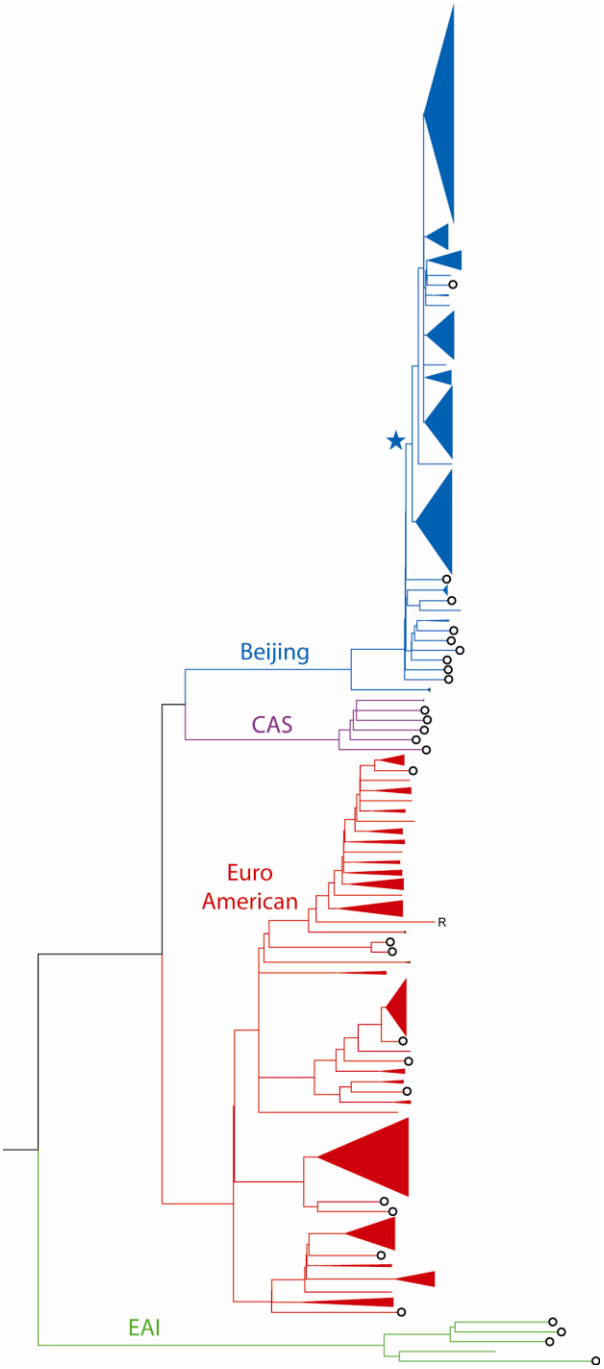
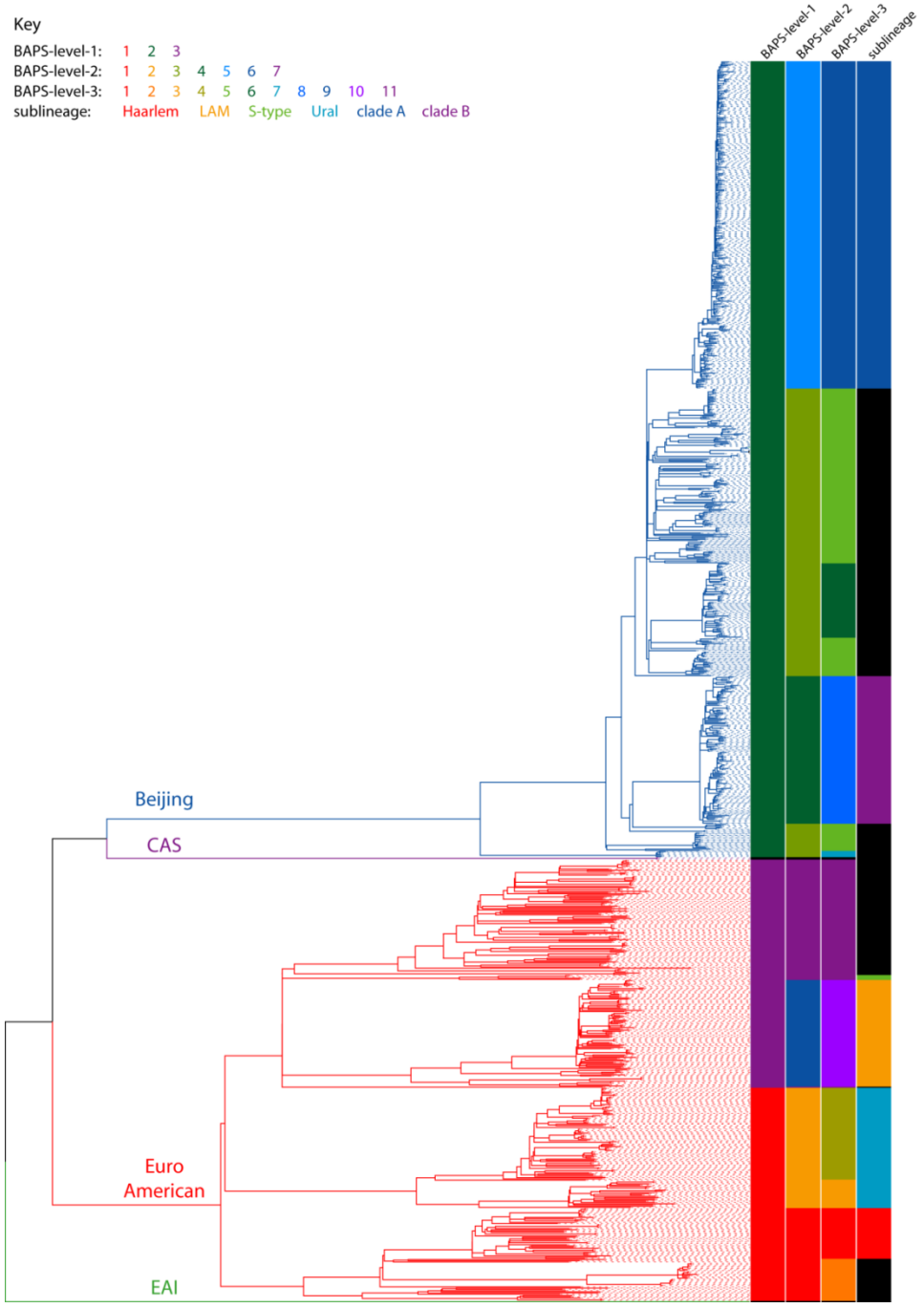
Supplementary Data 1. Maximum likelihood phylogeny of 1,035 *M. tuberculosis* isolates based on 32,445 variable sites.

The phylogeny illustrated in Figure 2 is provided in Newick format to show branch bootstrap support values.

see additional file

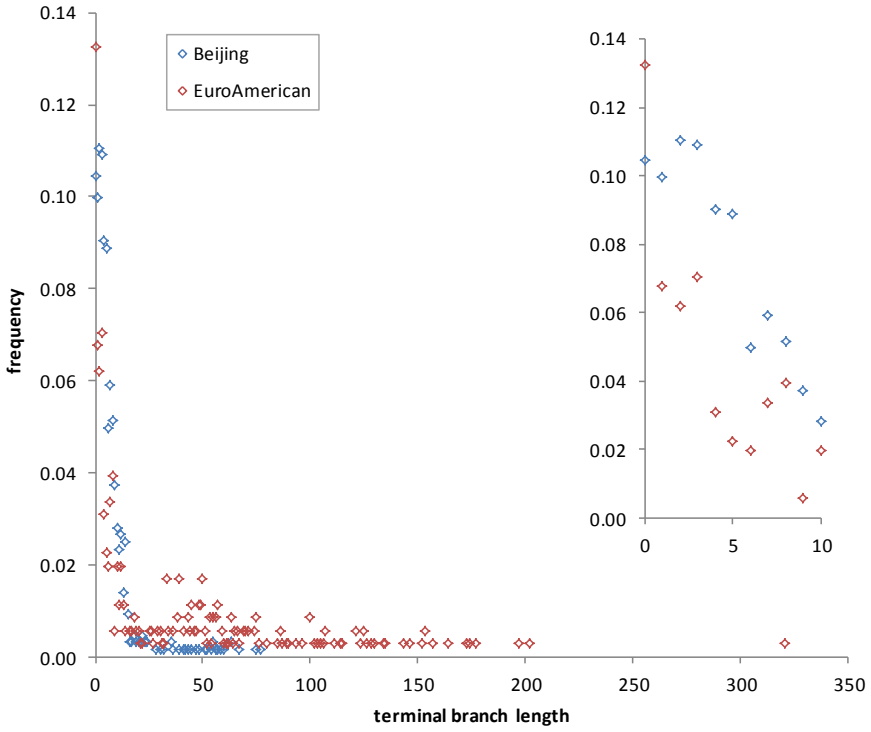


Supplementary Figure 1. Distribution of Samara isolates relative to global collection
The phylogeny in Figure 2 is illustrated with all clades that contained only Samaran or UK XDR isolates collapsed. The ancestral node of the Beijing East European sublineage is indicated with a star. Isolates from the UK, representing global diversity, are marked with white circles. The position of the reference sequence, H37Rv, is marked 'R'.

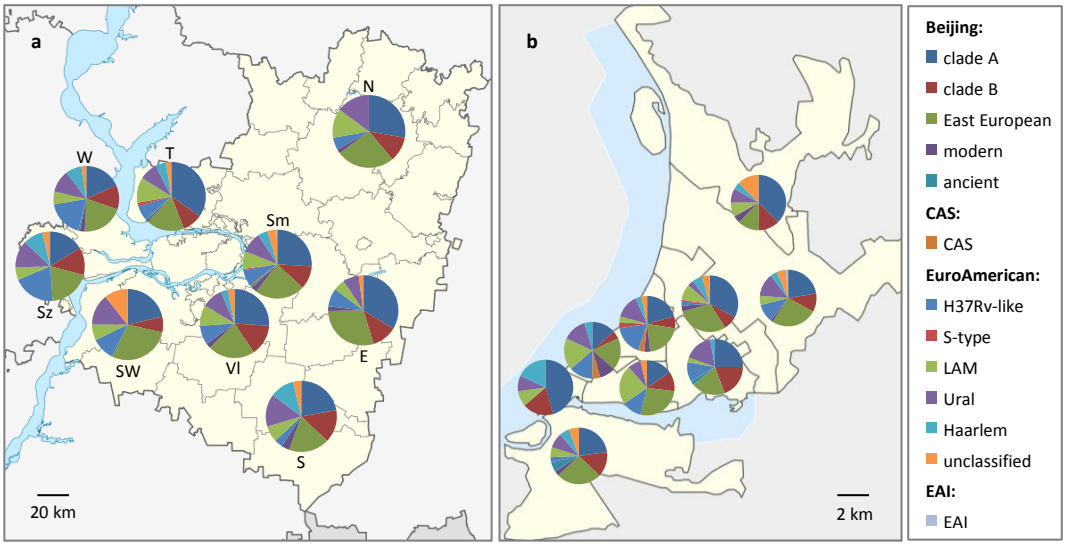


Supplementary Figure 2. Population genetic analysis

Genetic structuring of the data was investigated using hierBAPS, which delineates the population using nested clustering. The estimated mode of the posterior distribution had 3, 7 and 11 clusters at levels 1-3 of the hierarchy, respectively. All clusters in the mode were significantly supported when compared against alternative partitions (posterior probability of any cluster at least 100-fold higher than for the alternative). Sublineages are defined in the text.

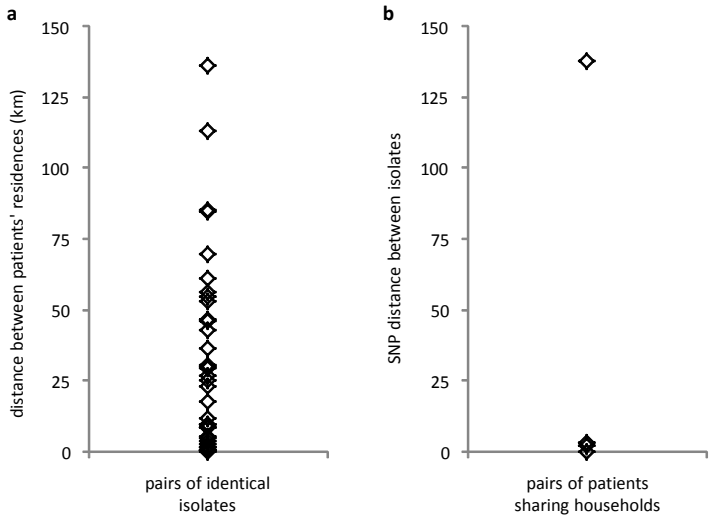


Supplementary Figure 3. Terminal branch lengths of Beijing versus EuroAmerican isolates
The number of SNPs between each isolate and it's last common ancestor was determined from the phylogeny illustrated in Figure 2. The frequency of each distance was calculated and corrected for the number of isolates in the lineage. All branch lengths are shown in the main panel and those less than 10 SNPs in the inset.

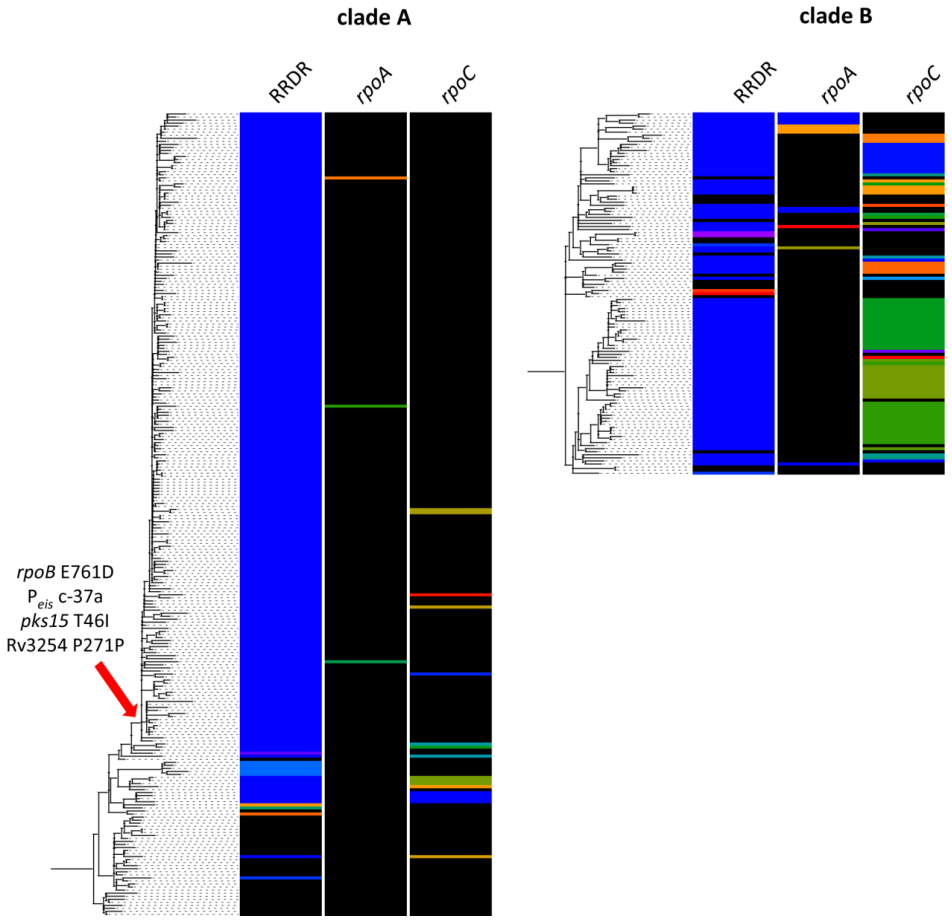


Supplementary Figure 4. Distribution of sublineages across Samara

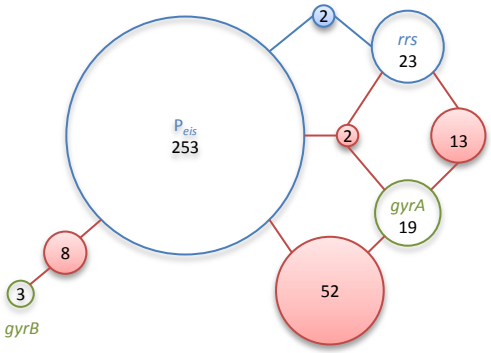
The geographic origin of the sequenced patient isolates is illustrated. Samara (a) is divided into regions: North (N), South (S), East (E), Southwest (SW) and West (W, West of the river Volga); cities: Samara City (Sm), Togliatti (T) and Syzran (Sz); and the region surrounding Samara City (Volzhskyy, Vl). Samara City (b) is divided into 9 districts.



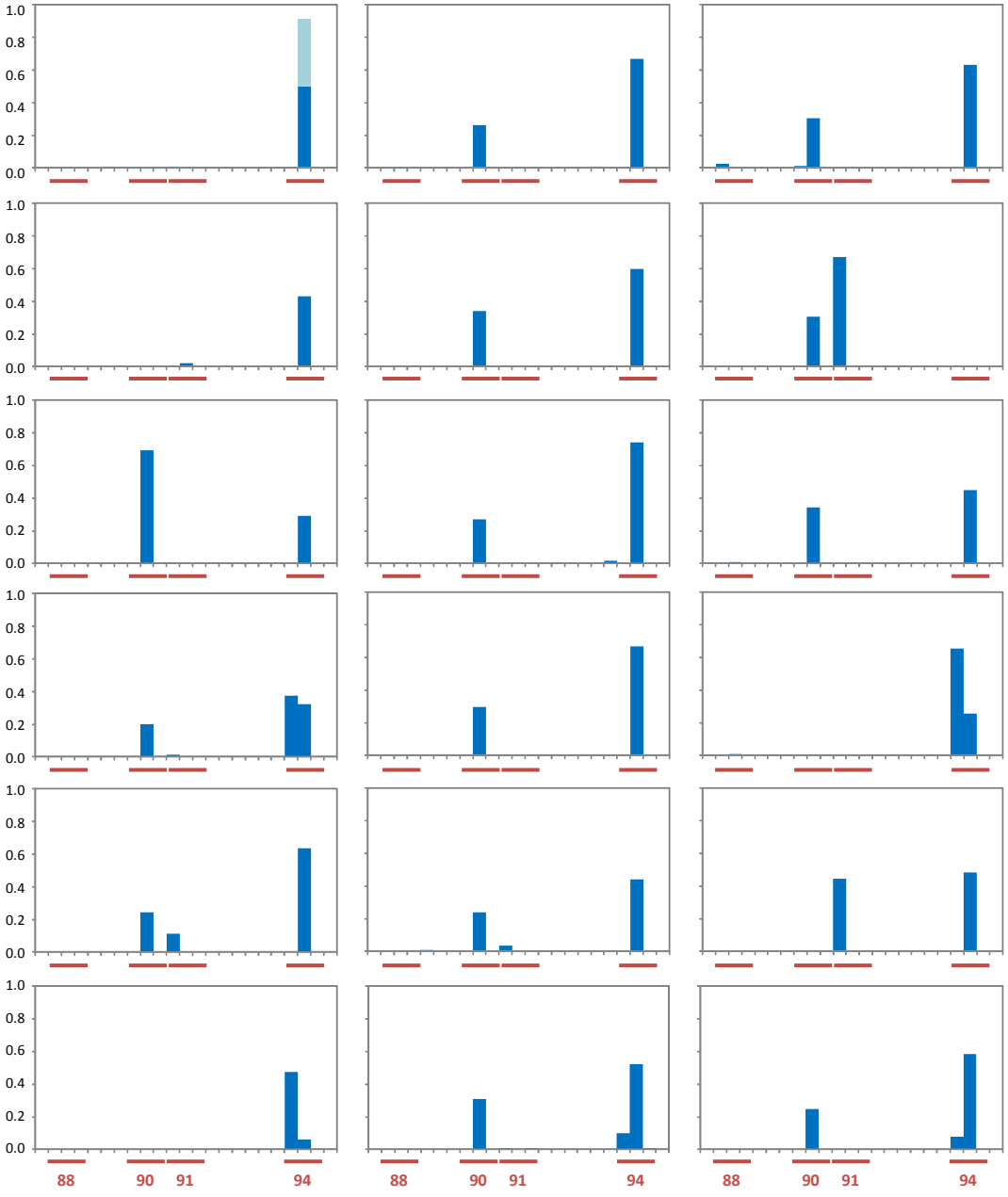
Supplementary Figure 5. Geographic versus genetic distance between isolates
a) The distance between the home addresses of patients sharing identical isolates. Where more than two isolates were identical, the distance to the geographically closest isolate is shown. b) The number of SNPs separating isolates from patients sharing households.



Supplementary Figure 6. Distribution of *rpoAC* compensatory mutations in clades A and B
The phylogenies of clades A and B are depicted on the left of each panel. Colored bands represent different polymorphisms. The first column shows nsSNPs in the rifampicin resistance determining region of *rpoB* and the right two columns nsSNPs in *rpoAC*. The genotypes illustrated are provided in full in Supplementary Table 4. The ancestral node of the clade with a significant absence of *rpoAC* nsSNPs is marked with an arrow. The four SNPs on this branch are shown.

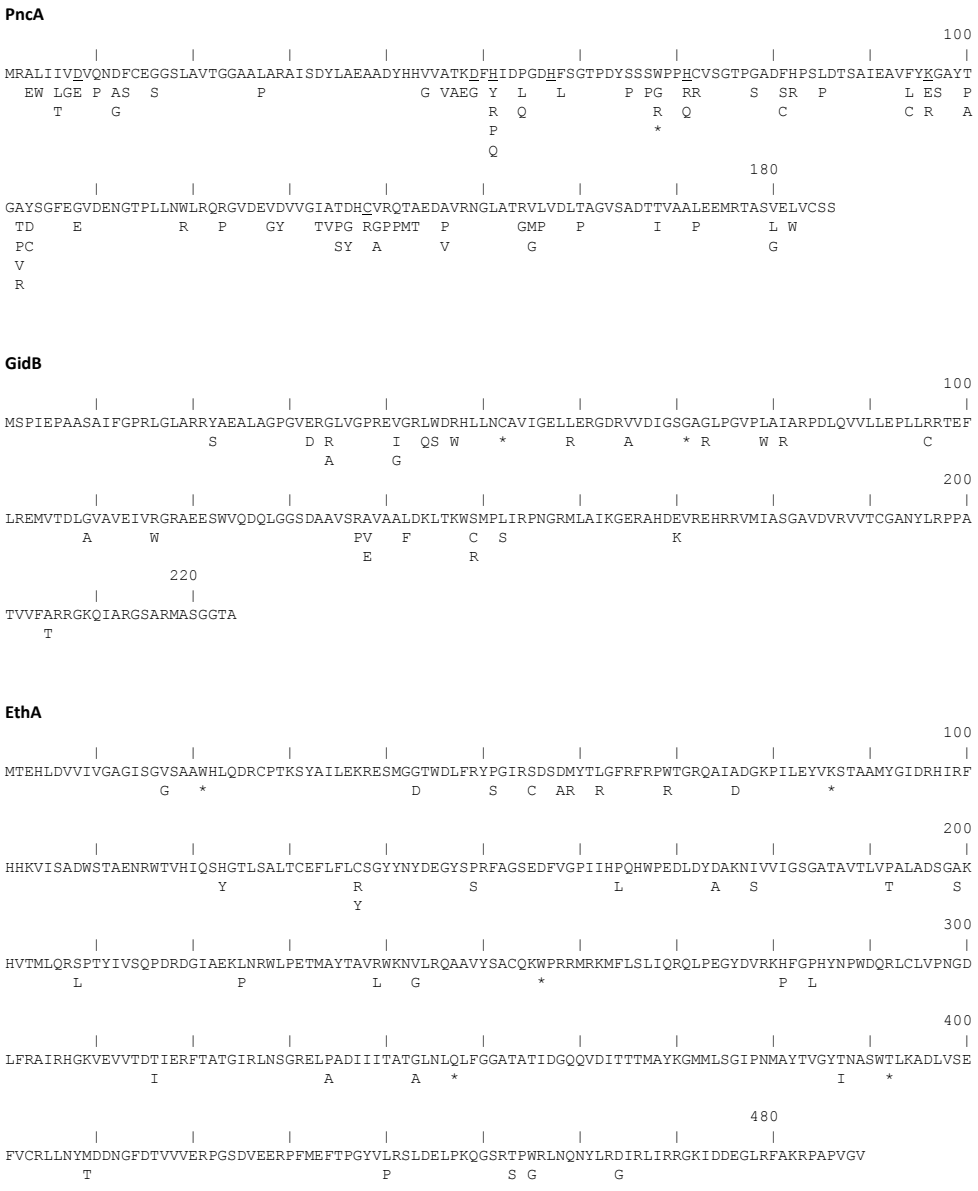


Supplementary Figure 7. Prevalence and co-occurrence of mutations conferring an XDR genotype
The number of isolates harbouring genotypes conferring resistance to fluoroquinolones (green) or second-line injectables (blue) are depicted as empty circles. Filled circles show the number of isolates carrying more than one of these mutations. Red filled circles are XDR.



Supplementary Figure 8. Heterogeneity at the *gyrA* QRDR locus

Each graph represents an isolate with ambiguous base calls between codons 88 to 94 in GyrA. The proportion of reads with alternate basecalls at each site are shown. The isolate depicted in the top left graph had two alternate alleles at one site.



Supplementary Figure 9. Distribution of non-synonymous and nonsense SNPs in three highly variable genes implicated in drug resistance: PncA, GidB and EthA

The amino acid sequence of the reference is shown; amino acid substitutions are shown below the sequence and stop codons marked with an asterisk. In PncA, residues that comprise the iron binding site (D49, H51, H57, H71) and catalytic triad (C138, D8, K96) are underlined[1].

Supplementary Table 1. Comparison of patient data for sequenced sample with the remainder of the population

a) All patients [a]

	Sequenced			Non-sequenced					Attributive	95% confidence
Parameter	Number	Total	%	Number	Total	%	P	χ-square	Risk	interval
New case [b]	658	996	66%	1196	1766	68%	0.373	0.795	-0.017	-0.054 to 0.021
Male	739	996	74%	1366	1766	77%	0.062	3.493	-0.032	-0.066 to 0.003
Imprisonment	240	991	24%	423	1752	24%	0.965	0.002	0.001	-0.033 to 0.035
Recreational drug use [c]	159	976	16%	267	1744	15%	0.499	0.456	0.010	-0.020 to 0.039
Alcohol use [c]	180	962	19%	278	1683	17%	0.152	2.056	0.022	-0.009 to 0.053
Homelessness	61	994	6%	102	1766	6%	0.699	0.149	0.004	-0.016 to 0.023
Employment	516	993	52%	911	1748	52%	0.939	0.006	-0.002	-0.041 to 0.038
HIV-positive	189	996	19%	274	1766	16%	0.019	5.466	0.035	0.004 to 0.065
MDR	456	996	46%	748	1766	42%	0.081	3.043	0.040	0.000 to 0.079
XDR	47	996	5%	72	1766	4%	0.425	0.636	0.006	-0.010 to 0.023

b) MDR and XDR TB sub-group [d]

	Sequenced			Non-sequenced					Attributive	95% confidence
Parameter	Number	Total	%	Number	Total	%	P	χ-square	Risk	interval
New case [b]	232	456	51%	390	748	52%	0.671	0.181	-0.013	-0.073 to 0.047
Male	339	456	74%	576	748	77%	0.294	1.102	-0.027	-0.079 to 0.025
Imprisonment	115	453	25%	213	745	29%	0.228	1.455	-0.032	-0.085 to 0.021
Recreational drug use [c]	85	444	19%	150	741	20%	0.646	0.211	-0.011	-0.059 to 0.037
Alcohol use [c]	65	439	15%	126	718	18%	0.223	1.486	-0.027	-0.073 to 0.018
Homelessness	24	455	5%	59	748	8%	0.083	3.007	-0.026	-0.056 to 0.004
Employment	236	455	52%	340	743	46%	0.040	4.217	0.061	0.001 to 0.121
HIV-positive	93	456	20%	130	748	17%	0.191	1.707	0.030	-0.017 to 0.078

c) HIV-positive sub-group [e]

	Sequenced			Non-sequenced					Attributive	95% confidence
Parameter	Number	Total	%	Number	Total	%	P	χ-square	Risk	interval
New case [b]	122	189	65%	182	274	66%	0.677	0.174	-0.019	-0.111 to 0.074
Male	138	189	73%	214	274	78%	0.208	1.588	-0.051	-0.135 to 0.034
Imprisonment	66	189	35%	94	273	34%	0.914	0.012	0.005	-0.088 to 0.098
Recreational drug use [c]	117	179	65%	150	263	57%	0.079	3.089	0.083	-0.013 to 0.180
Alcohol use [c]	35	180	19%	38	261	15%	0.175	1.840	0.049	-0.028 to 0.125
Homelessness	17	187	9%	21	274	8%	0.584	0.299	0.014	-0.042 to 0.071
Employment	70	187	37%	89	274	32%	0.272	1.206	0.050	-0.044 to 0.143
MDR	93	189	49%	130	274	47%	0.709	0.139	0.018	-0.079 to 0.115
XDR	12	189	6%	15	274	5%	0.693	0.156	0.009	-0.040 to 0.057

[a] no significant differences in any patient characteristics apart from HIV positivity

[b] versus previously treated

[c] including using in the past

[d] no significant differences in any patient characteristics apart from employment status

[e] no significant differences

Supplementary Table 2. Comparison of sublineage distribution between the isolated Western region and the rest of Samara

		West			Rest			Unknown	
Lineage	**Sublineage/ Clade**	**Number**	**Proportion**	**95% CI**	**Number**	**Proportion**	**95% CI**	**Number**	**P**
Beijing	clade A	23	18.3%	12.0-26.1	226	27.7%	24.7-30.9	15	0.025
Beijing	clade B	15	11.9%	6.8-18.9	92	11.3%	9.2-13.7	12	1.055
Beijing	other East European	26	20.6%	13.9-28.8	194	23.8%	20.9-26.9	12	0.867
Beijing	other modern	3	2.4%	0.5-6.8	17	2.1%	1.2-3.3	2	1.141
Beijing	ancient	1	0.8%	0.02-4.3	4	0.5%	0.1-1.3	0	0.663
CAS	-	0	0.0%	0-2.9	2	0.2%	0.03-0.9	0	0.578
EuroAmerican	H37Rv-like[a]	23	18.3%	11.9-26.1	63	7.7%	6.0-9.8	4	<0.001
EuroAmerican	S-type	0	0.0%	0-2.9	4	0.5%	0.1-1.3	0	0.431
EuroAmerican	LAM	8	6.3%	2.8-12.1	73	9.0%	7.1-11.1	5	0.331
EuroAmerican	Ural	14	11.1%	6.2-17.9	79	9.7%	7.8-11.9	4	0.620
EuroAmerican	Haarlem	10	7.9%	3.9-14.1	29	3.6%	2.4-5.1	2	0.022
EuroAmerican	unclassified[b]	3	2.4%	-	31	3.8%	-	3	-
EAI	-	0	0.0%	0-2.9	1	0.1%	0-0.7	0	0.694
	Total	*126*			*815*			*59*	

[a] The H37Rv-like sublineage refers to the clade containing the reference sequence (Figure 2)
[b] Diverse strains that are not part of any sublineage

Supplementary Table 3. Genetic distance between UK XDR isolates

		Closest related isolate		
UK Isolate	**Patient country of birth**	**SNP distance**	**Patient country of residence**	**Phylogenetic group**
114060062	Lithuania	16	Estonia	Beijing, East European sublineage, cladeB
112260020	Lithuania	35	Russia	Beijing, East European sublineage, cladeB
111880072	Latvia	13	Russia	Beijing, East European sublineage, clade A
112460232	Latvia	26	Russia	Beijing, East European sublineage
114540002	China	154	UK	Beijing

Supplementary Table 4. Polymorphisms at drug resistance loci

Isolates are listed in the order of the phylogenetic tree shown in Figure 3. Amino acid substitutions are shown in upper case. Nucleotide substitutions in intergenic regions and rRNA genes are in lower case. Ambiguous basecalls are marked 'n'. Nucleotide positions in promoter regions 'P', are given relative to the start codon. Insertions (ins) and deletions (del) are given with their position in the gene, followed by size in brackets. Where deletions span all or most of the locus they are marked 'lg' (large). Columns labelled RIF (rifampicin), INH (isoniazid), PZA (pyrazinamide), EMB (ethambutol), STR (streptomycin), CAP (capreomycin), AMI (amikacin), MOX (moxifloxacin), OFL (ofloxacin) and PRO (prothionamide) show phenotypic susceptibility results: resistant (R), sensitive (S).

see additional file

Supplementary Table 5. Correlation of polymorphisms at drug resistance loci with resistance phenotypes [a]

a) Rifampicin

rpoB RRDR	Sensitive	Resistant
L430P	10	0
L430P;D435G	0	1
L430P;M434V	1	0
Q432P	1	0
D435G	1	0
D435V	0	6
D435Y	2	0
D435Y;G456S	0	1
del1292(12)	0	2
H445D	0	4
H445L	0	4
H445N	1	0
H445N;L452P	0	3
H445R	0	1
H445Y	0	3
S450L	7	369
S450P	0	1
S450W	0	2
L452P	2	1
-	416	48
any RRDR	**25**	**398**

b) Pyrazinamide

P_{pncA}	*pncA*	Sensitive	Resistant
c-13a	-	1	0
t-11a	-	1	0
t-11c	-	3	7
t-11g	-	2	3
a-7g	-	0	3
-	A3E	0	1
-	L4W	0	1
-	I6L	120	8
-	I6T	0	1
-	V7G	1	4
-	D8E	0	2
-	Q10P	0	0
-	D12A	0	1
-	D12G	1	0
-	F13S	1	0
-	G17S	1	0
-	L27P	1	0
-	ins102(1)	0	1
-	ins122(1)	0	1
-	del127(131)	0	1
-	V44G	0	1
-	A46V	1	0
-	T47A	2	1
-	K48E	1	1
-	D49G	0	1
-	del150(1)	0	1
-	H51Y	0	0
-	H51P	1	1
-	H51R	2	0
-	H51Q	0	1
-	P54L	1	0
-	P54Q	0	3
-	F58L	2	0
-	del186(1)	0	3
-	ins193(7)	0	1
-	S65P	1	0
-	S67P	0	2
-	W68G	1	6
-	W68R	1	2
-	W68*	0	0
-	H71R	3	1
-	H71Q	0	1
-	C72R	0	1
-	G78S	0	0
-	ins234(1)	0	0
-	F81C	0	1
-	F81S	1	0
-	H82R	0	3
-	L85P	1	0
-	ins257(2)	1	1
-	F94L	1	1
-	F94C	1	0
-	K96E	0	3
-	K96R	0	1
-	G97S	0	3
-	T100A	0	1
-	T100P	0	2
-	A102P	1	0
-	A102T	0	1
-	A102R	1	0
-	A102V	1	0
-	Y103D	0	1
-	Y103C	1	0
-	G108E	0	0
-	del333(1)	0	1
-	del342(1)	0	2
-	ins350(1)	0	2
-	W119R	0	1
-	R123P	0	0
-	ins383(2)	0	1
-	V128G	1	6
-	D129Y	0	1
-	ins392(1)	0	1
-	ins392(2)	0	4
-	I133T	0	1
-	A134V	0	1
-	T135P	0	2
-	T135S	1	0
-	D136Y	0	3
-	D136G	1	0
-	ins408(1)	1	3
-	C138R	0	2
-	V139A	6	1
-	V139G	1	0
-	R140P	0	1
-	Q141P	0	1
-	T142M	0	0
-	A143T	1	0
-	A146P	0	1
-	A146V	0	1
-	R154G	0	1
-	V155M	0	2
-	V155G	0	1
-	L156P	0	1
-	T160P	0	1
-	ins481(1)	0	0
-	T168I	2	0
-	del515(2)	0	0
-	L172P	1	1
-	V180L	0	0
-	V180G	0	4
-	L182W	1	0
-	del(lg)	1	3
-	-	537	46
	any promoter SNP	**7**	**14**
	any KO mutation[b]	**3**	**26**
	nsSNP (excluding I6L)	**44**	**82**

Correlation of polymorphisms at drug resistance loci with resistance phenotypes (cont.)

c) Streptomycin

rpsL	rrs_{str}	*gidB*	Sensitive	Resistant
K43R	-	-	5	218
K43R	c517t	-	0	4
K43R	-	S149C	0	2
K88M	-	-	0	1
K88R	-	-	0	48
-	a514c	-	2	13
-	a514c	L44Q	0	5
-	a514t	-	0	1
-	c517t	-	33	156
-	c517t	Y22S	0	1
-	c517t	del629(1)	0	1
-	a908c	-	0	1
-	a906g	-	0	1
-	a906g	G71*	1	3
-	-	del53(1)	2	3
-	-	Y22S		1
-	-	del88(1)	2	0
-	-	E32D	1	0
-	-	G34R	0	1
-	-	G34A	1	0
-	-	del103(1)	0	2
-	-	del116(1)	0	2
-	-	V41I	1	0
-	-	V41G	1	0
-	-	W45S	1	0
-	-	R47W	1	0

rpsL	rrs_{str}	*gidB*	Sensitive	Resistant
-	-	del142(1)	1	0
-	-	C52*	1	0
-	-	L59R	0	1
-	-	ins194(1)	1	0
-	-	V65A	2	0
-	-	G73R	1	0
-	-	L79W	0	1
-	-	I81R	2	0
-	-	R96C;A205T	1	1
-	-	G109A	2	0
-	-	R116W	1	1
-	-	del352(1)	0	1
-	-	del397(364)	0	1
-	-	R137P	0	1
-	-	A138V	1	0
-	-	A138E	0	1
-	-	del414(1)	1	0
-	-	S149R	0	2
-	-	L152S	1	0
-	-	E170K	1	0
-	-	A205T	1	0
-	-	-	315	39
		any *rpsL*	**5**	**273**
		any *rrs*	**36**	**186**
		any *gidB* KO mutation[b]	**12**	**13**
		any *gidB* nsSNP	**19**	**18**

d) Ethambutol

P_{embA}	*embB*	Sensitive	Resistant
c-16a	-	1	0
c-16t	-	4	4
c-15g	-	1	0
c-11a	-	0	1
c-12t	-	2	2
-	M306L	0	4
-	M306V	34	63
-	M306I	25	16
-	Y319C	3	0
-	Y319S	1	2
-	D354A	83	78
-	E378A	0	1
-	G406A	2	10
-	Q497K	2	4
-	G406D	11	2
-	Q497R	9	10
-	H1002R	1	1

P_{embA}	*embB*	Sensitive	Resistant
-	D1024N	0	2
-	M306V;D354A	0	2
-	M306I;Q497R	0	3
-	M306V;D1024N	0	1
c-16t	D354A	0	3
c-16t	M306I	0	2
c-16a	M306I	0	1
c-16a	D354A	0	4
c-16g	D354A	0	1
c-12t	M306V	0	1
c-12t	D354A	0	2
c-12t	G406A	0	1
c-12t	Q497R	0	1
c-8a	D354A	1	5
-	-	443	46
	any single SNP	**176**	**199**
	any double SNP	**1**	**27**

e) Prothionamide

P_{ethA}	*ethA*	Sensitive	Resistant
a-7g	-	116	56
-	del104(2627)	0	1
-	V17G	1	1
-	W21*	0	1
-	del111(1)	8	16
-	G43D	1	0
-	W45*	0	2
-	P51S	0	3
-	M59R	3	4
-	L62R	3	0
-	W69R;P334A	0	1
-	A76D	2	1
-	K86*	1	0
-	ins282(5)	0	1
-	C137R	1	3
-	C137Y	0	1
-	P149S	0	1
-	del480(1)	3	11
-	P164L	0	1
-	A199S;R239L	1	0
-	S208L	0	1
-	del704(1)	1	3
-	W256*	4	4

P_{ethA}	*ethA*	Sensitive	Resistant
-	del769(1)	0	3
-	H281P	0	2
-	ins865(1)	0	1
-	del887(1)	2	0
-	T314I	9	3
-	P334A	2	2
-	del1011(1)	11	15
-	G343A	0	2
-	Q347*	0	2
-	del1113(1)	1	0
-	T387I	8	6
-	W391*	0	1
-	del1182(1)	0	1
-	ins1269(1)	1	0
-	ins1293(1)	0	1
-	L440P	4	0
-	W455G	1	0
-	D464G	1	0
-	ins1429(2)	2	0
-	-	78	32
	any *ethA* KO mutation[b]	**34**	**62**
	any nsSNP	**37**	**32**
	promoter SNP	**116**	**56**

Correlation of polymorphisms at drug resistance loci with resistance phenotypes (cont.)

f) Amikacin and capreomycin

rrs$_{inj}$	amiS capS	amiS capR	amiR capS	amiR capR
a1401g	2	0	4	27
c1402t	0	1	0	0
g1484t	0	0	0	1
-	367	26	14	16
any SNP	**2**	**1**	**4**	**28**

g) Fluoroquinolones (moxifloxacin and ofloxacin)

gyrA	*gyrB*	moxS ofIS	moxR ofIS	moxS ofIR	moxR ofIR
A74S	-	1	0	0	1
A90V	-	2	0	1	6
S91P	-	0	0	0	5
D94A	-	2	1	0	6
D94G	-	1	0	1	26
D94N	-	0	0	0	6
D94Y	-	0	0	0	3
-	N499Y	0	0	0	1
-	N499S	2	0	0	0
-	E501D	1	0	0	2
-	D461N	1	0	0	1
-	A504V	1	0	0	0
-	A504T	1	0	0	0
-	-	261	24	24	13
	any gyrA nsSNP	**6**	**1**	**2**	**53**
	any gyrB nsSNP	**6**	**0**	**0**	**4**

[a] Dashes indicate a wild-type sequence at that locus
[b] Knockout mutation: large deletion, frameshifting indel or nonsense SNP

Supplementary Table 6. Summary statistics for correlation of drug resistance genotypes and phenotypes

		Mutation		No mutation					
Drug	**Resistance locus**	**Sensitive**	**Resistant**	**Sensitive**	**Resistant**	**Sensitivity**	**Specificity**	**Positive predictive value**	**Negative predictive value**
Rifampicin	*rpoB* RRDR	25	398	416	48	94.1%	89.7%	89.2%	94.3%
Pyrazinamide	P_{pncA}	7	14	734	162	66.7%	81.9%	8.0%	99.1%
	pncA KO mutation	3	26	738	150	89.7%	83.1%	14.8%	99.6%
	pncA nsSNP (excluding I6L)	75	82	666	94	52.2%	87.6%	46.6%	89.9%
	pncA I6L	119	8	622	168	6.3%	78.7%	4.5%	83.9%
Ethambutol	P_{embA}	9	28	614	245	75.7%	71.5%	10.3%	98.6%
	embB	172	220	451	53	56.1%	89.5%	80.6%	72.4%
	P_{embA} or *embB*	176	199	444	73	53.1%	85.9%	73.2%	71.6%
	P_{embA}/*embB* double SNP	1	27	619	245	96.4%	71.6%	9.9%	99.8%
Streptomycin	*rpsL*	5	273	378	240	98.2%	61.2%	53.2%	98.7%
	rrs_{str}	36	186	347	327	83.8%	51.5%	36.3%	90.6%
	rpsL or rrs_{str}	41	455	342	58	91.7%	85.5%	88.7%	89.3%
	gidB KO mutation	9	13	374	500	59.1%	42.8%	2.5%	97.7%
	gidB nsSNP	19	18	364	495	48.6%	42.4%	3.5%	95.0%
Prothionamide	P_{ethA}	116	56	149	126	32.6%	54.2%	30.8%	56.2%
	ethA KO mutation	34	62	231	120	64.6%	65.8%	34.1%	87.2%
	ethA nsSNP	37	32	228	150	46.4%	60.3%	17.6%	86.0%
Amikacin/capreomycin[a]	rrs_{inj}	2	33	367	56	94.3%	86.8%	37.1%	99.5%
Fluoroquinolones[b]	*gyrA*	6	56	267	65	90.3%	80.4%	46.3%	97.8%
	gyrB	6	4	267	117	40.0%	69.5%	3.3%	97.8%

[a] Classified as resistant if not susceptible to either amikacin or capreomycin
[b] Classified as resistant if not susceptible to either moxifloxacin or ofloxacin

Supplementary Table 7: Association of mutations at drug resistance loci with Beijing and EuroAmerican lineages

	Beijing (n=642)			EuroAmerican (n=355)			
Locus	**Number**	**Proportion**	**95% CI**	**Number**	**Proportion**	**95% CI**	**P [a]**
katG	478	0.74	0.71-0.78	107	0.30	0.25-0.35	<0.001
P_{inhA}	33	0.05	0.04-0.07	37	0.10	0.07-0.14	0.002
rpoB	430	0.67	0.63-0.71	67	0.19	0.15-0.23	<0.001
P_{pncA}	19	0.03	0.02-0.05	2	0.01	0.00-0.02	0.012
pncA	306	0.48	0.44-0.52	29	0.08	0.06-0.12	<0.001
P_{embA}	35	0.05	0.04-0.08	6	0.02	0.01-0.04	0.004
embB	408	0.64	0.60-0.67	50	0.14	0.11-0.18	<0.001
rpsL	263	0.41	0.37-0.45	51	0.14	0.11-0.18	<0.001
rrs_{str}	244	0.38	0.34-0.42	17	0.05	0.03-0.08	<0.001
gidB	16	0.02	0.01-0.04	49	0.14	0.10-0.18	<0.001
rrs_{inj}	33	0.05	0.04-0.07	7	0.02	0.01-0.04	0.015
P_{eis}	299	0.47	0.43-0.51	18	0.05	0.03-0.08	<0.001
gyrA	81	0.13	0.10-0.15	5	0.01	0.00-0.03	<0.001
gyrB	11	0.02	0.01-0.03	0	0.00	0.00-0.01	0.010
P_{ethA}	215	0.33	0.30-0.37	0	0.00	0.00-0.01	<0.001
ethA	176	0.27	0.24-0.31	51	0.14	0.11-0.18	<0.001

[a] Mutations at all loci are significantly associated with the Beijing lineage, except P_{inhA} and *gidB* mutations which are significantly associated with the EuroAmerican lineage

Supplementary Table 8. Co-occurrence of RpoABC nsSNPs with RRDR genotypes

			RRDR Genotype of Isolates			Likelihood Compensatory Mutation[a]
Gene	Substitution	Isolates/ Cluster	WT	S450L	Other nsSNP	
rpoA	G31A/S	3,1	0	4	0	high
	R153W	1	1	0	0	low
	K177M	1	0	1	0	-
	T181A	1	0	1	0	-
	V183G	3	0	3	0	-
	E184D	1	0	1	0	-
	T187P/A	1,1,1,8,4,2,1	1	17	0	high
	D190G	1,1	0	2	0	-
	H270N	1	1	0	0	low
	L304R	1	0	1	0	-
	G305S	6	6	0	0	low
	S307L	1	0	1	0	-
rpoC	G332S/R/C	1,1,1	0	3	0	high
	N416T/S	2,1	0	3	0	high
	S428A	2	0	2	0	-
	V431M	4,1	0	5	0	high
	G433S/C	1,3	0	4	0	high
	P434R/Q	1,1	0	2	0	high
	K445R	1,14	0	15	0	high
	L449V	1,3,1	0	5	0	high
	F452C	1	0	1	0	-
	V483G/A	3,8,1,2,1,4,2,1	0	22	0	high
	W484G	3,1	0	4	0	high
	D485Y/N	1,5,13	0	19	0	high
	I491V/T	2,1,1,3,11	0	18	0	high
	L516P	1	0	1	0	-
	V517L	1,1	0	2	0	high
	G519D	1,1,1	0	2	0	high
	A521D	1,1	0	2	0	high
	H525Q	2	0	2	0	-
	L527V	1,17	0	18	0	high
	R572H	1	0	1	0	-
	G594E	75	61	10	4	low
	P601L	1	1	0	0	low
	T667M	3	3	0	0	low
	P678R	4	0	4	0	-
	H689R/H/S/K	1,1,1,1,1,1,1,1,1	1	8	0	high
	A734V	1	0	1	0	-
	D747A	4	0	3	0	-
	Q761R	1	0	1	0	-
	R770H	10	0	10	0	-
	E791Q	1	1	0	0	-
	N826K	1	0	0	1	-
	I832V	1	0	1	0	-
	S838C	1	0	1	0	-
	E842A	1	1	0	0	-
	L847R	1	0	1	0	-
	Y849C	1	0	1	0	-
	I885V	3,1,1	0	4	1	high
	D943G	3	0	3	0	-
	G945V	1	0	1	0	-
	P1040T/S/R	1,1,1,1,1	0	5	0	high
	I1046M	1	0	1	0	-
	E1092D	495	171	288	36	low
	S1100A	6	3	1	2	low
	V1130M	1	0	0	1	-
	V1252L	1	0	1	0	-
rpoB	S12T	5	5	0	0	low
	L42F	1	0	1	0	-
	P45S	1	0	1	0	-
	E82G	1	0	1	0	-
	D103E	1	0	0	1	-
	V170F	1	1	0	0	-
	D265G	1	0	0	1	-
	T399A	1	0	1	0	-
	P479T	1	0	1	0	-
	N487S	1	0	0	1	-
	I488V	1	0	1	0	-
	I491V	1,1	0	2	0	high
	V496M/L	2,7	0	9	0	high
	F503S	3	0	3	0	-
	D571A	1	0	1	0	-
	M707T	1	1	0	0	low
	H723Y	1	0	1	0	-
	L731P	1	0	1	0	-
	E761D	207	0	207	0	-
	I783V	1	1	0	0	low
	R827C	1	0	1	0	-
	H835R/P	1,1,1	0	2	1	high
	I925V	3	3	0	0	low
	D1006G	1	0	0	1	-
	S1124A	1	1	0	0	low

[a] Codons were considered to have a high likelihood of compensatory function if they emerged independently multiple times in isolates with RRDR mutations and a low likelihood if they occurred in isolates with WT RRDR.

Supplementary Table 9. Heterogeneity at drug resistance loci

Locus	Isolates with resistance allele	Isolates with heterogeneous alleles	Rate of heterogeneity[a]	95% CI	p [b]
katG	585	0	0.000	0.000-0.006	
rpsL	314	0	0.000	0.000-0.012	
P_{inhA}	70	0	0.000	0.000-0.051	
gidB	45	0	0.000	0.000-0.079	
rrs_{str}	261	1	0.004	0.000-0.021	
embB	457	2	0.004	0.001-0.016	
P_{eis}	317	2	0.006	0.001-0.023	
P_{ethA}	217	2	0.009	0.001-0.033	
rpoB	495	5	0.010	0.003-0.023	
ethA	113	2	0.017	0.002-0.061	
pncA	303	6	0.019	0.007-0.042	
P_{embA}	41	1	0.024	0.001-0.126	0.015
rrs_{inj}	40	1	0.024	0.001-0.129	0.016
P_{pncA}	21	2	0.087	0.011-0.280	0.527
gyrA	85	18	0.175	0.107-0.262	
gyrB	11	3	0.214	0.047-0.508	

[a] Proportion of isolates with a mutant allele that have heterogeneous alleles
[b] Pearson's chi-square test comparison to *gyrA*

Supplementary Table 10. Homoplasies in the phylogeny of Samara isolates [a]

Position in reference	Reference base	SNP base	Gene	Substitution
698	G	A	dnaA	R233Q
6620	G	A	gyrB	D461N
6742	A	C	gyrB	E501D
7563	G	T	gyrA	G88C
7570	C	T	gyrA	A90V
7572	T	C	gyrA	S91P
7581	G	A,T	gyrA	D94N,Y
7582	A	C,G	gyrA	D94A,G
13624	G	A	intergenic [c]	-
13636	T	C	intergenic [c]	-
13638	T	C,G	intergenic [c]	-
18147	C	T	pknA	V206I
36853	G	C	acpA	G83R
55549	G	T	ponA1	P629P
75233	C	A	intergenic	-
113793	C	G	nrp	A1265G
122334	T	C	Rv0104	F7L
128620	C	T	ctpI	A641T
147250	C	T	fusA2	P174P
160817	C	T	intergenic	-
164315	T	C	cyp138	V317A
218599	T	C	intergenic	-
222486	C	T	Rv0191	A66A
244552	A	G	mmpL3	R923R
255980	G	A	intergenic	-
284783	G	A	aftD	R691W
300817	C	T	intergenic	-
304904	C	T	nirB	R679R
332737	A	C	vapC25	S134S
332918	T	G	vapC25	N74T
333046	A	G	vapC25	L31L
333088	A	G	vapC25	H17H
333211	A	G	vapB25	V70A
342091	T	C	intergenic	-
390986	C	T	Rv0323c [c]	G90S
448551	C	T	Rv0371c	R29Q
471671	A	G	ndhA	M325T
479779	C	T	intergenic	-
485307	G	T	fadD30 [c]	R442S
497388	C	T	glnH	T306T
506572	C	T	lpqM	A494A
569846	C	T	intergenic	-
612556	G	A	Rv0520	P99P
617698	G	C	ccdA	W67C
631400	A	G	Rv0538	P452P
631403	G	C	Rv0538	T453T
684381	T	C,G	intergenic	-
690172	C	T	mce2C	P370L
731903	T	G	intergenic	-
761100	T	C	rpoB	L430P
761114	G	T	rpoB	D435Y
761115	A	T,G	rpoB	D435V,G
761144	C	A,T,G	rpoB	H445N,Y,D
761145	A	T,G	rpoB	H445L,R
761160	C	T	rpoB	S450L
761166	T	C	rpoB	L452P
761282	A	G	rpoB	I491V
762315	A	G	rpoB	H835R
764665	G	A	rpoC	V431M
764708	A	G	rpoC	K445R
764719	C	G	rpoC	L449V
764822	T	C,G	rpoC	V483A,G
764824	T	G	rpoC	W484G
764827	G	A	rpoC	D485N
764845	A	G	rpoC	I491V
764846	T	C	rpoC	I491T
764930	G	A	rpoC	G519D
764936	C	A	rpoC	A521D
764953	T	G	rpoC	L527V
765466	A	C	rpoC	N698H
765467	A	G	rpoC	N698S
765468	C	A	rpoC	N698K
766027	A	G	rpoC	I885V
766492	C	T	rpoC	P1040S
766493	C	G	rpoC	P1040R
781692	A	G	rpsL	K43R
781827	A	G	rpsL	K88R
817532	G	T	sppA	R622L
841331	T	C	vapC31	A33A
841382	G	A	vapC31	L50L
841383	G	A	vapC31	A51T
841453	C	A	vapC31	T74N
841487	T	C	vapC31	G85G
841634	G	T	vapC31	S134S
841655	G	A	vapC31	P141P
841656	C	A	vapC31	L142I
842049	A	C	Rv0750	I4I
842056	G	A	Rv0750	D7N
842061	C	T	Rv0750	C8C
842062	G	T	Rv0750	V9F
842063	T	A	Rv0750	V9D
842068	C	G	Rv0750	H11D
842070	C	G	Rv0750	H11Q
842116	C	G	Rv0750	L27V
902554	G	A	purF	A147T
908191	T	C	sseC2	T100A
911522	C	T	Rv0818	D182D
916768	C	T	Rv0823c	G295D
941757	C	T	Rv0845	A188V
972799	G	A	Rv0874c	L305L
974009	C	T	Rv0875c	Q97Q
1041451	C	T	pstB	T61M
1048110	A	C	ligD	S657R
1094544	T	C,G	intergenic	-
1094713	T	C	Rv0979c	D53G
1103338	T	G	Rv0987	L264R
1164577	A	G	intergenic	-
1254138	T	C	Rv1129c [c]	E135G
1338071	T	C	Rv1194c	I151V
1340665	A	G	intergenic	-
1340675	A	G	esxK	S3S
1341072	A	T	esxL	Q20L
1341081	T	C	esxL	L23S
1341168	G	A	esxL	G52E
1341182	T	C	esxL	L57L
1341262	G	A	esxL	A83A
1341303	A	T,G	intergenic	-
1342728	T	C,G	intergenic	-
1349383	T	G	fadD6	I15S
1357122	C	T	intergenic	-

Homoplasies in the phylogeny of Samara isolates (cont.)

Position in reference	Reference base	SNP base	Gene	Substitution
1380580	T	A	sugB	S240T
1404177	T	G	cyp130	K145N
1441541	C	G	Rv1288	D62E
1472367	A	C,T	rrs	514
1472370	C	T	rrs	517
1472759	A	G	rrs	906
1473254	A	G	rrs	1401
1490812	C	T	glgB	P503P
1552555	G	A	Rv1378c	R37W
1576489	T	G	lipI	T106P
1614824	A	G	pgk	K163R
1673433	C	T	intergenic	-
1673440	T	A,C,G	intergenic	-
1674056	G	A	fabG1	L203L
1674489	T	G	inhA [c]	S94A
1675181	C	T	hemZ	P53S
1761773	G	A	mmpL6	G8S
1774087	G	T	Rv1566c	P181Q
1791608	T	C	Rv1591	S11P
1813611	C	T	lgt	L145F
1840310	A	G	Rv1634	I379V
1841444	T	G	Rv1635c	I269I
1902386	T	C	intergenic	-
1934022	C	T	intergenic	-
2030363	A	G	esxM	S3S
2030529	T	C	esxM	*59Q
2030856	A	G	esxN	E52G
2030950	G	A	esxN	A83A
2074466	G	C	intergenic	-
2108611	C	G	apa	P290A
2122403	C	T	lldD2	V253M
2123141	C	T	lldD2	V7M
2123153	C	T	lldD2	V3I
2123154	C	T	lldD2	A2A
2123168	C	T	intergenic	-
2123170	A	T	intergenic	-
2123177	T	G	intergenic	-
2123189	C	T	Rv1873	S3L
2123190	A	C	Rv1873	S3S
2132768	C	T	Rv1882c	A148A
2142407	C	T	intergenic	-
2155176	C	T,G	katG	S315N,T
2158913	C	A	fadB5	G63*
2174224	A	G	Rv1922	V50V
2187314	G	A	intergenic	-
2195930	T	C	Rv1944c	T5A
2195931	A	C	Rv1944c	D4E
2207533	C	T	intergenic	-
2212447	T	G	mce3C	A271A
2277281	A	G	Rv2030c	Y405Y
2288736	A	G	pncA	L172P
2288748	G	A	pncA	T168I
2288848	T	G	pncA	T135P
2288868	A	C	pncA	V128G
2288947	C	T	pncA	A102T
2288953	T	G	pncA	T100P
2288971	A	G	pncA	F94L
2289039	T	C	pncA	H71R
2289049	A	C	pncA	W68G
2289109	T	C	pncA	K48E
2289112	T	C	pncA	T47A
2289231	A	C	pncA	V7G
2289240	A	C	pncA	L4W
2289261	T	A,C,G	intergenic	-
2296051	G	C	pks12	P3649A
2300246	A	G	pks12	A2250A
2300561	T	G	pks12	P2145P
2300564	A	G	pks12	D2144D
2306315	A	G	pks12	A227A
2338686	T	C	intergenic	-
2338688	C	T	intergenic	-
2338921	A	C	Rv2082	R68R
2339269	C	T	Rv2082	P184P
2339308	T	G	Rv2082	S197S
2339309	A	C	Rv2082	M198L
2339310	T	C	Rv2082	M198T
2363691	C	A	Rv2102	P98T
2372502	C	G	dop	R26P
2372559	G	C	dop	P7R
2381038	G	T	intergenic	-
2439213	A	G	intergenic	-
2444629	T	G	Rv2182c	D237A
2531618	A	G	Rv2258c	R96R
2563472	C	T	sseB	D97D
2625836	C	T	intergenic	-
2625932	T	C	esxO	A83A
2626113	A	G	esxO	L23S
2626116	C	G	esxO	G22A
2626118	G	C	esxO	A21A
2626686	A	C	Rv2348c	I101M
2643278	G	A	Rv2361c	L67L
2700247	G	A	Rv2402	V571M
2703923	C	T	intergenic	-
2703959	G	A	intergenic	-
2703971	C	T	intergenic	-
2703980	T	C	intergenic	-
2715350	C	T	intergenic	-
2715352	G	A	intergenic	-
2715354	G	A	intergenic	-
2715355	G	C	intergenic	-
2715377	C	A	intergenic	-
2726149	C	T	intergenic	-
2726153	G	A	intergenic	-
2747159	T	C	folC [c]	S150G
2747203	T	G	folC [c]	D135A
2747479	A	G	folC [c]	I43T
2747488	T	C	folC [c]	E40G
2765896	G	A	lipP	T78T
2778964	G	A	gdh	L1103L
2828497	C	G	intergenic	-
2828498	G	A	intergenic	-
2828501	G	A	intergenic	-
2829926	T	C,G	intergenic	-
2863716	C	T	aroF	G235G
2867764	T	C	lppB	I211I
2881463	A	G	Rv2561	Y16C
2881480	T	G	Rv2561	W22G
2886547	G	A	Rv2566	R56H
2932443	G	A	tesB2	D236D
2953260	G	A	intergenic	-

Homoplasies in the phylogeny of Samara isolates (cont.)

Position in reference	Reference base	SNP base	Gene	Substitution
2982964	A	C,G	Rv2665	P86P,P
2986835	G	A	Rv2670c	A5V
2995856	G	A	echA15	G245E
3034771	T	C	Rv2722	Y43Y
3035380	G	A	Rv2723	S155N
3074238	T	C	thyA	D81G
3075294	G	A	Rv2765	A217A
3089259	G	A	pepR	R371W
3095501	T	C	Rv2787	I128T
3098318	A	G	fadE21	L210P
3112885	G	A	Rv2804c	A73V
3130012	C	T	Rv2823c	L590L
3135644	T	C	intergenic	-
3136343	G	A	Rv2828A	R89W
3173115	G	A	intergenic	-
3175449	G	A	intergenic	-
3177552	T	C	relF	I3T
3189531	C	T	intergenic	-
3224892	G	T	ffh	R468S
3247324	C	G	ppsA	D624E
3248082	G	A	ppsA	R877H
3248083	C	T	ppsA	R877R
3251428	G	A	ppsB	A117T
3253560	G	A	ppsB	E827E
3266777	A	C	ppsD	I1508L
3297348	G	A	pks15 [c]	T167T
3343419	C	T	hupB	R137K
3430088	G	A	Rv3064c	A57A
3446707	C	G	Rv3081	F220L
3448009	T	C	virS	G142G
3477274	T	C	Rv3108	R69R
3500157	G	A	Rv3134c	L201L
3534196	G	A	Rv3165c	L70F
3538160	G	A	Rv3169	K305K
3555465	G	A	Rv3190c	P411S
3568767	T	G	intergenic	-
3600338	G	A	Rv3224	T160T
3625949	G	A	mtrB	L225L
3640324	A	G	intergenic [c]	-
3640416	T	G	intergenic [c]	-
3679971	C	T	lhr	G1063G
3680940	G	A	lhr	P1386P
3681535	G	A	nei	G70S
3690955	A	G	intergenic	-
3691011	C	A	intergenic	-
3691013	A	G	intergenic	-
3691069	A	C	intergenic	-
3700229	G	A	Rv3312c	T37I
3774269	C	T	Rv3363c	V82M
3793110	G	A	Rv3378c	D49D
3798070	G	T	idsB	T143K
3823870	A	G	Rv3403c	M1T
3831629	G	A	intergenic	-
3855220	T	C	glmS	I560V
3877946	T	C	rpoA	D190G
3877956	T	C,G	rpoA	T187A,P
3895891	G	T	Rv3479	G22V
4001629	T	C	intergenic	-
4016427	A	G	intergenic	-
4030699	G	C	mutY	V67L
4060596	T	C	esxW	T2A
4087503	A	G	intergenic	-
4214150	T	C	Rv3768	L25P
4243197	G	C	intergenic	-
4243224	C	A,T,G	intergenic	-
4243228	C	T	intergenic	-
4247436	A	C,G	embB	M306L,V
4247438	G	A,C,T	embB	M306I,I,I
4247476	A	G	embB	Y319C
4247581	A	C	embB	D354A
4247653	A	C	embB	E378A
4247737	G	A,C	embB	G406D,A
4248009	C	A	embB	Q497K
4248010	A	G	embB	Q497R
4249525	A	G	embB	H1002R
4249590	G	A	embB	D1024N
4252886	C	T	intergenic	-
4269096	C	T	ubiA [c]	A249T
4269305	A	G	ubiA [c]	I179T
4313163	G	A	Rv3839	R131Q
4322607	G	A	Rv3848	S92N
4327072	A	G	ethA	C137R
4327487	A	G	intergenic	-
4338372	A	G	whiB6 [c]	C53R
4338607	A	C	intergenic [c]	-
4338609	T	A	intergenic [c]	-
4359172	G	C	espK	T206T
4371255	G	A	eccD2	A146V
4398755	G	A	Rv3910	G718S

[a] Sites for which >5% isolates had an ambiguous basecall were excluded.
[b] Novel regions putatively associated with drug resistance identified by Farhat et al.[14]
[c] Novel regions putatively associated with drug resistance identified by Zhang et al.[15]

Supplementary Table 11: Comparison of clustering rate of drug resistance SNPs

	Number of isolates		
Locus	**unclustered**	**clustered**	**P**
katG	27	558	<0.001 [a]
P_{inhA}	19	51	<0.001 [a]
rpoB	62	448	<0.001 [a]
P_{pncA}	11	10	0.758
pncA	67	238	<0.001 [a]
P_{embA}	16	25	0.047 [a]
embB	49	415	<0.001 [a]
rpsL	19	295	<0.001 [a]
rrs_{str}	7	255	<0.001 [a]
gidB	17	30	0.007 [a]
rrs_{inj}	18	22	0.371
P_{eis}	38	284	<0.001 [a]
gyrA	52	34	0.006 [b]
gyrB	7	4	0.395
P_{ethA}	0	216	<0.001 [a]
ethA	20	95	<0.001 [a]

[a] SNPs were significantly more likely to be found in phylogenetic clusters
[b] SNPs were significantly less likely to be found in phylogenetic clusters

References

1. Petrella, S. *et al.* Crystal structure of the pyrazinamidase of *Mycobacterium tuberculosis*: insights into natural and acquired resistance to pyrazinamide. *PLOS ONE* **6,** e15785 (2011).