nature genetics

Whole-genome sequencing of rifampicin-resistant *Mycobacterium tuberculosis* strains identifies compensatory mutations in RNA polymerase genes

Iñaki Comas[1,8], Sonia Borrell[2,3], Andreas Roetzer[4], Graham Rose[1], Bijaya Malla[2,3], Midori Kato-Maeda[5], James Galagan[6,7], Stefan Niemann[4] & Sebastien Gagneux[2,3]

© 2012 Nature America, Inc. All rights reserved.

Epidemics of drug-resistant bacteria emerge worldwide, even as resistant strains frequently have reduced fitness compared to their drug-susceptible counterparts[1]. Data from model systems suggest that the fitness cost of antimicrobial resistance can be reduced by compensatory mutations[2]; however, there is limited evidence that compensatory evolution has any significant role in the success of drug-resistant bacteria in human populations[3–6]. Here we describe a set of compensatory mutations in the RNA polymerase genes of rifampicin-resistant *M. tuberculosis*, the etiologic agent of human tuberculosis (TB). *M. tuberculosis* strains harboring these compensatory mutations showed a high competitive fitness *in vitro*. Moreover, these mutations were associated with high fitness *in vivo*, as determined by examining their relative clinical frequency across patient populations. Of note, in countries with the world's highest incidence of multidrug-resistant (MDR) TB[7], more than 30% of MDR clinical isolates had this form of mutation. Our findings support a role for compensatory evolution in the global epidemics of MDR TB[8].

The worldwide emergence of MDR strains of *M. tuberculosis* is threatening to make one of mankind's most pervasive infectious diseases incurable[8]. MDR strains of *M. tuberculosis* are resistant to isoniazid and rifampicin, the two most commonly used and effective drugs for the treatment of TB. Theoretical studies have predicted that one of the key factors driving the current MDR TB epidemics is the relative fitness of drug-resistant strains compared to drug-susceptible ones[9,10]. Experimental work has shown that drug resistance in bacteria is often associated with a fitness cost[1,2,11–13], but some drug-resistance mutations cause little or no loss of fitness[12,14,15]. Furthermore, fitness cost linked to drug resistance mutations can be reduced by compensatory evolution[2,14], although little data exist on the clinical relevance of this phenomenon[3,6]. In *M. tuberculosis*, compensatory mechanisms were identified for fitness defects related to isoniazid and aminoglycoside resistance[16,17], but the corresponding compensatory mutations are rare in clinical strains[18], suggesting that they have a minor role in the epidemiology of MDR TB. Compensatory evolution has been reported to occur in resistance to rifampicin in *Escherichia coli*[14], but little is known with respect to compensatory evolution in rifampicin-resistant *M. tuberculosis*.

Rifampicin binds to the β subunit of the RNA polymerase encoded by *rpoB* and inhibits transcription. More than 95% of *M. tuberculosis* clinical strains resistant to rifampicin harbor a mutation in an 81-bp region of *rpoB* known as the rifampicin resistance–determining region (RRDR)[18], and these mutations are associated with a high level of resistance to rifampicin. We have previously shown that all laboratory-generated mutants of *M. tuberculosis* with a rifampicin resistance–conferring mutation in the RRDR have reduced fitness compared to their drug-susceptible ancestors when grown in the absence of rifampicin[12]. By contrast, some *M. tuberculosis* clinical strains isolated from individuals with TB who developed rifampicin resistance during treatment showed no fitness cost compared to their rifampicin-susceptible counterparts, despite carrying the same *rpoB* mutation as some of the laboratory-derived strains[12]. At the time, we hypothesized that these clinical strains might have acquired compensatory mutations during the course of treatment.

Here we tested this hypothesis by comparing the genome sequences of ten paired clinical rifampicin-resistant isolates to the genomes of the corresponding rifampicin-susceptible isolates recovered from the same infected individual at an earlier time point (**Supplementary Table 1**)[12]. We identified all nonsynonymous and intergenic mutations found only in the rifampicin-resistant genomes (**Supplementary Table 2**). In addition, we experimentally evolved six laboratory-derived rifampicin-resistant mutants from rifampicin-susceptible ancestors[12] during 45 weeks of serial subculture in the absence of rifampicin (**Supplementary Table 3**). Comparison of the

[1]Division of Mycobacterial Research, Medical Research Council, National Institute for Medical Research, London, UK. [2]Department of Medical Parasitology and Infection Biology, Swiss Tropical and Public Health Institute, Basel, Switzerland. [3]University of Basel, Basel, Switzerland. [4]Molecular Mycobacteriology, Research Centre Borstel, Borstel, Germany. [5]Department of Medicine, San Francisco General Hospital, University of California, San Francisco, San Francisco, California, USA. [6]The Broad Institute of MIT and Harvard University, Cambridge, Massachusetts, USA. [7]Department of Microbiology, Boston University, Boston, Massachusetts, USA. [8]Current address: Genomics and Health Unit, Centre for Public Health Research, Valencia, Spain. Correspondence should be addressed to S.G. (sebastien.gagneux@unibas.ch).

Received 10 June; accepted 16 November; published online 18 December 2011; doi:10.1038/ng.1038

© 2012 Nature America, Inc. All rights reserved.

whole-genome sequences of the *in vitro*–evolved strains to their respective rifampicin-susceptible ancestors allowed us to identify putative compensatory mutations, as well as mutations likely to represent adaptations to growth in the laboratory (**Supplementary Table 4**). Of note, all of the *in vitro*–evolved rifampicin-resistant strains maintained their original *rpoB* mutation, which is consistent with a greater number of potential mutational targets resulting in compensation rather than reversion[2,19].

After combining our clinical and *in vitro* data and excluding mutations representing laboratory adaptations or phylogenetic markers (**Supplementary Tables 4** and **5**), we identified 54 putative compensatory mutations in 38 genes and 10 intergenic regions (**Supplementary Table 6**). *rpoA* and *rpoC* were notable in that they harbored multiple mutations in the laboratory-evolved strains (one strain) or the paired clinical strains (four strains; **Fig. 1** and **Table 1**). These genes encode the α and β′ subunits of the RNA polymerase, respectively. Based on the known interactions between the RpoA, RpoB and RpoC subunits[20], we reasoned that nonsynonymous changes in *rpoA* and *rpoC* occurring only in rifampicin-resistant genomes were likely to be compensatory. Mapping these amino acid substitutions encoded by these mutations onto the three-dimensional structure of the *E. coli* RNA polymerase[20] showed that they localized to the interface between the α and β′ subunits (**Fig. 2**), indicating that they potentially affect the interaction between these subunits.

In addition to these plausible effects on RNA polymerase structure, we expected compensatory mutations in rifampicin-resistant *M. tuberculosis* (i) to occur frequently in MDR clinical isolates and not in rifampicin-susceptible isolates, (ii) to be associated with mutations in the RRDR and (iii) to occur only in rifampicin-resistant strains with *rpoB* mutations. Because *M. tuberculosis* is genetically monomorphic[21] with no ongoing horizontal gene transfer[22,23], rates of convergent evolution in this microbe are in general extremely low[24,25]. Drug resistance–conferring mutations, however, undergo convergent evolution, as drug pressure selects for the same mutations across the different phylogenetic lineages of *M. tuberculosis*[26]. According to this rationale, we expected compensatory mutations to also show convergent evolution. Furthermore, because *M. tuberculosis* is genetically homogeneous, the occurrence of more than two amino acid variants at the same codon position is rare[27], except in the context of drug resistance[18]. Thus, we also expected particular codons involved in compensation to harbor multiple alleles, as several alternative amino acid substitutions might have similar compensatory effects.

To test these predictions, we screened four complementary panels of clinical *M. tuberculosis* strains for nonsynonymous changes in *rpoA* and *rpoC*. The first panel comprised 117 MDR strains from global sources, representing five of the six major lineages of human-adapted *M. tuberculosis*[28] (**Supplementary Table 7**). The second panel served as a control and included 131 rifampicin-susceptible strains representing the global diversity of *M. tuberculosis*[29,30]. The third panel consisted of 212 MDR clinical isolates from Abkhazia/Georgia, Uzbekistan and Kazakhstan (**Supplementary Table 7**)[31–33], regions that are among those with the highest MDR TB incidence in the world[8]. The fourth panel comprised 40 pan-susceptible isolates from Uzbekistan (**Supplementary Table 8**). All of the 329 MDR strains included in panels 1 and 3 had phenotypically confirmed rifampicin resistance, and 321 of 332 (99.7%) harbored at least one nonsynonymous mutation in the RRDR.

After excluding phylogenetic markers and mutations likely to have been caused by laboratory adaptation, we found that 89 of 329 of all MDR strains (27.1%) had a nonsynonymous mutation in *rpoA* or *rpoC*. In addition to the mutations already observed in our clinically paired or experimentally evolved strains, we found 28 previously unidentified nonsynonymous changes in these genes (**Table 1**). By contrast, none of the 171 rifampicin-susceptible control strains had any of these mutations. Furthermore, all MDR strains having an *rpoA* or *rpoC* mutation also had a mutation in the RRDR, whereas none of the 11 rifampicin-resistant strains without *rpoB* mutations had a mutation in *rpoA* or *rpoC*.

When combining these data, we found that 11 codon positions in *rpoA* and *rpoC* had the same putative compensatory mutations in more than one phylogenetic lineage of *M. tuberculosis*, and 8 codon positions were found to encode more than one amino acid change (**Fig. 1**). Such occurrences of multiple alterations in a single codon are only rarely observed in *M. tuberculosis* outside of drug resistance. It is particularly unlikely that positions would by chance show convergent evolution across different *M. tuberculosis* lineages and at the same time harbor multiple allelic variants. Hence, we focused the rest of our investigation on the mutations falling in codon positions that satisfied both of these criteria (**Table 1**).

We computationally predicted the effect of these high-probability compensatory mutations (HCMs) on protein function by comparing the degree of evolutionary conservation in other bacteria of the

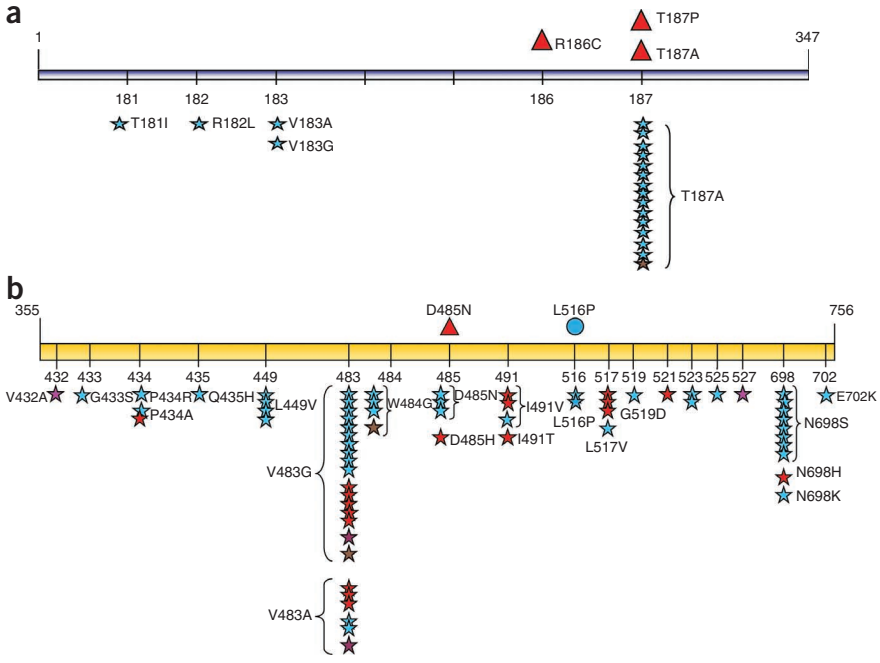

Figure 1 Putative compensatory mutations in *rpoA* and *rpoC* of *M. tuberculosis*. (**a**,**b**) Mutations identified after genome sequencing of experimentally evolved strains (circle) or paired clinical isolates (triangles) are indicated above the gene diagrams of *rpoA* (**a**) and *rpoC* (**b**). Mutations identified by screening a global and a high-burden collection of MDR strains are indicated by stars below the gene diagrams. Colors indicate the respective strain lineage (blue, lineage 2; red, lineage 4; brown, lineage 5; pink, lineage 1). Some of these mutations occurred in multiple lineages or affect the same codon position.

Table 1 Putative compensatory alterations in rifampicin-resistant *M. tuberculosis*

Gene	Encoded amino acid change	Experimentally evolved strains	Clinical paired isolates	Global MDR strains	High-burden MDR strains	Total strains	High-probability compensatory mutation	SIFT score
rpoA	R182L	0	0	1	0	1	No	0.00
rpoA	R186C	0	1	0	0	0	No	0.00
rpoA	T187P	0	1	0	0	0	Yes	0.00
rpoA	T181I	0	0	0	1	1	No	0.00
rpoA	T187A	0	1	5	10	15	Yes	0.00
rpoA	V183A	0	0	1	0	1	No	0.00
rpoA	V183G	0	0	1	0	1	No	0.00
rpoC	A521D	0	0	1	0	1	No	0.00
rpoC	D485H	0	0	0	1	1	Yes	0.00
rpoC	D485N	0	1	0	3	3	Yes	0.00
rpoC	E702K	0	0	0	1	1	No	0.00
rpoC	G433S	0	0	0	1	1	No	0.00
rpoC	G519D	0	0	0	1	1	No	0.00
rpoC	H525N	0	0	0	1	1	No	0.00
rpoC	I491T	0	0	0	1	1	Yes	0.00
rpoC	I491V	0	0	0	3	3	Yes	0.00
rpoC	L449V	0	0	0	5	5	No	0.00
rpoC	L516P	1	0	0	2	2	No	0.00
rpoC	N698H	0	0	0	1	1	Yes	0.00
rpoC	N698K	0	0	0	1	1	Yes	0.00
rpoC	N698S	0	0	1	7	8	Yes	0.00
rpoC	P434A	0	0	0	2	2	No	0.00
rpoC	P434R	0	0	1	0	1	Yes	0.00
rpoC	Q435H	0	0	0	1	1	No	0.00
rpoC	Q523E	0	0	0	1	1	No	0.00
rpoC	Q523K	0	0	0	1	1	No	0.00
rpoC	G433S	0	0	1	0	1	No	0.00
rpoC	V432A	0	0	1	0	1	No	0.00
rpoC	V483A	0	0	1	5	6	Yes	0.00
rpoC	V483G	0	0	6	11	17	Yes	0.00
rpoC	V517L	0	0	1	3	4	No	0.00
rpoC	L527V	0	0	1	0	1	No	0.00
rpoC	W484G	0	0	1	3	4	No	0.00

orthologous protein positions encoded by these mutations using SIFT scores[34]. For comparison, we used 13 publicly available mycobacterial genomes not belonging to the *M. tuberculosis* complex. As a proof of concept, we first tested whether we could correctly predict that mutations in *rpoB* that are known to confer rifampicin resistance (**Supplementary Table 7**) were more likely to be functional changes than were phylogenetic markers in the same gene (**Supplementary Table 5**) and found that we could accurately make this prediction ($P < 0.01$, Mann-Whitney *U*-test). In testing the HCMs in *rpoA* and *rpoC* (**Table 1**), we found that these mutations were also predicted to be more functional than phylogenetic markers found in the same genes ($P < 0.01$, Mann-Whitney *U*-test; **Supplementary Table 5**).

To test whether the predicted functional effects of HCMs correlated with strain fitness, we combined our new data on the occurrence of these mutations with our older data on the relative fitness of rifampicin-resistant *M. tuberculosis* strains[12]. We found that three of four clinical MDR strains with no competitive fitness cost harbored an HCM (**Fig. 3a**). By contrast, none of the six clinical strains with a statistically significant reduction in fitness had any HCM ($P < 0.05$, Fischer's exact test). Furthermore, when calculating the difference in fitness between the laboratory-derived rifampicin-resistant strains and the clinical strains with the same rifampicin resistance–conferring mutation and belonging to the same phylogenetic lineage, we found

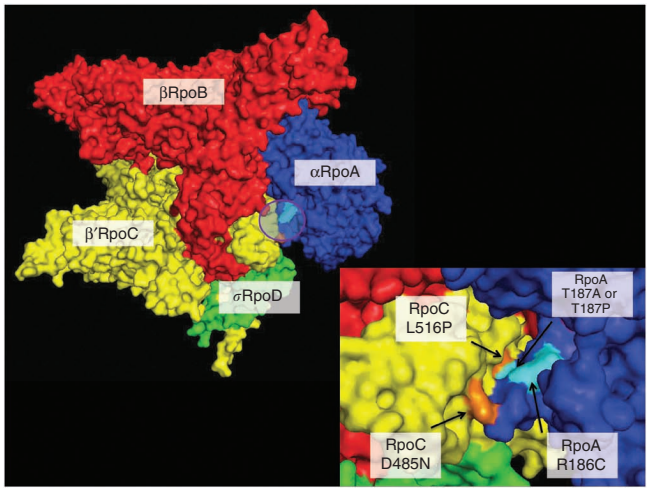

Figure 2 Putative compensatory mutations in *rpoA* and *rpoC* fall in regions encoding the interface of the RNA polymerase subunits. Amino acid substitutions identified in rifampicin-resistant experimentally evolved isolates and paired clinical isolates were mapped onto the structure of the *E. coli* RNA polymerase. The alterations are localized to residues of RpoA (light blue) and RpoC (orange) that are predicted to have roles in RNA polymerase subunit interaction. Residue numbers are indicated according to *M. tuberculosis* coordinates. RpoA (α subunit), blue; RpoB (β subunit), red; RpoC (β′ subunit), yellow; RpoD (σ subunit), green.

© 2012 Nature America, Inc. All rights reserved.

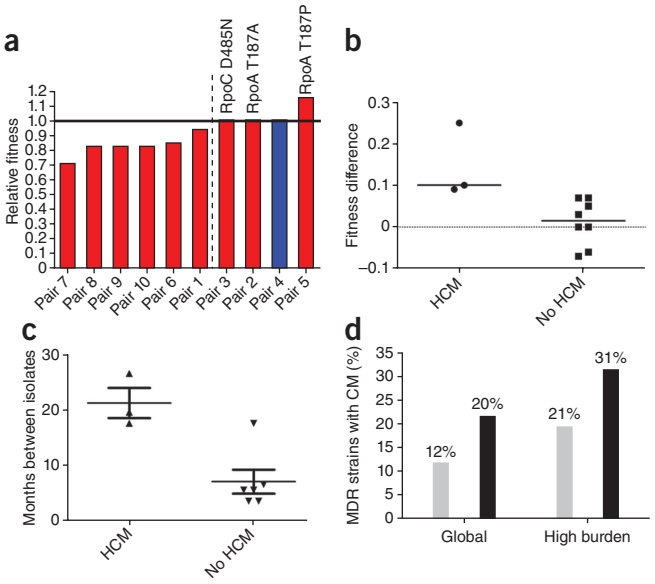

Figure 3 Experimental and clinical relevance of putative compensatory mutations. (**a**) Experimental competitive fitness of ten clinical isolates that acquired rifampicin resistance over the course of treatment compared to their susceptible counterparts. The amino acid changes encoded by HCMs are indicated in the pair in which they were identified. Bar colors indicate strain lineage (blue, lineage 2; red, lineage 4). (**b**) Difference in relative fitness between ten rifampicin-resistant paired clinical isolates compared to laboratory-generated mutants carrying the same rifampicin resistance–conferring mutation and with the same genetic background as defined by strain lineage. Data are shown for clinical strains with or without an HCM. Horizontal lines indicate median fitness differences. (**c**) Time in months between the isolation of the first and the second strain of each clinical pair. Horizontal lines indicate the median time intervals. (**d**) Percentage of MDR strains with putative compensatory mutations in *rpoA* or *rpoC*. Gray bars, the percentage of strains carrying HCMs; black bars, strains carrying any putative compensatory mutation. Data for a global collection of strains and for regions of Abkhazia/Georgia, Uzbekistan and Kazakhstan with high MDR TB burden are shown.

that there was always a bias toward increased fitness in the clinical strains with an HCM and that this difference was greater than the median difference in fitness among the other clinical strains ($P < 0.05$, Mann-Whitney U test; **Fig. 3b**). Finally, we found that the median time between the isolation of the susceptible clinical isolate and the rifampicin-resistant isolate from the same individual was longer for strains with an HCM than for strains carrying only an RRDR mutation (20 versus 6 months; $P < 0.05$, Mann-Whitney U-test; **Fig. 3c**). Taken together, these data show that HCMs in *rpoA* and *rpoC* are associated with high *in vitro* fitness of MDR clinical strains of *M. tuberculosis* and that the emergence of HCMs is time-dependent.

One could argue that because the three clinical strains harboring HCMs have accumulated additional mutations (**Supplementary Table 2**), the high fitness of these strains cannot be directly attributed to HCMs. Although not discarding the possibility of alternative compensatory mechanisms, at least for clinical pair 3 (**Fig. 3a**), we made several observations that support a causal relationship between the HCM found in *rpoC* in this isolate and increased fitness. Specifically, this strain contains only one additional nonsynonymous mutation compared to its rifampicin-susceptible ancestor (**Supplementary Table 2**). However, this additional mutation occurs in *aroG*, a gene that does not belong to the same functional class of transcriptional regulators, and no interactions of the AroG protein with RNA polymerase subunits are known according to the latest version of the STRING database[35]. Moreover, SIFT analysis[34] predicted that the *aroG* mutation encoding an M311T substitution would have no functional consequence (SIFT score = 1.00). Of note, several environmental mycobacteria also encode this same amino acid change, providing additional evidence against a role for this *aroG* mutation in compensatory evolution. In summary, the HCM in *rpoC* seems to be sufficient to confer the observed high fitness in this strain.

We and others have shown that, in the context of rifampicin resistance, *in vitro* competitive fitness for *M. tuberculosis* correlates with *in vivo* fitness as measured by the frequency of different rifampicin resistance–conferring mutations in clinical settings[12,13]. In other words, rifampicin resistance–conferring mutations associated with no or low fitness cost *in vitro* are the most frequent in clinical strains. Hence, in the context of rifampicin resistance, the clinical frequency of mutations can be used as a proxy measure for the *in vivo* fitness of drug-resistant *M. tuberculosis* strains among different patient populations[4,15]. When we determined the frequency with which HCMs occurred in MDR clinical strains, we found that 12.0% of our global panel of MDR isolates carried such a mutation (**Fig. 3d**). This proportion increased to 21.3% in our panel of strains from regions with a high MDR TB burden ($\chi^2 = 4.5$, $P < 0.05$). When we relaxed our selection criteria and repeated this analysis with all putative compensatory mutations in *rpoA* and *rpoC* (**Table 1**), we found that 19.7% of the global MDR strains carried such a mutation, compared to 31.3% of MDR strains from regions with high MDR TB burden ($\chi^2 = 5.14$, $P < 0.05$). The high frequency of compensatory mutations in strains from Abkhazia/Georgia, Uzbekistan and Kazakhstan is consistent with the success of MDR strains in these regions, where up to 50% of individuals with TB are estimated to carry MDR strains compared to a global average of only 3% (ref. 7).

In conclusion, our results suggest that the acquisition over time of particular mutations in *rpoA* and *rpoC* in rifampicin-resistant *M. tuberculosis* strains leads to the emergence of MDR strains with high fitness. Furthermore, our data show that these mutations occur at high frequencies in clinical settings, particularly in hotspot regions of MDR TB[9]. Additional studies are needed to determine whether MDR strains of *M. tuberculosis* with mutations in *rpoA* or *rpoC* have increased transmission rates and how these mutations contribute to the success of these strains. Use of targeted genotyping of these mutations will enable TB control programs to focus on the most transmissible MDR strains. Our findings also suggest that mathematical models that aim at predicting the future of the global MDR TB epidemic should take into account the effects of compensatory mutations as well as the time necessary for such mutations to emerge.

URLs. Sanger Institute unpublished sequencing data, http://www.sanger.ac.uk/resources/downloads/bacteria/mycobacterium.html.

METHODS

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

Accession numbers. Sequence data can be found in the Sequence Read Archive at the EBI (SRP001097).

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

ACKNOWLEDGMENTS

We thank P. Small and C. Davis Long for stimulating discussions at the onset of this project, D. Young and M. Coscolla for reviewing the manuscript, A. Candel for advice in the interpretation of the RNA polymerase molecular structure and

© 2012 Nature America, Inc. All rights reserved.

T. Van for technical support. We acknowledge the Wellcome Trust Sanger Institute for making available unpublished DNA sequence data (see URLs). We would like to thank T. Ubben, I. Razio and the other members of the German National Reference Center for Mycobateria for technical assistance and all partners that have contributed to previous studies in regions of high MDR TB incidence and collected some of the strains analyzed here. This project was funded wholly or in part by the US National Institute of Allergy and Infectious Disease, US National Institutes of Health and US Department of Health and Human Services (contract HHSN266200400001C, grants AI090928 and AI034238). This work was also supported by the Medical Research Council, UK (MRC_U117588500), the Swiss National Science Foundation (PP00A-119205) and the European Community LONG-DRUG (QLK-CT-2002-01612) and TB PAN-NET (FP7-223681) projects.

AUTHOR CONTRIBUTIONS

I.C., S.B. and S.G. planned the experiments. I.C., S.B., A.R., B.M., G.R., M.K.-M., J.G. and S.G. performed the experiments. I.C., S.B., A.R., G.R., S.N. and S.G. analyzed the data. I.C., S.B. and S.G. wrote the manuscript. All authors critically reviewed 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://www.nature.com/reprints/index.html.

© 2012 Nature America, Inc. All rights reserved.

1. Andersson, D.I. & Levin, B.R. The biological cost of antibiotic resistance. *Curr. Opin. Microbiol.* **2**, 489–493 (1999).
2. Andersson, D.I. & Hughes, D. Antibiotic resistance and its cost: is it possible to reverse resistance? *Nat. Rev. Microbiol.* **8**, 260–271 (2010).
3. Björkholm, B. *et al.* Mutation frequency and biological cost of antibiotic resistance in *Helicobacter pylori*. *Proc. Natl. Acad. Sci. USA* **98**, 14607–14612 (2001).
4. Böttger, E.C., Springer, B., Pletschette, M. & Sander, P. Fitness of antibiotic-resistant microorganisms and compensatory mutations. *Nat. Med.* **4**, 1343–1344 (1998).
5. Borrell, S. & Gagneux, S. Infectiousness, reproductive fitness and evolution of drug-resistant *Mycobacterium tuberculosis*. *Int. J. Tuberc. Lung Dis.* **13**, 1456–1466 (2009).
6. Nagaev, I., Bjorkman, J., Andersson, D.I. & Hughes, D. Biological cost and compensatory evolution in fusidic acid-resistant *Staphylococcus aureus*. *Mol. Microbiol.* **40**, 433–439 (2001).
7. World Health Organization. *Global Tuberculosis Control—Surveillance, Planning, Financing*. (WHO, Geneva, Switzerland, 2010).
8. Gandhi, N.R. *et al.* Multidrug-resistant and extensively drug-resistant tuberculosis: a threat to global control of tuberculosis. *Lancet* **375**, 1830–1843 (2010).
9. Blower, S.M. & Chou, T. Modeling the emergence of the 'hot zones': tuberculosis and the amplification dynamics of drug resistance. *Nat. Med.* **10**, 1111–1116 (2004).
10. Cohen, T. & Murray, M. Modeling epidemics of multidrug-resistant *M. tuberculosis* of heterogeneous fitness. *Nat. Med.* **10**, 1117–1121 (2004).
11. Mariam, D.H., Mengistu, Y., Hoffner, S.E. & Andersson, D.I. Effect of *rpoB* mutations conferring rifampin resistance on fitness of *Mycobacterium tuberculosis*. *Antimicrob. Agents Chemother.* **48**, 1289–1294 (2004).
12. Gagneux, S. *et al.* The competitive cost of antibiotic resistance in *Mycobacterium tuberculosis*. *Science* **312**, 1944–1946 (2006).
13. Billington, O.J., McHugh, T.D. & Gillespie, S.H. Physiological cost of rifampin resistance induced *in vitro* in *Mycobacterium tuberculosis*. *Antimicrob. Agents Chemother.* **43**, 1866–1869 (1999).
14. Reynolds, M.G. Compensatory evolution in rifampin-resistant *Escherichia coli*. *Genetics* **156**, 1471–1481 (2000).
15. Sander, P. *et al.* Fitness cost of chromosomal drug resistance–conferring mutations. *Antimicrob. Agents Chemother.* **46**, 1204–1211 (2002).
16. Sherman, D.R. *et al.* Compensatory *ahpC* gene expression in isoniazid-resistant *Mycobacterium tuberculosis*. *Science* **272**, 1641–1643 (1996).
17. Shcherbakov, D. *et al.* Directed mutagenesis of *Mycobacterium smegmatis* 16S rRNA to reconstruct the *in vivo* evolution of aminoglycoside resistance in *Mycobacterium tuberculosis*. *Mol. Microbiol.* **77**, 830–840 (2010).
18. Ramaswamy, S. & Musser, J.M. Molecular genetic basis of antimicrobial agent resistance in *Mycobacterium tuberculosis*: 1998 update. *Tuber. Lung Dis.* **79**, 3–29 (1998).
19. Levin, B.R., Perrot, V. & Walker, N. Compensatory mutations, antibiotic resistance and the population genetics of adaptive evolution in bacteria. *Genetics* **154**, 985–997 (2000).
20. Opalka, N. *et al.* Complete structural model of *Escherichia coli* RNA polymerase from a hybrid approach. *PLoS Biol.* **8**, e1000483 (2010).
21. Achtman, M. Evolution, population structure, and phylogeography of genetically monomorphic bacterial pathogens. *Annu. Rev. Microbiol.* **62**, 53–70 (2008).
22. Hirsh, A.E., Tsolaki, A.G., DeRiemer, K., Feldman, M.W. & Small, P.M. Stable association between strains of *Mycobacterium tuberculosis* and their human host populations. *Proc. Natl. Acad. Sci. USA* **101**, 4871–4876 (2004).
23. Supply, P. *et al.* Linkage disequilibrium between minisatellite loci supports clonal evolution of *Mycobacterium tuberculosis* in a high tuberculosis incidence area. *Mol. Microbiol.* **47**, 529–538 (2003).
24. Comas, I., Homolka, S., Niemann, S. & Gagneux, S. Genotyping of genetically monomorphic bacteria: DNA sequencing in *Mycobacterium tuberculosis* highlights the limitations of current methodologies. *PLoS ONE* **4**, e7815 (2009).
25. Hershberg, R. *et al.* High functional diversity in *Mycobacterium tuberculosis* driven by genetic drift and human demography. *PLoS Biol.* **6**, e311 (2008).
26. Hazbón, M.H. *et al.* Convergent evolutionary analysis identifies significant mutations in drug resistance targets of *Mycobacterium tuberculosis*. *Antimicrob. Agents Chemother.* **52**, 3369–3376 (2008).
27. Comas, I. *et al.* Human T cell epitopes of *Mycobacterium tuberculosis* are evolutionarily hyperconserved. *Nat. Genet.* **42**, 498–503 (2010).
28. Gagneux, S. & Small, P.M. Global phylogeography of *Mycobacterium tuberculosis* and implications for tuberculosis product development. *Lancet Infect. Dis.* **7**, 328–337 (2007).
29. Gagneux, S. *et al.* Variable host-pathogen compatibility in *Mycobacterium tuberculosis*. *Proc. Natl. Acad. Sci. USA* **103**, 2869–2873 (2006).
30. Wirth, T. *et al.* Origin, spread and demography of the *Mycobacterium tuberculosis* complex. *PLoS Pathog.* **4**, e1000160 (2008).
31. Kubica, T. *et al.* The Beijing genotype is a major cause of drug-resistant tuberculosis in Kazakhstan. *Int. J. Tuberc. Lung Dis.* **9**, 646–653 (2005).
32. Cox, H.S. *et al.* The Beijing genotype and drug resistant tuberculosis in the Aral Sea region of Central Asia. *Respir. Res.* **6**, 134 (2005).
33. Pardini, M. *et al.* Characteristics of drug-resistant tuberculosis in Abkhazia (Georgia), a high-prevalence area in Eastern Europe. *Tuberculosis (Edinb.)* **89**, 317–324 (2009).
34. Ng, P.C. & Henikoff, S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31**, 3812–3814 (2003).
35. Szklarczyk, D. *et al.* The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.* **39**, D561–D568 (2011).

ONLINE METHODS

Data sets. To identify the most likely mutations involved in the compensatory evolution of fitness cost associated with rifampicin resistance, we analyzed different data sets that allowed us to address different aspects of the evolution of these compensatory mutations and control for mutations unrelated to rifampicin resistance and compensation. The microbiological and molecular methods used to analyze these data sets are as follows.

Experimentally evolved strains. This data set is composed of 6 spontaneous rifampicin-resistant mutants obtained in the laboratory[12] and experimentally evolved during 45 weeks of growth in the absence of rifampicin to select for compensatory mutations. The genomes of the evolved strains were sequenced in this study.

Paired clinical strains. This data set is composed of 10 rifampicin-resistant clinical strains from 10 individuals with TB who developed resistance to rifampicin during treatment. The genomes of the 10 rifampicin-resistant strains, as well their rifampicin-susceptible counterparts, were sequenced in this study.

Global MDR strain collection. We analyzed 117 MDR strains from different geographic regions and representative of 5 of the 6 *M. tuberculosis* complex (MTBC) lineages[28] to look at the frequency of compensatory mutations at a global level and their occurrence across lineages. For these strains, we sequenced the complete *rpoA* gene as well as the region of *rpoC* encoding residues that interact with RpoA.

High-burden MDR-TB strain collection. We analyzed 212 MDR strains from three areas with a high burden of MDR-TB (104 strains from Nukus, Uzbekistan; 92 from Almaty, Kazakhstan; and 16 from Abkhazia/Georgia). For these strains, the complete *rpoA* gene and the region of *rpoC* encoding interaction sites with RpoA were sequenced.

Control data sets. To identify mutations unrelated to rifampicin resistance and compensation, we used four control data sets.

Global collection of drug-susceptible MTBC strains. To control for mutations representing phylogenetic markers, we used 21 previously published drug-susceptible MTBC genomes representative of all 6 major MTBC lineages[27]. In addition we used DNA sequence data from 110 drug-susceptible strains for which the whole genomes are currently being analyzed and extracted the sequence data for *rpoA*, *rpoB* and *rpoC*.

Laboratory-evolved, drug-susceptible strains. To control for mutations associated with adaptation to the laboratory, we experimentally evolved the two drug-susceptible ancestors of the *in vitro*-generated rifampicin-resistant mutants over a period of 45 weeks. Whole-genome sequencing was performed on these strains.

Rifampicin-susceptible paired clinical isolates. The 10 paired rifampicin-susceptible clinical strains from the individuals with TB who developed rifampicin resistance during treatment served as additional controls for common SNPs and other mutations unrelated to rifampicin resistance or compensation.

Local collection of drug-susceptible strains. To control for mutations unrelated to rifampicin resistance but associated with autochthonous lineages circulating in central Asia and to show that putative compensatory mutations do not occur in drug-susceptible strains, we sequenced the whole *rpoA* gene and the region of *rpoC* encoding the interaction sites with RpoA in 40 strains from Uzbekistan.

Experimental evolution. Spontaneous rifampicin-resistant mutants generated from the two drug-susceptible clinical MTBC strains CDC1551 (lineage 4) and T85 (lineage 2) were isolated as described before[12]. Six of these mutant clones, as well as the two ancestor strains, CDC1551 and T85, were cultured in 1l roller bottles with 100 ml of Middlebrook 7H9 broth (Difco) containing albumin-dextrose-catalase supplement (ADC; Difco) and 0.1% Tween at 37 °C. Cultures were subcultured 15 times every 3 weeks. For each subculture, 0.1 ml of growth culture (corresponding to ~5×10^7 bacteria) was transferred into 100 ml of fresh medium, corresponding to a dilution factor of 1:1,000. During the 45 weeks of the experiment, the overall number of bacterial generations was ~200. Samples were stored periodically at −80 °C and later revived for DNA extraction and genome sequencing.

Paired clinical isolates. Serial isolates were obtained from 10 individuals with TB who developed resistance to rifampicin during treatment as described[12]. For each isolate, single colonies were subcultured, and aliquots were prepared and stored at −80 °C. DNA was extracted and used for genome sequencing. For each pair of isolates, head-to-head competitions were carried out, the results of which have previously been published[12].

Use of clinical isolates. The clinical isolates analyzed in this study are part of several existing strain collections which were compiled in the past following standard-of-care procedures. Given the retrospective nature of the work involving only anonymized bacterial isolates, informed consent was not necessary for this particular study, according to the guidelines of the relevant Institutional Review Boards (IRBs) of the Swiss Tropical and Public Health Institute/University of Basel, the Medical Research Council UK, the Research Centre Borstel/University of Lübeck and University of California, San Francisco. A few strains included in this study were collected prospectively in Nepal. Informed consent was obtained from participants, and IRB approval for this prospective strain collection came from the Nepal Health Research Council and the Ethikkommission beider Basel (EKBB).

Genome sequencing. Bacterial strains were grown from single colonies. Genomic DNA was extracted using a standard kit (Qiagen), and DNA (2 µg) was sequenced using an Illumina Genome Analyzer. Sequencing libraries were constructed using standard kits from Illumina according to the manufacturer's instructions. Libraries for each strain were loaded into a single lane of a flow cell. SYBR green assays were used to test flow cells for optimal cluster density. Paired-end read sequencing was performed with read lengths of 76 bp. The mean number of reads generated per run was 17.6 million (range 4.4–26.7 million), which translated to a mean sequencing depth of 302 (range 77–512).

SNP calls from individual strains. We used mapping and assembly with qualities (MAQ) software to map the reads produced by the Illumina sequencer to the reference genome, which was that of the most recent common ancestor as determined by our previous work (note that this ancestor is H37Rv-like in its structure, but H37Rv alleles were substituted by those present in the inferred common ancestor of all MTBC lineages)[27]. For each strain, we used the same default values for mapping and SNP calling, which removed sequences that had base calls with Phred quality values of <30 or depth coverage of <5 or those that occurred in reads that mapped to nonunique sequences across the genome. Finally, we removed those cases that involved heterozygous calls and those that occurred in genes annotated as belonging to the PE or PPE gene families or related to mobile elements of the genome.

SNPs in the experimentally evolved strains. We produced individual high-confidence SNP calls for the mutants generated in the CDC1551 and T85 backgrounds. We pooled the SNPs from each background strain and generated a list of nonredundant positions, which we checked against those individual SNP calls that were filtered out during mapping to the reference genome. In this way, we could detect false negative SNPs that might have been removed during the initial filtering process.

SNPs in the paired clinical strains. For each isolate in a pair, we generated a list of high-confidence SNPs and checked all unique SNPs (those that were not shared by both isolates) against those that were previously filtered out to control for false negative SNP calls.

In total, we generated two final SNP lists for the experimentally evolved mutants (one for each background strain) and one SNP list for each of the ten rifampicin-resistant paired clinical strains. SNPs were annotated using the H37Rv genome annotation at NCBI (RefSeq NC_000962) and classified as synonymous, nonsynonymous or intergenic using ANNOVAR[36].

Mapping of the *rpoA*, *rpoB* and *rpoC* mutations onto the RNA polymerase molecular structure. The mutations identified in *rpoA* and *rpoC* in the experimentally evolved and paired clinical strains, as well as the rifampicin resistance–conferring mutations in *rpoB*, were mapped onto the three-dimensional structure of the RNA polymerase complex of *E. coli* (Protein Data Bank; 3LUO)[20]. To perform this mapping, an alignment of the corresponding nucleotide sequences in *M. tuberculosis* and the *E. coli* genome was generated

© 2012 Nature America, Inc. All rights reserved.

such that the homologous positions could be determined. The mutations were mapped and visualized using PyMOL (The PyMOL Molecular Graphics System, Version 1.2r3pre; Schrödinger). To determine the residues most likely to interact between the three subunits, we used the FoldX[37] program.

Predicting the functional effect of mutations. We used SIFT[34] to predict the mutations most likely to affect protein function. Briefly, SIFT looks for homologs in other bacteria for the gene of interest and (i) scores the conservation of the positions where mutations are found and (ii) weights this score by the nature of the amino acid change. These measures are then incorporated into a proxy measure of the impact of a specific mutation on protein function. As a bacterial database, we used the available non–*M. tuberculosis* complex complete mycobacterial genomes ($N = 13$).

DNA sequencing of *rpoA*, *rpoB* and *rpoC*. Identification of rifampicin resistance–conferring mutations in the RRDR of *rpoB* was performed using GenoType MTBDRplus kits (HAIN Lifesciences) according to the manufacturer's recommendations or by DNA sequencing as described before[12]. Oligonucleotide primers were designed for PCR amplification and sequencing of the entire *rpoA* locus (Rv3457c; 1044 bp; 347 aa), and of the region encoding the RpoA-RpoC in *rpoC* (Rv0668; aa 356–756). DNA was amplified in a 96-well format and 50-µl reactions were carried out on a Tprofessional thermocycler (Biolabo). PCR products were purified and sequenced commercially (Macrogen Korea). Sequence chromatogram files were analyzed using the Staden package[38]. In order to identify sequence polymorphisms, the consensus sequence for each strain was compared to the corresponding gene sequence of the H37Rv reference genome using MEGA 5 (ref. 39) software.

36. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
37. Guerois, R., Nielsen, J.E. & Serrano, L. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.* **320**, 369–387 (2002).
38. Bonfield, J.K., Smith, K. & Staden, R. A new DNA sequence assembly program. *Nucleic Acids Res.* **23**, 4992–4999 (1995).
39. Tamura, K., Dudley, J., Nei, M. & Kumar, S. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol. Biol. Evol.* **24**, 1596–1599 (2007).

© 2012 Nature America, Inc. All rights reserved.

doi:10.1038/ng.1038