Evolution and transmission of drug-resistant tuberculosis in a Russian population

Nicola Casali1, Vladyslav Nikolayevsky1, Yanina Balabanova1, Simon R Harris2, Olga Ignatyeva3, Irina Kontsevaya3, Jukka Corander4, Josephine Bryant2, Julian Parkhill2, Sergey Nejentsev5, Rolf D Horstmann6, Timothy Brown1 & Francis Drobniewski1,2

The molecular mechanisms determining the transmissibility and prevalence of drug-resistant tuberculosis in a population were investigated through whole-genome sequencing of 1,000 prospectively obtained patient isolates from Russia. Two-thirds belonged to the Beijing lineage, which was dominated by two homogeneous clades. Multidrug-resistant (MDR) genotypes were found in 48% of isolates overall and in 87% of the major clades. The most common rpoB mutation was associated with fitness-compensatory mutations in rpoA or rpoC, and a new intragenic compensatory substitution was identified. The proportion of MDR cases with extensively drug-resistant (XDR) tuberculosis was 16% overall, with 65% of MDR isolates harboring eis mutations, selected by kanamycin therapy, which may drive the expansion of strains with enhanced virulence. The combination of drug resistance and compensatory mutations displayed by the major clades confers clinical resistance without compromising fitness and transmissibility, showing that, in addition to weaknesses in the tuberculosis control program, biological factors drive the persistence and spread of MDR and XDR tuberculosis in Russia and beyond.

Tuberculosis is the second leading cause of death from an infectious disease after HIV. In 2011, there were an estimated 8.7 million new cases and 1.4 million deaths from the disease1. The increasing prevalence of drug resistance is a major public health concern that threatens the progress made in controlling drug-sensitive tuberculosis2. Globally, 4% of new cases and 20% of previously treated cases are infected with extensively drug-resistant (XDR) tuberculosis, which is MDR with resistance to any fluoroquinolone and at least one second-line injectable agent (kanamycin, amikacin or capreomycin), have now been found in every region of the world1. The high incidence of MDR and XDR tuberculosis in Samara3,16, a region with over 3 million citizens in Russia, afforded an opportunity to study the emergence and spread of antibiotic resistance within a population. In this prospective study, the largest of its kind through chromosomal mutation, typically resulting in a fitness cost seen as a reduced growth rate in vitro7. Fitness cost generally inversely correlates with the frequency of a mutation in clinical isolates8. Compensatory mutations mitigating the deleterious effects of resistance-conferring mutations are also important in determining the transmissibility of specific genotypes9. Studies using molecular epidemiological clustering rates to assess the transmission cost of resistance-associated genotypes report varying results10,11, suggesting that fitness costs may be affected by epistasis: that is, the phenotypic effect of a mutation depends on the presence or absence of other mutations in the same genome12.

Whole-genome sequencing offers the power to track the evolutionary mechanisms that promote the development and transmission of drug resistance within pathogen populations with unparalleled resolution. Evidence for adaptive selection in response to antibiotic therapy can be found by identifying homoplases (mutations arising independently multiple times within a phylogeny) or loci that are subject to frequent mutation13–15. By identifying all changes that occur in a genome, it is possible to uncover co-occurring polymorphisms that contribute to resistance phenotypes or signify epistatic effects12.

Received 16 September 2013; accepted 2 January 2014; published online 26 January 2014; doi:10.1038/ng.2878

1Public Health England (PHE) National Mycobacterium Reference Laboratory, Clinical TB and HIV Group, Blizard Institute, Queen Mary University of London, London, UK. 2Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK. 3Samara Oblast Tuberculosis Dispensary, Samara, Russian Federation. 4Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland. 5Department of Medicine, University of Cambridge, Cambridge, UK. 6Department of Molecular Medicine, Bernhard Nocht Institute for Tropical Medicine, Hamburg, Germany. 7Department of Infectious Diseases, Imperial College, London, UK. Correspondence should be addressed to F.D. (f.drobniewski@qmul.ac.uk).
yet reported, we used whole-genome sequencing to investigate the molecular mechanisms underlying resistance, fitness compensation and positive epistasis that together determine the transmissibility and prevalence of drug-resistant strains. Comparative analysis with XDR isolates from the UK addressed the origin of highly drug-resistant tuberculosis in this low-prevalence country.

RESULTS

Population structure

During a 2-year period (2008–2010), 2,348 \textit{M. tuberculosis} isolates were prospectively collected from individual patients living in Samara, Russia. The genomes of 1,000 isolates were sequenced. Comparison of patient epidemiological data indicated that this sample was representative of the entire population and covered the distribution of patients with tuberculosis across the whole region (Fig. 1 and Supplementary Table 1). For comparison, we selected 28 sequences from a study of over 2,000 London-based patients with tuberculosis, originating from 90 different countries, as representatives of the global population\cite{17,18}. We included five phenotypically XDR strains originating from 90 different countries, as representatives of the global population\cite{17,18}. We included five phenotypically XDR strains isolated from UK patients in 2011, as well as an isolate from Estonia representing the clone that dominates the \textit{M. tuberculosis} population of this country\cite{19}.

Mapping reads for each isolate against the reference sequence H37Rv identified a total of 32,445 SNPs in nonrepetitive regions of the genome, including 238 SNPs associated with drug resistance. These variable sites were used to reconstruct a maximum-likelihood phylogeny (Fig. 2 and Supplementary Data Set 1). Tree topology was consistent with the global phylogenetic structure of \textit{M. tuberculosis} sensu stricto comprising four main lineages\cite{20,21}. Of the 1,000 Samaran isolates, 642 belonged to the Beijing lineage, 355 belonged to the Euro-American lineage, 2 belonged to the Central Asian (CAS) lineage and a single isolate belonged to the East African–Indian (EAI) lineage, reflecting the proportions seen in the whole Samaran patient population.

In comparison to the disparate Beijing isolates from the UK, the majority of Samaran Beijing sequences formed a monophyletic group that we term the ‘East European sublineage’\cite{18} (Supplementary Fig. 1), consistent with a single relatively recent expansion of this lineage in the region. Bayesian population genetic clustering\cite{22} defined the two largest clones nested within the Beijing lineage—clade A and clade B—comprising 264 and 119 isolates, respectively (Supplementary Fig. 2). Outside of these clades, the majority of the remaining Beijing isolates were members of smaller clusters; 60\% (387/642) of all Beijing isolates differed from their last common ancestor by 5 or fewer SNPs (Supplementary Fig. 3). The western part of Samara is geographically isolated by the River Volga. Comparison of the isolate population in the west to that in the rest of the region showed a significant reduction in the proportion of clade A isolates (18\% versus 28\%, \textit{P} = 0.04; Supplementary Fig. 4 and Supplementary Table 2).

The Euro-American lineage showed significantly higher nucleotide diversity (\textit{\pi} = 0.0042, s.d. = 0.00018, 95\% confidence interval (CI) = 0.0038–0.0045) than the Beijing lineage (\textit{\pi} = 0.0022, s.d. = 0.00039, 95\% CI = 0.0015–0.0030). Fewer Euro-American isolates were members of closely related clusters; 39\% (137/355) differed from their last common ancestor by 5 or fewer SNPs (\textit{P} < 0.001). Although global diversity within the Euro-American lineage has not yet been well characterized, Homolka \textit{et al.}\cite{23} identified eight SNP-based sublineages that were broadly concordant with molecular fingerprint–based classifications. We identified four of these sublineages in Samara (Haarlem, LAM, Ural and S-type), whereas 36\% of isolates could not be classified by this scheme (Fig. 2).

Short genetic distances between isolates have been used to infer the likelihood of direct transmission\cite{24}. In our data set, four pairs of patients with sequenced isolates shared households. Within three

Figure 1 Coverage of the population of patients with tuberculosis. \textit{a} The locations of Samara Oblast in Russia (red) and the Baltic States (Lithuania (Li), Latvia (Lv) and Estonia (E)) are shown. \textit{b,c} The number of sequenced patient isolates from each territory (green) and city (blue; Samara City (Sm), Togliatti (T) and Syzran (Sz)) of Samara Oblast \textit{b} or district of Samara City \textit{c} is shown inside the corresponding circle. The area of each circle reflects coverage of the region (the number of isolates sequenced relative to the number of tuberculosis cases notified).

Figure 2 Maximum-likelihood phylogeny of 1,035 \textit{M. tuberculosis} isolates based on 32,445 variable sites. The four \textit{M. tuberculosis} lineages—Beijing, CAS, Euro-American and EAI—are indicated. The Euro-American SNP-defined sublineages\cite{25,26} and the major Beijing clades are shaded. The ancestral node of the Beijing East European sublineage is indicated with a star. Radial dotted lines show the positions of isolates from the UK; those with an XDR phenotype are marked with white circles. The Estonian strain is indicated by a filled blue circle. The position of the reference sequence H37Rv is marked by “R.” The East European sublineage, clade A and clade B had 100\% bootstrap support (Supplementary Data Set 1).
of these households, the patient isolates were almost identical, with zero, two and three SNPs separating each pair, consistent with intrahousehold transmission or infection from a common source. Patients in the fourth household were infected with unrelated isolates (183 SNPs different). In addition to the household pair, 30 further pairs of isolates, 3 clusters of 3 isolates and 1 cluster of 5 isolates had no SNP differences within the cluster. Patients with identical isolates were resident in the same region of Samara in only 20 of these 35 clusters. By mapping patient addresses, we found that patients with identical strains lived up to 136 km apart (Supplementary Fig. 5).

XDR tuberculosis isolates from four UK patients who originated from the Baltic States were members of the East European sublineage, with two belonging to clade B and one belonging to clade A, suggesting that infections may have been acquired during stays in their country of origin or within sympatric communities in the UK (Fig. 2). In support of this conjecture, the fifth XDR isolate, from a Chinese patient, was not a member of the East European sublineage. The Estonian isolate was also a member of clade B. Correlation of variable-number tandem repeat (VNTR) fingerprint data indicated that this strain is the same one that predominates across northwestern Russia. The East European isolates were remarkably closely related, with isolates from different countries separated by as few as 13 SNPs (Supplementary Table 3).

Prevalence of MDR genotypes

A maximum-likelihood phylogeny of the 1,000 Russian isolates was reconstructed and annotated with drug resistance genotypes (Fig. 3 and Supplementary Tables 4–6). The most commonly mutated drug resistance locus, in codon 315 of \( \text{katG} \), which confers high-level resistance to isoniazid, was substituted in 74% (478/642) of Beijing isolates and in 30% (106/355) of Euro-American isolates (\( P < 0.001 \); Fig. 4 and Supplementary Table 7). A new nonsense SNP in codon 668 also mediated resistance, consistent with the requirement of KatG for activation of the pro-drug. All (478) of the Beijing \( \text{katG} \) mutants encoded a p.Ser315Thr substitution, whereas 11% (12/106) of Euro-American isolates encoded 1 of 3 alternative substitutions in codon 315 (\( P < 0.001 \)).

A total of 70 isolates had mutations in the promoter of the \( \text{fabG1-inhA} \) operon that confer low-level cross-resistance to isoniazid and its structural analogs, ethionamide and prothionamide. As determined using the phylogenetic reconstruction, the \( \text{inhA} \) mutation arose in a \( \text{katG} \) mutant in 44 of the isolates. It is improbable that isoniazid therapy would select for mutations conferring low-level resistance in the presence of a SNP conferring high-level resistance, indicating that the majority of these promoter SNPs were acquired in response to therapy with ethionamide or prothionamide.

Mutations in the 81-bp rifampicin resistance–determining region (RRDR) of \( \text{rpoB} \) are accurate predictors of rifampicin resistance in \( M. \text{tuberculosis} \). Within this region, we identified 20 nonsynonymous SNPs and 2 small deletions, which affected 70% (430/642) of Beijing and 19% (67/355) of Euro-American isolates (\( P < 0.001 \)). The most common rifampicin resistance genotype, encoding a p.Ser450Leu substitution, was found in 90% (390/435) of Beijing isolates with RRDR mutations and in 67% (45/67) of Euro-American isolates (\( P < 0.001 \)).
emerging significantly more frequently in this genetic background (16/435 versus 4/565, \( P = 0.001 \)). Mutations affecting the most commonly substituted residue, Thr187, were acquired independently at least seven times, providing strong evidence that \( \text{rpoA} \) is subject to positive selection.

On the basis of these genotypes, 66% (422/642) of Beijing and 17% (61/355) of Euro-American isolates have a predicted MDR phenotype (\( P < 0.001 \)). The proportion of isolates that were MDR in clade A and clade B combined was significantly higher than for the rest of the Beijing lineage (332/383 versus 90/259, \( P < 0.001 \)).

Compensatory mutations in MDR isolates

We investigated the occurrence of compensatory mutations in \( \text{rpoA} \) and \( \text{rpoC} \) in isolates carrying rifampicin resistance mutations. We identified 14 different nonsynonymous SNPs in \( \text{rpoA} \), 11 of which were found in isolates containing the \( \text{rpoB} \) mutation encoding p.Ser450Leu (equivalent to the p.Ser531Leu substitution in \( \text{Escherichia coli} \)).

![Figure 4](image-url)

Figure 4 Prevalence of drug resistance mutations and association with lineage. The proportion of isolates harboring polymorphisms at each drug resistance locus was subdivided by lineage. Asterisks indicate significant differences between lineages (\( P < 0.05 \); Supplementary Table 7). Data are based on the polymorphisms detailed in Supplementary Table 4.

![Figure 5](image-url)

Figure 5 Distribution of rifampicin resistance and compensatory amino acid substitutions in the RNA polymerase subunits RpoA, RpoB and RpoC. Resistance-conferring mutations are clustered in the RRDR region of \( \text{rpoB} \). All other substitutions depicted are putative compensatory alterations that co-occurred with the p.Ser450Leu alteration in \( \text{rpoB} \).
to positive selective pressure. Clustering of SNP sites within three small regions suggests that the encoded residues are important for interaction with the rifampicin-binding pocket (Fig. 5). Eighteen of the 58 nonsynonymous SNPs identified in rpoC were homoplasic. The nonsynonymous SNPs in rpoC were significantly more likely to arise in isolates harboring an rpoB mutation encoding p.Ser450Leu than in those with wild-type RRDR or other resistance-conferring mutations (76/435 versus 11/565, \( P < 0.001; \) Supplementary Table 8). In total, 36 of the 59 amino acid substitutions in RpoA or RpoC found to be associated with the p.Ser450Leu alteration in RpoB had not previously been reported\(^{18,29}\).

Overall, 47\% (170/360) of isolates harboring the mutation encoding p.Ser450Leu had a putative compensatory mutation in rpoA or rpoC. However, such compensatory mutations were remarkably infrequent in clade A (Fig. 3 and Supplementary Fig. 6): 89\% (150/169) of non-clade A Beijing isolates harboring p.Ser450Leu had a nonsynonymous SNP in rpoA or rpoC compared to 9\% (20/221) of clade A isolates (\( P < 0.001 \)). This difference in frequency was unexpected given the obvious epidemiological success of this clade. Thus, we predicted that the large distal cluster of clade A isolates carrying an rpoB mutation encoding p.Ser450Leu harbored alternative compensatory mutations that restored fitness. By inspection of the phylogeny, we deduced the branch on which this putative mutation likely occurred; one of the four SNPs on this branch resulted in a p.Glu761Asp substitution in RpoB (Supplementary Fig. 6). Intragenic compensatory mutations in rpoB harboring resistance-conferring mutations have been observed in experimentally evolved \( E.\ coli\)\(^{30}\), \( Pseudomonas aeruginosa\)\(^{31}\) and \( Salmonella enterica\)\(^{32}\). We surmise that the p.Glu761Asp substitution provides an analogous fitness benefit in \( M.\ tuberculosis\) strains carrying the rpoB mutation encoding p.Ser450Leu.

In addition to the p.Glu761Asp substitution and excluding RRDR polymorphisms, a further 26 nonsynonymous SNPs were identified in rpoB. Sixteen of these co-occurred with a mutation encoding p.Ser450Leu (Fig. 5) and were significantly more likely to be found in isolates without alternative compensatory mutations (24/37 versus 2/396, \( P < 0.001 \)). Multiple substitution events in codons 496 and 835 provide confirmation that regions of rpoB other than the RRDR are under selective pressure. Including all putative compensatory SNPs, 97\% (421/435) of isolates carrying an rpoB mutation encoding p.Ser450Leu harbored additional rpoABC mutations, which did not appear in any isolates with wild-type RRDR and may mitigate the deleterious effect of this resistance-conferring mutation.

Sherman et al.\(^{33}\) proposed that loss of catalase-peroxidase function in isoniazid-resistant \( kat\)G mutants was compensated by upregulation of alkyl hydroperoxidase, \( ahp\)C. We identified four polymorphic sites within the \( ahp\)C regulatory region, including two homoplasics. Of the nine isolates harboring \( ahpC\) SNPs, four carried a \( kat\)G mutation encoding p.Ser315Thr, four had a rare \( kat\)G mutation (encoding p.Ser315Gly or p.Trp668*) and one had wild-type \( kat\)G. The \( ahpC\) SNPs were significantly more likely to arise in isolates with unusual \( kat\)G mutations (2/12 versus 3/572, \( P = 0.004 \)), supporting the theory that the p.Ser315Thr alteration has low or no fitness cost\(^{34}\).

**Prevalence of XDR genotypes**

We ascertained that 17\% (71/422) of Beijing MDR and 7\% (4/61) of Euro-American MDR isolates had genotypes predicting an XDR phenotype (\( P = 0.046 \)). Mutations in the \( eis\) promoter that confer kanamycin resistance\(^{35}\) were found in 66\% (317/483) of all MDR isolates. Eight different sites were polymorphic, including five homoplasics. \( eis\) mutations were significantly more common in isolates belonging to clade A or clade B compared to the rest of the Beijing lineage (275/332 versus 21/90, \( P < 0.001 \)). Mutations in the drug target \( rrs\), which confer resistance to amikacin and capreomycin, as well as to kanamycin\(^{36}\), were found in only 40 isolates, 4 of which had a preexisting SNP in the \( eis\) promoter.

Fluoroquinolones target the DNA gyrase \( GyrA\) and \( GyrB\), and resistance is conferred by mutations affecting the quinolone resistance–determining regions (QRDRs) that interact with the drugs\(^{37}\). Eighty-six isolates harbored mutations affecting the QRDR of \( GyrA\), and 11 Beijing isolates had SNPs affecting the \( GyrB\) QRDR. Substitutions in \( gyrAB\) arose relatively more often in isolates with \( rrs\) mutations than in those with \( eis\) promoter mutations (15/40 versus 62/317, \( P = 0.009; \) Supplementary Fig. 7).

In addition to isolates with fluoroquinolone resistance genotypes, we noted a significant number of isolates with ambiguous base calls within the sequences encoding QRDRs (Supplementary Table 9). This phenomenon was almost never observed at other drug resistance loci. For 18 isolates with ambiguous \( gyrA\) genotypes, ambiguity was often apparent at more than one of the codons (codons 90, 91 and 94) that most commonly confer resistance (Supplementary Fig. 8). In inspecting raw sequencing reads mapping to this region, we did not find multiple substitutions that were present on a single read, indicating that multiple fluoroquinolone-resistant clones coexisted in a single patient. Fluoroquinolone treatment of non-tuberculous infections could drive the acquisition of fluoroquinolone resistance in patients chronically infected with tuberculosis\(^{38}\). If this were the case, we would expect to see resistance in non-MDR isolates; however, we found no resistant or heterogeneous QRDR genotypes in non-MDR isolates, indicating that fluoroquinolone exposure occurred as part of tuberculosis therapy.

**Adaptive selection at other drug resistance loci**

Repeated independent acquisition of SNPs, identified by phylogenetic homoplasy, provides strong evidence of selection\(^{13}\) (Supplementary Table 10). Farhat et al.\(^{14}\) recently identified 22 new genomic regions that were targets of positive selection in drug-resistant strains (excluding repetitive regions), including 4 that harbored homoplaxies in our data set. Using a complementary approach, Zhang et al.\(^{15}\) identified 98 new regions that were enriched for SNPs in drug-resistant versus drug-sensitive strains, including 11 in which we identified homoplaxies.

\( emb\)B harbored homoplasic mutations in codons 306, 406 and 497 that are commonly associated with ethambutol resistance, although discordance with susceptibility testing is reported\(^{39}\). We identified homoplasic mutations in five additional \( emb\)B codons (Table 1). Surprisingly, the most frequent homoplasic \( emb\)B substitution was p.Asp354Ala, which affected a large cluster of clade A isolates as well as two unrelated isolates. This unusual alteration was associated with phenotypic ethambutol resistance in 50\% (83/166) of the isolates tested. Multiple acquisitions of SNPs in the region upstream of \( emb\)B provide evidence that operon upregulation also confers resistance\(^{40}\). Promoter and coding SNPs frequently co-occurred, and isolates with two mutations were more often phenotypically resistant (27/28 versus 199/375, \( P < 0.001; \) Supplementary Table 5), offering an explanation for the poor concordance between \( emb\)B codon 306 mutations and phenotypic resistance\(^{39}\).

The pyrazinamide resistance gene \( pncA\)\(^{41}\) was the most variable gene in the genome (Table 2 and Supplementary Fig. 9). In addition, its promoter harbored five different mutations. \( gid\)B was the second most polymorphic gene, indicating that it is a target of selective
Table 1 Ethambutol resistance mutations

| Locus | Mutation | Substitution     | Number of isolates | Number of acquisitions | Additional alterations |
|-------|----------|------------------|--------------------|------------------------|-----------------------|
| P<sub>embB</sub> | c.–16C>T, c.–16C>A, c.–16C>G | – | 15,7 | 9, 3, 1 | p.Met306Ile (4), p.Aspp354Ala (9) |
| | c.–15C>G | – | 1 | 1 | – |
| | c.–12C>T | – | 9 | 8 | p.Met306Val (1), p.Aspp354Ala (2), p.Gln497Arg (1) |
| | c.–11C>A | – | 1 | 1 | – |
| | c.–8C>A | – | 7 | 1 | p.Aspp354Ala (7) |
| embB | c.916A>G, c.916A>C | p.Met306Ile, p.Asp354Ala | 23 | 14, 9 | p.Gln497Arg (4), c.–12C>T (3), c.–16C>A (1) |
| | c.918G>A, c.918G>C, c.918G>T | p.Met306Val | 114 | 29, 2 | c.–12C>T (1), p.Aspp354Ala (2) |
| | c.956A>C, c.956A>G | p.Tyr319Ser, p.Tyr319Cys | 3 | 1, 2 | – |
| | c.1061A>C | p.Aspp354Ala | 213 | 3 | – |
| | c.1133A>C | p.Glu378Ala | 2 | 2 | – |
| | c.1217G>A, c.1217G>C | p.Gly406Asp, p.Gly406Ala | 16 | 6, 4 | – |
| | c.1489C>A | p.Gln497Ly | 9 | 3 | – |
| | c.1490A>G | p.Gln497Arg | 27 | 11 | – |
| | c.3005A>G | p.His1002Arg | 3 | 3 | – |
| | c.3070G>A | p.Asp1024Asn | 3 | 2 | – |

<sup>a</sup>High-confidence mutation in TB Drug Resistance Mutation Database (see URLs). <sup>b</sup>Number of times the mutation independently arose. <sup>c</sup>The number of isolates with the additional alteration is given in parentheses.

Transmissibility of drug resistance

Previous studies relied upon fingerprint clustering to estimate the transmission dynamics of drug resistance genotypes. Employing a similar principle but applying the improved resolution of whole-genome sequencing, we investigated transmissibility by using the phylogeny to estimate the number of isolates that independently acquired a SNP versus the number of isolates that inherited that SNP from an inferred common ancestor (indicating primary resistance).

SNPs conferring resistance to rifampicin, isoniazid, streptomycin and ethambutol were significantly more likely to be found in phylogenetic clusters than not (P < 0.05; Fig. 6 and Supplementary Table 11).

Table 2 Mutations in highly polymorphic genes implicated in drug resistance

| Locus | Drug resistance | Gene length (bp) | Nonsynonymous SNPs | Nonsense SNPs | Indels | Large deletions |
|-------|-----------------|-----------------|--------------------|---------------|--------|----------------|
| pncA  | Pyrazinamide    | 561             | 79                 | 1             | 16     | 5              |
| gidB  | Streptomycin    | 675             | 28                 | 2             | 9      | 1              |
| ethA  | Ethionamide     | 1,470           | 33                 | 6             | 19     | 2              |

Mutation frequency was determined by calculating the number of nonsynonymous SNPs per gene, adjusted for gene length.

<sup>a</sup>All indels resulted in frameshifts.

By inferring the order of SNP acquisition events from the phylogenetic tree, we determined that, for 97% (481/495) of isolates with RRDR SNPs, a katG mutation affecting Ser315 occurred before or on the same branch as the RRDR SNP. Hence, phylogenetic clustering of rpoB SNPs is essentially a surrogate for the clustering of MDR genotypes.

The most frequently acquired resistance genotype was for pyrazinamide; the modal number of isolates harboring each pncA coding or promoter polymorphism was 1 (65 of 106 clusters), and only 1 mutation was shared by a phylogenetic cluster of more than 7 isolates. The common pncA nonsynonymous SNP, encoding the conservative substitution Ile69Leu, was found in 157 clade A isolates. This SNP was not associated with phenotypic pyrazinamide resistance in vitro (Supplementary Table 5); however, no secondary pncA mutations were identified in these isolates.

SNPs in gyrA were significantly less likely to be found in clusters (P = 0.006; Supplementary Table 11); the largest cluster contained 5 isolates, and 59% (52/88) of resistant isolates did not cluster. In contrast, eis promoter SNPs were typically found in large clusters, the most notable of which was a cluster of 207 clade A isolates.
DISCUSSION
In the largest bacterial whole-genome sequencing project reported so far, we provide a region-wide snapshot of the *M. tuberculosis* population in Samara, Russia. Circulating strains belonged mainly to two lineages: Beijing and Euro-American. In concordance with other Russia-based studies, the Beijing lineage was dominant and accounted for two-thirds of isolates. Relative to isolates from the Beijing lineage, Euro-American isolates were phylogenetically diverse, and tree topology supports the division of this lineage into multiple sublineages. Samaran Beijing isolates were essentially monophyletic with respect to isolates representing the global population, forming a group we term the East European sublineage, which was dominated by two clades of extremely limited diversity. Short genetic distance between tuberculosis isolates has been used to infer transmission links. In this population, we found large geographic distances within even identical clusters, suggesting that short SNP distances did not always reflect transmission events or that the importance of casual contact may be underestimated.

Genotypes conferring drug resistance were extremely common; 48% of isolates had an MDR genotype, and 16% of these were XDR. In comparison to a small microbiologically based study conducted in Samara in 2001 (ref. 45), the proportion of MDR isolates resistant to amikacin has remained relatively stable (7.2% versus 8.5%), whereas the frequency of fluoroquinolone resistance has risen substantially from 4.3% to 23.8%. However, fluoroquinolone resistance was significantly more likely to be acquired than other resistance genotypes, indicating that gyrAB mutants may have impaired transmission fitness, impeding the spread of XDR clones. Current rates of resistance support the continued use of fluoroquinolones together with amikacin or capreomycin, rather than kanamycin, for MDR therapy. Whereas a fluoroquinolone in combination with prothionamide was highly effective in the treatment of MDR strains in Lithuania, which has a comparable MDR tuberculosis problem and a similar historical treatment strategy, the frequency of ethA mutations in Russian isolates means that thioamides may be ineffective in this population.

Drug resistance has previously been associated with the Beijing lineage. Here we provide evidence that Beijing isolates are more likely to harbor the isoniazid resistance genotype, conferred by a *katG* mutation encoding p.Ser315Thr, that has a negligible fitness cost, and the rifampicin resistance genotype, conferred by an *rpoB* mutation encoding p.Ser450Leu, that we find strongly associated with putative compensatory mutations within the RNA polymerase genes. Other studies have shown that *rpoC* mutations restored fitness in competitive growth assays and were significantly more common in isolates belonging to a fingerprint cluster than in non-clustered isolates. These observations explain, at least in part, the predominance of Beijing MDR isolates. The widespread use of kanamycin in MDR tuberculosis therapy may further exacerbate the spread of MDR strains by selecting strains with *eis* mutations that both confer resistance and increase bacterial multiplication in host macrophages by disrupting the protective immune response. Notably, *eis* mutations were significantly associated with the dominant Beijing clades.

The epidemiological success of clade A strains was particularly striking, and the ‘comb-like’ structure of the tree, apparent in the distal portion of the clade, suggests a highly infectious clone. The incomplete dispersion of clade A to western Samara supports its recent and rapid spread. From this success, we deduced the presence of unidentified fitness-enhancing mutations, which led to the discovery of new compensatory mutations in *rpoB* associated with rifampicin resistance. In addition, we surmised that a new SNP upstream of *ethA* likely confers low-level clinical resistance to ethionamide, not detected in vitro, through promoter disruption or enhanced binding of the EthR repressor. Typically, the most common resistance-conferring SNPs are associated with the least fitness cost, implying that successful clones would carry these SNPs. Thus, the discovery of this new *ethA* promoter SNP and the rare *embB* nonsynonymous SNP encoding p.Asp354Ala in clade A was unexpected.

The majority of clade A isolates encoded a conservative p.Ile6Leu substitution in the pyrazinamide resistance gene *pncA*, which does not confer resistance in vitro and is the only nonsynonymous SNP in *pncA* found in a large cluster. The deduced frequency of acquired pyrazinamide resistance in the sequenced population, together with the small cluster size associated with primary resistance, suggests that mutants with non-functional PncA have impaired transmission efficiency. Given the prevalence of pyrazinamide resistance in the population, it is difficult to reconcile the epidemiological success of clade A with a pyrazinamide-sensitive phenotype. We speculate that the p.Ile6Leu substitution results in intermediate PncA activity that manifests as pyrazinamide sensitivity in vitro but as clinical resistance in vivo and allows retention of sufficient nicotinamidase activity for efficient transmission.

We propose that the unusual combination of resistance-conferring and compensatory mutations acquired by clade A comprise a ‘perfect storm’, providing clinical drug resistance without compromising fitness and transmissibility.

Preventing tuberculosis transmission relies on accurate and rapid diagnosis. Molecular methods that reduce the time taken to detect drug-resistant tuberculosis strains expedite the institution of effective therapy and efficient infection control measures, thus minimizing the infectious period. Whole-genome sequencing offers the potential for rapid, unambiguous determination of all existing clinically relevant drug resistance mutations, and decreasing costs have neutralized a major argument over the value of targeted sequencing relative to whole-genome sequencing. We have reported the existence of multiple mutations in some resistance loci, as well as lineage- and clade-specific association of certain mutations that are suggestive of undiscovered epistatic interactions. These observations indicate that drug resistance may be more multifactorial than previously appreciated, which may in some cases explain discordance between phenotypes and genotypes. As more resistance-conferring loci are identified and the phenotypic effects of multiple mutations and strain background are elucidated, the public health value of routine whole-genome sequencing for the diagnosis of drug resistance will increase, although it may vary depending on prevalence and likely exposure to resistant strains.

We have reported, as have others, the extensive prevalence of MDR Beijing isolates in Russia and the former Soviet states of Eastern Europe, which shared a common treatment and Bacille Calmette–Guérin (BCG) vaccination strategy. In addition to programmatic and clinical weaknesses, we have identified plausible biological mechanisms contributing to the devastating MDR tuberculosis situation in the region. The current dominance of clade B across Eastern Europe and the isolation of both clade A and clade B strains in the UK indicate that the situation throughout the European Union could follow that in the East.

**URLs.** TB Drug Resistance Mutation Database, [http://www.tbdreamdb.com/](http://www.tbdreamdb.com/); SMALT, [http://www.sanger.ac.uk/resources/software/smalt](http://www.sanger.ac.uk/resources/software/smalt); FigTree, [http://tree.bio.ed.ac.uk/software/figtree](http://tree.bio.ed.ac.uk/software/figtree).

**METHODS**

Methods and any associated references are available in the online version of the paper.
Accession codes. Raw sequence data have been submitted to the European Nucleotide Archive under accession ERP000192.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We are grateful to members of the Public Health England National Mycobacterium Reference Laboratory and the Samara Regional Tuberculosis Laboratory for bacteriological work, particularly M. Stone, X. Gonzalez and A. Broda. We would also like to thank the Samara Tuberculosis Service, particularly I. Fedorin, as well as V. Kulichenko. We thank R. Hooper for expert statistical advice, S. Hofner for bacteriological advice and S. Bentley for sequencing advice. This study was supported by European Union Framework Programme 7 (grant 201483; TB-EUROGEN), with sequencing funded by the Wellcome Trust (grant 098051) and EUROGEN. S.N. is a Wellcome Trust Senior Research Fellow in Basic Biomedical Science (095198/Z/10/Z) and is also supported by European Research Council Starting Grant StG-260477.

AUTHOR CONTRIBUTIONS

N.C., V.Y. and F.D. designed the study. O.I., I.K., V.N. and Y.B. recruited patients and collected epidemiological data. O.I. and F.D. performed laboratory work. N.C., S.R.H. and J.C. conducted sequence analysis. N.C., T.B., V.N. and F.D. drafted the manuscript. T.B., V.N., Y.B., O.I., I.K., S.R.H., J.P., I.B., S.N. and R.D.H. provided critical analysis and reviewed the manuscript. All authors approved the final draft.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. World Health Organization. Global Tuberculosis Report (World Health Organization, Geneva, 2012).
2. Gandhi, N.R. et al. Multidrug-resistant and extensively drug-resistant tuberculosis: a threat to global control of tuberculosis. Lancet 375, 1830–1843 (2010).
3. Balabanova, Y. et al. Survival of civilian and prisoner drug-sensitive, multi- and extensively drug-resistant tuberculosis cohorts prospectively followed in Russia. PLoS ONE 6, e20531 (2011).
4. Balabanova, Y. et al. Survival of drug resistant tuberculosis patients in Lithuania: retrospective national cohort study. BMJ Open 1, e000351 (2011).
5. Health Protection Agency. Tuberculosis in the UK. 2012 Report (Health Protection Agency, London, 2012).
6. Udwadia, Z.F., MDR, XDR, TDR tuberculosis: ominous progression. Thorax 67, 286–298 (2012).
7. Andersson, D.I. & Hughes, D. Antibiotic resistance and its cost: is it possible to reverse resistance? Nat. Rev. Microbiol. 8, 260–271 (2010).
8. Böttger, E.C. & Springe, B. Tuberculosis drug resistance, fitness, and strategies for global control. Eur. J. Pediatr. 167, 141–148 (2008).
9. de Vos, M. et al. Putative compensatory mutations in the rpsO gene of rifampicin-resistant Mycobacterium tuberculosis are associated with ongoing transmission. Antimicrob. Agents Chemother. 57, 827–832 (2013).
10. Dye, C., Williams, B.G., Espinal, M.A. & Ravignione, M.C. Erasing the world’s slow work. N. Engl. J. Med. 362, 142–151 (2012).
11. Dye, C., Williams, B.G., Espinal, M.A. & Ravignione, M.C. Erasing the world’s slow work. N. Engl. J. Med. 362, 142–151 (2012).
12. Klevmark, J. & Wren, B.W. Bacterial epidemiology and biology—lessons from genome sequencing. Genome Biol. 12, 230 (2011).
13. Farhat, M.R. et al. Genomic analysis identifies targets of convergent positive selection in drug-resistant Mycobacterium tuberculosis. Nat. Genet. 45, 1183–1189 (2013).
14. Zhang, H. et al. Genome sequencing of 161 Mycobacterium tuberculosis isolates from China identifies genes and intergenic regions associated with drug resistance. Nat. Genet. 45, 1255–1260 (2013).
15. Drobniowski, F. et al. Drug-resistant tuberculosis, clinical virulence, and the dominance of the Beijing strain family in Russia. J. Am. Med. Assoc. 293, 2726–2731 (2005).
16. Brown, T., Nikolayevskyy, V., Velji, P. & Drobniowski, F. Associations between Mycobacterium tuberculosis strains and phenotypes. Emerg. Infec. Dis. 16, 272–280 (2010).
17. Casali, N. et al. Microevolution of extensively drug-resistant tuberculosis in Russia. Genome Res. 22, 735–745 (2012).
**ONLINE METHODS**

**Study population and whole-genome sequencing.** From October 2008 to 2010, 2,348 patients with pulmonary disease and culture-proven tuberculosis were recruited from all 18 civilian tuberculosis dispensaries located across Samara. *M. tuberculosis* isolates were prospectively archived at the Samara Tuberculosis Service. Anonymized epidemiological data were stored on a password-protected Access database. Informed consent was obtained from all subjects. The study was approved by the Samara Medical Ethics Committee, the Queen Mary Research Ethics Committee and the University of Cambridge Research Ethics Committee.

Isolates were cultured on Middlebrook medium for 4–6 weeks at 37 °C. Sweeps of colonies were harvested and lysed by vortexing with glass beads, and genomic DNA was purified using a DNeasy Blood and Tissue kit (Qiagen). Paired-end multiplex libraries with a mean insert size of 200 bp were prepared as previously described50. Sequencing was performed at the Wellcome Trust Sanger Institute on the Illumina Genome Analyzer GAII or HiSeq 2000 platform, generating reads of 54 bp, 75 bp or 100 bp.

**Sequence analysis.** Sequence reads were aligned to the corrected H37Rv reference genome18,57 with SMALT (see URLs), and GATK indel realignment was applied52. PinDel53 was used to predict the positions of indels and structural variants; these variants were visually checked in the mapping files. Candidate SNPs were identified using SAMtools54. At each mapped position, alleles were considered to be valid if they were supported by greater than 70% of mapped reads, including at least five in each direction and a minimum mapping quality of 45. SNPs located within repetitive regions were excluded from analysis18.

Mixed base calls in nonrepetitive regions of the genome were considered valid if they had mapping quality of ≥25, both calls were supported by at least five reads on each strand, and P values for strand bias, base quality bias, map quality bias and tail distance bias were >20.001.

To assess data consistency, nine isolates that were sequenced with 54-bp paired-end reads were recultured, and 100-bp paired-end sequence was generated. For each of these technical replicates, there were no inconsistencies in the bases called at variant sites that passed quality filters in both sequences.

**Phylogenetic and population genetic analyses.** A maximum-likelihood phylogeny was reconstructed with RAxML59 using a general time-reversible model with gamma correction for among-site rate variation. Calculation of 100 bootstrap replicates provided support for nodes on the tree. The phylogenetic tree was visualized with FigTree (see URLs). Ancestral sequences were reconstructed onto each node of the phylogeny using PAML56. From these ancestral sequences, SNPs were reconstructed onto branches of the tree.

To statistically define population structure, we used BAPS (Bayesian Analysis of Population Structure) software55,58, in particular its hierBAPS module22, which delineates population structure using nested clustering. Three nested levels of molecular variation were fitted to the data using 10 independent runs of the stochastic optimization algorithm with the a priori upper bound of the number of clusters varying over the interval of 50–300 across the runs.

To estimate nucleotide diversity π (ref. 59) for the Euro-American and Beijing clades, the functions available in the PGEToolbox60 were used in parallel on a cluster computing environment. The very large number of sequences in each clade would require excessive amounts of computing resources when analyzing all the sequences in a single process, and, hence, to allow for more economical calculations, 50 random subsets of 100 strains were sampled from each clade, and inference was performed using 100 bootstraps for each of them, as described by Cai et al.60, and averaging the results. Confidence intervals for π estimates were computed using the normal distribution approximation with s.d. derived by the bootstrap procedure.

**Molecular fingerprinting and microbiological testing.** Isolates were characterized by spoligotyping according to standard methods61.

Susceptibility to the first-line drugs rifampicin, isoniazid, streptomycin, ethambutol and pyrazinamide was determined using the absolute concentration method on Lowenstein-Jensen slopes62 or by using the automated Mycobacterial Growth Indicator Tube (MGIT) 960 system (Becton, Dickinson)63. For MDR isolates, susceptibility to the second-line drugs amikacin, capreomycin, ofloxacin, moxifloxacin and prothionamide was also determined using the MGIT system64.

A quality assurance procedure was implemented to ensure that isolate metadata corresponded to the appropriate sequence. Data for isolates belonging to the SNP-defined Beijing lineage that did not exhibit the characteristic Beijing spoligotype and, conversely, for isolates in other SNP-defined lineages that shared the Beijing spoligotype were excluded from further analysis (n = 45). The specificity of the katG mutation encoding pSer315Thr for the prediction of isoniazid resistance is >99% (ref. 64); thus, microbiological data for isolates with this genotype but a sensitive phenotype were also excluded (n = 58).

**Statistical methods.** Simple descriptive statistics were used to compare patient data for the sample population and the remaining population and to characterize the prevalence of mutations and the geographic distribution of sublineages; 95% confidence intervals (CIs) were established. In addition, the sampled population was evaluated by attributable risk analysis. The significance of differences between studied groups of variables was calculated using two-sample tests of proportions: Pearson χ² or Fisher’s exact test when any expected group size was less than five. Statistical tests were two-sided at α = 0.05. Analysis was performed using STATA (version 12.1, StataCorp).

---

50. Harris, S.R. et al. Evolution of MRSA during hospital transmission and intercontinental spread. Science 327, 469–474 (2010).
51. Cole, S.T. et al. Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. Nature 393, 537–544 (1998).
52. DePristo, M.A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43, 491–498 (2011).
53. Ye, K., Schulz, M.H., Long, Q., Apweiler, R. & Ning, Z. PinDel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics 25, 2865–2871 (2009).
54. Li, H. et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079 (2009).
55. Stanafakis, A., Ludwig, T. & Meier, H. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics 21, 456–463 (2005).
56. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24, 1586–1591 (2007).
57. Corander, J., Marttinen, P., Sirén, J. & Tang, J. Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. BMC Bioinformatics 9, 539 (2008).
58. Tang, J., Hanage, W.P., Fraser, C. & Corander, J. Identifying currents in the gene pool for bacterial populations using an integrative approach. PLoS Comput. Biol. 5, e1000455 (2009).
59. Nei, M. Molecular Evolutionary Genetics (Columbia University Press, New York, 1987).
60. Cai, J.J. PGEToolbox: a Matlab toolbox for population genetics and evolution. J. Hered. 99, 438–440 (2008).
61. Kamerbeek, J. et al. Simultaneous detection and strain differentiation of Mycobacterium tuberculosis for diagnosis and epidemiology. J. Clin. Microbiol. 35, 907–914 (1997).
62. Canetti, G. et al. Mycobacteria: laboratory methods for testing drug sensitivity and resistance. Bull. World Health Organ. 29, 565–578 (1963).
63. Krüüner, A., Yates, M.D. & Drobniewski, F.A. Evaluation of MGIT 960–based molecular fingerprinting and microbiological testing. J. Clin. Microbiol. 44, 811–818 (2006).
64. Ling, D.I., Zwerling, A.A. & Pai, M. GenoType MTBDR assays for the diagnosis of multidrug-resistant tuberculosis: a meta-analysis. Eur. Respir. J. 32, 1165–1174 (2008).