nature genetics

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

Christopher B Ford[1,11], Philana Ling Lin[2,11], Michael R Chase[1], Rupal R Shah[1], Oleg Iartchouk[3], James Galagan[4–6], Nilofar Mohaideen[7], Thomas R Ioerger[8], James C Sacchettini[7], Marc Lipsitch[1,9], JoAnne L Flynn[10] & Sarah M Fortune[1]

Tuberculosis poses a global health emergency, which has been compounded by the emergence of drug-resistant *Mycobacterium tuberculosis* (Mtb) strains. We used whole-genome sequencing to compare the accumulation of mutations in Mtb isolated from cynomolgus macaques with active, latent or reactivated disease. We sequenced 33 Mtb isolates from nine macaques with an average genome coverage of 93% and an average read depth of 117×. Based on the distribution of SNPs observed, we calculated the mutation rates for these disease states. We found a similar mutation rate during latency as during active disease or in a logarithmically growing culture over the same period of time. The pattern of polymorphisms suggests that the mutational burden *in vivo* is because of oxidative DNA damage. We show that Mtb continues to acquire mutations during disease latency, which may explain why isoniazid monotherapy for latent tuberculosis is a risk factor for the emergence of isoniazid resistance[1,2].

In active tuberculosis, affected individuals harbor large numbers of replicating organisms and are treated with multiple antibiotics to prevent the emergence of new 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[1,2]. 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 sought 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.

Conventional approaches for measuring bacterial mutation rates cannot be applied to Mtb *in vivo*. However, high-density whole-genome sequencing (WGS) allowed us to assess the capacity of Mtb for mutation over the course of infection with minimal bias and maximum sensitivity[3–5]. As the nonhuman primate is the only animal model that mimics the broad range of disease seen in human tuberculosis[6,7], 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[7]) (**Fig. 1**). Genome coverage averaged 93% across these isolates, and the average read depth was 117× across the genomes (**Supplementary Table 1**). We identified putative polymorphisms using both a scaffolded approach[8,9] and a *de novo* assembly method[10], and we validated polymorphic sites using Sanger sequencing or through independent identification by WGS. Through these analyses, we identified 14 unique SNPs (**Fig. 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. Although we have used WGS previously to detect insertions and deletions[9], 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.

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 nonhuman primates is unknown. *In vitro*, Mtb has a generation time of approximately 20 hours under nutrient-rich conditions[11]. In mice, the bacterial organ burden increases at roughly this rate during the first weeks of infection, but the subsequent onset of

[1]Department of Immunology and Infectious Diseases, Harvard School of Public Health, Boston, Massachusetts, USA. [2]Department of Pediatrics, Children's Hospital of Pittsburgh of the University of Pittsburgh Medical Center, Pittsburgh, Pennsylvania, USA. [3]Partners Healthcare Center for Personalized Genetic Medicine, Harvard Medical School, Cambridge, Massachusetts, USA. [4]Department of Biomedical Engineering, Boston University, Boston, Massachusetts, USA. [5]Department of Microbiology, Boston University, Boston, Massachusetts, USA. [6]The Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [7]Department of Biochemistry and Biophysics, Texas A&M University, College Station, Texas, USA. [8]Department of Computer Science and Engineering, Texas A&M University, College Station, Texas, USA. [9]Center for Communicable Disease Dynamics, Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts, USA. [10]Department of Microbiology and Molecular Genetics, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania, USA. [11]These authors contributed equally to this work. Correspondence should be addressed to S.M.F. (sfortune@hsph.harvard.edu).

Received 15 December 2010; accepted 24 March 2011; published online 24 April 2011; doi:10.1038/ng.811

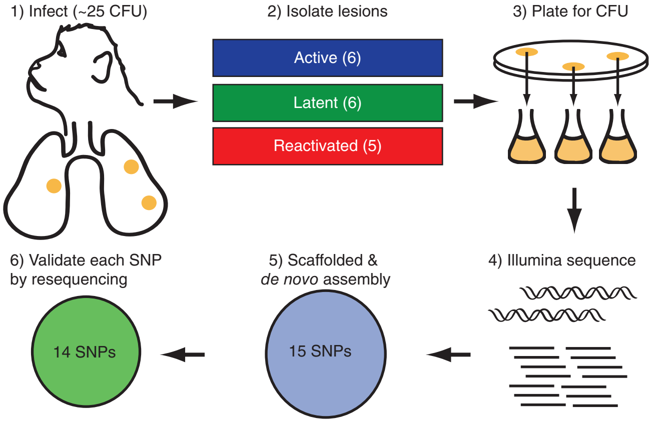

Figure 1 Experimental protocol for assessing mutational capacity in different disease states. 1) Cynomolgus macaques were infected with ~25 colony forming units (CFU) of Mtb Erdman using bronchoscopy. 2) Animals were killed in the indicated stages of disease for strain isolation. 3) Eighteen pathologic lesions were plated for bacterial colonies. Thirty-three strains were isolated for WGS. 4) Genomic DNA was isolated from these strains and then analyzed using Illumina sequencing. 5) Reads were assembled using both *de novo* and scaffolded approaches. Fifteen 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.

the adaptive immune response causes bacterial replication to slow substantially or cease entirely[12,13]. 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 show a range of histopathologies and bacterial burdens, suggesting that a spectrum of bacterial physiologies may occur in latency[14,15].

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 h), calculating the rate that would be required to generate the number of polymorphisms identified by WGS (**Fig. 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 defined as the predicted mutation rate per generation if the *in vivo* generation time were equivalent to the *in vitro* generation time of 20 h, μ (20 h). Although Mtb is likely to have a much longer generation time *in vivo*, especially during prolonged latent infection, we used μ (20 h) as a highly conservative boundary estimate of the *in vivo* mutation rate, which allowed us to directly compare the mutational capacity of the bacterium in different *in vivo* conditions. Notably, we found that the bacterial population's capacity for mutation, μ (20 h), during latency (2.71×10^{-10}) and reactivated disease (3.03×10^{-10}) is equivalent to that of Mtb from animals with active disease (2.01×10^{-10}) (**Table 1**). The 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 (**Fig. 2d**). 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*. A Luria and Delbrück fluctuation analysis measures the rate at which coding polymorphisms conferring a selectable phenotype arise under stable *in vitro* conditions[16]. Although 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[17], which is 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 2.1×10^{-9} (**Supplementary Fig. 1a**,**b**), consistent with or slightly higher than previously published values for Mtb[18,19]. 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[20] (**Supplementary Fig. 1c**). Dividing

Clinical course			ACTIVE															LATENT								REACTIVATED									
Animal			A				B	C				D						E		F			G			H			I						
Duration (days)			91				265	293				355						281		297			299			447			507						
Anatomical site of lesion			LLL		ACL	RML	ACL	RLL				RML			ACL			CN		ACL	RLL	RML	LLL		RLL	RLL			ACL	RUL	BN			LLL	
strain			1	2	3	4	1	1	2	3	4	1	2	3	4	5	6	1	2	1	2	3	1	2	3	1	2	3	1	2	3	4	5	6	7
Coordinate	Gene	Inoculum																																	
635497	*Rv0542c*	C	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	T	•	•	•	•	•	•	•	•
682043	*Rv0585c*	C	•	•	•	•	•	•	•	•	•	•	•	•	G	G	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
690264	*Rv0592*	G	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	C	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
693453	*Rv0594*	C	•	•	•	•	T	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
766229	*Rv0668*	G	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	A	•	•	•	•
975906	*Rv0876c*	T	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	C	•	•	•	•	•	•	•	•	•	•	•	•
1256717	*Rv1131*	G	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	A	•	•	•	•	•	•	•	•
1854208	*Rv1644*	G	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	T	•	•	•
1861203	*Rv1650*	G	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	A	•	•
2350697	*Rv2092c*	G	•	•	•	•	•	•	•	•	•	T	T	T	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
2448250	*Rv2187*	G	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	A	•	•	•	•	•
3655598	*Rv3273*	G	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	T	•	•	•	•	•	•	•	•	•	•	•	•	•
4183984	*Rv3732*	A	•	•	•	•	•	•	•	•	•	•	•	•	G	G	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
4346906	*Rv3870*	C	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	A	A	A	•	•	•	•	•	•	•

Figure 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 with 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). 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. Coordinates (Coord.) are given for H37Rv. Inoculum (Inoc) represents the sequence at the given coordinate of the inoculating strain, Mtb Erdman.

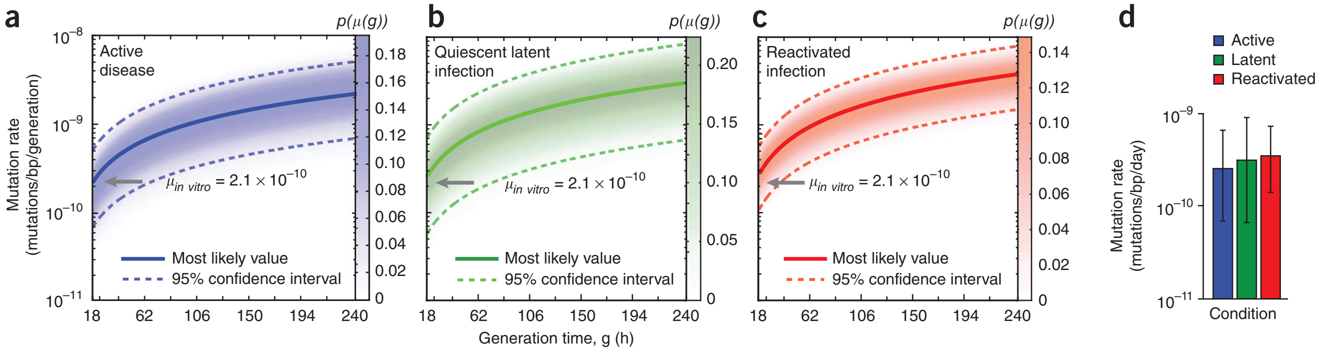

Figure 3 The mutational capacity of strains from latency and reactivated disease is similar to that of strains from active disease or *in vitro* growth. (**a**–**c**) The mutation rate (μ) was estimated based on the number of unique SNPs (m) observed in each condition (4 active, 3 latent and 7 reactivated). This calculation was performed over a range of generation times (g, 18–240 h per generation) to allow for the uncertainty in growth rate *in vivo*. We determined the probability of observing μ when g is fixed at any given time to build the probability distribution function around each estimate and to define the 95% confidence intervals. We determined the single-base mutation rate of the bacterium during *in vitro* growth ($\mu_{in\ vitro}$) by fluctuation analysis (**Supplementary Fig. 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 $\mu_{in\ vitro}$. 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 rates per day regardless of the disease status of the host. Error bars represent 95% confidence intervals.

the phenotypic rate by the target size, we determined that the *in vitro* mutation rate of the inoculating strain (Erdman) is $\mu_{in\ vitro} = 2.1 \times 10^{-10}$. Thus μ (20 h), 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[12–15]. 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 replicates on average every 135 h, as in mice after ten weeks of chronic infection[12], the bacterial population must have an average mutation rate per generation of 1.80×10^{-9}, nearly an order of magnitude greater than the *in vitro* mutation rate (**Table 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) (**Fig. 4a**). This is consistent with the model that Mtb faces an oxidative environment in the macrophage phagolysosome[21,22] and data indicating that genes involved in the repair of oxidative damage are essential for bacterial survival during mouse infection[23]. In addition, we found that the pattern of polymorphisms in Mtb from cynomolgus macaques is similar to the pattern of neutral polymorphisms that emerged during the evolution of extensively drug-resistant Mtb strains in affected individuals from South Africa (**Fig. 3a**)[9]. 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[24] (**Fig. 4b**). This may occur because the immune response that results in latent infection causes more oxidative damage to the bacterial DNA[14,25] or because a portion of the bacteria may enter a metabolically quiescent state in which DNA repair is diminished[26,27].

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

Table 1 The predicted mutation rate for biologically relevant generation times

Gen. time (h) (g)	Growth condition	$\mu(g)$active (95% CI)[a]	$\mu(g)$latent (95% CI)[a]	$\mu(g)$reactivated (95% CI)[a]
20	Rich media	2.01×10^{-10} (8.09×10^{-11} to 4.15×10^{-10})	2.71×10^{-10} (5.57×10^{-11} to 7.89×10^{-10})	3.03×10^{-10} (1.22×10^{-10} to 6.24×10^{-10})
45	Macrophage	4.77×10^{-10} (1.30×10^{-10} to 1.22×10^{-9})	5.99×10^{-10} (1.23×10^{-10} to 1.75×10^{-9})	6.71×10^{-10} (2.70×10^{-10} to 1.38×10^{-9})
135	Mouse infection at 10 weeks	1.43×10^{-9} (3.90×10^{-10} to 3.66×10^{-9})	1.80×10^{-9} (3.70×10^{-10} to 5.25×10^{-9})	2.01×10^{-9} (8.09×10^{-10} to 4.15×10^{-9})

The generation time (g) was varied from 18–240 h, 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[11,12,30]. Gen., generation.

[a] Mutation rates were estimated using the equation shown: $\mu = m/[N^*(t/g)]$, over g = 18–240 h.

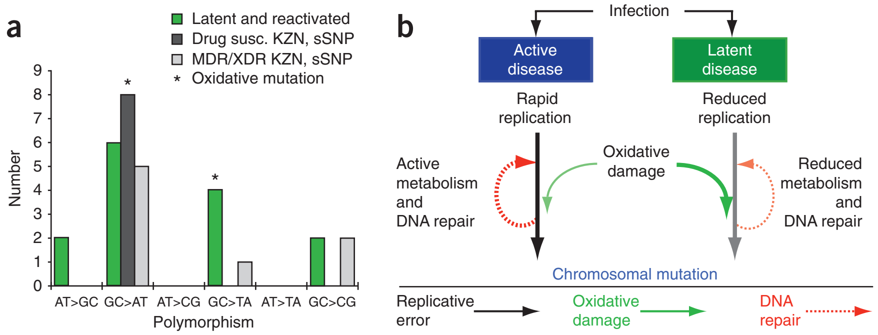

Figure 4 Mutations in Mtb isolated from macaques with latent infection and related human isolates are putative products of oxidative damage. (**a**) Ten of the 14 mutations observed in this study could be the product of oxidative damage: the deamination of cytosine (GC>AT) or the production of 7,8-dihydro-8-oxoguanine (GC>TA) by the oxidation of guanine. We saw one of each type of mutation observed 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 (sSNP) identified by WGS of a set of closely related strains from South Africa[9]. Susc., susceptible; MDR, multidrug resistant; XDR, extensively resistant. (**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.

resistance after IPT in immunocompetent populations[1,2], some of which may be attributable to selective killing of susceptible bacterial populations, leaving only resistant populations to reactivate[28]. Our results suggest that in addition to these mechanisms, part of the increased risk of INH resistance after IPT may be caused by the selection of monoresistant mutants that arise during latency. IPT is now being recommended globally for HIV-positive individuals with clinically latent tuberculosis where bacterial burden and the rate of treatment failure may be higher because of immunocompromise[29]. If our data from the macaque model are predictive of the mutational capacity of Mtb in HIV-positive individuals, INH monoresistance could arise at a substantial rate. These findings emphasize the importance of drug resistance testing and careful monitoring for treatment failure in these populations.

URLs. *Mycobacterium tuberculosis* Sequencing Project, Broad Institute of Harvard and MIT, http://www.broadinstitute.org/; tuberculosis database, http://www.tbdb.org/.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

Accession codes. Short reads have been posted in NCBI-SRA, and accession numbers for each of the 34 *M. tuberculosis* Erdman strains are presented in **Supplementary Table 1**.

Note: Supplementary information is available on the Nature Genetics website.

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 S.M.F., by a subcontract from National Institute for Allergy and Infectious Diseases (NIAID) U19 AI076217 to S.M.F., by the US National Institutes of Health (NIH) RO1 HL075845 to J.L.F. and by the Bill and Melinda Gates Foundation (J.L.F.). The genome sequencing has been funded in part with federal funds from the National Institute of Allergy and Infectious Disease, US NIH, US Department of Health and Human Services, under contract no. HHSN266200400001C. The project described was supported in part by Award Number U54GM088558 to M.L. 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

C.B.F. performed molecular studies, conducted the data analyses, prepared the figures and drafted the manuscript; P.L.L. and J.L.F. conducted the infection of the cynomolgus macaques, determined clinical state and acquired bacterial strains on necropsy; M.R.C. analyzed sequence data and directed validation of SNPs; R.R.S. performed molecular and fluctuation analyses; O.I. oversaw sequencing of isolates sent to Partners Healthcare Center for Personalized Genetic Medicine (PHCPGM); J.G. oversaw sequencing of isolates sent to the Broad Institute; N.M., T.R.I. and J.C.S. oversaw sequencing and analysis of isolates sent to Texas A&M University; M.L. supervised and advised statistical analyses; S.M.F. initiated the project, performed molecular studies, supervised preparation and analysis of the data and drafted the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturegenetics/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Balcells, M.E., Thomas, S.L., Godfrey-Faussett, P. & Grant, A.D. Isoniazid preventive therapy and risk for resistant tuberculosis. *Emerg. Infect. Dis.* **12**, 744–751 (2006).
2. Cattamanchi, A. *et al.* Clinical characteristics and treatment outcomes of patients with isoniazid-monoresistant tuberculosis. *Clin. Infect. Dis.* **48**, 179–185 (2009).
3. 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).
4. Haag-Liautard, C. *et al.* Direct estimation of per nucleotide and genomic deleterious mutation rates in *Drosophila*. *Nature* **445**, 82–85 (2007).
5. Lynch, M. *et al.* A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proc. Natl. Acad. Sci. USA* **105**, 9272–9277 (2008).
6. Capuano, S.V. III *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).
7. Lin, P.L. *et al.* Quantitative comparison of active and latent tuberculosis in the cynomolgus macaque model. *Infect. Immun.* **77**, 4631–4642 (2009).
8. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18**, 1851–1858 (2008).
9. Ioerger, T.R. *et al.* Genome analysis of multi- and extensively-drug-resistant tuberculosis from KwaZulu-Natal, South Africa. *PLoS ONE* **4**, e7778 (2009).
10. Hernandez, D., Francois, P., Farinelli, L., Osteras, M. & Schrenzel, J. *De novo* bacterial genome sequencing: millions of very short reads assembled on a desktop computer. *Genome Res.* **18**, 802–809 (2008).
11. 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).
12. Gill, W.P. *et al.* A replication clock for *Mycobacterium tuberculosis*. *Nat. Med.* **15**, 211–214 (2009).
13. Muñoz-Elías, E.J. *et al.* Replication dynamics of *Mycobacterium tuberculosis* in chronically infected mice. *Infect. Immun.* **73**, 546–551 (2005).
14. Barry, C.E. *et al.* The spectrum of latent tuberculosis: rethinking the biology and intervention strategies. *Nat. Rev. Microbiol.* **7**, 845–855 (2009).
15. Lin, P.L. & Flynn, J.L. Understanding latent tuberculosis: a moving target. *J. Immunol.* **185**, 15–22 (2010).
16. 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).
17. Lang, G.I. & Murray, A.W. Estimating the per-base-pair mutation rate in the yeast *Saccharomyces cerevisiae*. *Genetics* **178**, 67–82 (2008).

18. 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).
19. Werngren, J. & Hoffner, S.E. Drug-susceptible *Mycobacterium tuberculosis* Beijing genotype does not develop mutation-conferred resistance to rifampin at an elevated rate. *J. Clin. Microbiol.* **41**, 1520–1524 (2003).
20. Telenti, A. *et al.* Detection of rifampicin-resistance mutations in *Mycobacterium tuberculosis*. *Lancet* **341**, 647–650 (1993).
21. 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).
22. 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).
23. Sassetti, C.M. & Rubin, E.J. Genetic requirements for mycobacterial survival during infection. *Proc. Natl. Acad. Sci. USA* **100**, 12989–12994 (2003).
24. Boshoff, H.I., 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).
25. 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).
26. Saint-Ruf, C., Pesut, J., Sopta, M. & Matic, I. Causes and consequences of DNA repair activity modulation during stationary phase in *Escherichia coli*. *Crit. Rev. Biochem. Mol. Biol.* **42**, 259–270 (2007).
27. Bjedov, I. *et al.* Stress-induced mutagenesis in bacteria. *Science* **300**, 1404–1409 (2003).
28. Cohen, T., Lipsitch, M., Walensky, R.P. & Murray, M. 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).
29. 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).
30. 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).

ONLINE METHODS

Preparation of isolates. Animals were infected as described previously[7] using 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[7] 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 (**Fig. 2**)[7]. Colonies from necropsy were subsequently streaked onto Lowenstein-Jensen slants and expanded for extraction of genomic DNA. Minimal expansion occurred between the isolation of strains and the extraction of genomic DNA.

Illumina sequencing. Two micrograms of genomic DNA from each isolate were used for sequencing with the Illumina Genome Analyzer (Illumina). Single-read and 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 quantitative PCR (qPCR) protocol with specific probes for the ends of the adapters. The qPCR assay measures the quantity of fragments properly adaptor ligated that are appropriate for sequencing. Based on the qPCR quantification, libraries were normalized to 2 nM and then denatured using 0.1 N NaOH. Cluster amplification of denatured templates occurred according to the manufacturer's protocol using V2 Chemistry and V2 Flowcells (with 1.4 mm 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 the 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, percent passing filter reads and total Gb produced were used to characterize process performance before downstream analysis. Paired-end reads of 51 bases were acquired and analyzed as described previously[9]. Short sequence read data is available on the NCBI sequence read archive (accession numbers are listed in **Supplementary Table 1**) and on an independent site hosted by the Broad Institute and linked through the tuberculosis database (see URLs).

Data filtering and assembly. Two read lengths were generated (**Supplementary Table 1**). Prior to mapping or assembly, the 2 × 75-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 (ref. 10) 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. N50 is defined as the contig length such that using contigs of that length or greater accounts for half the bases of the genome. Each assembly was analyzed by pairwise comparison using the MUMmer script dnadiff[31]. 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 (ref. 8). 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 the remaining parameters were defaults from the script easyrun. For 51-bp paired-end reads, a minimum read depth of five was required to call SNPs. For 36-bp 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**) or by independent identification by WGS.

Statistical analysis and estimation of the *in vivo* mutation rate from WGS data. The mutation rate was estimated from the number of SNPs observed in each clinical condition. Our equations assume that both the mutation rate and growth rate are parameters that, although 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 1**.

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

The 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 mutations observed, N is determined based on 91% coverage of a 4.4-Mb genome ($N = 4 \times 10^6$), 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.

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

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 h to 240 h to capture the maximum range of biologically plausible generation times. All calculations were performed in Matlab (MathWorks). 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, and 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[18]. Briefly, 20 independent cultures of 1.08×10^9 cells each in 4 ml 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 2 μg/ml of rifampicin. The number of mutations per culture (m_{MSS}) was calculated from the distribution of mutants using the Ma, Sarkar, Sandri (MSS) method[16] calculated by the Matlab scripts previously described[17]. The MSS method for determining m has been shown previously to be the most robust over a broad range of values for m[16]. The phenotypic mutation rate was estimated by dividing m_{MSS} by the number of cells plated ($N_t = 1.08 \times 10^9$). 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[9]. 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.

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