control checkpoint after cluster amplification to ensure optimal cluster densities on the flowcells. Flowcells were sequenced on a Genome Analyzer II, using V3 Sequencing-by-Synthesis kits and analyzed with the Illumina v1.3.4 pipeline. Standard quality control metrics including error rates, % passing filter reads, and total Gb produced were used to characterize process performance prior to downstream analysis. Paired-end reads of 51 bases were acquired and analyzed as described previously[18]. Short sequence read data is available on the NCBI SRA (accession numbers in **Supplementary Table 2.1**) and on an independent site hosted by the Broad Institute and linked through the TB Database (see **URLs**).

Data filtering and assembly

Two read lengths were generated (**Supplementary Table 2.1**). Prior to mapping or assembly, the 2x75 bp reads were trimmed to 48 bases and filtered. Any read containing an unknown base was discarded. Reads with homopolymeric runs of A/Ts greater than nine bases or G/Cs greater than ten bases were discarded. Reads with an average quality score of less than 20 were removed. On average greater than 8,000,000 reads were retained after filtering. The 36 bp reads were not filtered. For *de novo* assembly, reads were processed with Edena v2.1.1[59] in overlapping mode with the default parameters to allow for the detection of insertions and deletions as well as SNPs. In assembly mode, "strict" was enforced and independent assemblies were generated with length overlaps ranging from position 23 to 37 bases. Assemblies generating the largest N50 values were selected for polymorphism discovery. Each assembly was analyzed by pair wise comparison using the MUMmer script, dnadiff[80]. Polymorphisms were further processed from the 'SNP' files with Perl scripts and mapped to the reference genome H37Rv (GenBank

Accession AL123456). For scaffolded assembly, Illumina reads were mapped to the reference genome Haarlem with MAQ v0.7.1[58]. Illumina fastq files for each pair were converted with sol2sanger, individually mapped to the reference and merged together for each isolate. Three mismatches in the alignment seed were allowed during mapping. A minimum read depth of ten was required to call SNPs and remaining parameters were defaults from the script easyrun. For 51bp paired end reads, a minimum read depth of five was required to call SNPs. For 36bp reads, reads were aligned to the reference using easyrun defaults except that we allowed up to three mismatches in the seed. For the Erdman inoculum, four runs were merged to generate the assembly. As this represented the first WGS of the Erdman strain of Mtb, multiple Mtb finished genomes were used as references in preliminary alignments. The Haarlem sequence resulted in the fewest number of SNPs and was selected as the reference sequence for the remainder of the alignments (see **URLs**). Only sites of difference between the experimental isolates were pursued for further analysis. A master list of sites was created and calls for each site from all samples were extracted with the MAQ command "pileup", combined into a table and inspected manually. All polymorphic loci were validated either by Sanger sequencing using the indicated primers (**Supplementary Table 2.2**) or by independent identification by WGS.

Statistical analysis and estimation of *in vivo* mutation rate from WGS data

Mutation rate was estimated from the number of SNPs observed in each clinical condition. Our equations assume that both mutation rate and growth rate are parameters that, while potentially dynamic, can be averaged across the lifetime of the bacterium. Additionally, we assume that the number of mutations (m) is an accurate assessment of mutation rate during the life of the cell. SNPs observed multiple times within the same lesion were assumed to have arisen once and then

replicated; as such they were each only counted once. Equation (**1**) describes the estimation of the mutation rate of a single strain as described in **Table 2.1**.

$$\mu = m/[N*(t/g)] \qquad (\mathbf{1})$$

Mutation rate (μ) is determined by dividing the number of SNPs (m) by the genome size (N) times the number of generations (t/g). m is defined by the number of SNPs observed, N is determined based on 91% coverage of a 4.4Mb genome (N=4 x 10^6), t is the total duration of each infection in hours, and g is the generation time in hours. The application of this equation to a clinical condition is described by Equation (**2**). Samples were binned according to clinical condition and a representative mutation rate was estimated for each condition. Binning allows us to conservatively assess the distribution of mutations in each condition.

$$\mu = \frac{\sum_{i=1}^{n} m_i}{N * \sum_{i=1}^{n} (t_i / g)} \qquad (\mathbf{2})$$

In Equation (2), the sum of the SNPs observed (m_i) in a condition is divided by the genome size (N) multiplied by the sum of the number of replications possible (t_i / g). The number of replications possible is calculated by dividing the total length of infection (in hours, t_i) for strain *i* by the generation time in hours. The generation time, g, was varied from 18 hours to 240 hours to capture the maximum range of biologically plausible generation times. All calculations were performed in Matlab (Mathworks, Natick MA, USA). Estimates of mutation rate and 95% confidence intervals were determined using the *poissfit* function. Additional probability values were generated for each value of g using the *binopdf* function. The *binopdf* parameters and values matched exactly those produced by *poissfit,* reflective of the ability of the Poisson distribution to approximate the binomial distribution when $N_{poisson}$ is large and $p_{poisson}$ is small.

Thus, *binopdf* was used to calculate the probability density function for the observed number of mutations given a mutation rate and a fixed value for g, while *poissfit* was used to calculate the estimates of $\mu_{in\ vivo}$ and the 95% confidence intervals.

Determination of the *in vitro* mutation rate

To determine the *in vitro* mutation rate, we performed fluctuation analysis as previously described[67]. Briefly, 20 independent cultures of 1.08 x 10^9 cells each in 4mL of 7H9 supplemented with OADC, 0.05% Tween-80 and 0.5% glycerol were plated onto 7H10 plates supplemented with OADC, 0.05% Tween-80, 0.5% glycerol and 2µg/mL of rifampicin. The number of mutations per culture (m_{MSS}) was calculated from the distribution of mutants using the MSS method[65] calculated by the Matlab scripts described by Lang et al[66]. Phenotypic mutation rate was estimated by dividing m_{MSS} by the number of cells plated (N_t = 1.08 x 10^9). The number of *rpoB* mutations conferring rifampicin resistance in our assay was determined by amplifying the resistance region of *rpoB* from 96 independent isolates from fluctuation analysis. A single base mutation rate ($\mu_{in\ vitro}$) was calculated by dividing the rifampicin resistance rate by the number of mutations conferring rifampicin resistance.

Mutational spectrum of synonymous SNPs in the KwaZulu Natal Mtb isolates

We identified the synonymous mutations that distinguished the sequenced drug susceptible, MDR and XDR strains of Mtb from KwaZulu Natal, South Africa from each other using previously published data[18]. Polymorphisms were called in reference to the sequenced F11 strain of Mtb and polymorphisms in repetitive and IS elements, PE, PPE and PGRS genes and

pks12 were excluded from this analysis because these genes contain large, near perfect repeats that create a high likelihood of sequencing and assembly error.

URLs:

Mycobacterium tuberculosis Sequencing Project, Broad Institute of Harvard and MIT

(http://www.broadinstitute.org/)

Genomic Data for the 34 *M. tuberculosis* Erdman strains sequenced:

(http://www.broadinstitute.org/annotation/genome/mtb_monkey_reseq.1)

Tuberculosis Database (http://www.tbdb.org)

Acknowledgments. This work was supported by a New Innovator's Award, DP2 0D001378, from the Director's Office of the National Institute of Health to SMF, by a subcontract from NIAID U19 AI076217 to Sarah M Fortune, by NIH RO1 HL075845 to Joanne Flynn, and by the Bill and Melinda Gates Foundation (Joanne Flynn). The genome sequencing has been funded in part with federal funds from the National Institute of Allergy and Infectious Disease, US National Institutes of Health (NIH), US Department of Health and Human Services, under contract no. HHSN266200400001C. The project described was supported in part by Award Number U54GM088558 to Marc Lipsitch from the National Institute Of General Medical Sciences. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute Of General Medical Sciences or the National Institutes of Health. We thank D. Gurgil and J. Xu of the Enterprise Research IS group at Partners Healthcare for their support and for provision of the HPC facilities, and E. Klein for necropsy and pathology of the infected monkeys, as well as the veterinary technical staff for care of the animals. We also thank E. Rubin, C. Sassetti, B. Bloom, T. Rosebrock, and B. Aldridge for helpful feedback.

Author Contributions. Christopher B Ford performed molecular studies, conducted the data analyses, prepared the figures and drafted the manuscript; Philana Ling Lin and Joanne Flynn conducted the infection of the cynomolgus macaques, determined clinical state, and acquired bacterial strains on necropsy; Michael Chase analyzed sequence data and directed validation of SNPs; Rupal R. Shah performed molecular and fluctuation analyses; Oleg Iartchouk oversaw sequencing of isolates sent to PHCPGM; James Galagan oversaw sequencing of isolates sent to the Broad Institute; Nilofar Mohaideen, Thomas Ioerger., and James Sacchettini oversaw sequencing and analysis of isolates sent to Texas A&M; Marc Lipsitch supervised and advised statistical analyses; Sarah M. Fortune initiated the project, performed molecular studies, supervised preparation and analysis of the data, and drafted the manuscript.

Chapter 3 – Differences in the rate of mutation between strains of *Mycobacterium tuberculosis*

3.1 Introduction

Recently, strains of Mycobacterium tuberculosis (Mtb) have emerged that are resistant to most or all effective antibiotics[6,81-83]. Given the low mutation rate of Mtb[21,84] and its slow replication rate, it is unclear how the bacterium acquires resistances to multiple antibiotics, especially in the face of multi-drug treatment[85]. The most commonly cited risk factors for treatment failure due to antibiotic resistance are patient noncompliance[86-88], inappropriate drug regimens and dosing[81,88,89] and transmission of drug resistant strains[6,81-83,90,91]. In this debate, the relative importance of bacterial determinants of treatment failure has been unclear. However, recent evidence suggests that certain strains of Mtb may preferentially acquire drug resistances[21,84,92]. Given that all drug resistances in Mtb occur through chromosomal mutation, these data suggest that the mutational capacity of the bacterium may be an important determinant of the likelihood of drug resistance.

Defined genetically, Mtb forms phylogeographic lineages based on human demography[17,39,85,93,94]. Though less genetically diverse than many other pathogens, there is both experimental and clinical evidence that Mtb strains from different lineages vary in their capacity to cause disease[86-88,93-96] and acquire drug resistances[81,88-93,97-100]. Specifically, strains within the East Asian lineage of Mtb have been epidemiologically associated with an increased risk of drug resistance in several cross-sectional studies. Strains from this lineage have polymorphisms in DNA replication, recombination, and repair genes as compared to Euro-American strains, raising the possibility that they are more mutable than other Mtb strains[101]. However, these epidemiologic observations might also reflect social and programmatic factors correlating with

the phylogeography of the East Asian strains. Indeed, studies comparing the rate of drug resistance in the East Asian and Euro-American strains have produced conflicting results[68,102]. We demonstrate that Mtb strains from the East Asian lineage acquire drug resistances *in vitro* more quickly than Mtb strains from the Euro-American lineage. Their higher drug resistance rate *in vitro* reflects a higher basal mutation rate. Moreover, the *in vitro* mutation rate correlates well with the bacterial mutation rate in humans as determined by whole genome sequencing of clinical isolates. Finally, using an agent-based model, we show that the observed differences in mutation rate predict a significantly higher probability of multi-drug resistance in patients infected with East Asian lineage strains of Mtb.

3.2 Results

3.2.1 Effect of mutation and genetic background on drug resistance in *M. tuberculosis*

To measure the drug resistance rate of strains from the different Mtb lineages, we performed Luria and Delbrück fluctuation analysis on a panel of drug sensitive strains containing both laboratory and clinical isolates from the East Asian and Euro-American lineages[103,104]. Within both lineages, there was some strain-to-strain variation in the rate at which rifampicin resistance was acquired (**Figure 3.1**). However, strains from the East Asian lineage acquired resistance to rifampicin (2µg/mL) at significantly higher rates (1.78-37.07 fold) than the Euro-American strains (**Figure 3.1, Supplementary Table 3.1**).

The higher rate of rifampicin resistance could reflect three possible factors: 1) differences in the ability to survive and mutate after exposure to antibiotic 2) inherent differences in the number of *rpoB* mutations conferring rifampicin resistance (target size) or 3) differences in the basal mutation rate. To test these hypotheses, we chose well-characterized representatives from the Euro-American lineage, CDC-1551, and the East Asian lineage, HN878, for further study.

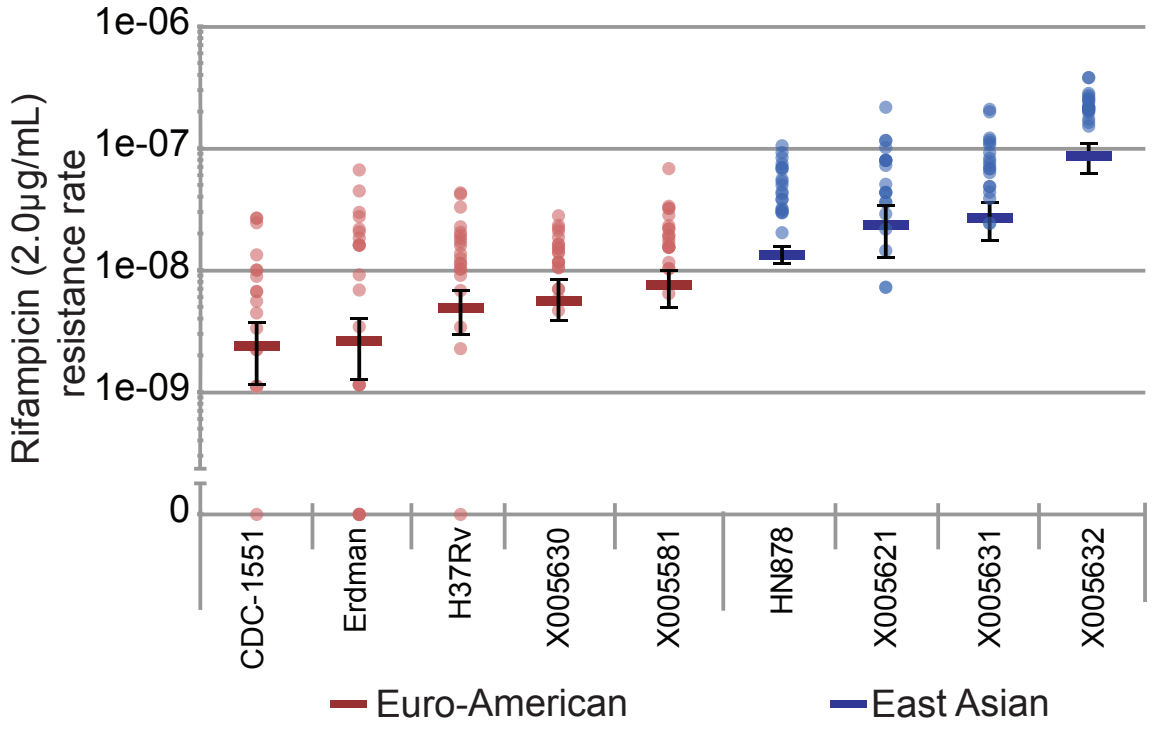


Figure 3.1 East Asian strains more rapidly acquire resistance to rifampicin (2µg/mL). Fluctuation analysis was used to determine the rifampicin (2µg/mL) resistance rate of clinical and laboratory strains from both the Euro-American and East Asian lineage. Euro-American strains are in red, East Asian strains are in blue. Circles represent mutation frequency (number of mutants per cell in a single culture), where darker circles represent multiple cultures with the same frequency. Bars represent the estimated mutation rate, with error bars representing the 95% confidence interval. Strains are displayed on the x-axis and the rifampicin resistance rate is displayed on the y-axis in log-scale. Values are listed in **Supplementary Table 3.1**.

3.2.2 Differences in response to antibiotic

We first sought to determine whether the East Asian and Euro-American strains differed in their ability to acquire drug resistance after exposure to drug. Fluctuation analysis assumes that all mutations occur prior to selection[104]; however, if a strain is capable of surviving and mutating in the presence of drug, it will produce a greater number of drug resistant mutants. Building on similar studies in *Saccharomyces cerevisiae*[66], we reasoned that if mutations occur in the presence of drug, then lowering the concentration of drug might allow strains from both lineages to grow and acquire mutations post-exposure, while increasing the concentration of drug might abrogate the ability of both strains to survive and acquire mutations in the presence of drug. However, varying rifampicin concentration 10-fold (0.5µg/mL – 5µg/mL) did not alleviate the statistically significant increase in rifampicin resistance rate found in the East Asian strain, HN878, relative to the Euro-American strain, CDC1551 (**Figure 3.2, Supplementary Table 3.1**).

To extend these findings, we analyzed the distribution of mutations observed in each fluctuation assay using analytical tools developed by Lang et al[66]. This analysis takes advantage of the fact that mutations occurring in culture, prior to antibiotic exposure, occur according to a Luria-Delbrück distribution. Mutation itself occurs according to a Poisson distribution, while the subsequent outgrowth of mutants during broth culture before selection generates a Luria-Delbrück distribution. In contrast, mutations occurring after plating on drug will occur according to a Poisson distribution without the expansion in culture that results in a Luria-Delbrück distribution[104-106]. Thus, any additional mutants resulting from acquisition of resistance post-exposure will increase both the estimated drug resistance rate and impose a Poisson distribution on the typical Luria-Delbrück distribution.

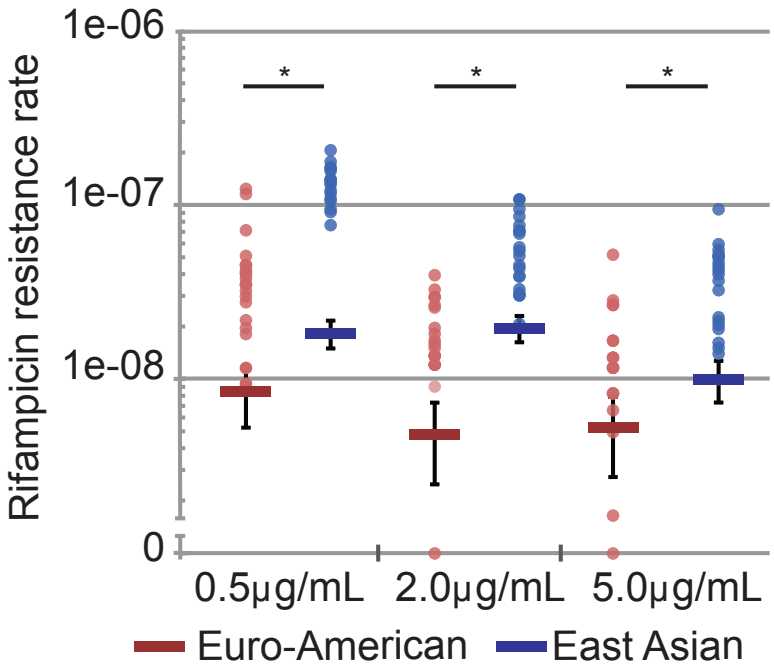


Figure 3.2 East Asian strains more rapidly acquire rifampicin resistance across multiple concentrations of antibiotic. Fluctuation analysis was used to determine the rifampicin (0.5, 2, 5µg/mL) resistance rate of representative strains from both the Euro-American and East Asian lineage (CDC-1551 and HN878, respectively). The Euro-American strain CDC-1551 is in red, and the East Asian strain HN878 is in blue. Circles represent mutation frequency (number of mutants per cell in a single culture), where darker circles represent multiple cultures with the same frequency. Bars represent the estimated mutation rate, with error bars representing the 95% confidence interval. Significance was determined by comparing strain pairs using the Wilcoxon rank-sum test. Strains are displayed on the x-axis and the rifampicin resistance rate is displayed on the y-axis in log-scale. Values are listed in **Supplementary Table 3.1**.

We therefore used a curve fitting approach to determine whether the distribution of mutant frequencies in the two strains is better fit using a one parameter, Luria-Delbrück model or a two parameter, Luria-Delbrück and Poisson model (**Figure 3.3a-f**). We used the Akaike information criterion with correction for sample size (AIC_C), to determine which model best fit the data[107,108]. The AIC_C quantifies the fit of a model to observed data, with a lower AIC_C reflecting better relative fit. In all conditions analyzed, $\Delta AICc$ (AIC_C(one parameter) - AIC_C(two parameter)) was less than zero, demonstrating that there is not a significant Poisson component in the distributions (**Figure 3.3g**). This indicates that post-exposure mutation is not responsible for the higher rifampicin resistance rate in HN878, the East Asian lineage strain. In addition, this analytic approach also suggests that the difference in rifampicin resistance rates is not due to strain based differences in the fitness effects of the drug resistance mutations[104,106]. If the drug resistant mutants in either strain suffered a strong fitness cost, the outgrowth of mutants in culture prior to selection would have been slower than drug susceptible cells, driving the Luria-Delbrück distribution back towards the underlying Poisson distribution of mutation. However, our data suggest that in both strains, drug resistant mutants occur solely according to a Luria-Delbrück distribution.

3.2.3 Differences in Target size

Rifampicin resistance is encoded by multiple mutations in the rifampicin resistance-determining region (RRDR) of *rpoB*. Strains in which there are a greater number of potential mutations in the RRDR that produce drug resistance would more rapidly acquire resistance to rifampicin. Therefore, we sought to determine whether Mtb strains of different lineages differ in the total number of mutations that confer rifampicin resistance, accounting for differences in the

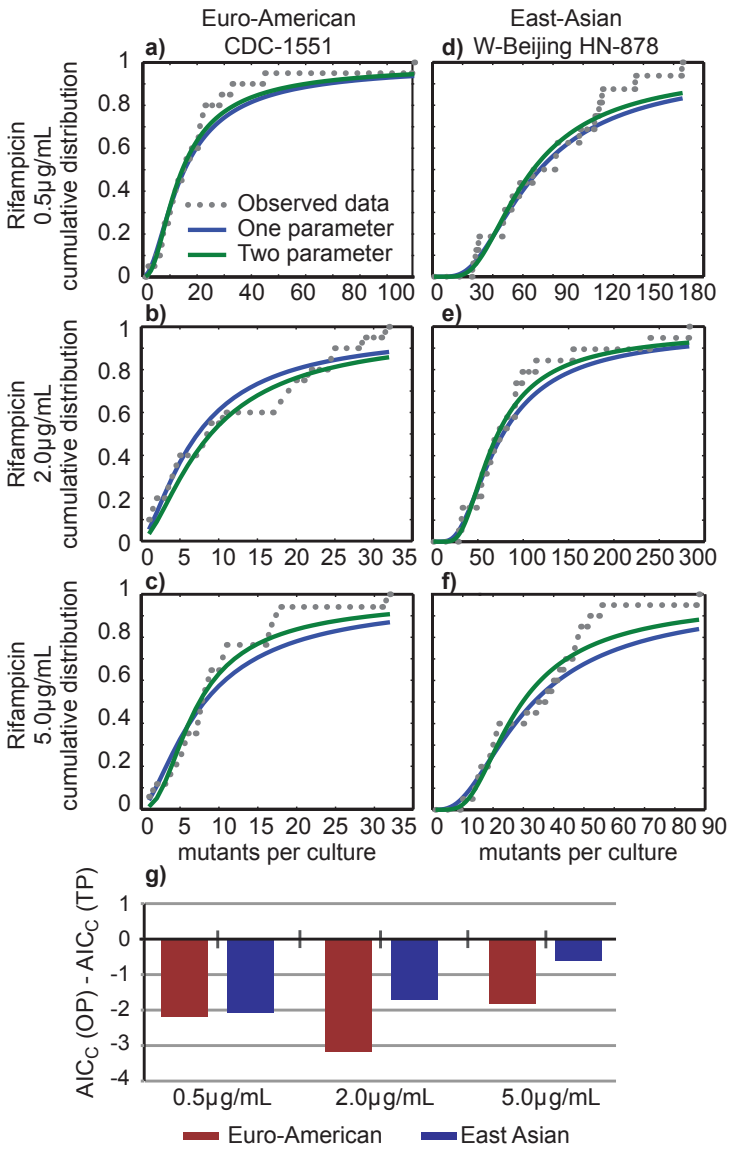


Figure 3.3 Differences in the rate of drug resistance are not due to differences in fitness of drug resistant mutants or the ability to survive and mutate in the presence of drug. Curve fitting analysis was performed to determine if the cumulative distribution of the fluctuation analysis data in Figure 2 better fit a one parameter, Luria-Delbrück model or a two parameter, Luria-Delbrück and Poisson model. **(a-f)** A dotted grey line represent the cumulative distribution function of the observed data, a solid blue line represents the cumulative distribution function of the one-parameter model, and a solid green line represents the cumulative distribution function of the two-parameter model. The number of mutants per culture is displayed on the x-axis, and the probability of observing (x) mutants per culture or fewer is shown on the y-axis. **(g)** To determine which model best fit each data set, we determined the Akaike Information Criterion, corrected for small sample size (AIC_C). A smaller AIC_C represents a better fit, given a penalty for more parameters in a model. If the AIC_C (one parameter) is smaller than the AIC_C (two parameter), then the resulting value will be negative, reflecting a better fit for the one parameter model.

rifampicin resistance rates. We sequenced the RRDR of *rpoB* to determine the number of mutations conferring rifampicin resistance in both strains under each condition tested above. 100 independent mutants from each fluctuation assay in Figure 2 were sequenced to give a total of 600 RRDR sequences (**Figure 3.4a, Supplementary Table 3.2**). All of the mutations that we identified correspond to mutations seen clinically[109]. For both strains, the target size became smaller as drug concentration increased, indicating that some *rpoB* mutations generate lower level rifampicin resistance. We found small differences in target size between the two strains at two of the three drug concentrations tested, such that the number of mutations conferring resistance at both 0.5 and 25μg/mL was higher in the East Asian strain. These data suggest that target size differences between strains contribute to strain based differences in the acquisition of rifampicin resistance. However, correcting for target size, the mutation rate of the East Asian strain, HN878, remained significantly higher than the Euro-American strain, CDC-1551, at each of the drug concentrations tested (**Figure 3.4b, Supplementary Table 3.3**). Therefore, it is likely differences in basal mutation rate that lead to differences in the acquisition of rifampicin resistance.

3.2.4 Resistance to other antibiotics

If the mutation rate of HN878 were higher than that of CDC1551, the East Asian Mtb strain should also acquire resistances to other antibiotics at a higher rate. We therefore assessed the rate at which HN878 and CDC1551 acquire resistance to ethambutol (5μg/mL) and isoniazid (1μg/mL). For both antibiotics, the rate of resistance was nearly 3-fold (2.51 and 2.75, respectively) higher in the East Asian strain, HN878, consistent with the increased rate of rifampicin resistance (**Figure 3.5, Supplementary Table 3.1**). Taken together, these results suggest

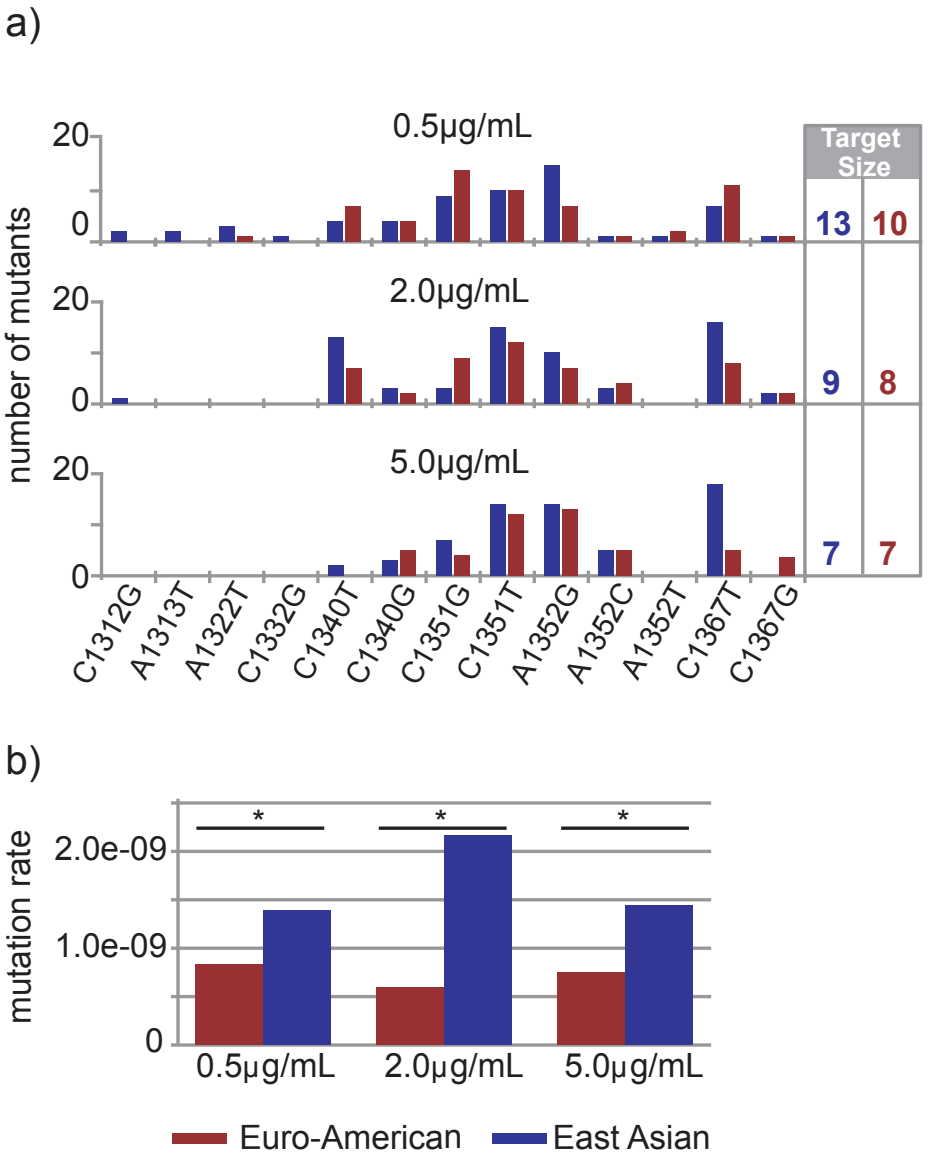


Figure 3.4 Differences in target size exist between the East Asian isolate HN878 and the Euro-American isolate CDC-1551 but do not explain the difference in drug resistance rate. **(a)**The target size (the number of mutations conferring rifampicin resistance) of each strain under each condition was determined by sequencing the rifampicin resistance-determining region of 100 isolates from each strain in each condition. Each mutation is shown on the x-axis, with coordinates representing position within *rpoB* (Rv0667). The number of mutants per strain uniquely formed within a culture is shown on the y-axis. Euro-American strains are shown in red; East Asian strains are shown in blue. The box to the right displays target size, the number of unique mutations conferring rifampicin resistance. **(b)** The per base pair mutation rate is determined by normalizing the drug resistance rate by target size. Drug concentration is shown on the x-axis, mutation rate per base pair is shown on a linear scale on the y-axis. Euro-American strains are shown in red; East Asian strains are shown in blue. Significance was determined by comparing strain pairs using the Wilcoxon rank-sum test. Values are found in **Supplementary Table 3.3**.

that Mtb strains from the East Asian lineage have a higher basal mutation rate than strains from the Euro-American lineage.

3.2.5 The *in vitro* mutation rate correlates with the *in vivo* mutation rate

We then sought to understand how these *in vitro* measures of mutation translate to the *in vivo* environment. In our previous work, we determined that in nonhuman primates, Mtb mutates at a relatively fixed rate over time and this *in vivo* per-day mutation rate is well-approximated by the *in vitro* per-day mutation rate as measured by fluctuation analysis and adjusted for target size[21]. To determine if the *in vitro* mutation rate is similarly concordant with the mutation rate of Mtb during human infection, we analyzed the whole genome sequences of Mtb isolates derived from an outbreak of a Euro-American strain in British Columbia, Canada[20] and determined both the number of SNPs and the time of isolation relative to a historical isolate (**Figure 3.6a**). By reconstructing the phylogeny of these strains through Bayesian Markov chain Monte Carlo analysis [110,111] (**Supplemental Figure 3.1**), and informing the phylogeny with dates for each isolate, we have estimated the base substitution rate (equivalent to the mutation rate under a neutral model of evolution[112]) in this outbreak. Strikingly, we found that the British Columbia strains acquired mutations at approximately the same rate over time as the Mtb strains isolated from macaques with active and latent disease irrespective of disease course (**Figure 3.6b, Supplemental Table 3.4**). In addition, the rate at which these strains acquired mutations *in vivo* was well approximated by the *in vitro* per-day mutation rate of the lineage 4 strain (Erdman) used in the macaque infections as defined by fluctuation analysis. Finally, these data indicate that a molecular clock of 0.3-0.5 mutations/genome/year may be applicable to analysis of Mtb genetic diversity over the time scales assessed here.

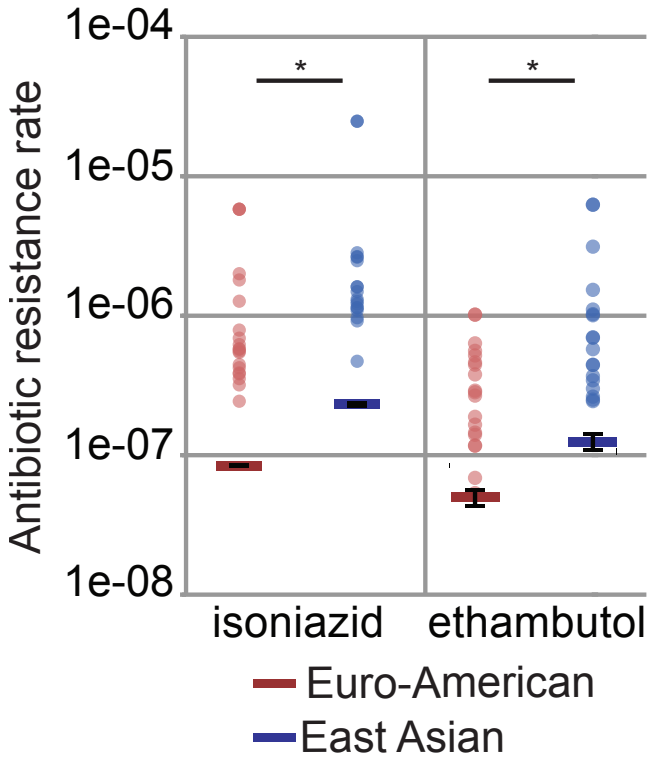


Figure 3.5 East Asian strains more rapidly acquire resistance to multiple antibiotics. Fluctuation analysis was used to determine the isoniazid (1µg/mL) and ethambutol (5µg/mL) resistance rate for the Euro-American strain CDC-1551 (shown in red) and the East Asian strain (HN878) shown in blue. Circles represent mutation frequency (number of mutants per cell in a single culture), where darker circles represent multiple cultures with the same frequency. Bars represent the estimated mutation rate, with error bars representing the 95% confidence interval. Significance was determined by non-overlapping 95% confidence interval. Values are listed in **Supplementary Table 3.1**.

3.2.6 A time-based model of mutation and drug-resistance predicts MDR before treatment.

Given these data, we developed an agent-based model of the evolution of drug resistance within a patient in order to assess the potential clinical impact of the observed differences in mutation rate between the Euro-American and East Asian strains. Our model of drug resistance utilizes a stochastic mutation parameter in which mutation occurs at a constant rate over time and we informed this parameter with the *in vitro* mutation rates for CDC1551 and HN878 as a proxy for their mutation rates in the human host (**Supplementary Figure 3.2a, Supplementary Table 3.5**). We used this model to simulate the emergence of MDR within an infected individual prior to diagnosis and treatment (**Supplementary Figure 3.2b**).

As a result of the differences in mutation rate, patients infected with the East Asian strain, HN878, are at a significantly increased risk of MDR as compared to patients infected with the Euro-American strain, CDC1551 (**Figure 3.7a**). When all other parameters (birth, death, fitness and bacterial burden at the time of diagnosis) are kept equal, the difference in the probability of MDR before diagnosis and treatment is approximately 22-fold. We find similar results when using an alternative model of drug resistance developed by Colijn et al in which mutation is replication rather than time dependent (**Supplementary Figure 3.2c**)[113]. We assessed the sensitivity of our model to fluctuations in both growth rate and fitness (**Figure 3.7b & c**). Varying these parameters does not alter our principle conclusion that patients infected with East Asians strains of Mtb are at a significantly higher risk of the *de novo* acquisition of multidrug resistance, reflecting the multiplicative effects of an increased risk of acquiring each individual drug resistance due to a higher basal mutation rate.

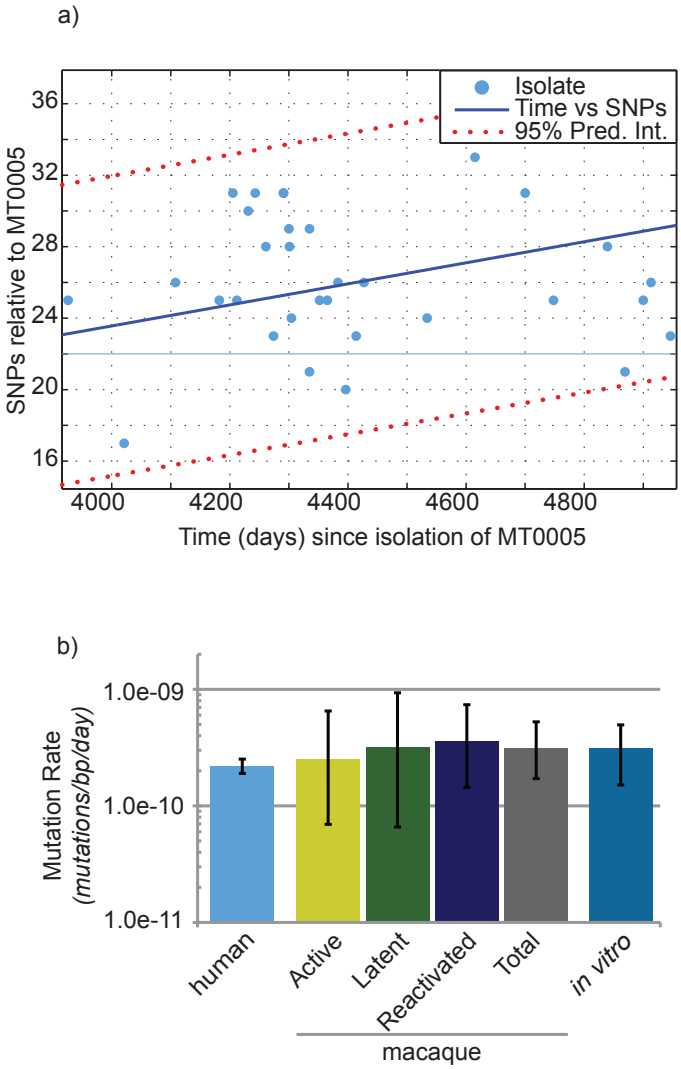


Figure 3.6 Estimate of mutation rate derived from clinical isolates (a) The number of SNPs and the number of days separating the clinical isolate and MT0005 are plotted. The data are fit to a first order polynomial to illustrate the trend. **(b)** Estimates of mutation rate in human isolates were derived by reconstructing the phylogeny from the isolates represented in (a). Mutation rate is shown on the y-axis in log scale. Estimates of mutation rate from the macaque model and the infecting strain, Erdman (*in vitro*) were determined previously[21].

3.3 Discussion

Here we demonstrate that strains from the East Asian lineage of Mtb acquire drug resistances *in vitro* more rapidly than strains from the Euro-American lineage. This is likely not the result of an enhanced ability of these strains to survive and mutate in the presence of drug and we find no evidence of strong fitness effects that would explain the observed differences in drug resistance rates. Interestingly, we do find evidence that the genetic context of a given Mtb strain can impact the rate at which drug resistances arise. In our analysis, the East Asian isolate, HN878, was permissive for a broader range of *rpoB* mutations than the Euro-American isolate CDC1551. However, the difference in target size is not sufficient to explain the observed difference in rifampicin resistance rates, suggesting that a basal difference in mutation drives the accelerated rate of drug resistance in HN878. In support of this, we find that HN878 more rapidly acquires drug resistance to multiple antibiotics. We expect that differences in mutation rate and differences in target size both contribute to the two to thirty-five fold differences in rifampicin resistance rates that we have measured in the other Mtb strains and in future work will establish the relative contribution of these factors to the drug resistance rate of each strain.

To establish the *in vivo* relevance of these findings, we sought to assess the concordance between mutation rates measured *in vitro* and *in vivo*. Strikingly, we found that the mutation rate of Mtb *in vitro* is very close to the mutation rate – assessed as mutation over time – in isolates from a human transmission chain. Thus, Mtb acquires mutations at a similar rate over time in people as it does in an actively growing culture. This is consistent with our previous findings that the *in vitro* mutation rate over time was similar to the rate of mutation over time in Mtb isolated from the macaques with latent and active disease.

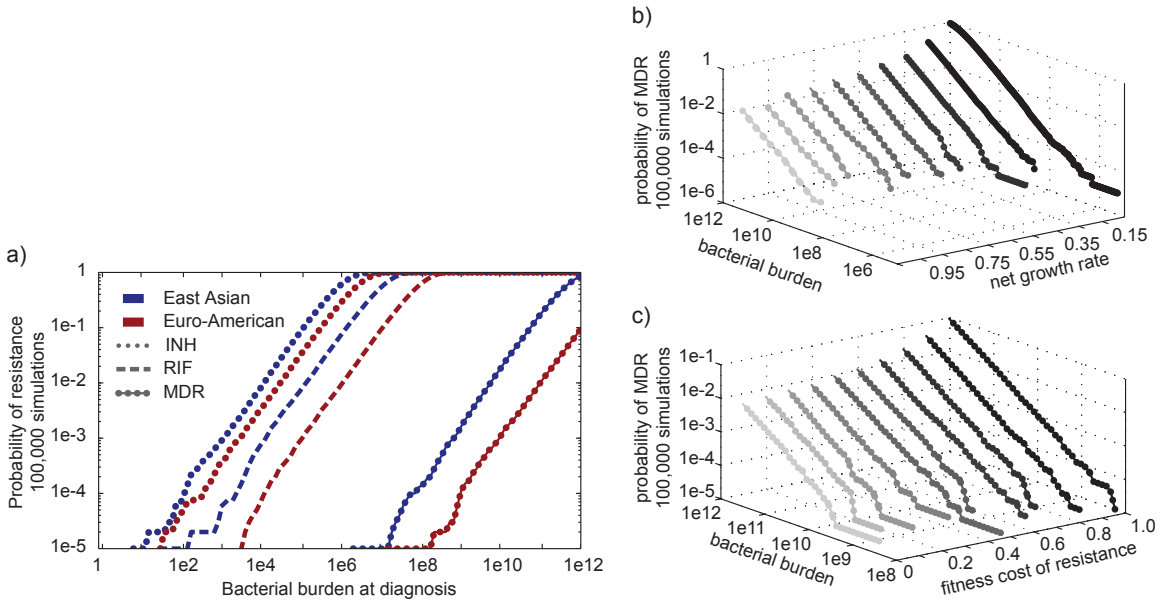


Figure 3.7 An agent based model of drug resistance predicts emergence of resistance before treatment (a) Estimates of the probability of observing drug resistance within a population were derived using an agent-based model of resistance in 200,000 simulations, 100,000 for each lineage. Model parameters are listed in **Supplementary Table 3.5**. Bacterial burden at diagnosis is shown on the x-axis, the probability of observing resistance is shown on the y-axis in log scale. **(b, c)** To determine the sensitivity of our model to variations in growth rate and fitness, we varied each parameter (see **Supplementary Table 3.5**) and determined the probability of observing resistance (z-axis, log scale) at any given bacterial burden (y-axis, log scale) for a specified parameter set (x-axis).

We propose two possible explanations for these data. First, there may be a population of Mtb replicating *in vivo* at a rate similar to the replication rate *in vitro.* These same bacteria may be over-represented in clinical isolates, suggesting that they are more likely to cause disease and be transmitted. Alternatively, it is possible that the mutation rate of Mtb is driven by a time dependent rather than a replication dependent factor. For example, the replicative error rate in Mtb could be very low relative to time and mutations may occur both *in vitro* and *in vivo* largely through DNA damage from endogenous metabolic processes or exogenous stressors.

In future work, we expect that these models may be resolved in part by elucidating the molecular basis of strain-based differences in mutation rate. The differences in mutation rate between clinical Mtb strains are more modest than the differences that distinguish clinical isolates of other bacteria such as *Pseudomonas aeruginosa* or *Escherichia coli*[114-116]. Clinical isolates of these pathogens may become orders of magnitude more mutable than wild type strains through the loss of mismatch repair[114]. However, mycobacteria, and all actinomycetes, lack mismatch repair entirely[117,118], and the molecular basis of replicative fidelity in mycobacteria remains unclear. While mutations in DNA replication and repair proteins are enriched in Mtb strains from the East Asian lineage, no single gene mutation has been found to accelerate the mutation rate of Mtb in isogenic strains. Importantly, East Asian strains also differ in important metabolic pathways from Euro-American Mtb strains[119,120]. Thus, it is possible that genetic differences outside DNA replication and repair contribute to the differences in mutation rate that we have measured.

While further study will be needed to decipher the role of specific biologic factors in driving mutation, we have leveraged the observation that Mtb mutates at a constant rate per unit time to develop a predictive model of the evolution of multi-drug resistance *in vivo*. This

demonstrates that is possible to see multi-drug resistance evolve before the onset of treatment. Moreover, differences in mutation rate have multiplicative effects leading to stark differences in the incidence of multidrug resistance. Indeed, we parameterized our model with data from HN878, which has a relatively modestly elevated rate of drug resistance. Strain X005632 has a 35 fold higher rate of rifampicin resistance as compared to CDC1551. With a similarly elevated rate of isoniazid resistance, the risk of multidrug resistance in an individual infected with X005632 would be three orders of magnitude higher than for a patient infected with CDC1551.

While we have focused on the evolution of rifampicin and isoniazid resistance, our findings are applicable to the evolution of resistances to any antibiotic. Effectively, the rate of resistance to any antibiotic or antibiotic combination is a product of the mutation rate and target size. While resistance to these new combinations may not yet be prevalent, the target size can be very large for some of the antibiotics in the new regimens[121,122]. Thus, resistance to these new combinations may be difficult to avoid especially in the context of East Asian strain infection.

Consistent with epidemiologic data suggesting that severe disease at diagnosis is associated with the acquisition of MDR[123,124] in new cases, our model also predicts that bacterial burden is a critical determinant of the probability of drug resistance. Smear microscopy is the most common primary diagnostic for Mtb around the world but is orders of magnitude less sensitive than both culture and molecular diagnostics[125,126]. If the number of organisms in a patient's sputum roughly reflects the bacterial burden in that individual, patients diagnosed via smear may be at a thousand fold greater risk of harboring drug resistant bacilli than those with access to more sensitive diagnostics. Here we show that this risk may be even higher in the setting of infection with East Asian strains of Mtb. Taken together, these data emphasize the importance of biologic factors in the development of drug resistance and suggest these should be

considered in efforts to limit the emergence of novel resistances to both existing antibiotics and new treatment regimens.

3.5 Materials and Methods

Fluctuation Analysis

Fluctuation analysis was performed as previously described[21]. For a single strain, starter cultures of *M. tuberculosis* were inoculated from freezer stocks of optical density (OD) 1.0 culture. Once at an OD of 1.0, 300,000 cells were used to inoculate 120mL of Middlebrook 7H9 supplemented with 10% Middlebrook OADC, 0.0005% tween 80, and 0.005% glycerol, giving a total cell count of 10,000 cells per 4mL culture. This volume was immediately divided to start 24 cultures of 4mL each in 30mL square PETG culture bottles (Nalgene, Rochester NY). Cultures were grown at 37°C with shaking for 11 to 14 days, until reaching an OD of 1.0. Once at an OD of 1.0, 20 cultures were transferred to 15mL conical tubes and spun at 4000 RPM for 10 minutes at 4°C. Cultures were then resuspended in 250-500μL of 7H9/OADC/tween/glycerol and spotted onto 7H10/OADC/tween/glycerol plates supplemented with 0.5, 2, or 5μg/mL rifampicin (Sigma, R3501), 1μg/mL isoniazid (Sigma, I3377), or 5μg/mL ethambutol (MP Biomedicals, 157949). Once spread using sterile glass beads (4mm diameter), plates were allowed to dry and subsequently incubated at 37°C for 28 days. Cell counts were determined by serial dilution of 4 cultures for each strain. The drug resistance rate was determined by calculating m (the estimated number of mutations per culture) based on the number of mutants (r) observed on each plate using the Ma, Sarkar, Sandri (mss) method as previously described[66,104]. Dividing m by N_t, the number of cells plated for each culture, gives an estimated drug resistance rate. 95% confidence intervals were estimated using equations (24) and (25) as described in Roshe and Foster[104,105]. For comparing pairs of fluctuation analysis data (**Figure 2**), the nonparametric two-sided Wilcoxon rank sum test (also known as the Mann-Whitney *U*-test) was performed using the

ranksum command in Matlab with alpha set to 0.05, comparing the frequency of drug resistant mutants in each culture.

Fluctuation Analysis Data Analysis

To estimate the extent to which the data met the assumptions of Luria-Delbrück fluctuation analysis, we performed a curve fitting analysis as described by Lang and Murray[66]. Briefly, data were fit using either a one-parameter model consistent with the Luria-Delbrück model, or a two-parameter model containing an additional parameter describing a Poisson distribution. The fit of each model was assessed using the least-squares methodology described by Lang and Murray, with AIC_C calculated as described previously[108]. A lower AIC_C reflects a better fit given a penalty for increasing the number of parameters, and a negative ΔAIC_C (ΔAIC_C = AIC_C (one-parameter) – AIC_C (two parameter)) indicates the one parameter model is a better approximation of the data.

Determination of Target Size

The number of *rpoB* mutations conferring resistance to 0.5, 2, and 5μg/mL of rifampicin was determined by isolating 100 colonies, five from each fluctuation analysis culture, into 100μL Middlebrook 7H9 supplemented with 10% Middlebrook OADC, 0.0005% tween 80, and 0.005% glycerol. Cultures were grown overnight at 37°C, and then heat-inactivated at 85°C for 2 hours. Heat-inactivated culture was then used as template for PCR and sequencing using primers previously described[127]. Sequences were analyzed for mutation relative to the reference sequence H37Rv, and totaled. For each culture, duplicate mutations were only counted once. The absolute number of unique mutations seen across cultures for a given condition was used to determine target size for each strain under each condition.

Estimate of mutation rate from human isolates. To determine the per base, per day mutation rate in human isolates, phylogenies were created using the concatenated SNP sequences reported by Gardy et al from a clonal outbreak of a Euro-American strain in British Columbia, Canada[20] using BEAST v.1.7.2 [110,111] to perform Bayesian MCMC analysis. Prior to phylogenetic analysis, SNPs located in repeat regions (PE_PGRSs, PPEs, and transposable elements) were excluded, consistent with our previous analysis of SNPs from whole genome sequencing to estimate mutation rate[21]. Concatenated SNP sequences were compiled and prepared using BEAUti v1.7.2 to select analysis parameters and construct the xml input file. Concatenated sequences were converted to NEXUS format, and loaded into BEAUti where time was added to each isolate. Time was defined in days based on time elapsed from symptom onset relative to isolation of the historical isolate, MT0005 (1995). A GTR substitution model was used with empirically determined base frequencies. Default priors were used for 10,000,000 chains. Output was analyzed in Tracer v1.5, and all parameters produced an effective sample size of 200 or greater. Phylogentic tree construction was completed using TreeAnnotator v1.7.2 with a posterior probability limit of 0.5 and a burnin of 1000 trees, leaving 9001 potential trees for construction. Tree visualization was completed using FigTree v1.3.1 and the tree was rooted on MT0005.

Mathematical simulation of drug resistance

We developed an agent-based mathematical model of the evolution of drug resistance within an individual according to the following set of equations:

$$(1) \qquad N_S(t) = [N_S(t-1)*(b-d_A)] - m_{R \cdot S} - m_{H \cdot S}$$

$$(2) \qquad N_R(t) = [N_R(t-1)*(b*(1-cr_R) - d_A)] + m_{R \cdot S} - m_{H \cdot R}$$

$$(3) \qquad N_H(t) = [N_H(t-1)*(b*(1-cr_H) - d_A)] + m_{H \cdot S} - m_{R \cdot H}$$

(4) $N_{MDR}(t) = [N_{MDR}(t-1)*(b*(1-cr_{MDR}) - d_A)] + m_{R \cdot H} + m_{H \cdot R}$

Where:

(5) $m_{R \cdot S} \sim \text{Poisson}(\mu_R * N_S(t-1))$

(6) $m_{H \cdot S} \sim \text{Poisson}(\mu_H * N_S(t-1))$

(7) $m_{H \cdot R} \sim \text{Poisson}(\mu_H * N_R(t-1))$

(8) $m_{R \cdot H} \sim \text{Poisson}(\mu_R * N_H(t-1))$

These equations were parameterized with the values displayed in Supplementary Table 5. All simulations were run in Matlab (Natick, MA). For all simulations and for both mutation parameter sets ($\mu_{H\text{-}W}$ & $\mu_{R\text{-}W}$, $\mu_{H\text{-}CDC}$ & $\mu_{R\text{-}CDC}$), simulations of the evolution of drug resistance were run 100,000 times to determine the probability of observing drug resistance with a given set of parameters. To determine the effect of varying birthrate, 10 simulations of 200,000 simulated patients (100,000 per simulated strain) each were run with b = (0.20:1.10 in increments of 0.10), giving a net birth rate of 0.05:0.95. To determine the effect of varying the fitness of drug resistance mutants, 10 simulations of 200,000 patients each (100,000 per simulated strain) were run with $cr_H = cr_R$ = (0.0 : 0.90 in increments of 0.10). For all simulations, bacterial burden was allowed to increase to 1e12 bacteria within a patient, and the probability of observing rifampicin resistance, isoniazid resistance, and MDR was determined by dividing the number of simulated patients with at least one resistant bacteria by the total number of simulated patients.

Author Contributions. Christopher B Ford designed the project, performed molecular studies, conducted the data analyses, designed and conducted the mathematical model, prepared the figures and drafted the manuscript; Rupal R. Shah conducted preliminary molecular studies; Midori Kato Maeda and Sebastien Gagneux supplied the isolates; Ted Cohen and Megan Murray advised the mathematical model; Jen Gardy and Jay Johnston supplied the whole genome sequencing data; Marc Lipsitch supervised and advised statistical analyses and mathematical modeling; Sarah M. Fortune initiated the project, supervised analysis of the data, and drafted the manuscript.

Chapter 4 – Mycobacteria do not use canonical mechanisms of proofreading to maintain DNA replicative fidelity.

4.1 Introduction

Mycobacterium tuberculosis (Mtb) infects over 1.8 billion people, with approximately 20 million cases of active disease and 2 million deaths annually. Effective treatment is compromised by the evolution of drug resistance[4,6,7,82], despite treatment with a multidrug regimen[85]. *In vitro*, resistance occurs at a rate of approximately 10^{-7} to 10^{-9} per replication cycle, depending on the antibiotic[21,128]. Unlike other pathogens, all known mediators of drug resistance in *Mtb* are chromosomally encoded, arising through single nucleotide polymorphisms in genes coding antibiotic targets or antibiotic processing pathways. In the absence of horizontal gene transfer, mutation as a result of DNA damage or replicative error becomes the dominant source of genetic diversity and antibiotic resistance. In this context, strains or isolates of Mtb with a higher mutation rate may have a fitness advantage. Indeed, in other human pathogens, such as *Escherichia coli* 0157:H7, *Salmonella enterica*[129] or *Pseudomonas aeruginosa*[116], isolates have been identified that have a ~100- fold increase in mutation frequency resulting from a suspension of post-replicative mismatch repair (MMR). As mycobacteria, and indeed, all actinomycetes, lack homologs of the major MMR proteins MutSHL[117,118], it is unclear how a mutator strain might arise.

In the majority of prokaryotic systems, DNA replication fidelity is accomplished using three primary mechanisms: (1) base pair selection by the primary replicative polymerase, (2) replication associated proofreading 3) and MMR. While mycobacteria lack MMR, and the contribution of replicative mechanisms to fidelity is unknown in mycobacteria, our work indicates that *Mtb* has a mutation rate comparable to *Eschericia coli* (*E. coli*) [21,130]. In the

absence of canonical post-replicative MMR, it is possible that mycobacteria rely solely on replication-associated mechanisms of fidelity. Though mycobacteria share many traits with gram-positives, the core DNA replication machinery most closely resembles systems from gram-negative systems[67,117]. In support of this, *Mtb* possesses a homolog for the gram-negative proofreading 3'-5' exonuclease subunit, polIIIε (DnaQ), and the α subunit of the polIII holoenzyme (DnaE1) does not possess any domains reflective of intrinsic proofreading activity. In model systems, DnaQ interacts directly with DnaE between the thumb and finger domains, positioning itself immediately distal to the site of polymerization[131]. From this position, DnaQ interacts directly with DNA to improve both processivity and fidelity by engaging in a series of electrostatic interactions with the newly synthesized strand and excising incorrectly paired bases through the action of an three highly conserved exonuclease motifs[132,133]. Oddly, unlike *E. coli, Haemophilus influenzae, or Streptococcus pneumoniae*, the epsilon subunit appears to be dispensable for normal growth *in vitro*[72,134-138]. Here, we seek to determine the role of proofreading in maintaining DNA replicative fidelity in mycobacteria.

4.2 Results

4.2.1 Identification and deletion of *dnaQ* homologs in mycobacteria.

The polIIIε exonuclease subunit of the replicative holoenzyme is characterized by the presence of three exonuclease motifs in their N-terminus, the first two of which (exoI & exoII) are shared by the larger family of 3'-5' exonucleases. In proofreading enzymes, the ExoIII motif is absent, replaced by an ExoIIIε motif found predominantly in polymerase-associated proofreading exoncleases[131-133,139]. To identify the potential proofreading subunit in mycobacteria, we searched for 3'-5' exonucleases containing this highly conserved motif

Figure 4.1

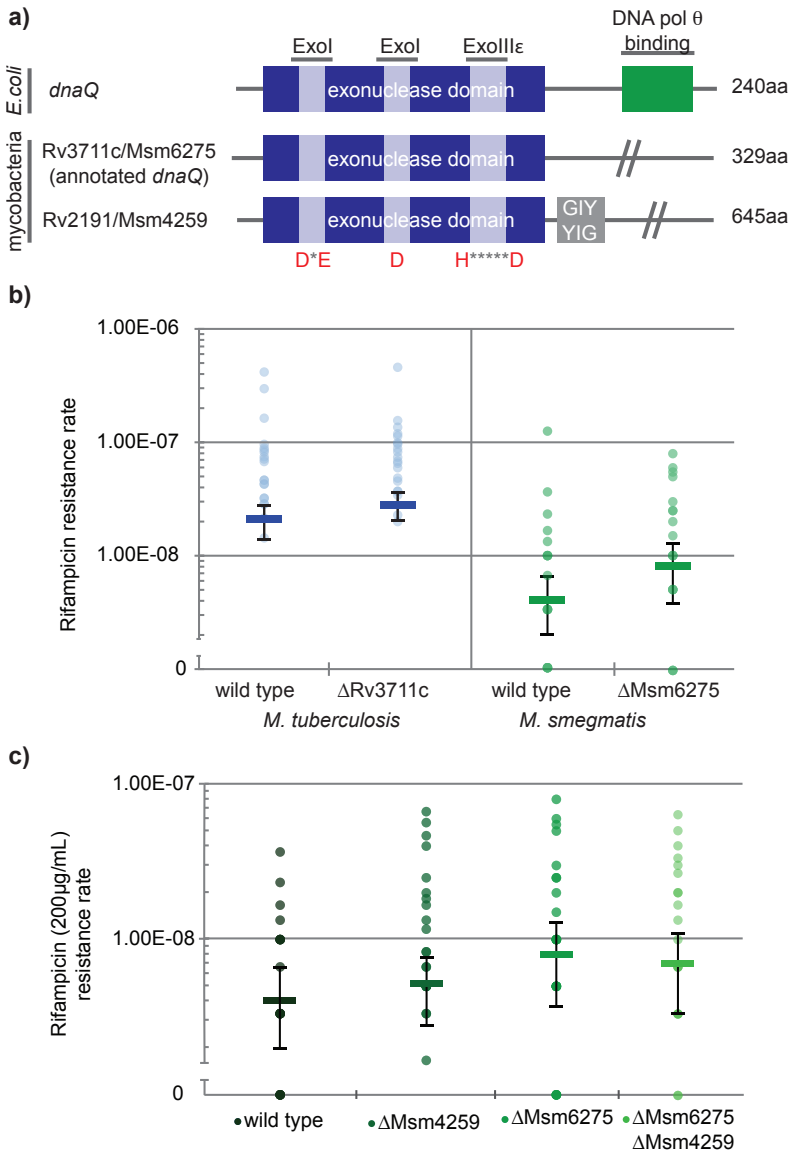

a)
ExoI
ExoI
ExoIIIε
DNA pol θ binding
E.coli
dnaQ
exonuclease domain
240aa
mycobacteria
Rv3711c/Msm6275 (annotated dnaQ)
exonuclease domain
329aa
Rv2191/Msm4259
exonuclease domain
GIY YIG
645aa
D*E
D
H*****D
b)
1.00E-06
1.00E-07
1.00E-08
0
Rifampicin resistance rate
wild type
ΔRv3711c
wild type
ΔMsm6275
M. tuberculosis
M. smegmatis
c)
1.00E-07
1.00E-08
0
Rifampicin (200μg/mL) resistance rate
wild type
ΔMsm4259
ΔMsm6275
ΔMsm6275 ΔMsm4259

Figure 4.1 (Continued) Deletion of two 3'-5' exonucleases with an ExoIIIε motif in mycobacteria. (a) Mycobacteria are unique in that they possess two potential *dnaQ* homologs: Rv3711c/Msm6275 and Rv2191/Msm4259. The exonuclease domain is shown in blue, with the ExoI,II, IIIε highlighted in a lighter shade of blue. For all motifs, the conserved residues are shown below in red. The DNA polIIIθ binding domain of *E. coli dnaQ* is shown in green. **(b)** Fluctuation analysis of ΔRv3711c and ΔMsm6275 deletion strains on rifampicin (2μg/mL and 200μg/mL respectively) reveals no significant change in mutation rate relative to wild type. Mtb is displayed in blue, Msm in green. Circles represent mutation frequency (number of mutants per cell in a single culture), where darker circles represent multiple cultures with the same frequency. Bars represent the estimated mutation rate, with error bars representing the 95% confidence interval. Significance was determined by comparing strain pairs using the Wilcoxon rank-sum test. Values are listed in **Supplementary Table 4.1. (c)** In Msm, deletion of a second gene containing a conserved ExoIIIε motif (Msm4259) either alone or in combination with deletion of Msm6275 does not lead to a significant increase in mutation rate in comparison to wild type by fluctuation analysis on rifampicin (200μg/mL).

structure. Our search yielded two results, Rv3711c (Msm6275) and Rv2191 (Msm4259) (**Figure 4.1a**). While Rv3711c resembles the domain structure found in *E. coli*, Rv2191 possesses an additional domain – a GIY-YIG motif[140] characteristically found at the N-terminus of UvrC proteins - suggesting it may be involved in nucleotide excision repair, similar to other UvrC homologs[141,142].

Therefore, we have made an unmarked deletion mutant of the annotated *dnaQ* (Rv3711c & Msm6275) from both *Mtb* and *M. smegmatis* (Msm) through homologous recombination. Mutation or deletion of *dnaQ* in *E. coli* has previously been reported to result in a 100-1000 fold increase in mutation rate[143-146]. We used Luria-Delbrück fluctuation analysis[103,104] to assess the rate at which our *dnaQ* deletion strains acquired rifampicin resistance. Surprisingly, in both Msm and Mtb, deletion resulted in only a mild, statistically insignificant (though repeatable) increase in rifampicin resistance rate (1.35 and 1.97 fold increase, respectively. $p>0.05$ by Wilcoxon Rank sum)(**Figure 4.1b, Supplementary Table 4.1**). As deletion was confirmed by both PCR and Southern blot, these data suggest that Rv3711c (and Msm6275), despite possessing an intact ExoIIIε motif, does not contribute significantly to DNA replicative fidelity.

To address the possibility that Msm4259, despite possessing atypical architecture, may be the primary replicative proofreading subunit, we made a unmarked deletion mutant of Msm4259 through homologous recombination in both wild type Msm and MsmΔMsm6275 and repeated our fluctuation analysis (**Figure 4.1c, Supplementary Table 4.1**). While deletion of Msm6275 led to a slight (though reproducible) increase in mutation rate, deletion of Msm4259 alone or in combination with Msm6275 does not lead to an increase in mutation. This suggests that Msm4259 is not responsible for proofreading nascent DNA, nor it is likely to serve in a redundant pathway with Msm6275.

To investigate the possibility that deletion of Msm 4259 or Msm6275 in Msm and Rv3711c in Mtb exert a subtle phenotype by altering the manner in which mycobacteria mutate in the presence to antibiotic, we analyzed the cumulative distribution of mutations observed in each fluctuation analysis (**Figure 4.2a-f**). As described in Chapter 3, if cells continue to grow and mutate in the presence of antibiotic, the distribution of mutants will deviate from the expected Luria Delbrück distribution. We used a curve fitting approach developed by Lang and Murray[66] to determine whether the distribution of mutant frequencies in the two strains is better fit using a one parameter, Luria-Delbrück model or a two parameter, Luria-Delbrück and Poisson model. We then used the Akaike information criterion with correction for sample size (AIC_C), to determine which model best fit the data[107,108]. The AIC_C quantifies the fit of a model to observed data, with a lower AIC_C reflecting better relative fit. In the deletion strains analyzed, ΔAICc (AIC_C(one parameter) - AIC_C(two parameter)) was less than zero, demonstrating that there is not a significant Poisson component in the distributions (**Figure 4.2g**). While the ΔAICc was slightly over 0 (~0.08) for wild type H37Rv, we attribute this to slight differences in methodology as repeated fluctuation analyses with this strain have consistently yielded ΔAICc<0. Taken together, this data suggests that neither ExoIIIε motif containing protein significantly contributes to the maintenance of genomic fidelity. However, the possibility remains that mycobacteria possess alternative pathways of genomic fidelity that can compensate for the loss of *dnaQ*. To test this hypothesis, we have pursued a forward genetic approach to identify alternative, noncanonical mechanisms for fidelity and replication in mycobacteria that are conditionally essential in the absence of *dnaQ*.

Figure 4.2

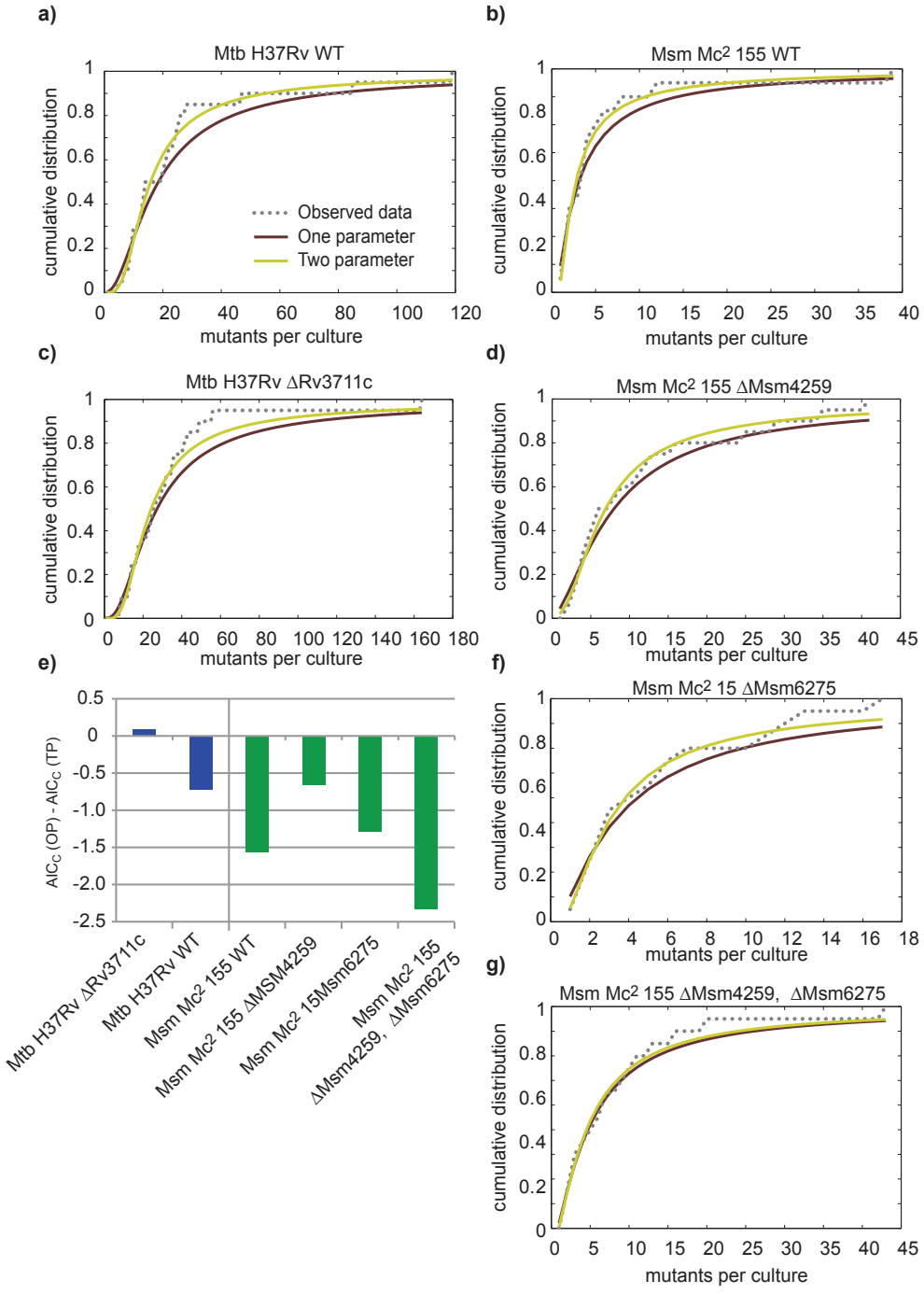

a)
Mtb H37Rv WT
cumulative distribution
mutants per culture
Observed data
One parameter
Two parameter
b)
Msm Mc2 155 WT
c)
Mtb H37Rv ΔRv3711c
d)
Msm Mc2 155 ΔMsm4259
e)
AICC (OP) - AICC (TP)
Mtb H37Rv ΔRv3711c
Mtb H37Rv WT
Msm Mc2 155 WT
Msm Mc2 155 ΔMSM4259
Msm Mc2 15Msm6275
Msm Mc2 155 ΔMsm4259, ΔMsm6275
f)
Msm Mc2 15 ΔMsm6275
g)
Msm Mc2 155 ΔMsm4259, ΔMsm6275

Figure 4.2 (Continued) The cumulative distribution of drug resistant mutants from deletion strains is best fit by a one parameter, Luria Delbrück distribution. Curve fitting analysis was performed to determine if the cumulative distribution of the fluctuation analysis data better fit a one parameter, Luria-Delbrück model or a two parameter, Luria-Delbrück and Poisson model. **(a-f)** A dotted grey line represent the cumulative distribution function of the observed data, a solid red line represents the cumulative distribution function of the one-parameter model, and a solid yellow line represents the cumulative distribution function of the two-parameter model. The number of mutants per culture is displayed on the x-axis, and the probability of observing (x) mutants per culture or fewer is shown on the y-axis. **(g)** To determine which model best fit each data set, we determined the Akaike Information Criterion, corrected for small sample size (AIC_C). A smaller AIC_C represents a better fit, given a penalty for more parameters in a model. If the AIC_C (one parameter) is smaller than the AIC_C (two parameter), then the resulting value will be negative, reflecting a better fit for the one parameter model.

4.2.2 Forward genetic search for genes essential in the absence of Rv3711c

Given the essentiality of proofreading in other systems and the absence of a mutator phenotype in H37RvΔRv3711c, are there genes that are conditionally essential in the absence of Rv3711c? To address this question, we employed transposon capture and sequencing (TraCS)[72,134] to identify genes that are conditionally essential in the absence of Rv3711c, reasoning that these gene products may act in concert with Rv3711c to maintain genomic fidelity. Fitness may be compromised through either a disruption of normal replicative processivity (resulting in slower growth), or through excessive error rates (error catastrophe) as is seen in the *mutD* strain in *E. coli*[143-146]. By either mechanism, a reduction in fitness resulting from interruption of genes coding redundant mechanisms of genomic fidelity can be quantified through TraCS.

Through transduction with the ϕMycoMar-T7 phage carrying the Himar-1 transposon[138], we have created triplicate Mtb transposon libraries in both the parent strain (H37Rv) to our deletion mutant and in the deletion mutant itself (H37RvΔRv3711c). Genomic DNA was extracted from each library, sheared, and prepared for sequencing. Briefly, an adapter was ligated to end-repaired DNA, and transposon junctions were amplified using primers homologous to the transposon and the adapter. Both primers contain Illumina attachment homology and read-primer homology. Sequencing of each library yielded at least ~4,000,000 valid, mappable reads that passed filter for each library. Pooling the biologic replicates, we observed insertions a total of 47,924 unique TN insertion sites in H37Rv and a total of 55,321 unique sites in H37RvΔRv3711c, suggesting excellent coverage across a genome with 74,602 TA insertion sites (**Figure 4.3a**). There was very little evidence of abundant PCR amplification disproportionate to template number due to stochastic early exponential amplification (PCR

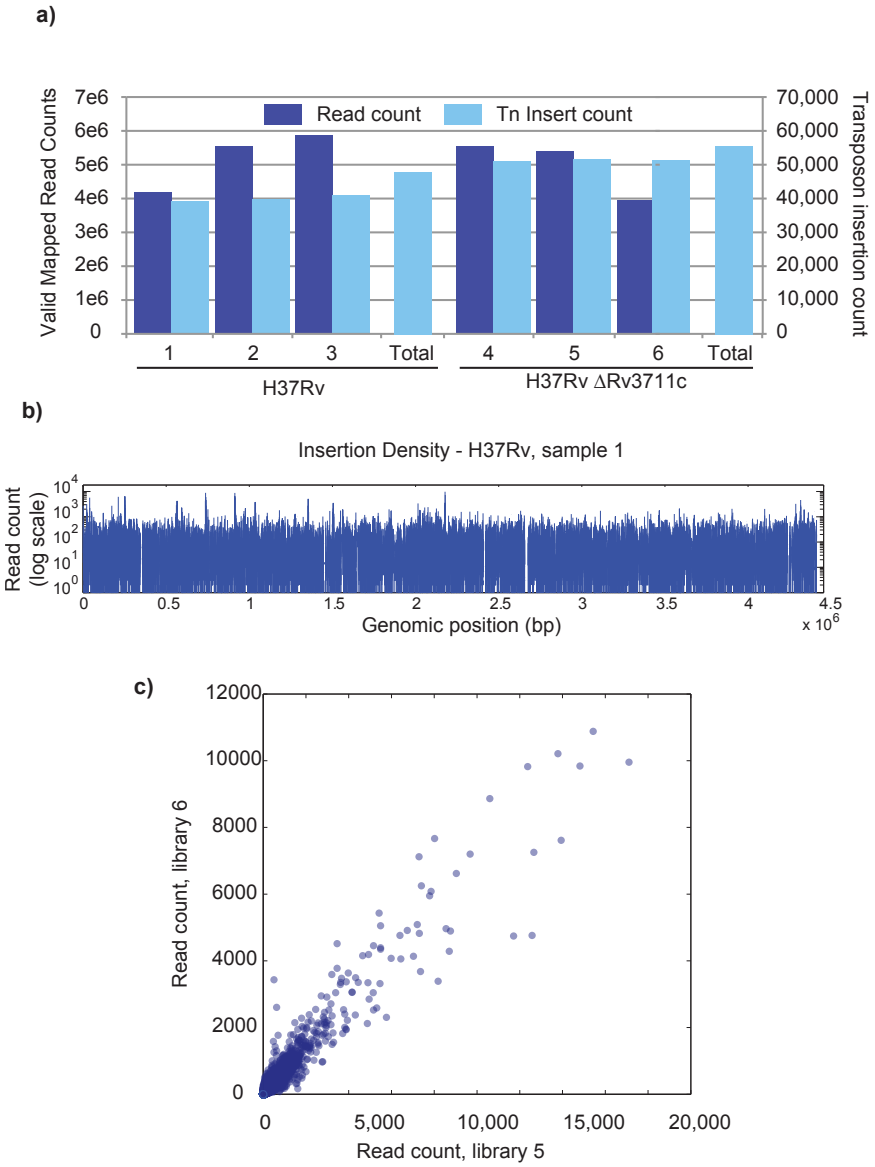


Figure 4.3 TraCS allows for the quantitative profiling of independent transposon insertion mutants. (a) The number of valid mapped reads per library is displayed in dark blue, and the number unique TN-TA insertion sites is displayed in light blue, with the total number of unique sites shown for each strain. **(b)** The number of reads per TA site for a single library is shown, with little evidence of PCR jackpotting. **(c)** The agreement in number of reads per TA site in two biologic replicates is shown by plotting the reads per TA in library 5 against the reads per TA in library 6.

jackpotting) (**Figure 4.3b & c**).

To identify genes that were significantly different between H37Rv and H37RvΔRv3711c, we performed pair-wise comparisons between each control and experimental sample, and tested for significance using a two-sided Wilcoxon rank sum test. We then generated a composite p-value for each gene using Fisher's method, and plotted p-value against normalized fold-change in read count to identify genes that were significantly different by more than two-fold between wild type H37Rv and H37RvΔRv3711c (**Figure 4.4a, Supplementary Table 4.2**). Here we have used strict conservative thresholds for both significance ($p<1.38e-5$) and fold change (fold change < 0.5, >2), in part to avoid stochastic differences between libraries. Only nine genes were under represented in the H37RvΔRv3711c library, including Rv3711c itself (**Figure 4.4b**). Of the remaining eight, three are hypothetical proteins with unknown function and no significant homology to proteins of known function or known domains. Of these three, two (Rv02164c and Rv3587c) have few reads in either library, suggesting they may be the result of stochastic differences in transposon insertion (**Supplementary Table 4.2**). Three may be related to metabolism, either carbon metabolism (Rv0126 and Rv0127) or lipid metabolism (Rv1592c). Interestingly, *treS* (Rv0127) and mak (Rv0126) are involved in the conversion of maltose to trehalose, and inactivation of either leads to the toxic accumulation of maltose-1-phosphate[147]. The remaining two are components of two-component signal transductions systems (TCST) (*phoP* and *blaR*). Here, we will focus on *phoP*, as the differences in read count are most pronounced (**Figure 4.4c**).

phoP encodes the transcriptional regulator of the TCST PhoPR, the inactivation of which leads to high attenuation in macrophages and BALB/c mice[148]. The PhoPR regulon has been described previously by profiling transcription levels in a clinical isolate (MT103) and in a *phoP*

Figure 4.4

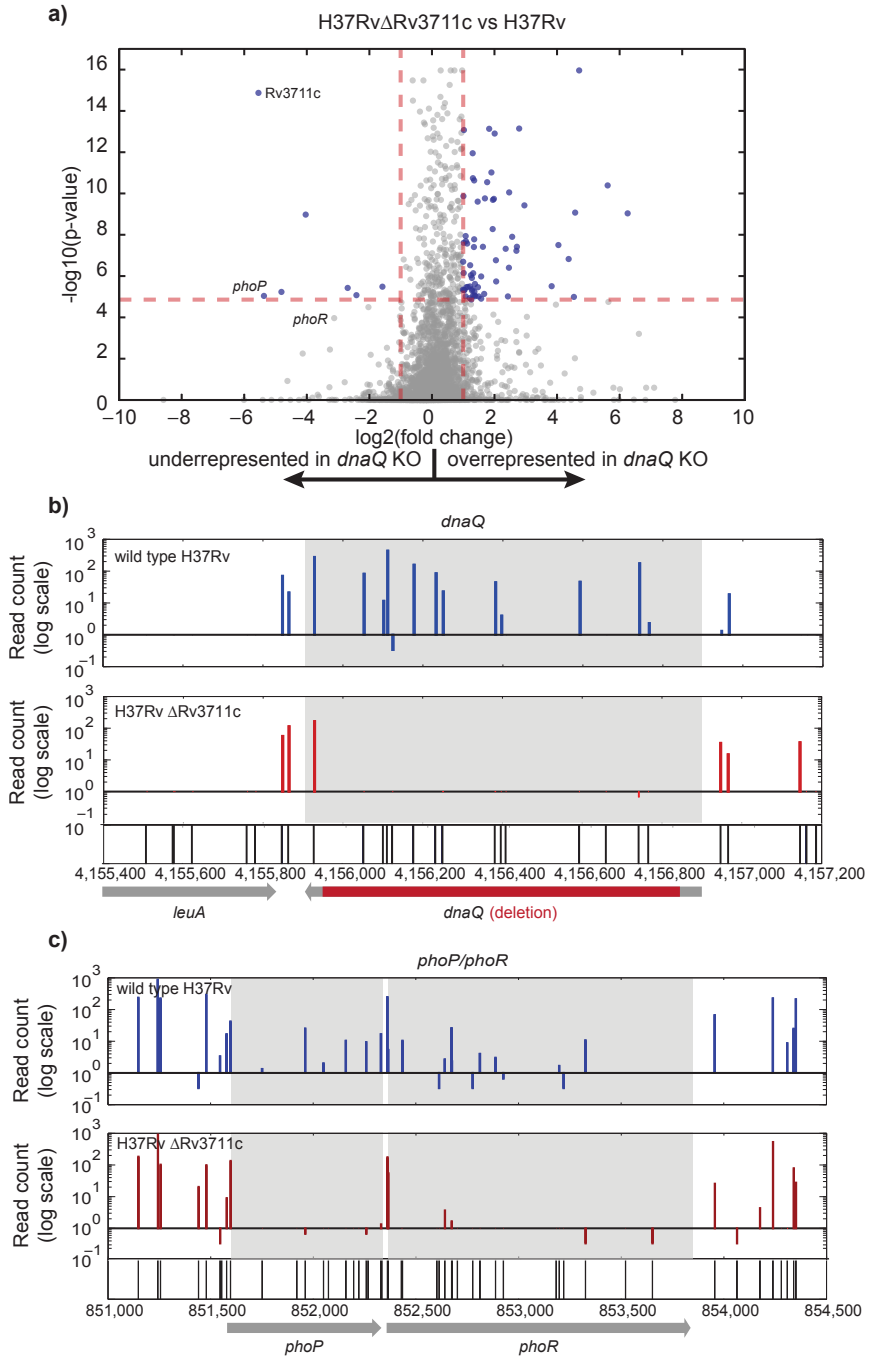

a)
H37RvΔRv3711c vs H37Rv
Rv3711c
phoP
phoR
-log10(p-value)
log2(fold change)
underrepresented in dnaQ KO
overrepresented in dnaQ KO
b)
dnaQ
wild type H37Rv
H37Rv ΔRv3711c
Read count (log scale)
4,155,400 4,155,600 4,155,800 4,156,000 4,156,200 4,156,400 4,156,600 4,156,800 4,157,000 4,157,200
leuA
dnaQ (deletion)
c)
phoP/phoR
wild type H37Rv
H37Rv ΔRv3711c
Read count (log scale)
851,000 851,500 852,000 852,500 853,000 853,500 854,000 854,500
phoP
phoR

Figure 4.4 (Continued) Analysis of TraCS data reveals genes that are significantly underrepresented in the H37RvΔRv3711c, including the Rv3711c and *phoP*. **(a)** For each gene, $\log_2$(fold change) was plotted on the x-axis was plotted against the $-\log_{10}$(p-value) as determined by two sided Wilcoxon rank sum. Vertical red dashed lines represent a fold change of two, and horizontal red-dashed lines represent a p-value of 1.38×10^{-5}, the Sidak-corrected p-value. Points in blue represent genes that fall outside these bounds, and are listed in Table 4.2. **(b)** The number of reads per TA is shown for the region surrounding and including Rv3711c. Insertions in H37Rv are shown in the top graph in blue; insertions in H37RvΔRv3711c are shown in the second panel in red. TA sites are shown in the third panel, and genes locations are displayed below the x-axis. **(c)** The same format is used to display read insertion density for *phoP* and *phoR*. All values are listed in **Supplementary Table 4.2.**

mutant generated by replacing a *EcoRV-BclI* restriction fragment with a hygromycin resistance cassette[149]. The majority of the 78-gene regulon of PhoPR is positively regulated, and can be divided into six major functional clusters. Five of these clusters are under positive regulation, including gene sets responsible for 1) hypoxia adaptation, 2) cellular respiration, 3) lipid metabolism, 4) virulence and 5) stress response (**Supplementary Table 4.2**). Only four genes are negatively regulated by PhoPR (*icl, fadB2, umaA1, PE*) three of which are operonic (*icl, fadB2, umaA1*) and may be involved in persistence.

To understand why *phoP* (and to a lesser extent, *phoR*) are underrepresented in our H37RvΔRv3711c library, we examined the significance and fold change of genes in the regulon (**Figure 4.5a, Supplementary Table 4.3**) and genes whose expression is correlated with *phoP* (**Figure 4.5b, Supplementary Table 4.4**). Interestingly, the most prominent gene, both in significance and magnitude, is *phoR*, which is known to be autoregulated by *phoP*. The strong signal from *phoR* supports the conclusion that disruption of this TCST leads to a fitness cost in H37RvΔRv3711c; however, no single gene in the PhoPR regulon explains the essentiality of this system in the absence of Rv3711c. It is likely that multigenic effects resulting from the disruption of the PhoPR lead to the underrepresentation of the system in the H37RvΔRv3711c library.

While no obvious DNA repair genes are present in our set of eight genes significantly underrepresented in the H37RvΔRv3711c library, we sought to determine how the essentiality of a suite of Mtb DNA repair genes in Mtb was altered by deletion of Rv3711c. Therefore, we determined the fold change and significance of this subset of genes (**Figure 4.6, Supplementary Table 4.4**). Strikingly, very few genes with annotated DNA replication or repair function were underrepresented H37RvΔRv3711c library; however three genes are significantly

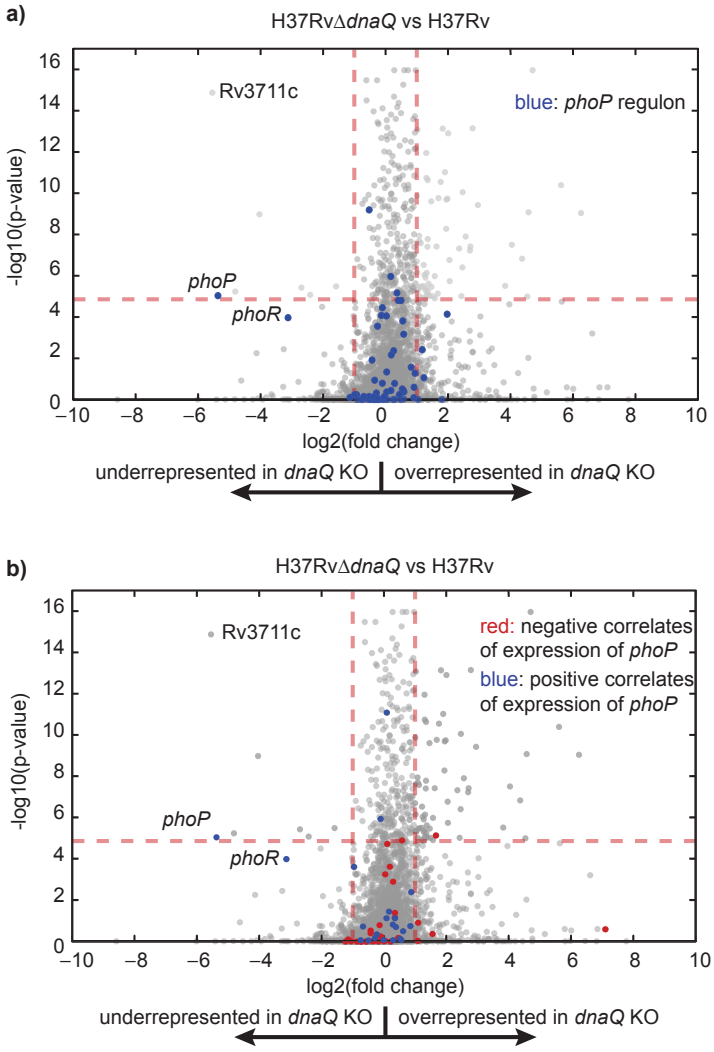


Figure 4.5 The fold change and significance of genes associated the PhoPR regulon is below threshold for both the *phoP* regulon and genes whose expression is correlated with *phoP*. **(a)** Members of the *phoP* regulon are shown in blue, including *phoP* and *phoR*. **(b)** Positive correlates of *phoP* expression are shown in blue, including *phoP* and *phoR*. Negative correlates of *phoP* expression are shown in red. All values are listed in **Supplementary Table 4.3 and 4.4.**

overrepresented. One possible interpretation of genes overrepresented in the H37RvΔRv3711c library is that these genes participate in a common pathway with Rv3711c. If deletion of Rv3711c leads to a fitness cost through accumulation of a toxic intermediate, then loss of other genes upstream of Rv3711c within the same pathway would have reduce the incurred fitness cost. In wild type cells, interruption of these genes would result in interruption of the pathway, resulting in some loss of fitness; thus the genes would be overrepresented in the H37RvΔRv3711c library. Amongst this set of three genes are two with known function: *mutT2* and Rv3204. *mutT2* is predicted to be involved in the removal of 8-oxo-guanine from the genome, and experimental evidence suggests that MutT2 from Msm is able to efficiently hydrolyze 8-oxo-dGTP, dGTP, and dTTP[150]. Rv3204 is short protein (101 amino acids) principally composed of a DNA binding domain characteristic of 6-O-methylguanine DNA methyltransferases. While its function is unknown, Rv3204 may coordinate DNA binding or damage recognition for a larger pathway. These results suggest that Rv3711c may participate in an alternative novel pathway of DNA repair, though further experimentation and analysis is needed to determine if such a pathway exists.

4.3 Discussion

In gram-negative bacteria, *dnaQ* codes for DNA pol-IIIε, an essential member of the replicative DNA pol-III holoenzyme, responsible for proofreading nascently polymerized DNA[131]. Loss or mutation of *dnaQ* leads to a strong mutator phenotype, resulting in enfeebled bacteria due to both mutation catastrophe and reduced polymerase processivity[143,146]. While mycobacteria share many features with gram-positive bacteria, their DNA repair machinery shares homology with gram-negatives, suggesting that *dnaQ* may play a similar role in mycobacteria. Here we have deleted the annotated *dnaQ* in both Mtb (Rv3711c) and Msm

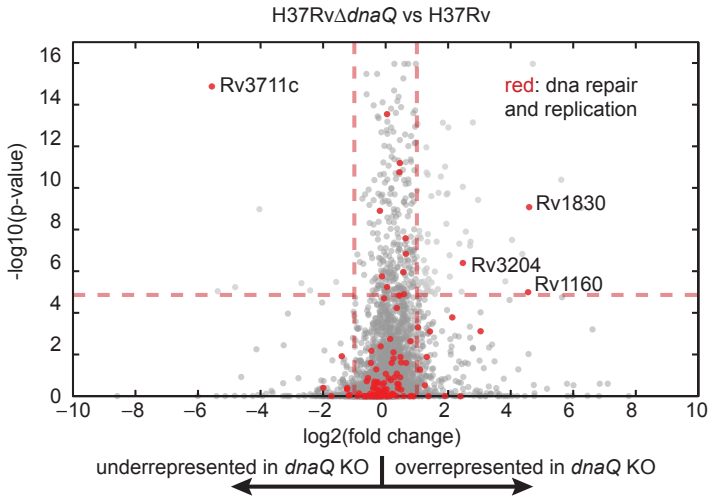


Figure 4.6 Three DNA repair associated genes are significantly overrepresented in the absence of Rv3711c. The fold change and significance of Genes associated with DNA repair or containing DNA binding and interaction domains are plotted as red circles. Three genes are significantly over represented in the Rv3711c deletion strain, suggesting that transposon mutants in these strains are more fit in the absence of Rv3711c. All values are listed in **Supplementary Table 4.5.**

(Msm6275) and found no significant change in mutation rate. Additionally, we have deleted a second potential DNA pol-IIIε subunit in Msm (Msm4259) and again found no significant change in mutation rate. These results suggest that mycobacteria do not use canonical mechanisms of proofreading, a conclusion supported by the observation that neither potential DNA pol-IIIε subunit is essential by transposon mutagenesis. However, we could not rule out the presence of a redundant system that mitigated both the essentiality of *dnaQ* and the mutator effect of deletion.

To pursue this hypothesis, we used TraCS to identify genes that are conditionally essential in the absence of Rv3711c in Mtb. Our analysis revealed only eight genes that are significantly underrepresented in the absence of Rv3711c, including *phoP*. Notably, we have used strict quantitative thresholds ($p<1.38e-5$, fold change < 0.5 or > 2) to reduce the effect of stochastic differences between libraries. While none of these eight possess domains suggesting function related to genomic fidelity, we cannot rule out the possibility that one or more are involved in proofreading. It is noteworthy that both *treS* (Rv0126) and *mak* (Rv0127) are underrepresented in the absence of Rv3711c. These play a critical role in one of the three possible trehalose synthesis pathways present in mycobacteria[147]. Additionally, PhoP (and to a lesser extent, PhoR) positively regulate the synthesis of methyl-branched fatty acid-containing acyltrehaloses found exclusively in pathogenic mycobacteria[151]. While the connection between trehalose and Rv3711c is unclear, it is possible that the protective effect of trehalose in the cytoplasm is necessary to compensate for any cellular stress generated by loss of Rv3711c.

Interestingly, amongst the genes that are overrepresented in the H37RvΔRv3711c library, we find three genes with annotated function suggestive of DNA repair or modification. It is possible that these genes are overrepresented as a result of participating in the same DNA repair

pathway as Rv3711c, a pathway that may not involve replicative fidelity. If Rv3711c is functioning as nuclease in a novel DNA repair pathway that includes *mutT2*, Rv3204, and Rv1160 it is possible that deletion of the exonuclease results in orphaned DNA repair intermediates, which can block replication by stalling replicative polymerases. By abrogating the function of upstream genes in the pathway, the formation of toxic DNA lesions may be avoided. A similar phenomenon is seen in base excision repair in *Salmonella typhimurium.* Fitness loss from the accumulation of apurinc sites following deletion of the endonucleases *xth* and *nfo* is ameliorated by deletion of the glycosylases *ung* and *fpg*[152]. Further work is needed to determine if these genes do participate in a novel repair pathway, and if so, the substrates and activity of that pathway.

Taken together, our results suggest that Mtb, and mycobacteria as a whole, do not rely on canonical methods of proofreading to maintain genomic integrity, leaving the question open: how do mycobacteria maintain genomic fidelity in the absence of homologous systems of proofreading and mismatch repair? There are two primary possibilities: (1) mycobacteria have evolved mechanisms of DNA replication that do not require proofreading by DNA pol III-ε for holoenzyme stability and fidelity, or (2) an unidentified nonhomologous protein has assumed this role in a manner independent of the *dnaQ* candidates investigated here. Further work focusing on the fidelity of DNAE1 and the polIII holoenzyme may resolve these two possibilities.

4.5 Materials and Methods

Creation of deletion mutants

Rv3711c, Msm6275, and Msm4259 were deleted from wild type Mtb (H37Rv) and Msm (Mc^2155) through homologous recombination with a suicide vector (pJM1) containing both a selectable hygromycin marker and a counterselectable *sacB* marker. Deletion was confirmed by PCR. Additionally, deletion of Rv3711c and Msm 6275 were confirmed by Southern blot.

Fluctuation Analysis

Fluctuation analysis was performed as previously described[21]. For a single strain, starter cultures of Mtb or M. smegmatis were inoculated from freezer stocks of optical density (OD) 1.0 culture. Once at an OD of 1.0, 300,000 cells were used to inoculate 120mL of Middlebrook 7H9 supplemented with 10% Middlebrook OADC, 0.0005% tween 80, and 0.005% glycerol, giving a total cell count of 10,000 cells per 4mL culture. For Mtb fluctuation analysis, this volume was immediately divided to start 24 cultures of 4mL each in 30mL square PETG culture bottles (Nalgene, Rochester NY). Mtb Cultures were grown at 37°C with shaking for 11 to 14 days, until reaching an OD of 1.0. Once at an OD of 1.0, 20 cultures were transferred to 15mL conical tubes and spun at 4000 RPM for 10 minutes at 4°C. For Msm fluctuation analysis, the initial culture was immediately divided to start 24 cultures of 4mL each in 15ml Falcon conical tubes (BD Biosciences.) Once at an OD of 1.0, 20 cultures were spun at 4000 RPM for 10 minutes at 4°C. Cultures were then resuspended in 250-500µL of 7H9/OADC/tween/glycerol and spotted onto 7H10/OADC/tween/glycerol plates supplemented with 2µg/mL rifampicin (Mtb) or 200µg/mL rifampicin (Msm) (Sigma, R3501). Once spread using sterile glass beads (4mm diameter), plates were allowed to dry and subsequently incubated at 37°C for 28 days (Mtb) or 7 days (Msm). Cell counts were determined by serial dilution of 4 cultures for each strain. The

drug resistance rate was determined by calculating m (the estimated number of mutations per culture) based on the number of mutants (r) observed on each plate using the Ma, Sarkar, Sandri (mss) method as previously described[66,104]. Dividing m by N_t, the number of cells plated for each culture, gives an estimated drug resistance rate. 95% confidence intervals were estimated using equations (24) and (25) as described in Roshe and Foster[104,105]. Pair wise comparison of fluctuation analysis data was done using the nonparametric two-sided Wilcoxon rank sum test (also known as the Mann-Whitney *U*-test) with the *ranksum* function in Matlab with alpha set to 0.05, comparing the frequency of drug resistant mutants in each culture.

Fluctuation Analysis Data Analysis

To estimate the extent to which the data met the assumptions of Luria-Delbrück fluctuation analysis, we performed a curve fitting analysis as described by Lang and Murray[66]. Briefly, data were fit using either a one-parameter model consistent with the Luria-Delbrück model, or a two-parameter model containing an additional parameter describing a Poisson distribution. The fit of each model was assessed using the least-squares methodology described by Lang and Murray, with AIC_C calculated as described previously[108]. A lower AIC_C reflects a better fit given a penalty for increasing the number of parameters, and a negative ΔAIC_C ($\Delta AIC_C = AIC_C$ (one-parameter) – AIC_C (two parameter)) indicates the one parameter model is a better approximation of the data.

Preparation transposon mutant libraries

Phage stocks were prepared as described previously[153]. 200mL of both H37Rv and H37RvΔ3711c were grown to an OD of 0.70, and then pelleted at 4000RPM for 10 minutes at 37°C. Supernatants were discarded and cultures were resuspended in an equal volume of MP buffer (preheated to 37°C, 50mM Tris 7.5, 150mM NaCl, 10mM $MgSO_4$, 2mM $CaCl_2$).

Cultures were spun again, and again washed in an equal volume of MP buffer. Washing was again repeated. Samples were resuspended in 20mLs of MP buffer (preheated to 37°C). At this point, 0.5mL of culture was removed and plated on 7H9/OADC/tween/glycerol supplemented with 25µg/mL kanamycin to determine the background rate of kanamycin resistance. Cultures were then transduced with 6mL of preheated (37°C) ϕMycoMar-T7 (1e11pfu/mL), and incubated at 37°C for 12 hours. Samples were then immediately plated on 7H9/OADC/tween/glycerol supplemented with 25µg/mL kanamycin and incubated at 37°C for 30 days. Libraries were collected by scraping, and aliquots were prepared for both genomic DNA extraction and freezing.

Extraction of genomic DNA

Genomic DNA was extracted by chloroform/methanol/phenol extraction. Scrapped colonies were resuspended in 20mL TE (0.1M Tris and 1mM EDTA, pH 8.0), and centrifuged at 4000 RPM for 10 minutes at 4°C. Cultures were resuspended in 20mL TE and 20mL chloroform: methanol (2:1) and mixed thoroughly for 5 minutes. The cell suspension was then centrifuged at 4000 RPM for 10 minutes at 4°C, both the supernatant (aqueous phase) and organic layer were discarded, leaving only the cell pellet. The cell suspension was again centrifuged at 4000 RPM for 10 minutes at 4°C, and the remainder of the aqueous and organic layers was discarded. Each pellet was allowed to dry for 1 hour before being resuspended in 20mL TE supplemented with lysozyme (final concentration 100µg/mL). Cell suspensions were then incubated for 12 hours at 37°C. Following incubation, 2mL 10% SDS and proteinase K (final concentration 100µg/mL) were added to each suspension, which were then mixed vigorously by vortexing. The suspension was then transferred to new 50mL Falcon tubes (BD Biosciences) containing an equal volume phenol: chloroform (1:1, phenol buffered to pH 8.0). This suspension was again mixed

vigorously before incubating at room temperature for 1 hour. Suspensions were then centrifuged at 10,000RPM for 15 minutes at 4ºC. The aqueous layer was removed and transferred to microcentrifuge tubes and supplemented with ½ volume of chloroform. After mixing, the aqueous layer was again separated by centrifuging at 14,000RPM for 15 minutes at 4ºC and supplemented with RNase A to a final concentration of 25 µg/mL). This suspension was incubated 37ºC for one hour, and then cooled for 10 minutes at 4ºC. An equal volume of phenol: chloroform (1:1) was again and after mixing, suspensions were centrifuged at 14,000RPM for 15 minutes at 4ºC. ½ volume of chloroform was added to the aqueous layer, and suspensions were spun at 14,000RPM for 15 minutes at 4ºC. The aqueous layer was removed, and mixed with one volume of isopropanol and $1/10^{th}$ volume of sodium acetate (pH 5.2). Precipitated DNA was removed by spooling and washed with 70% Ethos before being air dried and resuspended in 200uL sterile distilled H_2O.

Generation of transposon junction genomic library

Two aliquots of 100µL of purified genomic DNA at 50ng/µL per triplicate was sheared in the Covaris (Woburn, MA) E220 focused ultrasonicator using the following settings: duty cycle of 5%, intensity of 3, cycles per burst of 200, and a duration of 90 seconds. Sheared DNA was pooled and purified using Qiagen QIAQuick columns. DNA ends were repaired using the NEB blunting kit (E1201L, New England Biolabs, Ipswich, MA) according to protocol. A-overhangs were polymerized on to the ends of genomic DNA fragments by incubation with 2mM dATP, Choice-Taq polymerase (Denville, Metuchen, NJ) and Taq PCR buffer at 72ºC for 45 minutes. Adapters were prepared by annealing the following sequences in 80µM $MgCl_2$:

5’ – TACCACGACCA-NH2 – 3’ (adapter 1.1)

5’ – ATGATGGCCGGTGGATTTGTG**NNANNANNN**TGGTCGTGGTAT – 3’ (adapter 2.1)

where N represents a random nucleotide selected during oligonucleotide synthesis. The annealing mixture was incubated at 95ºC for 4 minutes, with a subsequent ramp down to 20ºC with a slope of descent of 1ºC per minute. Following annealing, 0.8µL of adapter mix adapters were ligated to 1.2µg of A-tailed genomic DNA fragments by T4 DNA ligase (New England Biolabs, Ipswich, MA) at 16ºC for 12 hours. Transposon junctions were then amplified from ligated genomic DNA. The following transposon primers were mixed in equimolar ratio and combined with the adapter primer, where NNNNNNNN represents the sequence of multiplex index.

Sol-Mar:

5'AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGAT CTCGGGGACTTATCAGCCAACC-3'

Sol Mar 1b:

5'AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGAT CTTCGGGGACTTATCAGCCAACC-3'

Sol Mar 4b:

5'AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGAT CTGATACGGGGACTTATCAGCCAACC-3'

Sol Mar 5b:

5'AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGAT CTATCTACGGGGACTTATCAGCCAACC-3'

Adapter primer:

5'CAAGCAGAAGACGGCATACGAGATNNNNNNNNGTGACTGGAGTTCAGACGTGTG CTCTTCCGATCTATGATGGCCGGTGGATTTGTG-3'

Amplification was performed using Phusion Polymerase (New England Biolabs, Ipswich, MA) supplemented with GC Phusion buffer. Samples were amplified using the following cycle conditions: (1) 98ºC for 1 minute; (2) 98ºC for 10 seconds; (3) 59ºC for 30 seconds; (4) 72ºC for 30 seconds; (5) repeat 2-4 19 additional times; (6) 72ºC for 10 minutes. Following amplification, samples were size selected by excising the 200-400bp fragment from a 2% agarose gel, purified by Qiagen QIAquick columns and sent for sequencing at the Broad Institute using a Illumina Genome Analyzer II. Reads were mapped to the recently re-annotated genome H37Rv, H37Rv_BD (GenBank ID: CP003248.1).

Analysis of TraCS data

Valid, mappable reads corresponding to insertion at a TA site were tabulated for each triplicate for both samples. Read counts were normalized by dividing by the total number of observed for a given library divided by the minimum number of total reads observed for any library of a given biological comparison. For pair wise comparison of read counts between a control and experimental triplicate, the number of reads at each TA site was compared for a given gene by two-sample Wilcoxon rank sum (Matlab, Natick, MA). Each possible pair wise comparison was performed, giving a total of nine p-values. A composite p-value was determined by Fisher's method to given a final p-value comparing experimental and control samples. Fold-change was calculated by dividing the total number of reads in the experimental sample by the total number of reads in the control samples for each gene.

Author Contributions. Christopher B Ford designed the project, performed molecular studies, conducted the data analyses, prepared the figures and drafted the manuscript; Rupal R. Shah conducted preliminary molecular studies; Jason Zhang and Chris Sassetti assisted with statistical analysis; Sarah M. Fortune initiated the project, supervised analysis of the data, and drafted the manuscript.

Chapter 5 – Concluding Remarks

5.1 The mutation rate of *M. tuberculosis* during the course of infection

Here we have used WGS and the cynomolgus macaque model to show that mutation occurs at a constant rate across time, regardless of disease state. These findings have distinct consequences for the evolution of drug resistance, particularly in latent infection, which is treated with only a single antibiotic[51,78,154]. Recent evidence suggests that the spectrum of latent infection overlaps with that of active disease[57,63,64], suggesting that bacterial burden in latently infected individuals may be higher than previously appreciated. If true, there is potential for the development of INH resistance during latency. Indeed, meta-analyses have found associations between IPT and subsequent INH resistance[51,52]; however resistance remains rare – which is perhaps not surprising given the rarity of reactivation. However, this may not hold true for the application of IPT to HIV patients with latent or subclinical tuberculosis[78,79] where bacterial burdens are likely higher. As IPT is deployed in this population, it will be critical to simultaneously deploy drug susceptibility testing to monitor INH resistance.

Why does Tb mutate at a roughly equal rate per day, regardless of disease state and replication dynamics? The observed spectrum of mutations suggests that the principle driver of mutation *in vivo* may be DNA damage rather than replicative error[74]. It remains to be determined if DNA damage is incurred as a result of host factors or as a byproduct of metabolism.

5.2 The consequences of variation in mutation rate

Strains of the East Asian, Beijing sublineage have been clinically associated with drug resistance in several regions[81,88-93,97-100]. Our data indicate that this increased association is at least in part due to an increased basal level of mutation rate, as opposed to differences in target size or the ability to survive and mutate in the presence of drug. In order to interpret the impact

of these *in vitro* measurements on the *in vivo* evolution of drug resistance, we turned to our previously published work[21] and published sequence data from a clonal outbreak in Vancouver[20]. From these data we determined that the per day mutation rate *in vivo* is well approximated by the mutation rate calculated *in vitro*. Finally, we developed a model of drug resistance grounded in our observations of the mutation dynamics *in vivo* and our *in vitro* measures of drug resistance. From our model, we can draw two principle predictions relating to the clinical incidence of MDR: (1) with a per day dynamic of drug resistance, it is possible to see MDR emerge before the onset of treatment; (2) modest differences in mutation rate exist between strains, and these modest differences can translate to large differences in the probability of acquiring drug resistance. Additionally, it is striking to note that the mutation rate estimated from clinical isolates, from the cynomolgus macaque model, and *in vitro* data converge on a single per day mutation rate – approximately $4.0 * 10^{-10}$, or 0.58 mutations per genome per year. These data are suggestive of a molecular clock[54,155,156] for *M. tuberculosis* (Mtb), which may be of future use in the analysis of genetic diversity.

Together, these data suggest two critical observations about the biology of mutation in Mtb. First, the mutation rate does not vary in accordance with replication. Though the exact dynamics of Mtb growth in human infection are unknown, the replication rate of Mtb is thought to vary between *in vitro* culture, active disease, and latent infection[61,62]. Despite this, our observation of a consistent rate across conditions suggests that mutation is driven in large part by a time-based factor, such as DNA damage. Secondly, our data indicate that strains of the East Asian lineage are able to more quickly acquire drug resistance, likely due to an increase in basal mutation rate. The observation that mutation rate varies between strains may be reflective of differences in the molecular mechanisms governing mutation and fidelity. The mechanisms that

drive mutation and regulate fidelity are largely unknown in Mtb, though it is clear that they may differ in important ways from other systems.

5.3 Mycobacteria do not employ canonical mechanisms of fidelity

In the majority of model prokaryotic systems, replicative fidelity is regulated by three mechanisms: (1) polymerase base selection, (2) polymerase associated proofreading, and (3) and post-replicative mismatch repair (MMR). Mycobacteria lack homologs of the core post-replicative repair proteins, MutSHL. To determine if mycobacteria compensate for a lack of MMR through enhanced proofreading, we deleted two potential polIIIε subunits. While both contain a canonical ExoIIIε motif[131,157], deletion of either gene alone or in combination did not result in a significant change in mutation rate, suggesting neither is necessary for maintaining genomic fidelity.

While Msm4259/Rv2191 possesses additional domains suggesting it may be involved in NER[142], Msm6275/Rv3711c bears striking homology to *dnaQ*. To address the possibility that Rv3711c acts in a redundant proofreading pathway with an unidentified gene, we utilized a forward genetic screen to search for genes significantly underrepresented in the absence of Rv3711c. While our TraCS screen failed to identify genes potentially involved in a redundant pathway that would obscure a mutator phenotype, we did uncover a pair of significant fold changes. The first is an enhanced requirement for PhoPR and a pair of genes involved in the TreS trehalose metabolism pathway. While none of the members of the PhoPR regulon were significantly underrepresented, PhoP is known to positively regulate the production of methyl-branched fatty acid-containing acyltrehaloses. Taken together with the underrepresentation of *treS* and *mak*, these results suggest an important role for trehalose metabolism in the absence of Rv3711c. The second is a relaxed requirement for a trio of genes associated with DNA repair. It

is possible that these genes are involved in a potentially novel DNA repair pathway, one in which Rv3711c would act as the terminal exonuclease. If deletion of Rv3711c resulted in a toxic orphan repair intermediate, subsequent deletion of these upstream genes may halt the pathway prior to formation of toxic intermediates[152]. Whether Rv3711c participates in a novel repair pathway, and what role trehalose plays in tolerating the loss of Rv3711c remain open questions. However, from this work it is clear that Rv3711c, while annotated as *dnaQ*, does not serve a proofreading function in the cell, suggesting that mycobacteria do not use canonical mechanisms of proofreading.

In the absence of homologs of both canonical MMR and proofreading, how do mycobacteria maintain genomic fidelity? It is possible that mycobacteria, including Mtb, have co-opted alternative systems to serve as mediators of replicative and post-replicative fidelity. In work not addressed here, we have undertaken a forward genetic screen for mutator alleles in *M. smegmatis*, the results of which may shed light on these novel systems. Alternatively, the primary mycobacterial replicative polymerase (DnaE1) may possess intrinsic mechanisms of fidelity superior to homologous systems in model organisms. While the polymerase does not possess any domains reflective of 3'-5' exonuclease activity, it is possible that it has evolved mechanisms to improve base selection. Indeed, antimutator alleles of the homologous *dnaE* from *E. coli* have been identified, though their effect was only a 30-fold reduction in mutation[146,158]. In *E. coli*[144,145,159,160], the fidelity of polIIIα is at most $1x10^{-6}$. This suggests that if fidelity in mycobacteria is solely determined by base-selection, then the process must be at least 10,000 times more specific in mycobacteria than *E. coli.*

One proposed benefit of dissociating mechanisms of fidelity from the polymerase through both polIIIε and post-replicative MMR is the ability to suspend these processes in times

of stress[131,161,162]. This allows for replication to continue despite DNA damage, and results in the rapid creation of genetic diversity – potentially leading to adaptation. In the absence of both MMR and proofreading, it is possible that Mtb accomplishes this through use of DnaE2, an alternative polymerase with homology to both mycobacterial DnaE1 and DnaE from *E. coli*[67]. Indeed, DnaE2 is required for persistence in a mouse model of infection, suggesting it is involved in tolerating host-induced stress. While the molecular determinants of fidelity in mycobacteria remain unknown, it is clear that mycobacteria are highly divergent from model systems and further work may reveal novel insights into this core process.

5.4 The future evolution of drug resistance

Ultimately, this work is motivated by the occurrence of mutation leading to drug resistance in a clinical setting. The scope of the drug resistance epidemic is staggering – approximately 5% of cases are multidrug resistant, a percentage that climbs to 22% in some regions of the globe[14]. Given the strong association of drug resistant tuberculosis with mortality[7,83], reversing the drug resistance epidemic is essential to reducing the morbidity and mortality burden of the disease.

As reports of MDR, XDR, and TDR tuberculosis become more common, the need for new antibiotics is pressing. Fortunately, several new compounds are in late stage clinical development, and new regimens, driven in part by the Global Alliance for Tb Drug Development may offer a shortened course of therapy[163]. As new antibiotics and new regimens emerge, strategies to avoid a new generation of MDR, XDR, and TDR isolates must be pursued. While adherence to properly prescribed regimens will be essential, our work and the work of others[113] suggests that the mutational capacity of Mtb is sufficient to develop MDR before the onset of treatment. This is largely a result of the large bacterial burden present at diagnosis and treatment,

a factor determined by the limited sensitivity of diagnostics. In the face of substantial capacity for mutation and resistance, early and active case detection with novel, sensitive point of care diagnostics remains our best hope of curbing the drug resistance epidemic.

Bibliography

1. Daniel, T. ScienceDirect - Respiratory Medicine : The history of tuberculosis. *Respiratory medicine* (2006).

2. Donoghue, H. D. Insights gained from palaeomicrobiology into ancient and modern tuberculosis. *Clin. Microbiol. Infect.* **17**, 821–829 (2011).

3. Zink, A. R. *et al.* Characterization of Mycobacterium tuberculosis complex DNAs from Egyptian mummies by spoligotyping. *Journal of Clinical Microbiology* **41**, 359–367 (2003).

4. Udwadia, Z. F. MDR, XDR, TDR tuberculosis: ominous progression. *Thorax* **67**, 286–288 (2012).

5. Udwadia, Z. F., Amale, R. A., Ajbani, K. K. & Rodrigues, C. Totally Drug-Resistant Tuberculosis in India. *Clin Infect Dis* **54**, 579–581 (2012).

6. Velayati, A. A. *et al.* Emergence of new forms of totally drug-resistant tuberculosis bacilli: super extensively drug-resistant tuberculosis or totally drug-resistant strains in Iran. *Chest* **136**, 420–425 (2009).

7. Gandhi, N. R. *et al.* Extensively drug-resistant tuberculosis as a cause of death in patients co-infected with tuberculosis and HIV in a rural area of South Africa. *Lancet* **368**, 1575–1580 (2006).

8. Comstock, G. W., Baum, C. & Snider, D. E. Isoniazid prophylaxis among Alaskan Eskimos: a final report of the bethel isoniazid studies. *Am. Rev. Respir. Dis.* **119**, 827–830 (1979).

9. YOUMANS, G. P. & KARLSON, A. G. Streptomycin sensitivity of tubercle bacilli; studies on recently isolated tubercle bacilli and the development of resistance to streptomycin in vivo. *Am Rev Tuberc* **55**, 529–535 (1947).

10. DUNNER, E., BROWN, W. B. & WALLACE, J. The Effect of Streptomycin with Para-Amino Salicylic Acid on the Emergence of Resistant Strains of Tubercle Bacilli. *Chest* **16**, 661–666 (1949).

11. Combs, D. L., O'Brien, R. J. & Geiter, L. J. USPHS Tuberculosis Short-Course Chemotherapy Trial 21: effectiveness, toxicity, and acceptability. The report of final results. *Ann. Intern. Med.* **112**, 397–406 (1990).

12. Bloom, B. R. & Murray, C. J. Tuberculosis: commentary on a reemergent killer. *Science* **257**, 1055–1064 (1992).

13. Edlin, B. R., Tokars, J. I. & Grieco, M. H. An outbreak of multidrug-resistant tuberculosis among hospitalized patients with the acquired immunodeficiency syndrome. *New England journal …* (1992).

14. Wright, A., Zignol, M. & Global Project on Anti-tuberculosis Drug Resistance Surveillance, W. I. Anti-Tuberculosis Drug Resistance in the world, Report No. 4. (2008).

15. Lipsitch, M. & Levin, B. R. Population dynamics of tuberculosis treatment: mathematical models of the roles of non-compliance and bacterial heterogeneity in the evolution of drug resistance. *Int J Tuberc Lung Dis* **2**, 187–199 (1998).

16. Filliol, I. *et al.* Global phylogeny of Mycobacterium tuberculosis based on single nucleotide polymorphism (SNP) analysis: insights into tuberculosis evolution, phylogenetic accuracy of other DNA fingerprinting systems, and recommendations for a minimal standard SNP set. *J Bacteriol* **188**, 759–772 (2006).

17. Hershberg, R. *et al.* High Functional Diversity in Mycobacterium tuberculosis Driven by Genetic Drift and Human Demography. *Plos Biol* **6**, e311 (2008).

18. Ioerger, T. R. *et al.* Genome analysis of multi- and extensively-drug-resistant tuberculosis from KwaZulu-Natal, South Africa. *PLoS ONE* **4**, e7778 (2009).

19. Ioerger, T. R. *et al.* The non-clonality of drug resistance in Beijing-genotype isolates of Mycobacterium tuberculosis from the Western Cape of South Africa. *BMC Genomics* **11**, 670 (2010).

20. Gardy, J. L. *et al.* Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. *N Engl J Med* **364**, 730–739 (2011).

21. Ford, C. B. *et al.* Use of whole genome sequencing to estimate the mutation rate of Mycobacterium tuberculosis during latent infection. *Nat Genet* (2011).doi:10.1038/ng.811

22. Saunders, N. J. *et al.* Deep resequencing of serial sputum isolates of Mycobacterium tuberculosis during therapeutic failure due to poor compliance reveals stepwise mutation of key resistance genes on an otherwise stable genetic background. *J. Infect.* **62**, 212–217 (2011).

23. Rozo-Anaya, J. C. & Ribón, W. Molecular tools for Mycobacterium tuberculosis genotyping. *Rev Salud Publica (Bogota)* **12**, 510–521 (2010).

24. Niemann, S. *et al.* Genomic diversity among drug sensitive and multidrug resistant isolates of Mycobacterium tuberculosis with identical DNA fingerprints. *PLoS ONE* **4**, e7407 (2009).

25. Wilgenbusch, J. C. & Swofford, D. Inferring evolutionary trees with PAUP*. *Curr Protoc Bioinformatics* **Chapter 6**, Unit 6.4 (2003).

26. García de Viedma, D., Marín, M., Ruiz, M. J. & Bouza, E. Analysis of clonal composition of Mycobacterium tuberculosis isolates in primary infections in children. *Journal of Clinical Microbiology* **42**, 3415–3418 (2004).

27. Martín, A., Herránz, M., Serrano, M., Bouza, E. & de Viedma, D. Rapid clonal analysis of recurrent tuberculosis by direct MIRU-VNTR typing on stored isolates. *BMC Microbiol* **7**, 73 (2007).

28. Bandera, A. *et al.* Molecular epidemiology study of exogenous reinfection in an area with a low incidence of tuberculosis. *Journal of Clinical Microbiology* **39**, 2213–2218 (2001).

29. Caminero, J. A. *et al.* Epidemiological evidence of the spread of a Mycobacterium tuberculosis strain of the Beijing genotype on Gran Canaria Island. *Am J Respir Crit Care Med* **164**, 1165–1170 (2001).

30. Chaves, F., Dronda, F., Alonso-Sanz, M. & Noriega, A. R. Evidence of exogenous reinfection and mixed infection with more than one strain of Mycobacterium tuberculosis among Spanish HIV-infected inmates. *AIDS* **13**, 615–620 (1999).

31. García de Viedma, D., Marín, M., Ruiz Serrano, M. J., Alcalá, L. & Bouza, E. Polyclonal and compartmentalized infection by Mycobacterium tuberculosis in patients with both respiratory and extrarespiratory involvement. *J INFECT DIS* **187**, 695–699 (2003).

32. Nardell, E., McInnis, B., Thomas, B. & Weidhaas, S. Exogenous reinfection with tuberculosis in a shelter for the homeless. *N Engl J Med* **315**, 1570–1575 (1986).

33. Small, P. M. *et al.* Exogenous reinfection with multidrug-resistant Mycobacterium tuberculosis in patients with advanced HIV infection. *N Engl J Med* **328**, 1137–1144 (1993).

34. Stavrum, R. *et al.* High diversity of Mycobacterium tuberculosis genotypes in South Africa and preponderance of mixed infections among ST53 isolates. *Journal of Clinical Microbiology* **47**, 1848–1856 (2009).

35. Kaplan, G. *et al.* Mycobacterium tuberculosis growth at the cavity surface: a microenvironment with failed immunity. *Infect Immun* **71**, 7099–7108 (2003).

36. Turett, G. S. *et al.* Exogenous reinfection with multidrug-resistant Mycobacterium tuberculosis. *Clin Infect Dis* **24**, 513–514 (1997).

37. Horn, D. L. *et al.* Superinfection with rifampin-isoniazid-streptomycin-ethambutol (RISE)-resistant tuberculosis in three patients with AIDS: confirmation by polymerase chain reaction fingerprinting. *Ann. Intern. Med.* **121**, 115–116 (1994).

38. Geldmacher, C. *et al.* Preferential infection and depletion of Mycobacterium tuberculosis-specific CD4 T cells after HIV-1 infection. *J. Exp. Med.* **207**, 2869–2881 (2010).

39. Comas, I. *et al.* Human T cell epitopes of Mycobacterium tuberculosis are evolutionarily hyperconserved. *Nat Genet* **42**, 498–503 (2010).

40. Ernst, J. D. *et al.* Meeting Report: NIH Workshop on the Tuberculosis Immune Epitope Database. *Tuberculosis (Edinb)* **88**, 366–370 (2008).

41. Kato-Maeda, M., Bifani, P. J., Kreiswirth, B. N. & Small, P. M. The nature and consequence of genetic variability within Mycobacterium tuberculosis. *J. Clin. Invest.* **107**, 533–537 (2001).

42. Brosch, R. *et al.* A new evolutionary scenario for the Mycobacterium tuberculosis complex. *Proc Natl Acad Sci USA* **99**, 3684–3689 (2002).

43. Tsolaki, A. G. *et al.* Functional and evolutionary genomics of Mycobacterium tuberculosis: insights from genomic deletions in 100 strains. *Proc Natl Acad Sci USA* **101**, 4865–4870 (2004).

44. Uplekar, S., Heym, B., Friocourt, V., Rougemont, J. & Cole, S. T. Comparative genomics of esx genes from clinical isolates of Mycobacterium tuberculosis provides evidence for gene conversion and epitope variation. *Infect Immun* (2011).doi:10.1128/IAI.05344-11

45. Domenech, P., Kolly, G. S., Leon-Solis, L., Fallow, A. & Reed, M. B. Massive Gene Duplication Event among Clinical Isolates of the Mycobacterium tuberculosis W/Beijing Family. *J Bacteriol* **192**, 4562–4570 (2010).

46. Roberts, D. M., Liao, R. P., Wisedchaisri, G., Hol, W. G. J. & Sherman, D. R. Two sensor kinases contribute to the hypoxic response of Mycobacterium tuberculosis. *Journal of Biological Chemistry* **279**, 23082–23087 (2004).

47. Pevzner, P. A., Tang, H. & Waterman, M. S. An Eulerian path approach to DNA fragment assembly. *Proc Natl Acad Sci USA* **98**, 9748–9753 (2001).

48. Chiang, D. Y. *et al.* High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat Meth* **6**, 99–103 (2009).

49. Eisen, J. A., Heidelberg, J. F., White, O. & Salzberg, S. L. Evidence for symmetric chromosomal inversions around the replication origin in bacteria. *Genome Biol* **1**, RESEARCH0011 (2000).

50. Pop, M., Phillippy, A., Delcher, A. L. & Salzberg, S. L. Comparative genome assembly. *Brief. Bioinformatics* **5**, 237–248 (2004).

51. Balcells, M. E., Thomas, S. L., Godfrey-Faussett, P. & Grant, A. D. Isoniazid preventive therapy and risk for resistant tuberculosis. *Emerging Infect Dis* **12**, 744–751 (2006).

52. Cattamanchi, A. *et al.* Clinical characteristics and treatment outcomes of patients with isoniazid-monoresistant tuberculosis. *Clin Infect Dis* **48**, 179–185 (2009).

53. Denver, D. R., Morris, K., Lynch, M. & Thomas, W. K. High mutation rate and predominance of insertions in the Caenorhabditis elegans nuclear genome. *Nature* **430**, 679–682 (2004).

54. Haag-Liautard, C. *et al.* Direct estimation of per nucleotide and genomic deleterious mutation rates in Drosophila. *Nature* **445**, 82–85 (2007).

55. Lynch, M. *et al.* A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proceedings of the National Academy of Sciences* **105**, 9272–9277 (2008).

56. Capuano, S. V. *et al.* Experimental Mycobacterium tuberculosis infection of cynomolgus macaques closely resembles the various manifestations of human M. tuberculosis infection. *Infect Immun* **71**, 5831–5844 (2003).

57. Lin, P. L. *et al.* Quantitative comparison of active and latent tuberculosis in the cynomolgus macaque model. *Infect Immun* **77**, 4631–4642 (2009).

58. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* **18**, 1851–1858 (2008).

59. Hernandez, D., François, P., Farinelli, L., Osterås, M. & Schrenzel, J. De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer. *Genome Res* **18**, 802–809 (2008).

60. GUTIERREZ-VAZQUEZ, J. M. Studies on the rate of growth of mycobacteria. I. Generation time of Mycobacterium tuberculosis on several solid and liquid media and effects exerted by glycerol and malachite green. *Am Rev Tuberc* **74**, 50–58 (1956).

61. Gill, W. P. *et al.* A replication clock for Mycobacterium tuberculosis. *Nat Med* **15**, 211–214 (2009).

62. Muñoz-Elías, E. J. *et al.* Replication dynamics of Mycobacterium tuberculosis in chronically infected mice. *Infect Immun* **73**, 546–551 (2005).

63. Barry, C. E. *et al.* The spectrum of latent tuberculosis: rethinking the biology and intervention strategies. *Nat Rev Micro* **7**, 845–855 (2009).

64. Lin, P. L. & Flynn, J. L. Understanding latent tuberculosis: a moving target. *J Immunol* **185**, 15–22 (2010).

65. Sarkar, S., Ma, W. T. & Sandri, G. H. On fluctuation analysis: a new, simple and efficient method for computing the expected number of mutants. *Genetica* **85**, 173–179 (1992).

66. Lang, G. I. & Murray, A. W. Estimating the per-base-pair mutation rate in the yeast Saccharomyces cerevisiae. *Genetics* **178**, 67–82 (2008).

67. Boshoff, H. I. M., Reed, M. B., Barry, C. E. & Mizrahi, V. DnaE2 polymerase contributes to in vivo survival and the emergence of drug resistance in Mycobacterium tuberculosis. *Cell* **113**, 183–193 (2003).

68. Werngren, J. Drug-susceptible Mycobacterium tuberculosis Beijing genotype does not develop mutation-conferred resistance to rifampin at an elevated rate. *Journal of Clinical Microbiology* (2003).

69. Telenti, A. *et al.* Detection of rifampicin-resistance mutations in Mycobacterium tuberculosis. *The Lancet* **341**, 647–651 (1993).

70. Nathan, C. & Shiloh, M. U. Reactive oxygen and nitrogen intermediates in the relationship between mammalian hosts and microbial pathogens. *Proc Natl Acad Sci USA* **97**, 8841–8848 (2000).

71. Ng, V. H., Cox, J. S., Sousa, A. O., MacMicking, J. D. & McKinney, J. D. Role of KatG catalase-peroxidase in mycobacterial pathogenesis: countering the phagocyte oxidative burst. *Mol Microbiol* **52**, 1291–1302 (2004).

72. Sassetti, C. M. & Rubin, E. J. Genetic requirements for mycobacterial survival during infection. *Proc Natl Acad Sci USA* **100**, 12989–12994 (2003).

73. Lee, J., Remold, H. G., Ieong, M. H. & Kornfeld, H. Macrophage apoptosis in response to high intracellular burden of Mycobacterium tuberculosis is mediated by a novel caspase-independent pathway. *J Immunol* **176**, 4267–4274 (2006).

74. Boshoff, H. I. M., Durbach, S. I. & Mizrahi, V. DNA metabolism in mycobacterium tuberculosis: implications for drug resistance and strain variability. *Scand J Infect Dis* **33**, 101–105 (2001).

75. Fenhalls, G. *et al.* In situ detection of Mycobacterium tuberculosis transcripts in human lung granulomas reveals differential gene expression in necrotic lesions. *Infect Immun* **70**, 6330–6338 (2002).

76. Saint-Ruf, C., Pesut, J., Sopta, M. & Matic, I. Causes and Consequences of DNA Repair Activity Modulation During Stationary Phase inEscherichia coli. *Crit Rev Biochem Mol Biol* **42**, 259–270 (2007).

77. Bjedov, I. Stress-Induced Mutagenesis in Bacteria. *Science* **300**, 1404–1409 (2003).

78. Cohen, T., Lipsitch, M., Walensky, R. P. & Murray, M. B. Beneficial and perverse effects of isoniazid preventive therapy for latent tuberculosis infection in HIV-tuberculosis coinfected populations. *Proc Natl Acad Sci USA* **103**, 7042–7047 (2006).

79. Perriëns, J. H. *et al.* Increased mortality and tuberculosis treatment failure rate among human immunodeficiency virus (HIV) seropositive compared with HIV seronegative patients with pulmonary tuberculosis treated with 'standard' chemotherapy in Kinshasa, Zaire. *Am. Rev. Respir. Dis.* **144**, 750–755 (1991).

80. Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biol* **5**, R12 (2004).

81. Udwadia, Z. F., Amale, R. A., Ajbani, K. K. & Rodrigues, C. Totally Drug-Resistant Tuberculosis in India. *Clin Infect Dis* (2011).doi:10.1093/cid/cir889

82. Migliori, G. B., De Iaco, G., Besozzi, G., Centis, R. & Cirillo, D. M. First tuberculosis cases in Italy resistant to all tested drugs. *Euro Surveill.* **12**, E070517.1 (2007).

83. Gandhi, N. R. *et al.* Multidrug-resistant and extensively drug-resistant tuberculosis: a threat to global control of tuberculosis. *Lancet* **375**, 1830–1843 (2010).

84. David, H. L. Probability distribution of drug-resistant mutants in unselected populations of Mycobacterium tuberculosis. *Applied microbiology* **20**, 810–814 (1970).

85. Weis, S. E. *et al.* The effect of directly observed therapy on the rates of drug resistance and relapse in tuberculosis. *N Engl J Med* **330**, 1179–1184 (1994).

86. Bradford, W. Z. *et al.* The changing epidemiology of acquired drug-resistant tuberculosis in San Francisco, USA. *The Lancet* **348**, 928–931 (1996).

87. Goble, M. *et al.* Treatment of 171 patients with pulmonary tuberculosis resistant to isoniazid and rifampin. *N Engl J Med* **328**, 527–532 (1993).

88. Pablos-Méndez, A., Knirsch, C. A., Barr, R. G., Lerner, B. H. & Frieden, T. R. Nonadherence in tuberculosis treatment: predictors and consequences in New York City. *Am. J. Med.* **102**, 164–170 (1997).

89. Udwadia, Z. F., Pinto, L. M. & Uplekar, M. W. Tuberculosis management by private practitioners in Mumbai, India: has anything changed in two decades? *PLoS ONE* **5**, e12023 (2010).

90. Johnson, R. *et al.* Drug-resistant tuberculosis epidemic in the Western Cape driven by a virulent Beijing genotype strain. *Int J Tuberc Lung Dis* **14**, 119–121 (2010).

91. Streicher, E. M. *et al.* Genotypic and phenotypic characterization of drug-resistant Mycobacterium tuberculosis isolates from rural districts of the Western Cape Province of South Africa. *Journal of Clinical Microbiology* **42**, 891–894 (2004).

92. European Concerted Action on New Generation Genetic Markers and Techniques for the Epidemiology and Control of Tuberculosis Beijing/W genotype Mycobacterium tuberculosis and drug resistance. *Emerging Infect Dis* **12**, 736–743 (2006).

93. Glynn, J. R., Whiteley, J., Bifani, P. J., Kremer, K. & Van Soolingen, D. Worldwide occurrence of Beijing/W strains of Mycobacterium tuberculosis: a systematic review. *Emerging Infect Dis* **8**, 843–849 (2002).

94. Kato-Maeda, M. *et al.* Beijing Sublineages of Mycobacterium tuberculosis Differ in Pathogenicity in the Guinea Pig. *Clin. Vaccine Immunol.* **19**, 1227–1237 (2012).

95. Drobniewski, F. *et al.* Drug-resistant tuberculosis, clinical virulence, and the dominance of the Beijing strain family in Russia. *JAMA* **293**, 2726–2731 (2005).

96. Coscolla, M. & Gagneux, S. Does M. tuberculosis genomic diversity explain disease diversity? *Drug Discovery Today: Disease Mechanisms* **7**, e43–e59 (2010).

97. Pang, Y. *et al.* Spoligotyping and Drug Resistance Analysis of Mycobacterium tuberculosis Strains from National Survey in China. *PLoS ONE* **7**, e32976 (2012).

98. Vadwai, V., Shetty, A., Supply, P. & Rodrigues, C. Evaluation of 24-locus MIRU-VNTR in extrapulmonary specimens: Study from a tertiary centre in Mumbai. *Tuberculosis* 1–9 (2012).doi:10.1016/j.tube.2012.01.002

99. Huang, H. Y. *et al.* Mixed Infection with Beijing and Non-Beijing Strains and Drug Resistance Pattern of Mycobacterium tuberculosis. *Journal of Clinical Microbiology* **48**, 4474–4480 (2010).

100. Anh, D. D. *et al.* Mycobacterium tuberculosis Beijing genotype emerging in Vietnam. *Emerging Infect Dis* **6**, 302–305 (2000).

101. Mestre, O. *et al.* Phylogeny of Mycobacterium tuberculosis Beijing strains constructed from polymorphisms in genes involved in DNA replication, recombination and repair. *PLoS ONE* **6**, e16020 (2011).

102. de Steenwinkel, J. E. M. *et al.* Drug Susceptibility of Mycobacterium tuberculosisBeijing Genotype and Association with MDR TB. *Emerging Infect Dis* **18**, 660–663 (2012).

103. Luria, S. E. & Delbrück, M. Mutations of Bacteria from Virus Sensitivity to Virus Resistance. *Genetics* **28**, 491–511 (1943).

104. Rosche, W. A. & Foster, P. L. Determining mutation rates in bacterial populations. *Methods* **20**, 4–17 (2000).

105. Stewart, F. M. Fluctuation tests: how reliable are the estimates of mutation rates? *Genetics* **137**, 1139–1146 (1994).

106. Stewart, F. M., Gordon, D. M. & Levin, B. R. Fluctuation analysis: the probability distribution of the number of mutants under different conditions. *Genetics* **124**, 175–185 (1990).

107. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Automat. Contr.* **19**, 716–723 (1974).

108. Burnham, K. P. Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research* **33**, 261–304 (2004).

109. Sandgren, A. *et al.* Tuberculosis Drug Resistance Mutation Database. *Plos Med* **6**, e2 (2009).

110. Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. Bayesian Phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol* **29**, 1969–1973 (2012).

111. Drummond, A. J. & Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7**, 214 (2007).

112. Kimura, M. & Ota, T. On the rate of molecular evolution. *Journal of Molecular Evolution* **1**, 1–17 (1971).

113. Colijn, C., Cohen, T., Ganesh, A. & Murray, M. B. Spontaneous Emergence of Multiple Drug Resistance in Tuberculosis before and during Therapy. *PLoS ONE* **6**, e18327 (2011).

114. Hall, L. M. C. & Henderson-Begg, S. K. Hypermutable bacteria isolated from humans--a critical analysis. *Microbiology (Reading, Engl)* **152**, 2505–2514 (2006).

115. Blázquez, J. Hypermutation as a factor contributing to the acquisition of antimicrobial resistance. *Clin Infect Dis* **37**, 1201–1209 (2003).

116. Oliver, A., Cantón, R., Campo, P., Baquero, F. & Blázquez, J. High frequency of hypermutable Pseudomonas aeruginosa in cystic fibrosis lung infection. *Science* **288**, 1251–1254 (2000).

117. Mizrahi, V. & Andersen, S. J. DNA repair in Mycobacterium tuberculosis. What have we learnt from the genome sequence? *Mol Microbiol* **29**, 1331–1339 (1998).

118. Springer, B. *et al.* Lack of mismatch correction facilitates genome evolution in mycobacteria. *Mol Microbiol* **53**, 1601–1609 (2004).

119. Fallow, A., Domenech, P. & Reed, M. B. Strains of the East Asian (W/Beijing) lineage of Mycobacterium tuberculosis are DosS/DosT-DosR two-component regulatory system natural mutants. *J Bacteriol* **192**, 2228–2238 (2010).

120. Huet, G. *et al.* A lipid profile typifies the Beijing strains of Mycobacterium tuberculosis: identification of a mutation responsible for a modification of the structures of phthiocerol dimycocerosates and phenolic glycolipids. *Journal of Biological Chemistry* **284**, 27101–27113 (2009).

121. Cheng, S. J., Thibert, L., Sanchez, T., Heifets, L. & Zhang, Y. pncA mutations as a major mechanism of pyrazinamide resistance in Mycobacterium tuberculosis: spread of a monoresistant strain in Quebec, Canada. *Antimicrobial Agents and Chemotherapy* **44**, 528–532 (2000).

122. Scorpio, A. & Zhang, Y. Mutations in pncA, a gene encoding pyrazinamidase/nicotinamidase, cause resistance to the antituberculous drug pyrazinamide in tubercle bacillus. *Nat Med* **2**, 662–667 (1996).

123. Gelmanova, I. Y. *et al.* Barriers to successful tuberculosis treatment in Tomsk, Russian Federation: non-adherence, default and the acquisition of multidrug resistance. *Bull. World Health Organ.* **85**, 703–711 (2007).

124. Seung, K. J. *et al.* The effect of initial drug resistance on treatment response and acquired drug resistance during standardized short-course chemotherapy for tuberculosis. *Clin Infect Dis* **39**, 1321–1328 (2004).

125. Helb, D. *et al.* Rapid detection of Mycobacterium tuberculosis and rifampin resistance by use of on-demand, near-patient technology. *Journal of Clinical Microbiology* **48**, 229–237 (2010).

126. Weyer, K. Laboratory services in tuberculosis control. Part II: microscopy. *World Health Organization, Geneva, Switzerland* (1998).

127. Gagneux, S. The Competitive Cost of Antibiotic Resistance in Mycobacterium tuberculosis. *Science* **312**, 1944–1946 (2006).

128. David, H. L. Probability distribution of drug-resistant mutants in unselected populations of Mycobacterium tuberculosis. *Applied microbiology* **20**, 810–814 (1970).

129. LeClerc, J. E., Li, B., Payne, W. L. & Cebula, T. A. High mutation frequencies among Escherichia coli and Salmonella pathogens. *Science* **274**, 1208–1211 (1996).

130. Drake, J. W. The distribution of rates of spontaneous mutation over viruses, prokaryotes, and eukaryotes. *Ann. N. Y. Acad. Sci.* **870**, 100–107 (1999).

131. Hamdan, S. M., Carr, P. D., Brown, S. E., Ollis, D. L. & Dixon, N. E. Structural basis for proofreading during replication of the Escherichia coli chromosome. *Structure* **10**, 535–546 (2002).

132. Barnes, M. H., Spacciapoli, P., Li, D. H. & Brown, N. C. The 3′–5′ exonuclease site of DNA polymerase III from Gram-positive bacteria: definition of a novel motif structure. *Gene* **165**, 45–50 (1995).

133. Taft-Benz, S. A. & Schaaper, R. M. Mutational analysis of the 3‘-->5’ proofreading exonuclease of Escherichia coli DNA polymerase III. *Nucleic Acids Research* **26**, 4005–4011 (1998).

134. Griffin, J. E. *et al.* High-resolution phenotypic profiling defines genes essential for mycobacterial growth and cholesterol catabolism. *PLoS Pathog* **7**, e1002251 (2011).

135. van Opijnen, T., Bodi, K. L. & Camilli, A. Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. *Nat Meth* **6**, 767–772 (2009).

136. Akerley, B. J. A genome-scale analysis for identification of genes required for growth or survival of Haemophilusinfluenzae. *Proceedings of the National Academy of Sciences* **99**, 966–971 (2002).

137. Gerdes, S. Y. *et al.* Experimental determination and system level analysis of essential genes in Escherichia coli MG1655. *J Bacteriol* **185**, 5673–5684 (2003).

138. Sassetti, C. M., Boyd, D. H. & Rubin, E. J. Comprehensive identification of conditionally essential genes in mycobacteria. *Proc Natl Acad Sci USA* **98**, 12712–12717 (2001).

139. Moser, M., Holley, W., Chatterjee, A. & Mian, I. The proofreading domain of Escherichia coli DNA polymerase I and other DNA and/or RNA exonuclease domains. *Nucleic Acids Research* **25**, 5110 (1997).

140. Dunin-Horkawicz, S., Feder, M. & Bujnicki, J. M. Phylogenomic analysis of the GIY-YIG nuclease superfamily. *BMC Genomics* **7**, 98 (2006).

141. Van Houten, B., Eisen, J. A. & Hanawalt, P. C. A cut above: discovery of an alternative excision repair pathway in bacteria. *Proc Natl Acad Sci USA* **99**, 2581–2583 (2002).

142. Moolenaar, G. F., van Rossum-Fikkert, S., van Kesteren, M. & Goosen, N. Cho, a second endonuclease involved in Escherichia coli nucleotide excision repair. *Proceedings of the ...* (2002).

143. Fijalkowska, I. J. & Schaaper, R. M. Mutants in the Exo I motif of Escherichia coli dnaQ: defective proofreading and inviability due to error catastrophe. *Proc Natl Acad Sci USA* **93**, 2856–2861 (1996).

144. Degnen, G. E. & Cox, E. C. Conditional mutator gene in Escherichia coli: isolation, mapping, and effector studies. *J Bacteriol* **117**, 477–487 (1974).

145. Horiuchi, T., Maki, H. & Sekiguchi, M. A new conditional lethal mutator (dnaQ49) in Escherichia coli K12. *Mol Gen Genet* **163**, 277–283 (1978).

146. Fijalkowska, I. J. & Schaaper, R. M. Effects of Escherichia coli dnaE antimutator alleles in a proofreading-deficient mutD5 strain. *J Bacteriol* **177**, 5979–5986 (1995).

147. De Smet, K. A., Weston, A., Brown, I. N., Young, D. B. & Robertson, B. D. Three pathways for trehalose biosynthesis in mycobacteria. *Microbiology (Reading, Engl)* **146 (Pt 1)**, 199–208 (2000).

148. Pérez, E. *et al.* An essential role for phoP in Mycobacterium tuberculosis virulence. *Mol Microbiol* **41**, 179–187 (2001).

149. Gonzalo-Asensio, J. *et al.* PhoP: a missing piece in the intricate puzzle of Mycobacterium tuberculosis virulence. *PLoS ONE* **3**, e3496 (2008).

150. Vultos, Dos, T., Blázquez, J., Rauzier, J., Matic, I. & Gicquel, B. Identification of Nudix hydrolase family members with an antimutator role in Mycobacterium tuberculosis and Mycobacterium smegmatis. *J Bacteriol* **188**, 3159–3161 (2006).

151. Gonzalo-Asensio, J. *et al.* The virulence-associated two-component PhoP-PhoR system controls the biosynthesis of polyketide-derived lipids in Mycobacterium tuberculosis. *J Biol Chem* **281**, 1313–1316 (2006).

152. Richardson, A. R. *et al.* The Base Excision Repair system of Salmonella enterica serovar typhimurium counteracts DNA damage by host nitric oxide. *PLoS Pathog* **5**, e1000451 (2009).

153. Bardarov, S. *et al.* Conditionally replicating mycobacteriophages: a system for transposon delivery to Mycobacterium tuberculosis. *Proc Natl Acad Sci USA* **94**, 10961–10966 (1997).

154. Aït-Khaled, N. *et al.* Isoniazid preventive therapy for people living with HIV: public health challenges and implementation issues. *Int J Tuberc Lung Dis* **13**, 927–935 (2009).

155. Bromham, L. & Penny, D. The modern molecular clock. *Nat Rev Genet* **4**, 216–224 (2003).

156. Korber, B. Timing the Ancestor of the HIV-1 Pandemic Strains. *Science* **288**, 1789–1796 (2000).

157. Patel, P. H., Suzuki, M., Adman, E., Shinkai, A. & Loeb, L. A. Prokaryotic DNA polymerase I: evolution, structure, and 'base flipping' mechanism for nucleotide selection1. *J Mol Biol* **308**, 823–837 (2001).

158. Schaaper, R. M. Antimutator mutants in bacteriophage T4 and Escherichia coli. *Genetics* **148**, 1579–1585 (1998).

159. Pham, P. T., Olson, M. W., McHenry, C. S. & Schaaper, R. M. The base substitution and frameshift fidelity of Escherichia coli DNA polymerase III holoenzyme in vitro. *J Biol Chem* **273**, 23575–23584 (1998).

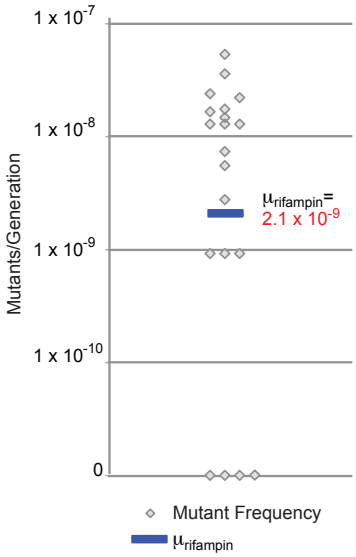
160. Heinrich, M. & Krauss, G. The fidelity of DNA polymerases during in vitro replication of a template containing 5-bromouracil at a specific site. *J Biol Chem* **264**, 119–123 (1989).

161. Studwell, P. S. & O'Donnell, M. Processive replication is contingent on the exonuclease subunit of DNA polymerase III holoenzyme. *J Biol Chem* **265**, 1171–1178 (1990).

162. Sutton, M. D., Murli, S., Opperman, T., Klein, C. & Walker, G. C. umuDC-dnaQ Interaction and its implications for cell cycle regulation and SOS mutagenesis in Escherichia coli. *J Bacteriol* **183**, 1085–1089 (2001).

163. Diacon, A. H. *et al.* 14-day bactericidal activity of PA-824, bedaquiline, pyrazinamide, and moxifloxacin combinations: a randomised trial. *Lancet* (2012).doi:10.1016/S0140-6736(12)61080-0

a



b

CFU	# of plates
0	4
1	3
3	1
6	1
8	1
14	3
16	1
18	1
19	1
24	1
26	1
39	1
58	1
Number of Cultures	20
MSS m value	2.27
Cell Count (N_t)	1.08×10^9
$\mu_{rifampin}$	2.1×10^{-9}
$\mu_{rifampicin}$ 95% CI	$1.01\text{-}3.30 \times 10^{-9}$

c

H37Rv Coordinate	Mutation	Amino Acid Change
761101	C>A	Q513E
761128	C>T	S522L
761128	C>G	S522W
761139	C>G	H526D
761139	C>T	H526Y
761140	A>G	H526R
761140	A>C	H526P
761149	G>T	R529L
761155	C>T	S531L
761155	C>G	S531W
Rifampin target size (bp)		10
$\mu_{in\ vitro}$		2.1×10^{-10}
$\mu_{in\ vitro}$ 95% CI		$1.01\text{-}3.30 \times 10^{-10}$

Supplementary Figure 2.1 The per base mutation rate of Mtb *in vitro*. (a,b) Luria and Delbrück fluctuation analysis was used to determine the rate of resistance to rifampicin. 20 independent cultures containing 1.08×10^9 cells each were plated and the resistance frequency was determined for each. The rate of resistance, $\mu_{rifampicin}$, was determined using the MSS method to calculate $m_{rifampicin}$, the representative number of mutations per culture. These data are representative of 4 biologically independent experiments. **(c)** The number of mutations conferring rifampicin resistance in our assay was determined using Sanger sequencing. Sequencing rifampicin resistant isolates from 96 independent cultures identified ten unique mutations. The amino acid changes represent the standard codon annotation used in *E. coli*. The per base mutation rate, $\mu_{in\ vitro}$, was determined by dividing $\mu_{rifampicin}$ by the target size.

Supplementary Table 2.1 Coverage and read depth for each sequenced isolate.

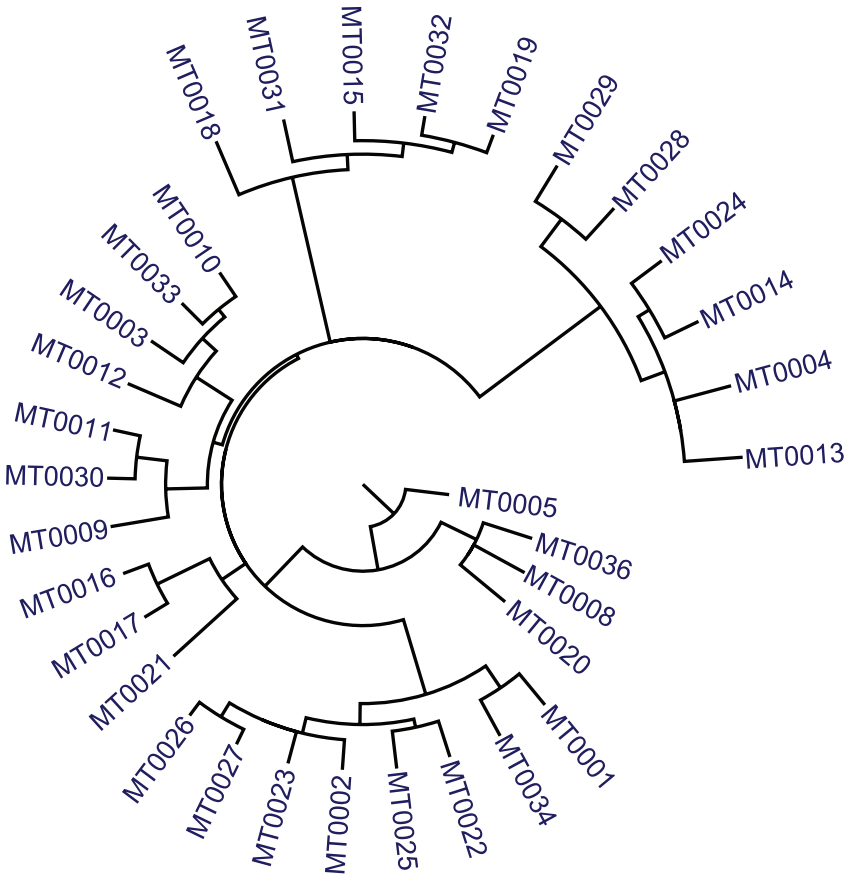
Animal & Isolate	Run Identifier[a]	Read Length (bp)[b]	Average Read Depth	Percent Coverage[c]
A – 2	15304-1B	75	278	96
A – 3	15304-3A	75	25	86
A – 7	15304_2A	36	15	61
A – 8	15304_4B	36	28	93
B – 1	7404-1	75	141	95
C – 1	7904-1	75	284	96
C – 2	7904-2	75	132	94
C – 3	7904-3	36	36	94
C – 3	7904-5	51	62	99
D – 4	11208-E4	51	15	97
D – 5	11208-E5	51	47	99
D – 7A	11208-E7	51	43	99
D – 3	11208-J3	51	114	99
D – 6	11208-J6	51	76	99
D – 7B	11208-J7	51	18	97
E – 2	7604-2	75	155	95
E – 4	7604-4	51	63	99
F – 1	8104-1C	75	148	89
F – 2	8104-2A	75	46	66
F – 4	8104_3B	36	22	91
G – 1	6404-1A	75	86	94
G – 2	6404-1B	75	183	95
G – 6	6404-3B	75	163	94
H-1	11105-1	51	62	99
H – 2	11105-2	75	227	95
H – 3	11105-3	75	208	96
I – 1	10403-1	75	228	95
I – 4	10403-4	75	121	90
I – 7	10403-7	75	224	94
I – 8	10403-8	75	205	95
I – 9	10403-9	51	80	98
I – 10	10403-10	75	189	96
I – 11	10403-11	75	207	96
Inoculum	JF-Erdman	36	38	94

Inoculum JF was sequenced four times, with the values shown representing total read depth from the four runs and total coverage. [a]Run Identifiers are based on the animal and strain identifiers used by Lin et al. [b]75bp reads were produced as paired-end reads by the Broad Institute of MIT and Harvard. Reads were subsequently trimmed to 48bp and pairing data was not used in assembly. 51 bp reads were produced by the Sacchettini lab at Texas A&M and were analyzed as paired end reads. 36bp reads were produced by Partners Healthcare Center for Personalized Genetic Medicine, Harvard Medical School. [c]Percent coverage values (defined as the fraction of sites covered in a genome) vary based on the method of assembly used. For 51bp reads, repetitive sequences were mapped using paired-end data to disambiguate the location of reads that map to multiple locations in the genome, allowing for a higher percent coverage value. A read depth of five was required to call a SNP. For 75 and 36 bp reads, repetitive sequences that could map to multiple locations were discarded resulting in a lower percent coverage for these isolates and a read depth of ten was required to call a SNP.

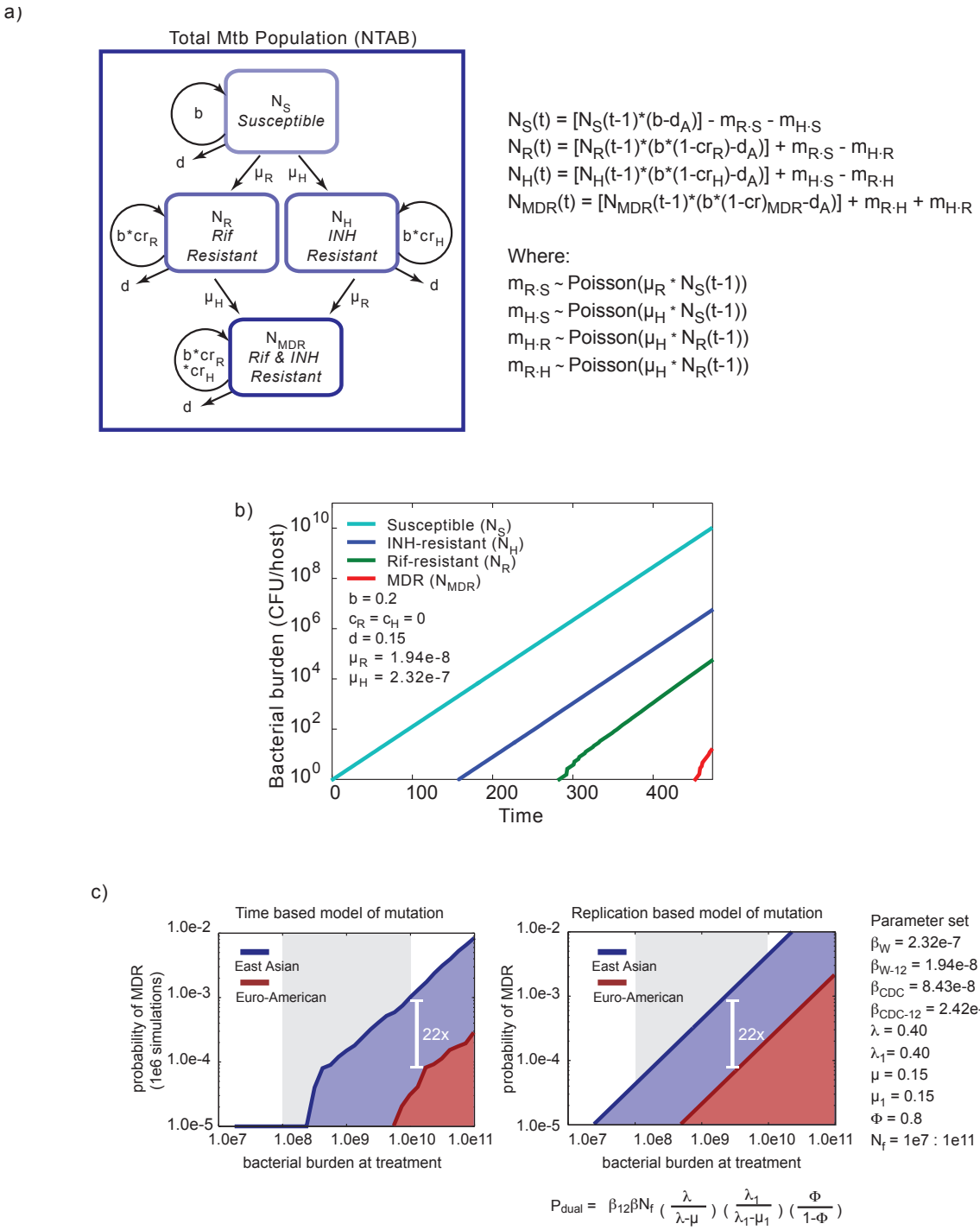
Supplementary Table 2.2 Primers used in PCR/Sequencing of validated SNPs.

H37Rv Coordinate	SNP	Forward Primer	Reverse Primer
635497	C>T	GACGTCGTCACTACCAGGG	CTCGGTGACTTCGACCAGAT
690264	G>C	GCGATGGTGGTCAATCTGCTGCC	TGCTACCTCCCGTCCCGTCAG
693453	C>T	TGGTGGTCCTTGGTTGGTAT	TAGTCGTGGTGATCGTCTGC
766229	G>A	GACCCGTACATCGAAACCTC	AGACCACCGGTGATGTCCTC
975906	T>C	GCAAGTACATCCGAGAACCC	TTCCAATACTGCCGGAAGAC
1256717	G>A	GCGATCAGCTATCTCGGTG	GTAGATGTCGGATTGGGTCG
1854208	G>T	ACCCATACAACGGCAAGTGT	CGAGCGCTCTCATACAGACA
1861203	G>A	GAGGTTTCTCCCACCCTTACCGAC	CTCGGGACACGTTGCGCAC
2448250	G>A	CCCTCTAGGCTTGACGACAG	CGAGGTCTCGTAGGTCGGTA
3655598	G>T	GTGTCGACAAGCTGCATCAC	ACGATGGTGATGGCGTAGAT
4346906	C>A	GTACATGTTGATGATGCCGC	ACATATGACTGACCGGCTCC

The following polymorphisms were identified multiple times by WGS and were not subjected to PCR resequencing: 682043, 2350697, 4183984.



Supplementary Figure 3.1 Phylogenetic analysis of clinical isolates Phylogenetic trees were constructed based on Bayesian MCMC analysis.



Supplementary Figure 3.2 Model structure and development (a) Model structure described both graphically and mathematically. **(b)** Simulation of bacterial growth and drug resistance within a simulated patient in which multidrug resistance evolved. **(c)** Comparison of time based model of mutation where probability of resistance is determined by simulation and a replication based model of resistance where probability is determined by the shown derived equation.

Supplementary Table 3.1 Rifampicin fluctuation analysis data

Strain	Lineage	Drug (μg/μL)	Total Cultures	m	Cell count	Mutation rate	95% Confidence Interval
CDC-1551	Euro-American	Rif 2	20	2.153	8.9e8	2.42e-9	1.14-3.82e-9
Erdman	Euro-American	Rif 2	20	2.274	8.64e8	2.63e-9	1.26-4.13e-9
H37Rv	Euro-American	Rif 2	20	4.337	8.73e8	4.97e-9	2.94-7.08e-9
X005581	Euro-American	Rif 2	20	5.830	7.7e8	7.57e-9	4.09-8.58e-9
X000630	Euro-American	Rif 2	20	4.815	8.50e8	5.67e-9	5.2-11.91e-9
HN878	East Asian	Rif 2	20	15.809	1.17e9	1.35e-8	1.095-1.605e-8
X005632	East Asian	Rif 2	20	7.898	8.8e7	8.97e-8	6.34-11.67e-8
X005631	East Asian	Rif 2	20	3.225	1.37e8	2.36e-8	1.26-3.519e-8
X005621	East Asian	Rif 2	20	5.523	2.04e8	2.71e-8	1.73-3.715e-8
CDC-1551	Euro-American	Rif 0.5	20	5.100	6.08e8	8.40e-9	5.24- 11.70e-9
CDC-1551	Euro-American	Rif 2	20	2.919	6.08e8	4.81e-9	2.48-7.28e-9
CDC-1551	Euro-American	Rif 5	17	3.171	6.08e8	5.22e-9	2.73-7.88e-9
HN878	East Asian	Rif 0.5	16	16.998	9.38e8	1.81e-8	1.49-2.14e-8
HN878	East Asian	Rif 2	19	18.223	9.38e8	1.94e-8	1.61-2.28e-8
HN878	East Asian	Rif 5	20	9.311	9.38e8	9.39e-9	7.29-11.26e-9
CDC-1551	Euro-American	INH 1	20	145.319	1.72e9	8.43e-8	8.14-8.728e-8
HN878	East Asian	INH 1	18	93.360	4.03e8	2.32e-7	2.19-2.43e-7
CDC-1551	Euro-American	ETH 5	20	19.602	3.93e8	4.98e-8	4.18-5.80e-8
HN878	East Asian	ETH 5	20	20.020	1.60e8	1.25e-7	1.05-1.45e-7

Supplementary Table 3.2 *rpoB* mutations

E. coli RpoB Coord. (amino acid)	Mtb RpoB Coord. (amino acid)	**Mtb *rpoB* coord. (nucleotide)**	**CDC -1551**			**HN878**		
			Drug Concentration (μg/mL)					
			0.5	*2*	*5*	*0.5*	*2*	*5*
Q513E	Q438E	**C1312G**	-	-	-	2	1	-
Q513L	Q438L	**A1313T**	-	-	-	2	-	-
D516V	D441V	**A1322T**	1	-	-	3	-	-
N519K	N444K	**C1332G**	-	-	-	1	-	-
S522L	S447L	**C1340T**	7	7	-	4	13	2
S522W	S447W	**C1340G**	4	2	5	4	3	3
H526D	H451D	**C1351G**	14	9	4	9	3	7
H526Y	H451Y	**C1351T**	10	12	12	10	15	14
H526R	H451R	**A1352G**	7	7	13	15	10	14
H526P	H451P	**A1352C****	1	4	5	1	3	5
H526L	H451L	**A1352T**	2	-	-	1	-	-
S531L	S456L	**C1367T**	11	8	5	7	16	18
S531W	S456W	**C1367G**	1	2	1	1	2	-
Total Target Size			**10**	**8**	**7**	**13**	**9**	**7**

** A1352C represents both single mutations found at this single site (quantity per strain : 1,2,-,1,1,1 respectively), as well as three clustered mutations: C1350G, A1352C, A1354C (quantity per strain: -, 2,5,-,2,4). These three mutations were found together, and the first of the three is a silent mutation.

Supplementary Table 3.3 Isoniazid and Ethambutol fluctuation analysis data

Strain	Drug Conc. (μg/mL)	Mutation Rate (mut./bp/gen.)	95% CI	Wilcoxon Rank Sum p-value
CDC-1551	0.5	8.40e-10	5.24e-10 – 1.17e-9	4.2866e-06
	2	6.01e-10	3.10e-10 – 9.10e-10	
	5	7.46e-10	3.91e-10 – 1.13e-9	1.8461e-05
HN878	0.5	1.39e-09	1.14e-9 - 1.65e-9	
	2	2.16e-09	1.79e-9 – 2.53e-9	3.8210e-04
	5	1.42e-09	1.04e-9 – 1.80e-9	

Supplementary Table 3.4 Estimates of *in vivo* mutation rate

Sample	Mutations (m)	Number of genomes	Average time per genome (days)	Mutation rate	95% CI
Human isolates*	n/a	32	n/a	2.21e-10	1.90 – 2.52 e-10
Active Disease, cynomolgus macaque**	4	15	261.85	4.45e-10	4.01 - 4.88 e-10
Latent infection, cynomolgus macaque**	3	8	293.52	2.55e-10	6.59 – 93.3 e-10
Reactivated Disease, cynomolgus macaque**	7	10	488.69	3.19e-10	1.44 - 7.38 e-10
All disease states, cynomolgus macaque**	14	33	338.31	3.14e-10	1.71 - 5.26 e-10
In vitro*, Erdman*	10	n/a	n/a	3.16e-10	1.51 - 4.95 e-10

* Estimated using BEAST v1.7.2

** Previously published, see reference (6).

Supplementary Table 3.5 Mathematical model parameter values

Variable	Descriptor	Figure 7a	Figure 7b	Figure 7c
$N_S(t)$	Susceptible bacterial population size at time (t)	Variable	Variable	Variable
$N_R(t)$	Rifampicin Resistant bacterial population size at time (t)	Variable	Variable	Variable
$N_H(t)$	Isoniazid Resistant bacterial population size at time (t)	Variable	Variable	Variable
$N_{MDR}(t)$	Multidrug resistant bacterial population size at time (t)	Variable	Variable	Variable
b	Bacterial growth rate, replications per day	0.40	0.20 : 1.10, increments of 0.10	0.40
d_A	Bacterial death rate	0.15	0.15	0.15
$\mu_{R\text{-}CDC}$	Rifampicin resistance rate, CDC-1551	2.42e-9	2.42e-9	2.42e-9
$\mu_{H\text{-}CDC}$	Isoniazid resistance rate, CDC-1551	2.32e-7	2.32e-7	2.32e-7
$\mu_{R\text{-}W}$	Rifampicin resistance rate, HN878	1.94e-8	1.94e-8	1.94e-8
$\mu_{H\text{-}W}$	Isoniazid resistance rate, HN878	2.32e-7	2.32e-7	2.32e-7
cr_H	Fitness cost of resistance, isoniazid resistance	0	0	0 : 0.9, increments of 0.10
cr_R	Fitness cost of resistance, rifampicin	0	0	0 : 0.9, increments of 0.10
cr_{MDR}	Fitness cost of resistance, multidrug resistance	cr_H*cr_R	cr_H*cr_R	cr_H*cr_R
Number of runs	Number of simulated patients	200,000 (100,000 per strain)	2,000,000 (100,000 per value of b, per strain)	2,000,000 (100,000 per value of cr, per strain)

Supplementary Table 4.1 Fluctuation analysis data

Species, Strain	Drug (µg/mL	Total cultures	m	Cell count	Mutation rate	95% Confidence Interval
Mtb H37Rv	Rif 2	20	5.86	2.81e8	2.10e-8	1.37-2.86e-8
Mtb H37RvΔRv3711c	Rif 2	20	7.91	3.50e8	2.83e-8	2.00-3.69e-8
Msm $Mc^2$155	Rif 200	20	1.22	3.00e8	4.07e-9	1.97-6.71e-9
Msm $Mc^2$155 ΔMsm4259	Rif 200	20	1.60	2.00e8	8.01e-9	3.65 – 13.0e-9
Msm $Mc^2$155 ΔMsm6275	Rif 200	20	3.13	6.00e8	5.21e-9	2.75-7.81e-9
Msm $Mc^2$155 ΔMsm4259 ΔMsm6275	Rif 200	20	2.12	3.00e8	7.05e-9	3.32-11.2e-9

Supplementary Table 4.2 Genes significantly two-fold above or below H37Rv in H37RvΔRv3711c, sorted by ratio.

Gene	p-value	Read Count, H37Rv	Read Count, H37Rv ΔRv3711c	Ratio	Function
Rv3711c	1.33E-15	2468.20	52.77	0.0214	exonuclease, DNA polymerase III, epsilon subunit
Rv0757	9.25E-06	118.62	2.87	0.0242	two-component system, OmpR family, response regulator
Rv0127	5.92E-06	20.44	0.73	0.0358	maltose kinase, mak
Rv1592c	1.05E-09	372.28	22.64	0.0608	triacylglycerol lipase
Rv1845c	3.73E-06	584.23	90.19	0.1544	sensor transducer, blaR
Rv0126	8.38E-06	50.25	9.46	0.1883	trehalose synthase, converts maltose to trehalose
Rv2164c	0	16.99	5.46	0.3214	unknown conserved hypothetical,proline rich membrane protein
Rv1203c	3.24E-06	397.25	132.71	0.3341	unknown conserved hypothetical
Rv3587c	0	1.89	0.73	0.3870	hypothetical membrane protein, possible fibronectin attachment protein
Rv2624c	0	14.89	389.64	26.1680	universal stress protein
Rv2026c	1.80E-06	44.08	183.27	4.1577	universal stress protein
Rv2944	1.73E-10	284.02	919.60	3.2378	transposase
Rv2810c	9.57E-06	626.49	1324.30	2.1139	transposase
Rv1377c	2.48E-10	329.84	909.86	2.7585	transferase
Rv2274c	1.11E-16	18.65	488.55	26.1970	toxin
Rv2764c	0	2.06	5.46	2.6530	thymidylate synthase
Rv2832c	4.05E-11	5.81	286.00	49.2060	sn-glycerol 3-phosphate transporter ATP-binding protein
Rv2253	8.41E-06	551.49	1380.10	2.5025	secreted protein
Rv1003	2.06E-08	283.87	603.72	2.1268	ribosomal RNA small subunit methyltransferase I
Rv2097c	0	0.94	2.92	3.0963	pup-protein ligase
Rv1574	2.48E-08	168.82	340.20	2.0152	phiRv1 phage protein
Rv2383c	0	1.39	74.35	53.6710	phenyloxazoline synthase MbtB
Rv1089	1.24E-08	69.18	410.36	5.9314	PE family protein
Rv3777	3.22E-06	466.76	1083.90	2.3222	oxidoreductase
Rv3389c	8.59E-06	485.74	1247.00	2.5672	oxidoreductase
Rv2322c	9.31E-10	4.87	372.76	76.5700	ornithine-oxo-acid transaminase
Rv0899	4.80E-06	304.30	626.58	2.0591	OOP family OmpA-OmpF porin
Rv0781	0	366.07	1029.90	2.8133	oligopeptidase B
Rv1912c	7.43E-14	562.81	2004.30	3.5613	NADPH2:quinone reductase
Rv1160	8.34E-10	18.88	450.44	23.8570	mutator mutT protein, mutT2
Rv0187	3.42E-06	270.24	581.47	2.1517	methyltransferase
Rv3342	4.99E-06	168.69	408.12	2.4194	methyltransferase
Rv3204	1.01E-05	4.98	115.83	23.2400	methyltransferase

Supplementary Table 4.2, continued

Gene	p-value	Read Count, H37Rv	Read Count, H37Rv ΔRv3711c	Ratio	Function
Rv0924c	2.34E-11	4251.80	10858.00	2.5537	metal ion (Mn2+/Fe2+) transporter (Nramp) family metal ion transporter
Rv0621	3.40E-06	547.70	1514.60	2.7653	membrane protein
Rv3282	1.18E-05	96.76	223.43	2.3092	maf-like protein
Rv2518c	0	1.35	3.71	2.7577	lipoprotein lppS
Rv2171	3.83E-08	77.20	507.01	6.5678	lipoprotein lppM
Rv1400c	1.84E-10	126.04	497.53	3.9476	lipase lipH
Rv2213	5.66E-06	360.11	924.86	2.5683	leucyl aminopeptidase
Rv1730c	0	320.53	985.85	3.0757	hypothetical protein
Rv1006	0	2832.00	6525.30	2.3041	hypothetical protein
Rv3643	1.24E-13	120.51	484.35	4.0191	hypothetical protein
Rv2901c	9.56E-12	130.69	489.67	3.7469	hypothetical protein
Rv1269c	8.84E-11	127.68	706.00	5.5293	hypothetical protein
Rv2309A	2.03E-10	487.82	1881.70	3.8573	hypothetical protein
Rv1551A	3.17E-08	27.49	452.74	16.4710	hypothetical protein
Rv2669	3.81E-08	76.16	235.41	3.0910	hypothetical protein
Rv2731	3.98E-08	1043.20	2637.60	2.5284	hypothetical protein
Rv1669	4.78E-08	233.01	1191.40	5.1131	hypothetical protein
Rv3562	6.02E-08	61.26	397.44	6.4878	hypothetical protein
Rv3658c	1.51E-07	12.00	247.76	20.6430	hypothetical protein
Rv2728c	2.98E-07	888.56	2071.10	2.3309	hypothetical protein
Rv3693	7.10E-07	672.29	1358.90	2.0212	hypothetical protein
Rv2227	7.32E-07	533.62	1332.90	2.4978	hypothetical protein
Rv1929c	9.07E-07	363.49	862.91	2.3740	hypothetical protein
Rv0310c	1.04E-06	276.37	823.16	2.9785	hypothetical protein
Rv1154c	1.25E-06	1033.80	2509.10	2.4270	hypothetical protein
Rv0312	2.39E-06	237.15	615.20	2.5941	hypothetical protein
Rv1719A	3.11E-06	25.52	361.57	14.1710	hypothetical protein
Rv0912Ac	4.28E-06	137.71	337.29	2.4493	hypothetical protein
Rv3657c	7.23E-06	50.71	161.20	3.1787	hypothetical protein
Rv0007	9.51E-06	224.37	587.96	2.6205	hypothetical protein
Rv3078	3.36E-06	69.14	152.91	2.2115	hydroxylaminobenzene mutase hab
Rv2715	1.12E-12	772.53	1903.60	2.4641	hydrolase
Rv1694	9.78E-06	149.09	802.13	5.3803	hemolysin TlyA family protein
Rv2031c	1.66E-08	509.61	1296.40	2.5439	heat shock protein hspX
Rv1925	2.03E-07	113000.00	226230.00	2.0020	fatty-acyl-CoA synthase
Rv1782	0	0.67	3.43	5.0900	ESX-5 secretion system protein eccB5
Rv1426c	8.56E-14	857.33	1739.70	2.0292	esterase lipO
Rv1142c	9.66E-06	579.06	1632.90	2.8199	enoyl-CoA hydratase
Rv0905	1.23E-05	432.13	1284.20	2.9719	enoyl-CoA hydratase

Supplementary Table 4.2, continued

Gene	p-value	Read Count, H37Rv	Read Count, H37Rv ΔRv3711c	Ratio	Function
Rv2748c	0	12.26	37.83	3.0864	DNA translocase ftsK
Rv0794c	1.34E-05	580.81	1393.00	2.3984	dihydrolipoamide dehydrogenase
Rv0695	3.77E-10	87.90	683.07	7.7708	creatinine amidohydrolase, mycofactocin system protein
Rv2583c	5.22E-09	2569.20	9846.30	3.8324	bifunctional (p)ppGpp synthase/hydrolase relA
Rv2262c	1.80E-11	792.13	1968.90	2.4855	apolipoprotein N-acyltransferase
Rv2638	3.97E-07	140.76	775.50	5.5094	anti-anti-sigma factor
Rv2380c	0	6.51	174.21	26.7670	amino acid adenylation domain-containing protein
Rv2379c	1.33E-10	44.26	88.91	2.0088	amino acid adenylation domain-containing protein
Rv2251	1.70E-07	263.38	1089.00	4.1348	alkyldihydroxyacetonephosphate synthase
Rv3501c	2.65E-08	361.11	791.10	2.1908	ABC transporter permease
Rv3569c	1.15E-08	241.92	512.13	2.1170	4,5-9,10-diseco-3-hydroxy-5,9,17-trioxoandrosta-1(10),2-diene-4-oate hydrolase
Rv0812	7.23E-14	10.51	72.81	6.9280	4-amino-4-deoxychorismate lyase
Rv2002	2.85E-11	246.27	836.00	3.3946	3-alpha-(or 20-beta)-hydroxysteroid dehydrogenase
Rv2918c	0	766.45	2629.90	3.4313	[protein-PII] uridylyltransferase
Rv2384	0	4.04	14.85	3.6751	(2,3-dihydroxybenzoyl)adenylate synthase

Supplementary Table 4.3 – Genes of the PhoPR regulon, sorted by ratio.

Gene	Z-Score*	p-value	Read Count, H37Rv	Read Count, H37Rv ΔRv3711c	Ratio
Rv1180	2.08	0.00E+00	0.00	0.00	0.0000
Rv2524	2.12	0.00E+00	1.42	0.00	0.0000
Rv0757	2.96	9.25E-06	118.62	2.87	0.0242
Rv0758	2.43	1.08E-04	123.57	14.15	0.1145
Rv0677	2.46	7.95E-01	2053.30	930.19	0.4530
Rv3141	2.11	5.78E-01	1459.50	732.28	0.5017
Rv2641	2.48	5.92E-01	1350.90	710.71	0.5261
Rv0968	2.19	9.98E-01	1895.90	1006.10	0.5307
Rv3880	3.33	1.00E+00	841.90	489.24	0.5811
Rv3477	3.20	9.99E-01	424.85	257.74	0.6067
Rv3129	3.44	7.77E-01	1004.20	617.81	0.6152
Rv3862	2.59	7.14E-01	4413.70	2725.40	0.6175
Rv3147	2.03	9.60E-01	1233.00	832.30	0.6750
Rv0821	2.25	7.01E-01	1491.20	1033.90	0.6933
Rv3148	2.01	6.57E-10	1542.50	1073.40	0.6959
Rv3879	2.62	9.69E-01	8601.60	6057.30	0.7042
Rv3878	2.89	1.00E+00	1722.60	1230.70	0.7144
Rv3132	2.19	1.23E-02	2261.70	1678.30	0.7420
Rv1812	2.17	7.59E-01	13047.00	9983.00	0.7652
Rv0847	2.65	1.00E+00	1893.10	1450.30	0.7661
Rv2389	2.47	9.75E-01	2745.90	2149.00	0.7826
Rv3270	2.36	9.85E-01	908.19	711.39	0.7833
Rv3822	2.72	1.16E-01	6605.30	5192.80	0.7862
Rv3877	2.40	9.88E-01	8736.60	6869.70	0.7863
Rv3866	2.11	9.81E-01	3210.20	2612.20	0.8137
Rv1218	2.21	4.86E-01	850.26	693.31	0.8154
Rv3487	3.17	9.87E-01	15382.00	12781.00	0.8309
Rv3825	2.23	2.79E-04	20684.00	17297.00	0.8363
Rv3876	2.46	9.00E-01	5485.70	4615.80	0.8414
Rv3864	2.13	9.97E-01	2323.30	1982.20	0.8532
Rv3197	2.54	9.99E-01	332.11	292.45	0.8806
Rv1185	3.32	8.58E-05	3525.00	3211.10	0.9109
Rv3873	2.37	1.64E-01	3722.70	3468.80	0.9318
Rv2780	2.19	3.53E-05	2392.10	2231.20	0.9327
Rv0251	2.14	8.73E-01	1088.40	1025.40	0.9421
Rv0250	2.64	9.87E-01	515.90	489.79	0.9494
Rv3881	2.41	7.35E-01	5421.30	5264.60	0.9711
Rv3135	4.08	9.56E-01	1826.50	1778.70	0.9738
Rv3127	2.51	9.09E-05	2679.10	2737.10	1.0217
Rv1996	2.53	4.68E-02	2056.70	2104.20	1.0231
Rv2137	2.61	4.25E-01	680.42	696.12	1.0231
Rv3143	2.50	9.83E-01	546.05	567.58	1.0394

Supplementary Table 4.3, continued

Gene	Z-Score*	p-value	Read Count, H37Rv	Read Count, H37Rv ΔRv3711c	Ratio
Rv0186	2.17	1.10E-06	3406.10	3814.20	1.1198
Rv3161	2.05	3.62E-01	2475.20	2785.10	1.1252
Rv2329	2.82	6.95E-03	4606.60	5215.40	1.1322
Rv2621	2.03	9.99E-01	235.33	277.76	1.1803
Rv1184	3.12	4.30E-03	2353.80	2806.20	1.1922
Rv3865	2.11	1.67E-01	970.41	1185.60	1.2217
Rv3867	2.18	9.69E-01	1122.50	1378.30	1.2279
Rv2590	3.14	6.86E-06	13295.00	17074.00	1.2842
Rv1219	2.33	1.59E-05	953.69	1260.30	1.3215
Rv2642	2.14	5.60E-01	643.49	874.63	1.3592
Rv2390	3.33	9.88E-01	2276.50	3187.90	1.4003
Rv3136	2.62	1.59E-05	1472.30	2076.30	1.4102
Rv3146	2.13	3.06E-01	2857.80	4144.60	1.4503
Rv1217	2.14	1.56E-04	857.20	1254.70	1.4638
Rv3133	2.25	3.51E-01	1075.80	1586.80	1.4750
Rv2396	4.08	9.48E-01	3242.30	4849.70	1.4957
Rv2630	2.08	6.81E-04	428.64	641.22	1.4959
Rv2628	2.24	4.14E-01	469.16	704.55	1.5017
Rv3849	3.36	2.71E-02	17.90	31.61	1.7656
Rv3155	2.14	8.12E-01	69.59	130.14	1.8700
Rv1986	2.51	2.53E-01	471.85	886.22	1.8782
Rv2744	2.21	5.28E-02	1074.70	2068.10	1.9243
Rv3269	2.18	9.98E-01	21.08	43.31	2.0548
Rv1639	3.75	3.86E-03	918.68	2071.70	2.2550
Rv2376	2.65	8.88E-02	177.09	413.81	2.3368
Rv0967	2.53	9.99E-01	56.63	197.85	3.4935
Rv1687	2.64	7.36E-05	221.02	864.15	3.9099
Rv3137	2.74	1.00E+00	0.00	0.73	Inf
Rv0440	2.32	0.00E+00	0.00	0.00	NaN
Rv2391	5.14	0.00E+00	0.00	0.00	NaN
Rv2392	3.82	0.00E+00	0.00	0.00	NaN
Rv2393	3.14	0.00E+00	0.00	0.00	NaN

*Z-scores obtained from

Supplementary Table 4.4 Genes whose expression is correlated with PhoP, sorted by correlation.

Gene	Corr.with *phoP*	p-value	Read Count, H37Rv	Read Count, H37RvΔ Rv3711c	Ratio	Function
Rv1843c	0.400	2.19E-01	8080.00	10557.00	1.3065	inosine-5-monophosphate dehydrogenase guaB1
Rv1699	0.381	0.00E+00	0.00	0.00	NaN	CTP synthase pyrG
Rv3059	0.368	1.23E-06	5516.40	5137.00	0.9312	cytochrome P450 136 cyp136
Rv2602	0.363	7.53E-01	2901.60	2374.30	0.8183	conserved hypothetical protein
Rv3734c	0.345	8.23E-02	5354.90	6852.20	1.2796	conserved hypothetical protein
Rv0758	0.343	1.08E-04	123.57	14.15	0.1145	two component system sensor kinase phoR
Rv1880c	0.342	8.16E-02	2471.20	2618.20	1.0595	cytochrome P450 140 cyp140
Rv1478	0.335	9.54E-01	1740.70	1041.60	0.5984	invasion-associated protein
Rv3865	0.335	1.67E-01	970.41	1185.60	1.2217	conserved hypothetical protein
Rv3867	0.333	9.69E-01	1122.50	1378.30	1.2279	conserved hypothetical protein
Rv0060	0.332	9.25E-01	35.50	52.40	1.4762	conserved hypothetical protein
Rv2867c	0.331	2.08E-01	1036.90	651.27	0.6281	conserved hypothetical protein
Rv2091c	0.317	2.62E-04	742.04	381.28	0.5138	membrane protein
Rv3773c	0.316	9.40E-01	1185.20	1176.90	0.9930	conserved hypothetical protein
Rv0673	0.313	4.40E-03	535.72	981.80	1.8327	enoyl-CoA hydratase echA4
Rv0038	0.310	7.85E-01	633.79	907.80	1.4323	conserved hypothetical protein
Rv2895c	0.307	3.54E-01	2352.90	3596.60	1.5286	mycobactin utilization protein viuB
Rv3765c	0.304	3.85E-02	2835.90	3185.20	1.1232	two component system transcriptional regulator
Rv3618	0.304	4.97E-01	3259.10	2783.30	0.8540	monooxygenase
Rv3583c	0.303	1.00E+00	0.67	0.00	0.0000	transcriptional regulator
Rv2153c	0.303	0.00E+00	0.00	0.00	NaN	UDP-N-acetylglucosamine-N-acetylmuramyl-(pentapeptide)pyrophosphoryl-undecaprenol-N-acetylglucosamine transferase murG
Rv0455c	0.303	1.97E-01	21.79	39.26	1.8015	conserved hypothetical protein
Rv1411c	0.301	9.97E-01	2846.90	2009.00	0.7057	lipoprotein lprG
Rv2496c	0.301	8.52E-12	1152.50	1229.10	1.0665	pyruvate dehydrogenase E1 component beta subunit pdhB
Rv0156	-0.441	0.00E+00	0.00	0.00	0.0000	NAD(P) transhydrogenase alpha subunit pntAb
Rv2647	-0.407	1.00E+00	572.02	244.26	0.4270	hypothetical protein
Rv2306c	-0.382	7.69E-01	452.63	204.48	0.4518	
Rv3123	-0.375	9.59E-01	1483.50	712.39	0.4802	hypothetical protein
Rv0403c	-0.371	9.88E-01	1710.80	938.42	0.5485	membrane protein mmpS1
Rv0316	-0.366	9.97E-01	1090.20	652.19	0.5982	muconolactone isomerase
Rv3560c	-0.354	1.00E+00	2614.60	1581.30	0.6048	acyl-CoA dehydrogenase fadE30
Rv3820c	-0.353	9.50E-01	6509.80	4367.10	0.6709	polyketide synthase associated protein papA2
Rv2066	-0.339	1.00E+00	6272.10	4269.20	0.6807	bifunctional cobI-cobJ fusion protein

Supplementary Table 4.4, continued

Gene	Corr.with *phoP*	p-value	Read Count, H37Rv	Read Count, H37RvΔ Rv3711c	Ratio	Function
Rv2960c	-0.335	3.85E-01	1310.70	974.60	0.7436	hypothetical protein
Rv1850	-0.333	2.85E-01	756.72	563.98	0.7453	urease alpha subunit ureC
Rv2600	-0.331	9.51E-01	3396.90	2614.20	0.7696	conserved membrane protein
Rv0204c	-0.331	9.27E-01	5640.80	4525.00	0.8022	conserved membrane protein
Rv3832c	-0.321	9.94E-01	2491.20	2210.40	0.8873	conserved hypothetical protein
Rv2290	-0.319	1.56E-01	3031.90	2759.40	0.9101	lipoprotein lppO
Rv1707	-0.317	5.52E-01	5878.90	5493.10	0.9344	conserved membrane protein
Rv1244	-0.313	9.48E-01	2791.40	2790.90	0.9998	lipoprotein lpqZ
Rv2272	-0.311	9.12E-01	170.02	174.72	1.0277	conserved membrane protein
Rv3539	-0.311	5.45E-04	2105.20	2166.60	1.0292	PPE family protein
Rv1900c	-0.310	1.85E-05	2143.10	2311.30	1.0785	lignin peroxidase lipJ
Rv1282c	-0.309	2.37E-04	977.89	1116.50	1.1417	oligopeptide-transport membrane protein ABC transporter oppC
Rv0008c	-0.309	9.98E-01	140.63	163.56	1.1631	membrane protein
Rv1901	-0.307	1.24E-03	5759.10	7085.30	1.2303	competence damage-inducible protein A cinA
Rv0366c	-0.306	4.05E-02	530.64	677.06	1.2759	conserved hypothetical protein
Rv2212	-0.306	6.75E-02	1320.80	1691.50	1.2807	conserved hypothetical protein
Rv0840c	-0.304	6.63E-01	1559.20	2149.90	1.3788	proline iminopeptidase pip
Rv2529	-0.304	1.26E-05	909.56	1359.50	1.4946	hypothetical protein
Rv1432	-0.303	9.18E-01	11.66	24.87	2.1323	dehydrogenase
Rv0100	-0.303	1.20E-01	557.90	1193.70	2.1397	conserved hypothetical protein
Rv2700	-0.303	4.26E-01	90.95	267.11	2.9368	conserved secreted protein
Rv3657c	-0.302	7.23E-06	50.71	161.20	3.1787	conserved alanine rich membrane protein
Rv2601	-0.301	2.52E-01	0.94	130.10	137.8100	spermidine synthase speE
Rv2421c	-0.301	0.00E+00	0.00	0.00	NaN	nicotinate-nucleotide adenylyltransferase nadD

Supplementary Table 4.5 Genes associated with DNA replication, recombination and repair, sorted by ratio

Rv number	Gene	p-value	Read Counts, H37Rv	Read Counts, H37RvDRv3711c	Ratio	Function
Rv1402	priA	0.00E+00	0.71	0.00	0.0000	putative primosomal protein n' pria (replication factor y)
Rv3711c	dnaQ	1.33E-15	2468.20	52.77	0.0214	probable dna polymerase iii (epsilon subunit) dnaq
Rv2592c	ruvB	3.99E-01	10.86	2.73	0.2515	probable holliday junction dna helicase ruvb
Rv3715c	recR	9.92E-01	301.87	90.37	0.2994	probable recombination protein recr
Rv0937c	mku	1.22E-02	1231.30	466.07	0.3785	dna end-binding protein, mku
Rv1696	recN	4.64E-01	1856.10	789.08	0.4251	probable dna repair protein recn (recombination protein n)
Rv3062	ligB	4.02E-01	3025.60	1291.80	0.4270	probable atp-dependent dna ligase ligb (polydeoxyribonucleotide synthase [atp]) (polynucleotide ligase [atp]) (sealase) (dna repair protein) (dna joinase)
Rv1287	Rv1287	9.44E-01	389.17	171.45	0.4405	conserved hypothetical protein
Rv1629	polA	8.55E-01	190.21	85.91	0.4517	probable dna polymerase i pola
Rv0330c	Rv0330c	9.99E-01	3146.40	1681.40	0.5344	hypothetical protein
Rv0767c	Rv0767c	7.33E-01	1444.40	800.72	0.5544	conserved hypothetical protein
Rv3674c	nth	9.95E-01	1572.30	963.09	0.6125	probable endonuclease iii nth (dna-(apurinic or apyrimidinic site)lyase) (ap lyase) (ap endonuclease class i) (endodeoxyribonuclease (apurinic or apyrimidinic)) (deoxyribonuclease (apurinic or apyrimidinic))
Rv3394c	Rv3394c	9.99E-01	838.46	531.35	0.6337	conserved hypothetical protein
Rv3735	Rv3735	1.00E+00	1672.60	1094.30	0.6543	
Rv0631c	recC	1.91E-01	1218.00	799.62	0.6565	probable exonuclease v (gamma chain) recc (exodeoxyribonuclease v gamma chain)(exodeoxyribonuclease v polypeptide)
Rv2528c	mrr	1.00E+00	2240.40	1474.70	0.6582	probable restriction system protein mrr
Rv0269c	Rv0269c	1.33E-01	2158.30	1459.40	0.6762	conserved hypothetical protein

Supplementary Table 4.5, continued

Rv number	Gene	p-value	Read Counts, H37Rv	Read Counts, H37RvDRv3711c	Ratio	Function
Rv3585	radA	9.97E-01	7223.40	5092.50	0.7050	dna repair protein rada (dna repair protein sms)
Rv1534	Rv1534	7.64E-01	927.46	656.94	0.7083	probable transcriptional regulator
Rv3201c	Rv3201c	2.51E-02	3247.60	2334.10	0.7187	probable atp-dependent dna helicase
Rv1020	mfd	9.74E-01	5301.70	3846.70	0.7256	probable transcription-repair coupling factor mfd (trcf)
Rv2917	Rv2917	6.65E-03	4854.80	3528.10	0.7267	conserved hypothetical alanine and arginine rich protein
Rv0195	Rv0195	9.99E-01	3655.10	2703.80	0.7397	possible two component transcriptional regulatory protein (probably luxr-family)
Rv0844c	narL	9.08E-01	6122.50	4630.70	0.7563	possible nitrate/nitrite response transcriptional regulatory protein narl
Rv1904	Rv1904	4.41E-01	4492.40	3402.50	0.7574	conserved hypothetical protein
Rv2132	Rv2132	1.00E+00	358.89	273.91	0.7632	conserved hypothetical protein
Rv2896c	Rv2896c	3.71E-01	1998.90	1529.30	0.7651	conserved hypothetical protein
Rv2985	mutT1	8.11E-01	1664.00	1274.20	0.7657	possible hydrolase mutt1
Rv3048c	nrdF2	0.00E+00	0.94	0.73	0.7741	ribonucleoside-diphosphate reductase (beta chain) nrdf2 (ribonucleotide reductase small subunit) (r2f protein)
Rv2807	Rv2807	2.58E-01	1523.90	1181.90	0.7756	conserved hypothetical protein
Rv1701	Rv1701	1.95E-01	283.18	220.01	0.7770	probable integrase/recombinase
Rv3732	Rv3732	9.99E-01	4704.00	3663.30	0.7788	
Rv3788	Rv3788	9.84E-01	2887.40	2250.10	0.7793	hypothetical protein
Rv2593c	ruvA	9.90E-01	18.16	14.46	0.7963	probable holliday junction dna helicase ruva
Rv3202c	Rv3202c	9.44E-01	146.90	118.33	0.8055	possible atp-dependent dna helicase
Rv1420	uvrC	5.09E-02	3821.40	3088.40	0.8082	probable excinuclease abc (subunit c - nuclease) uvrc
Rv0298	Rv0298	1.00E+00	1076.90	881.78	0.8189	hypothetical protein
Rv0413	mutT3	9.86E-01	2681.00	2203.80	0.8220	possible mutator protein mutt3 (7,8-dihydro-8-oxoguanine-triphosphatase) (8-oxo-dgtpase)
Rv0861c	ercc3	4.62E-01	1702.20	1402.00	0.8236	dna helicase ercc3
Rv2478c	Rv2478c	1.00E+00	1699.40	1401.40	0.8246	conserved hypothetical protein
Rv3750c	Rv3750c	9.49E-01	2247.30	1856.50	0.8261	possible excisionase

Supplementary Table 4.5, continued

Rv number	Gene	p-value	Read Counts, H37Rv	Read Counts, H37RvDRv3711c	Ratio	Function
Rv1107c	xseB	9.92E-01	675.28	560.95	0.8307	probable exodeoxyribonuclease vii (small subunit) xsea
Rv1108c	xseA	2.28E-01	891.96	744.55	0.8347	probable exodeoxyribonuclease vii (large subunit) xsea
Rv2415c	Rv2415c	2.10E-01	1892.80	1584.00	0.8369	conserved hypothetical protein
Rv3555c	Rv3555c	6.95E-01	972.60	823.35	0.8465	conserved hypothetical protein
Rv3731	ligC	4.97E-01	5275.10	4501.30	0.8533	possible atp-dependent dna ligase ligc (polydeoxyribonucleotide synthase [atp]) (polynucleotide ligase [atp]) (sealase) (dna repair protein) (dna joinase)
Rv2024c	Rv2024c	9.74E-01	7970.30	6877.20	0.8629	conserved hypothetical protein
Rv2756c	hsdM	1.24E-09	3353.70	2945.00	0.8782	possible type i restriction/modification system dna methylase hsdm (m protein) (dna methyltransferase)
Rv3198c	uvrD2	6.94E-01	3661.90	3233.30	0.8830	probable atp-dependent dna helicase ii uvrd2
Rv3714c	Rv3714c	1.00E+00	5859.70	5181.10	0.8842	conserved hypothetical protein
Rv3056	dinP	9.64E-01	3845.80	3417.70	0.8887	possible dna-damage-inducible protein p dinp (dna polymerase v) (pol iv 2) (dna nucleotidyltransferase (dna-directed))
Rv1537	dinX	4.07E-03	1474.30	1319.50	0.8950	probable dna polymerase iv dinx (pol iv 1) (dna nucleotidyltransferase (dna-directed))
Rv3733	Rv3733	9.96E-01	1077.10	973.21	0.9035	
Rv1179c	Rv1179c	1.76E-06	5943.00	5457.90	0.9184	hypothetical protein
Rv0427c	xthA	9.66E-01	1744.50	1612.20	0.9241	probable exodeoxyribonuclease iii protein xtha (exonuclease iii) (exo iii) (ap endonuclease vi)
Rv1633	uvrB	9.22E-01	10414.00	9686.20	0.9301	probable excinuclease abc (subunit b - helicase) uvrb
Rv2718c	Rv2718c	2.79E-01	606.58	564.53	0.9307	conserved hypothetical protein
Rv2464c	Rv2464c	5.30E-01	3702.90	3521.70	0.9511	possible dna glycosylase

Supplementary Table 4.5, continued

Rv number	Gene	p-value	Read Counts, H37Rv	Read Counts, H37RvDRv3711c	Ratio	Function
Rv0825c	Rv0825c	7.63E-01	5567.30	5343.20	0.9597	conserved hypothetical protein
Rv2894c	xerC	2.28E-01	254.95	245.56	0.9631	probable integrase/recombinase xerc
Rv2761c	hsdS	2.00E-05	1767.60	1704.70	0.9644	possible type i restriction/modification system specificity determinant hsds (s protein)
Rv0775	Rv0775	4.77E-01	3419.90	3323.70	0.9719	conserved hypothetical protein
Rv2362c	recO	8.41E-02	165.31	165.46	1.0009	possible dna repair protein reco
Rv0123	Rv0123	9.97E-01	199.66	202.14	1.0124	hypothetical protein
Rv1230c	Rv1230c	6.99E-01	3172.80	3223.60	1.0160	possible membrane protein
Rv2101	helZ	4.55E-01	5554.80	5650.10	1.0172	probable helicase helz
Rv0630c	recB	2.84E-14	2363.70	2426.00	1.0264	probable exonuclease v (beta chain)
Rv2191	Rv2191	5.77E-06	4861.80	5008.20	1.0301	conserved hypothetical protein
Rv2405	Rv2405	9.80E-01	633.91	654.56	1.0326	conserved hypothetical protein
Rv0668	rpoC	1.00E+00	255.90	264.31	1.0329	dna-directed rna polymerase (beta' chain) rpoc (transcriptase beta' chain) (rna polymerase beta' subunit).
Rv2248	Rv2248	1.39E-01	870.52	926.64	1.0645	conserved hypothetical protein
Rv0944	Rv0944	9.71E-01	790.36	869.02	1.0995	possible formamidopyrimidine-dna glycosylase (fapy-dna glycosylase)
Rv1277		1.77E-03	1673.80	1848.00	1.1040	
Rv3095	Rv3095	9.57E-01	1502.20	1667.10	1.1098	hypothetical transcriptional regulatory protein
Rv0629c	recD	1.70E-01	538.80	604.03	1.1211	probable exonuclease v (alpha chain) recd (exodeoxyribonuclease v alpha chain) (exodeoxyribonuclease v polypeptide)
Rv3586	Rv3586	2.53E-02	828.50	943.79	1.1392	conserved hypothetical protein
Rv2924c	fpg	7.52E-01	2449.70	2806.80	1.1457	probable formamidopyrimidine-dna glycosylase fpg (fapy-dna glycosylase)

Supplementary Table 4.5, continued

Rv number	Gene	p-value	Read Counts, H37Rv	Read Counts, H37RvDRv3711c	Ratio	Function
Rv2090	Rv2090	8.06E-01	2380.80	2799.10	1.1757	probable 5'-3' exonuclease
Rv1080c	greA	8.17E-01	603.75	714.92	1.1841	probable transcription elongation factor grea (transcript cleavage factor grea)
Rv0570	nrdZ	7.67E-03	2914.30	3453.40	1.1850	probable ribonucleoside-diphosphate reductase (large subunit) nrdz (ribonucleotide reductase)
Rv1329c	dinG	1.53E-02	2570.60	3056.70	1.1891	probable atp-dependent helicase ding
Rv3730c	Rv3730c	1.23E-01	3402.40	4126.10	1.2127	conserved hypothetical protein
Rv1828	Rv1828	0.00E+00	2.29	2.91	1.2686	conserved hypothetical protein
Rv2976c	ung	9.95E-01	369.34	469.33	1.2707	probable uracil-dna glycosylase ung (udg)
Rv3734	Rv3734	8.23E-02	5354.90	6852.20	1.2796	
Rv3263	Rv3263	5.77E-05	9144.30	11720.00	1.2816	probable dna methylase (modification methylase) (methyltransferase)
Rv3589	mutY	2.53E-01	1030.70	1329.80	1.2902	probable adenine glycosylase muty
Rv1041c	Rv1041c	4.20E-01	1503.00	1961.50	1.3051	probable is like-2 transposase
Rv1317c	alkA	2.73E-01	1298.90	1731.60	1.3331	probable ada regulatory protein alka (regulatory protein of adaptative response) (methylated-dna--protein-cysteine methyltransferase) (o-6-methylguanine-dna alkyltransferase) (o-6-methylguanine-dna methyltransferase) (3-methyladenine dna glycosylase ii)
Rv3370c	dnaE2	1.48E-05	1519.10	2041.20	1.3436	probable dna polymerase iii (alpha chain) dnae2 (dna nucleotidyltransferase)
Rv0938	ligD	1.77E-11	1758.00	2378.40	1.3529	atp dependent dna ligase (atp dependent polydeoxyribonucleotide synthase) (thermostable dna ligase) (atp dependent polynucleotide ligase) (sealase) (dna repair enzyme) (dna joinase)
Rv0003	recF	6.35E-12	500.92	683.82	1.3651	dna replication and repair protein recf (single-strand dna binding protein)

Supplementary Table 4.5, continued

Rv number	Gene	p-value	Read Counts, H37Rv	Read Counts, H37RvDRv3711c	Ratio	Function
Rv3856c	Rv3856c	1.24E-01	588.04	810.94	1.3790	conserved hypothetical protein
Rv1688	mpg	1.30E-02	930.04	1284.10	1.3807	possible 3-methyladenine dna glycosylase mpg
Rv0001	dnaA	0.00E+00	0.71	1.00	1.4043	chromosomal replication initiator protein dnaa
Rv1210	tagA	4.81E-01	7503.20	10550.00	1.4061	probable dna-3-methyladenine glycosylase i taga (tag i) (3-methyladenine-dna glycosylase i, constitutive) (dna-3-methyladenine glycosidase i)
Rv2973c	recG	2.36E-02	1542.60	2170.50	1.4071	probable atp-dependent dna helicase recg
Rv2515c	Rv2515c	1.09E-06	1240.40	1823.80	1.4702	conserved hypothetical protein
Rv2310	Rv2310	1.00E+00	188.90	279.81	1.4812	possible excisionase
Rv2529	Rv2529	1.26E-05	909.56	1359.50	1.4946	hypothetical protein
Rv1407	fmu	2.65E-08	2079.70	3216.00	1.5464	probable fmu protein (sun protein)
Rv0949	uvrD1	1.44E-07	934.37	1459.40	1.5620	probable atp-dependent dna helicase ii uvrd1
Rv1316c	ogt	2.51E-02	1237.80	1941.10	1.5682	probable methylated-dna--protein-cysteine methyltransferase ogt (6-o-methylguanine-dna methyltransferase) (o-6-methylguanine-dna-alkyltransferase)
Rv0500A	Rv0500A	1.00E+00	79.82	136.09	1.7049	conserved hypothetical protein
Rv2069	sigC	2.31E-03	763.37	1316.10	1.7241	probable rna polymerase sigma factor, ecf subfamily, sigc
Rv0142	Rv0142	9.64E-01	421.51	739.35	1.7540	conserved hypothetical protein
Rv0006	gyrA	1.00E+00	5.29	9.81	1.8536	dna gyrase (subunit a) gyra (dna topoisomerase (atp-hydrolysing)) (dna topoisomerase ii) (type ii dna topoisomerase)
Rv1259	udgB	4.89E-04	259.14	526.42	2.0314	
Rv2160A	Rv2160A	5.32E-02	105.09	224.25	2.1339	conserved hypothetical protein
Rv2327	Rv2327	3.04E-01	128.86	305.70	2.3724	conserved hypothetical protein
Rv0670	end	1.32E-02	441.26	1091.80	2.4744	probable endonuclease iv end (apurinase)

Supplementary Table 4.5, continued

Rv number	Gene	p-value	Read Counts, H37Rv	Read Counts, H37RvDRv3711c	Ratio	Function
Rv2737c	recA	9.98E-01	52.66	135.61	2.5751	reca protein (recombinase a) [contains: endonuclease pi-mtui (mtu reca intein)].
Rv3297	nei	7.89E-04	514.92	1364.60	2.6501	probable endonuclease viii nei
Rv2748c	ftsK	0.00E+00	12.26	37.83	3.0864	possible cell division transmembrane protein ftsk
Rv1981c	nrdF1	9.99E-01	99.75	368.45	3.6936	ribonucleoside-diphosphate reductase (beta chain) nrdf1 (ribonucleotide reductase small subunit) (r2f protein)
Rv2736c	recX	1.66E-04	31.83	138.46	4.3504	regulatory protein recx
Rv2898c	Rv2898c	1.00E+00	24.33	126.55	5.2013	conserved hypothetical protein
Rv2638	Rv2638	3.97E-07	140.76	775.50	5.5094	conserved hypothetical protein
Rv1638	uvrA	7.69E-04	19.63	160.16	8.1607	probable excinuclease abc (subunit a - dna-binding atpase) uvra
Rv3204	Rv3204	1.01E-05	4.98	115.83	23.2400	possible dna-methyltransferase (modification methylase)
Rv1160	mutT2	8.34E-10	18.88	450.44	23.8570	probable mutator protein mutt2 (7,8-dihydro-8-oxoguanine-triphosphatase) (8-oxo-dgtpase)
Rv1547	dnaE1	0.00E+00	0.00	8.70	Inf	probable dna polymerase iii (alpha chain) dnae1 (dna nucleotidyltransferase)
Rv1685c	Rv1685c	1.00E+00	0.00	0.71	Inf	conserved hypothetical protein
Rv1830	Rv1830	1.91E-01	0.00	175.01	Inf	conserved hypothetical protein
Rv2373c	dnaJ2	1.00E+00	0.00	2.14	Inf	probable chaperone protein dnaj2
Rv0002	dnaN	0.00E+00	0.00	0.00	NaN	dna polymerase iii (beta chain) dnan (dna nucleotidyltransferase)
Rv0005	gyrB	0.00E+00	0.00	0.00	NaN	dna gyrase (subunit b) gyrb (dna topoisomerase (atp-hydrolysing)) (dna topoisomerase ii) (type ii dna topoisomerase)
Rv0054	ssb	0.00E+00	0.00	0.00	NaN	probable single-strand binding protein ssb (helix-destabilizing protein)
Rv0058	dnaB	0.00E+00	0.00	0.00	NaN	probable replicative dna helicase dnab

Supplementary Table 4.5, continued

Rv number	Gene	p-value	Read Counts, H37Rv	Read Counts, H37RvDRv3711c	Ratio	Function
Rv0667	rpoB	0.00E+00	0.00	0.00	NaN	dna-directed rna polymerase (beta chain) rpob (transcriptase beta chain)
Rv1390	rpoZ	0.00E+00	0.00	0.00	NaN	probable dna-directed rna polymerase (omega chain) rpoz (transcriptase omega chain) (rna polymerase omega subunit)
Rv2343c	dnaG	0.00E+00	0.00	0.00	NaN	probable dna primase dnag
Rv2413c	Rv2413c	0.00E+00	0.00	0.00	NaN	conserved hypothetical protein
Rv2554c	Rv2554c	0.00E+00	0.00	0.00	NaN	conserved hypothetical protein
Rv2594c	ruvC	0.00E+00	0.00	0.00	NaN	probable crossover junction endodeoxyribonuclease ruvc (holliday junction nuclease) (holliday junction resolvase)
Rv2720	lexA	0.00E+00	0.00	0.00	NaN	repressor lexa
Rv2986c	hupB	0.00E+00	0.00	0.00	NaN	probable dna-binding protein hu homolog hupb (histone-like protein) (hlp) (21-kda laminin-2-binding protein)
Rv3014c	ligA	0.00E+00	0.00	0.00	NaN	probable dna ligase [nad dependent] liga (polydeoxyribonucleotide synthase [nad+])
Rv3051c	nrdE	0.00E+00	0.00	0.00	NaN	ribonucleoside-diphosphate reductase (alpha chain) nrde (ribonucleotide reductase small subunit) (r1f protein)
Rv3457c	rpoA	0.00E+00	0.00	0.00	NaN	probable dna-directed rna polymerase (alpha chain) rpoa (transcriptase alpha chain) (rna polymerase alpha subunit) (dna-directed rna nucleotidyltransferase)
Rv3596c	clpC1	0.00E+00	0.00	0.00	NaN	probable atp-dependent protease atp-binding subunit clpc1
Rv3644c	Rv3644c	0.00E+00	0.00	0.00	NaN	possible dna polymerase
Rv3646c	topA	0.00E+00	0.00	0.00	NaN	dna topoisomerase i topa (omega-protein) (relaxing enzyme)
Rv3648c	cspA	0.00E+00	0.00	0.00	NaN	probable cold shock protein a cspa
Rv3721c	dnaZX	0.00E+00	0.00	0.00	NaN	dna polymerase iii (subunit gamma/tau) dnaz/x
Rv3917c	parB	0.00E+00	0.00	0.00	NaN	probable chromosome partitioning protein parb