PNAS
www.pnas.org

Supplementary Information for

Pre-detection history of extensively drug-resistant tuberculosis in KwaZulu-Natal, South Africa

Tyler S. Brown, Lavanya Challagundla, Evan H. Baugh, Shaheed Vally Omar, Arkady Mustaev, Sara C Auld, N Sarita Shah, Barry N. Kreiswirth, James CM Brust, Kristin N. Nelson, Apurva Narechania, Natalia Kurepina, Koleka Mlisana, Richard Bonneau, Vegard Eldholm, Nazir Ismail, Sergios-Orestis Kolokotronis, D. Ashley Robinson, Neel R Gandhi, Barun Mathema

Barun Mathema
Email: bm2055@cumc.columbia.edu

This PDF file includes:

www.pnas.org/cgi/doi/10.1073/pnas.1906636116

Supplementary Methods

Whole genome sequence data processing and variant calling. Raw paired-end reads were filtered for length and trimmed for quality (Trim Galore, Babraham Bioinformatics) and duplicate reads were removed following alignment to the H37Rv reference genome (NC_000962.3) using the Burrows-Wheeler Aligner,(1) similar to the pre-processing pipeline described by O'Neill et al.(2) All isolates included in the analysis had reads covering >99% of the reference genome and average read depth >15x. SNPs were identified using Samtools v0.1.19 (3), and filtered for quality, read consensus (>75% reads supporting the alternate allele), and proximity to indels. Polymorphisms in or within 50 base pairs of hypervariable PPE/PE gene families, repeat regions, and mobile elements were excluded, similar to prior studies using WGS from *Mtb* (4). Drug resistance-conferring mutations were identified from whole genome sequence data in conjunction with targeted sequencing data described above. Genome assemblies were constructed *de novo* using ABySS (5).

Phylogenomic analysis and neutrality statistics. Although root-to-tip distance from an undated maximum-likelihood tree was positively correlated with increasing tip date in linear regression, time-scaled substitution rate estimates were not significantly different from those based on cluster-randomized tip dates(6) in most replicates, indicating that a strong temporal signal was not present in our sample (Figure S13). For this reason, we used a strict molecular clock and an informative prior on the mutation rate in all BEAST analyses, using the range of prior empiric estimates of the *Mtb* mutation rate derived from WGS data(7, 8) to define a normal distribution around 1.2E-7 (95%CI: 8.38E-8 – 1.56E-7) SNPs/site-year. To improve MCMC mixing and convergence, in some analyses we randomly downsampled the genetically monomorphic LAM4/KZN clade from 250 sequences to 50. Estimated sample sizes for all non-nuisance parameters in each BEAST run were > 200. We compared different population models in BEAST using via stepping-stone marginal likelihood estimation. We used DNASP v6(9) and the R package *pegas* to calculate neutrality statistics over the entire genome and by gene. We used a sublineage 2.2 isolate as the outgroup for analyses of LAM4/KZN and estimated p-values via coalescent simulation. We calculated Weir and Cockerham's F_{ST} for different subpopulations of interest using the R package *hierfstat*. We tested for differences between terminal branch lengths by clade using both the Mann-Whitney U test (one-sided with continuity correction) and a permutation testing comparing the mean terminal branch length against a null distribution generated by randomly permuting subpopulation assignments.

Biophysical modeling of rpoB mutations. We used Rosetta v 3.9(10) and VIPUR(11) to investigate the structural and energetic impact of *rpoB* mutations unique to LAM4/KZN. Rosetta has been used previously to interpret the energetic impact of nonsynonymous mutations(11, 12) and is capable of modeling both protein-RNA(13) and protein-protein interactions.(14) *Mtb* has only one RNA polymerase complex (RNAP) composed of several essential proteins, including the β, β', and α subunits encoded by *rpoB*, *rpoC*, and *rpoA* respectively. We used the Protein Data Bank (PBD) structure of the transcription initiation complex 5UH8(15) and removed unnecessary proteins (all but chain C). To assess the energetic impact of each mutation or combination of mutations, we ran Rosetta high resolution docking (10,000 trajectories) and quantified the energetic effect of each mutation on RNAP β subunit stability, RNAPβ–RNA interaction, and any

effect on the whole protein complex. Electrostatic surfaces for the *rpoB* active site were assessed using APBS through the PyMOL plugin.

To assess the energy of the protein-RNA interaction, we used Rosetta high resolution docking to refine the docking interface, eliminating potential artifacts or defects and providing an evaluation of the interaction energy in different conformations. When using Rosetta to predict structural models, the model with the lowest energy is usually determined to be most representative of the single, lowest energy structure though mutations can alter conformational sampling or the distribution of native-like states, which can be overlooked by focusing only on the best model. We use the average Rosetta energy across the 10,000 samples to represent the mutation effect.

While some methods assessing the energetic impact of a mutation focus only on local structural context, we have characterized the energetic impact of each mutant by evaluating the total energy of the RNAP β subunit. We have previously identified that there are many "long-range" mutational effects that can alter the structure and energetics of a protein far from the site of mutation, requiring assessment of the entire protein energy.(11) We attempted Rosetta docking with all nucleotide chains from 5UH8 but found that the additional constraint provided by the size of these chains and the lack of nucleotide-sampling in Rosetta prevented the RNAP β subunit from adopting diverse conformations during sampling. To focus on the interaction of the RNAP β subunit and RNA, we truncated the nascent RNA and template strand DNA to 10 nucleotides in the active site. We explored numerous Rosetta scoring schemes to account for possible RNA-protein molecular interactions and used the recently developed *rna res level energy7beta* energy function. This energy function is tuned to account for protein energetics while better accounting for electrostatics (from the nucleotide backbone) and delocalized p-orbital ring electrons, allowing for potential interaction between amino acid side-chains and the nucleotide bases. In Rosetta docking, the energies of the individual molecules and the total complex can be calculated. By removing the nucleotide chains from their docked positions and re-evaluating the Rosetta energy, we can calculate the apparent energy of interaction (the difference between the individual energies of the macromolecules). For each trajectory in the docking simulation we have a value for the total energy and the nucleotide-protein interaction.

Spatial clustering of rpoC compensatory mutations. We evaluated the spatial clustering of eight *rpoC* compensatory mutations using the recently developed K(t) distance metric.(16) The K(t) is measured as the fraction of mutations within a specified distance (t) and is compared to permutations of randomly selected positions in the same structure. We calculated K(t) using the alpha carbon coordinates for each residue in the protein and compared the eight compensatory mutations to 10,000 random permutations of size eight.

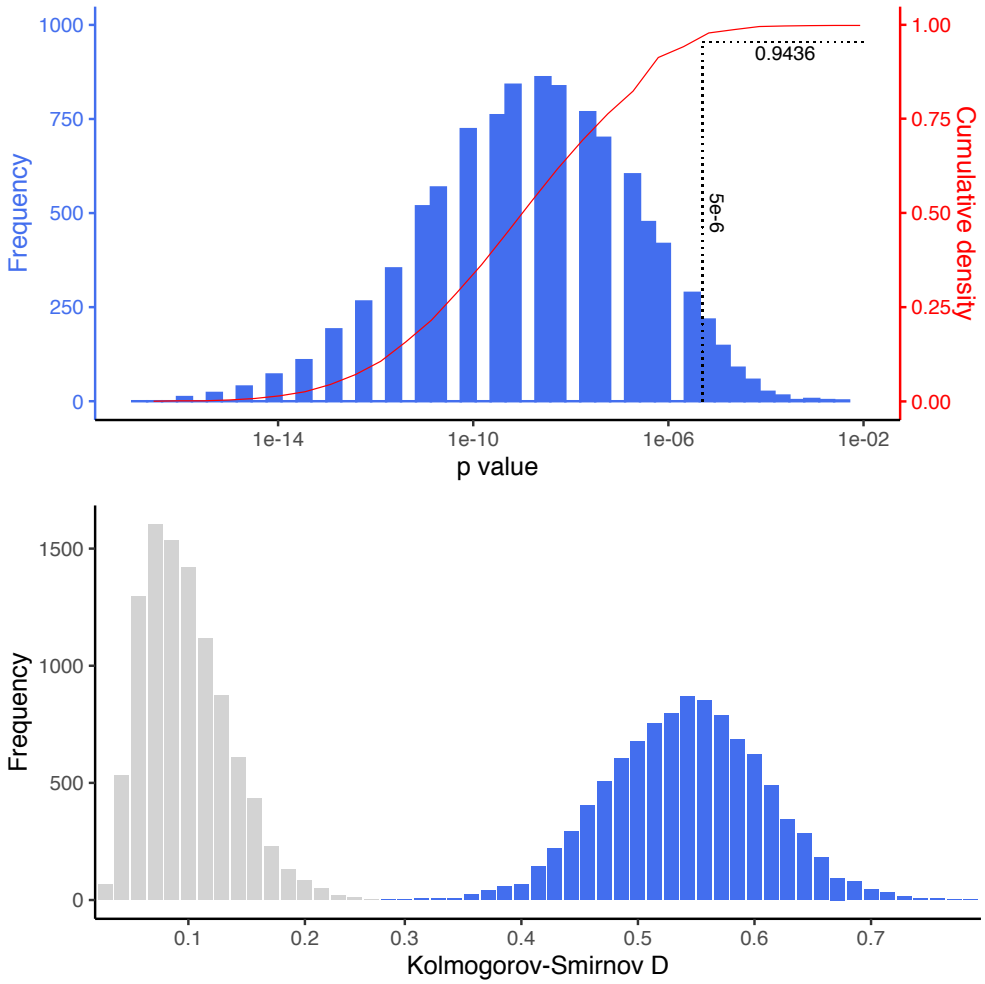


Fig. S1. Terminal branch length comparison for LAM4/KZN isolates versus non-LAM4/KZN isolates. Top panel: Distribution of p-values for Kolmogorov-Smirnov two-sample testing for 10,000 replicates comparing equal-sized samples from LAM4/KZN and non-LAM4/KZN isolates. Each replicate compares n=70 randomly selected isolates from each group, sampled with replacement. The red line displays the cumulative density of p-values, i.e. the proportion of all p-values that are smaller than the value given on the x-axis. All p-values in the distribution are < 0.005. The proportion of p-values smaller than 5E-6 (i.e. Bonferroni-corrected value for 10,000 tests) is labeled in black. Bottom panel: Observed (blue) and null (grey) distributions of the Kolmogorov-Smirnov test statistic (*D*), with 10,000 permutations in each distribution. The observed distribution is sampled as described for the two-sample test in the top panel. The null distribution was generated by taking 10,000 two-sided samples, each with 70 isolates, in which group labels (LAM4/KZN vs non-LAM4/KZN) are randomized across the two samples, and calculating *D*. The observed and null distributions are completely non-overlapping.

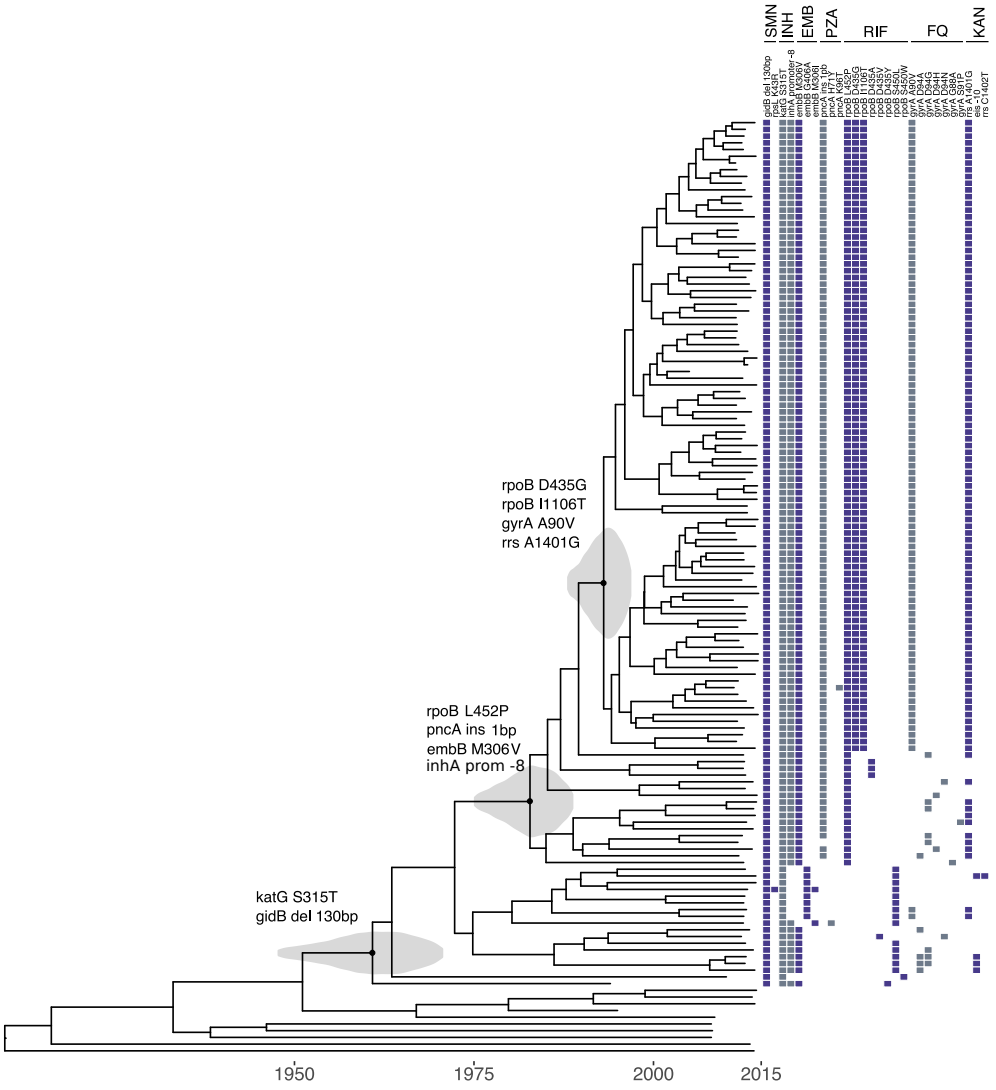


Fig. S2. Bayesian phylogenetic reconstruction for LAM4/KZN and closely related 4.3.2 isolates with estimated TMRCA for key drug resistance mutations. 95%HPD intervals for each TMRCA are indicated on corresponding nodes as violin plots. Estimated TMRCA (and 95%HPD intervals) for isolates carrying each mutation or set of mutations are: *katG* S315T, 1961 (1947-1970); *rpoB* L452P/*pncA* 1bp insertion/*embB* M306V/*inhA* promoter -8, 1983 (1975-1989); *gyrA* A90V/*rrs* a1401g/*rpoB* D435G/*rpoB* I1106T, 1993 (1988-1997).

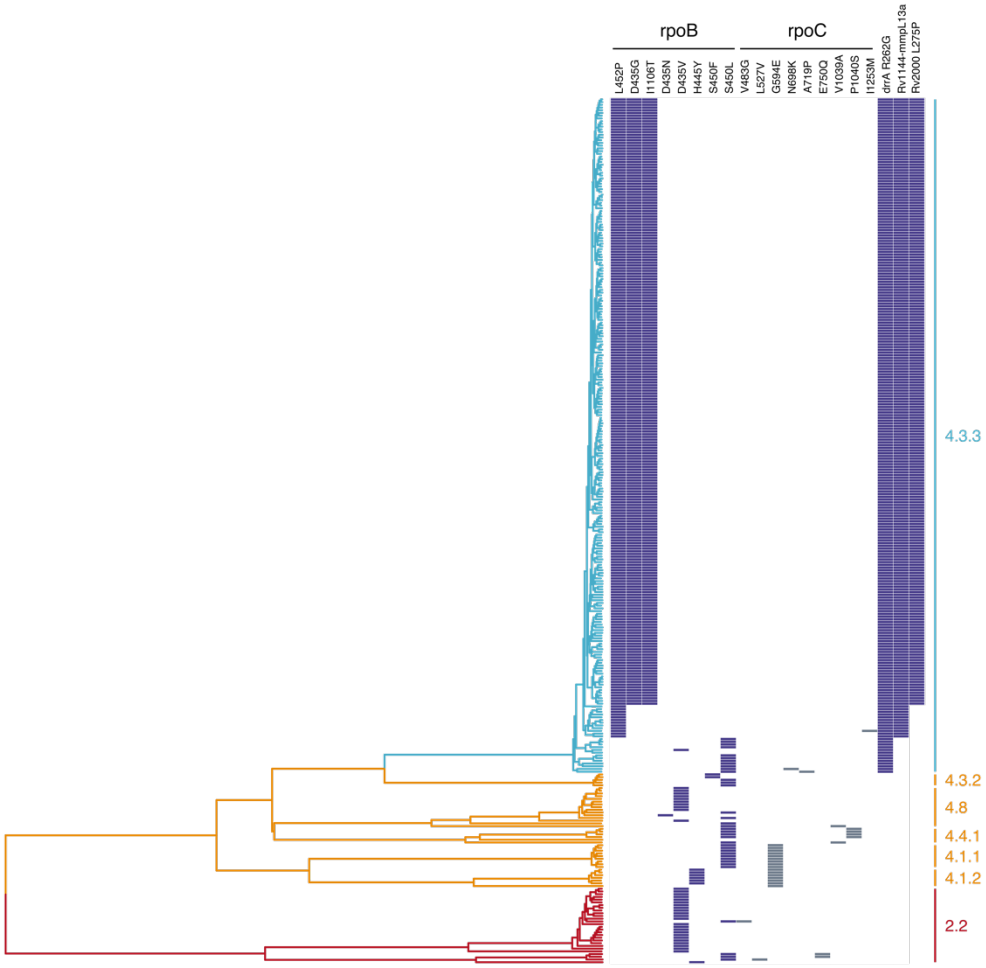


Fig. S3. Bayesian phylogenetic reconstruction for 318 XDR-TB isolates from KwaZulu-Natal, annotated with non-synonymous *rpoB* and *rpoC* mutations, plus *ddrA*, *Rv1144-mmpL13a* intergenic, and *Rv2000* mutations associated with XDR-TB phenotypes. Clades are colored by *Mtb* phylogeographic lineage (turquoise: LAM4/KZN/4.3.3; orange: non-LAM4/KZN lineage 4; red: lineage 2) and annotated using SNP-based sublineage classification per Coll *et al*.(17)

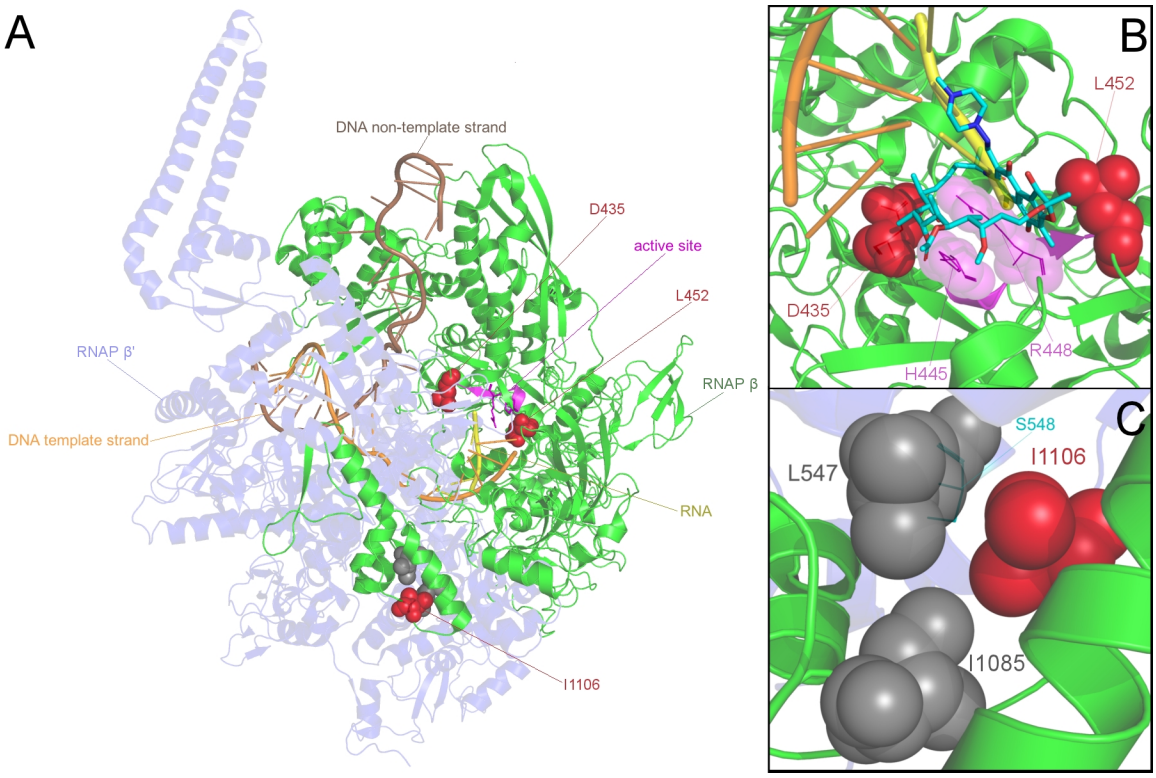


Fig. S4. *rpoB* mutations unique to LAM4/KZN occur within the RNAPβ and the RNAPβ-RNAPβ' interface. (A) All three of the mutations unique to LAM4/KZN (red) occur at important functional sites of RNAPβ. (B) RNAPβ L452P, corresponding to the first *rpoB* mutation acquired by LAM4/KZN, occurs adjacent to the protein active site in the so-called rifampin resistance-determining region, where it markedly reduces the stability of the protein (Rosetta change relative energy units, Δ_{REU} = +236, Table S5). Other *rpoB* mutations associated with decreased fitness in competitive growth assays(18) have similar destabilizing effects (Table S6). RNA docking analysis indicates that L452P still maintains favorable interaction with RNA that is nearly identical to wildtype. D435G, which we estimate was acquired approximately ten years after L452P, has a modest stabilizing effect on RNAPβ, partially mitigating the destabilization of RNAPβ L452P (Δ_{REU} = -6, relative to L452P single mutant). This stabilizing effect appears to result from reduced electric repulsion with the negatively charged nucleic acid backbone with the introduction of glycine at position 435 and may also restore flexibility to the region around the active site enhancing transcriptional efficiency.(19) (C) The third mutation, I1106T is far from the active site but occurs within the RNAPβ-RNAPβ' binding interface. This amino acid makes a specific contact (red and gray side-chains) to RNAPβ' and is spatially close to positions that are known to harbor compensatory mutations in RNAPβ' (Fig. S6) suggesting I1106T also favorably alters the RNAPβ-RNAPβ' interaction.

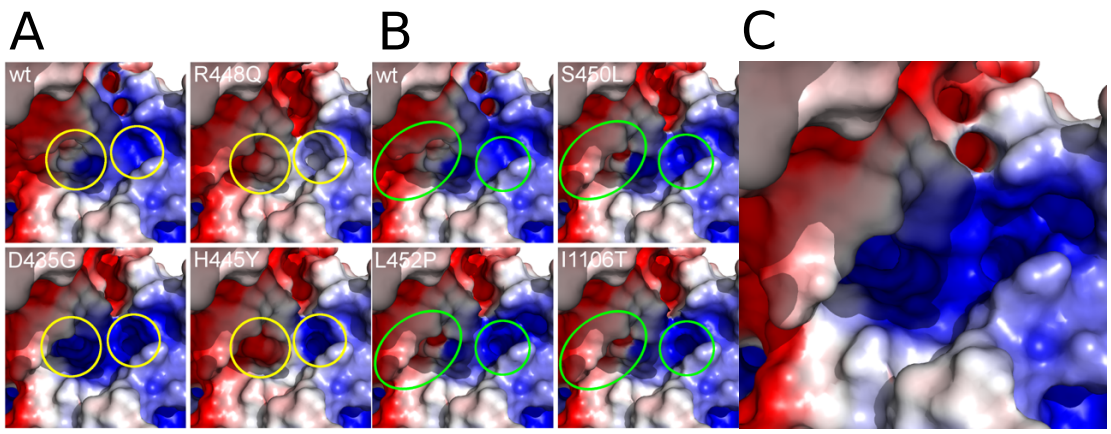


Fig. S5. APBS-derived electrostatic surfaces for charge-neutral (A, green circles) and charge-altering (B, yellow circles) mutations versus wildtype in the RNAPβ RNA active site. Positively charged regions are colored blue and negatively charged regions are colored red. D435G alters the distribution of charges in the active site both in isolation and in the presence of L452P (C), similar to prior observations on mutations at this site,(20) which may have an impact on activity or transcriptional targets.

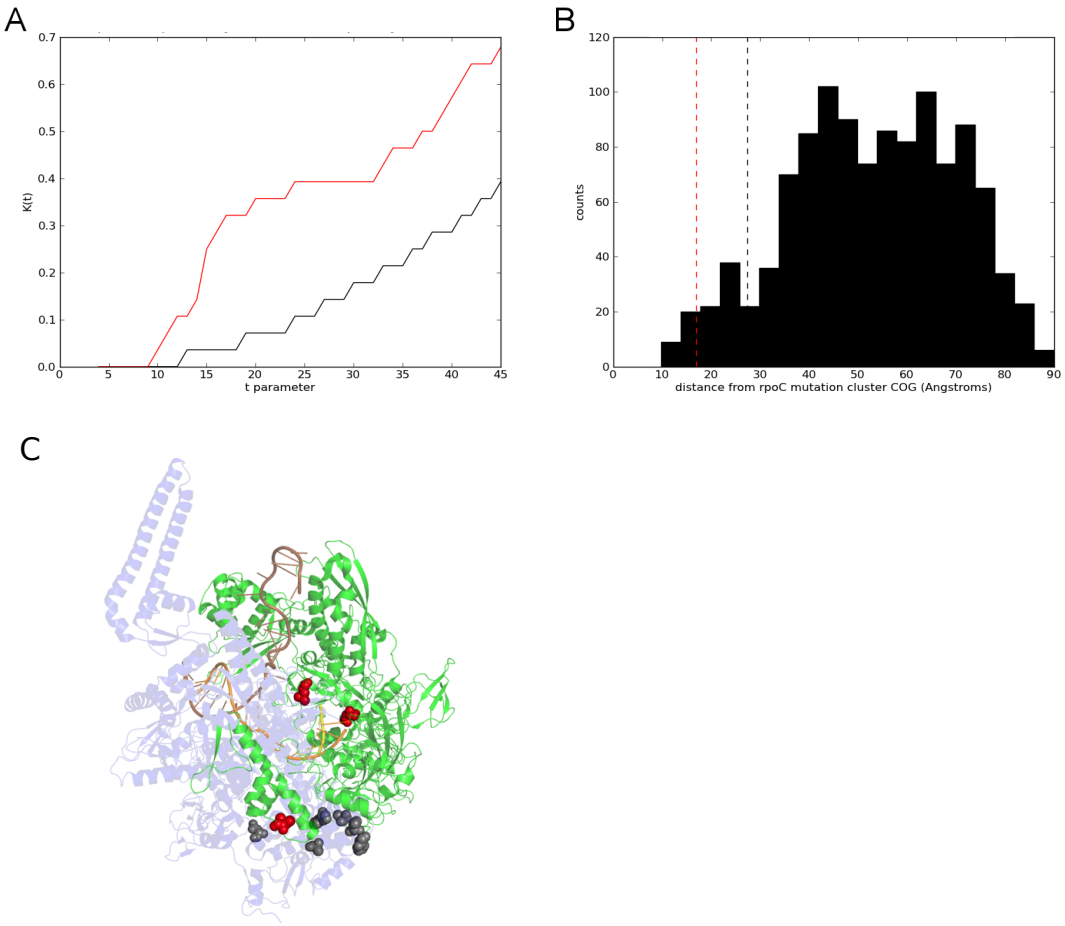


Fig. S6. I1106T is located near spatially clustered sites of compensatory mutations in *rpoC*. (A) Distances between eight compensatory mutations (red line) in *rpoC* are significantly closer together than random sets of the same size (black line). This median line (black) is derived from 10,000 random permutations. The area between this K(t) curve and the median curve is much higher than expected for random positions indicating these compensatory mutations in *rpoC* are clustered in space (p-value: 2.9E-4). (B) Many positions in *rpoB* are relatively close to the geometric center of the compensatory mutations in *rpoB*. The black dashed line is the approximate boundary of the mutation cluster and overlaps with many positions in *rpoB*. I1106T is very close to this cluster center and is within the 95th percentile (97.9%). (C) I1106T occurs along the RNAPβ-RNAPβ' protein binding interface. Although RNAPβ variants in this location are unique to LAM4/KZN, at least six putative compensatory mutations (grey spheres) have been identified in the adjacent region of RNAPβ'.

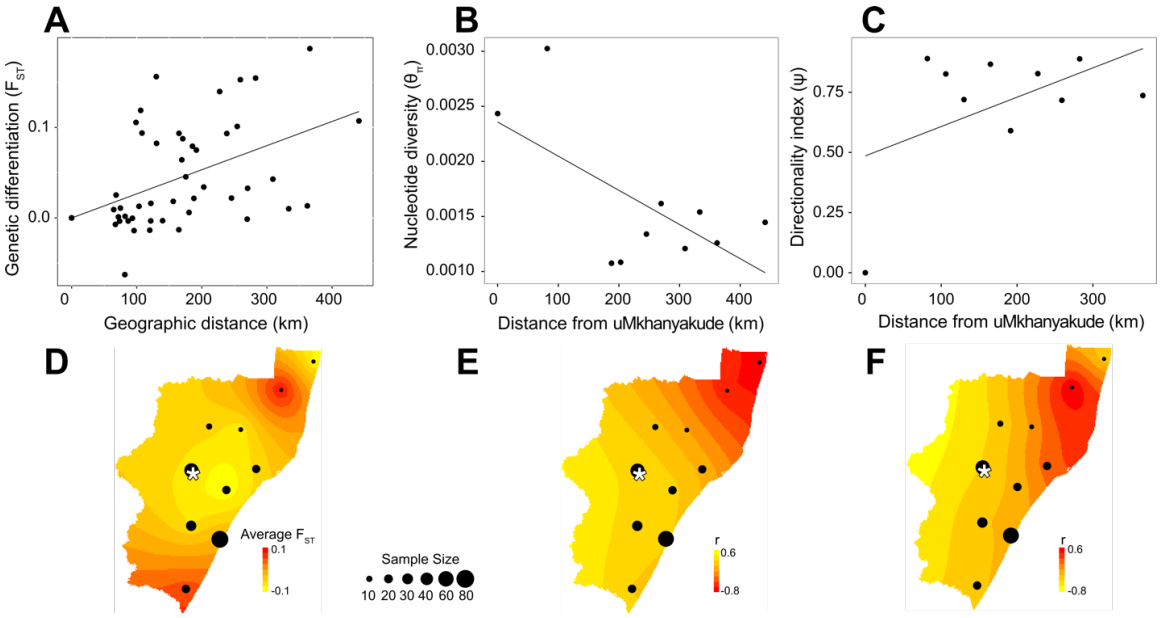
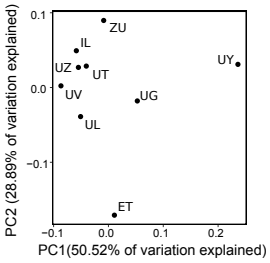


Fig. S7. Population genetic signatures of geographic range expansion from a common origin for LAM4/KZN isolates, using isolates geographically grouped by hierarchical clustering and haversine great-circle distances. (A) Pairwise F_{ST} vs geographic distance between isolates grouped by hierarchical clustering. (B and C) Linear regression of nucleotide diversity (π) or the directionality index (ψ) vs distance from uMkhanyakude district. (D) Average pairwise F_{ST} estimates for geographic clusters, with kriging interpolation between sampling points; red color indicates greater differentiation. (E and F) Spatial distribution of the correlations in B and C, with kriging interpolation between sampling points; red color indicates better evidence of origin. The location of Tugela Ferry is indicated with a star.

A



B

	ET	IL	UG	UL	UT	UV	UY	UZ	ZU
ET		0.472	0.033	0.164	0.017	0.414	**0.004**	0.047	**0.002**
IL	-0.008		0.19	0.223	0.179	0.409	0.289	0.288	0.628
UG	0.081	0.016		0.021	0.022	0.045	**0.005**	0.059	0.078
UL	0.02	0.018	0.08		0.129	0.173	**0.001**	0.279	0.104
UT	0.09	0.017	0.071	0.024		0.622	**0.003**	0.504	0.572
UV	-0.001	-0.002	0.056	0.017	-0.006		0.009	0.708	0.614
UY	0.174	0.011	0.119	0.126	0.123	0.106		0.007	0.055
UZ	0.079	0.005	0.057	0.004	-0.003	-0.008	0.117		0.519
ZU	0.138	-0.014	0.034	0.033	-0.007	-0.008	0.059	-0.006	

Fig. S8. (A) Principal component analysis and (B) pairwise F_{ST} values for LAM4/KZN subpopulations by district in KwaZulu-Natal. The lower triangular matrix in (B) shows pairwise F_{ST} values and upper triangular matrix shows p-values for corresponding F_{ST} values. p-values ≤ 0.005 are highlighted in bold text. Districts are abbreviated as follows: eThekwini (ET), iLembe (IL), Ugu (UG), uThukela (UL), uThungulu (UT), uMgungundlovu (UV), uMkhanyakude (UY), uMzinyathi (UZ), and Zululand (ZU).

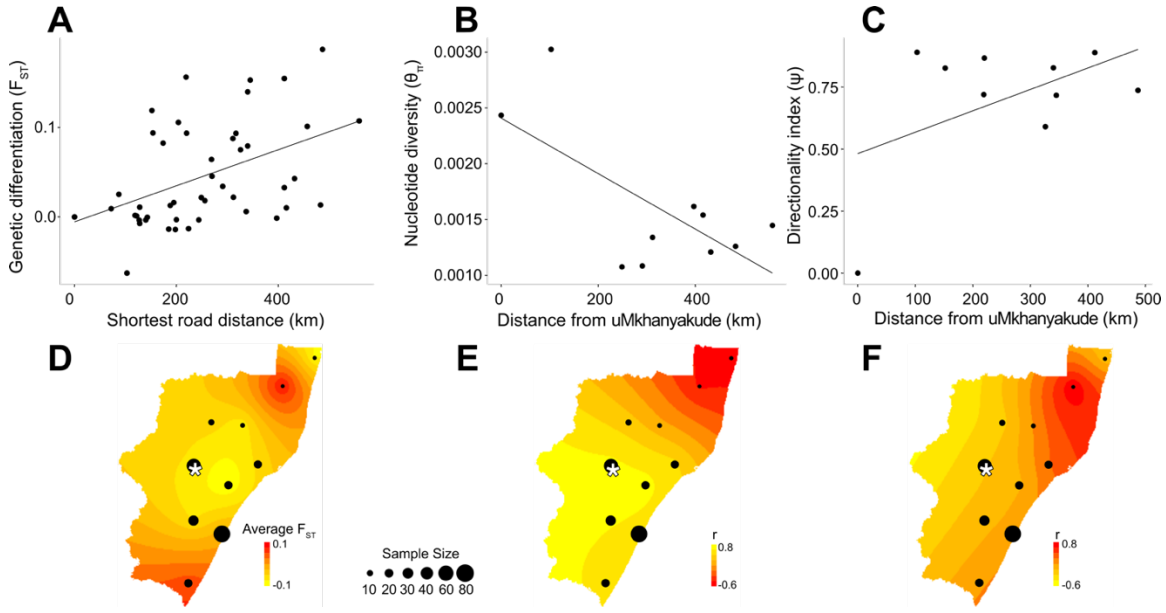


Fig. S9. Population genetic signatures of geographic range expansion from a common origin for LAM4/KZN isolates, using isolates geographically grouped by hierarchical clustering and shortest road distances**.** (A) Pairwise F_{ST} vs shortest road distance between isolates grouped by hierarchical clustering. (B and C) Linear regression of nucleotide diversity (π) or the directionality index (ψ) vs distance from uMkhanyakude district. (D) Average pairwise F_{ST} estimates for geographic clusters, with kriging interpolation between sampling points; red color indicates greater differentiation. (E and F) Spatial distribution of the correlations in B and C, with kriging interpolation between sampling points; red color indicates better evidence of origin.

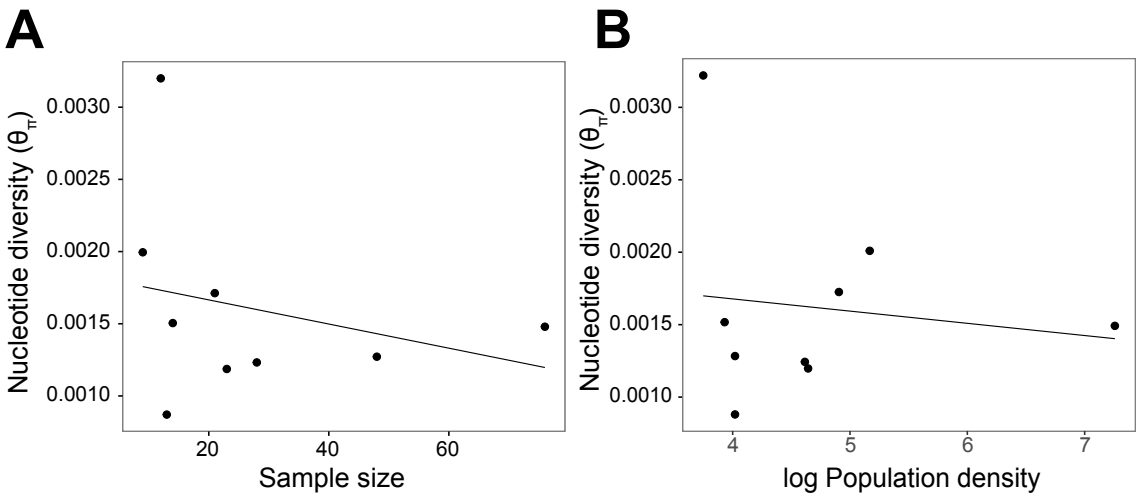


Fig. S10. Nucleotide diversity vs sample size (A) and log population density (B) for isolates grouped by district. Nucleotide diversity for isolates groups are not correlated with either sample size (r=-0.27, P=0.485) or log-transformed population density (r=-0.13, P=0.739).

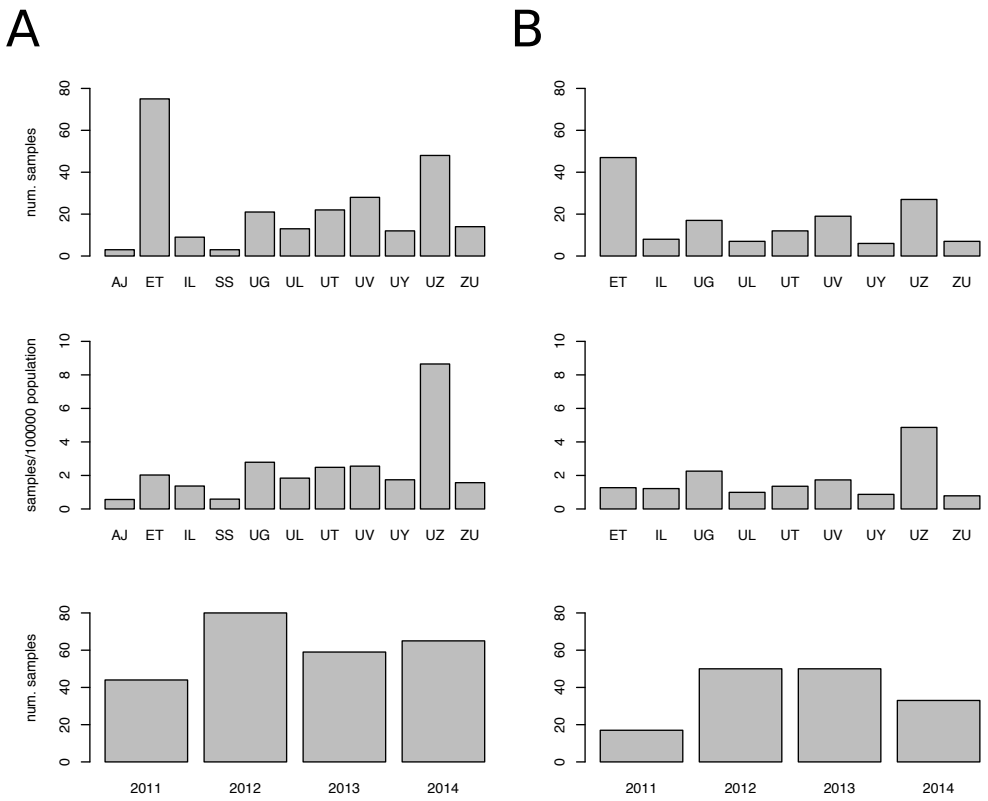


Fig. S11. Count and count per 10,000 population for LAM4/KZN XDR-TB isolates by district and by year. (A) Complete set of 250 isolates, (B) Down-sampled set of 50 isolates used in some analyses.

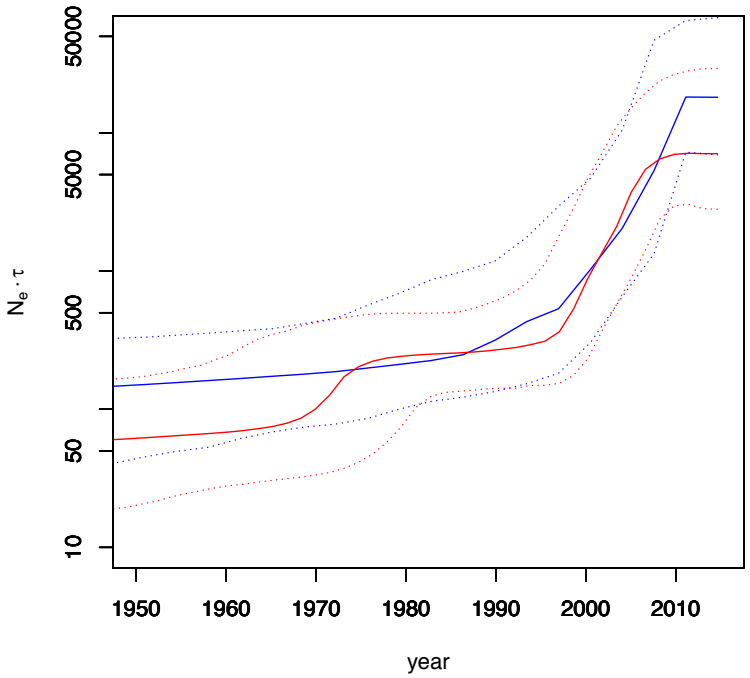


Figure S12. Bayesian skyline analysis for sequence alignment including all LAM4/KZN isolates (blue) and sequence alignment sampled to include only 50 LAM4/KZN isolates (red). Solid lines represent median values and dashed lines represent boundaries of the 95%HPD interval.

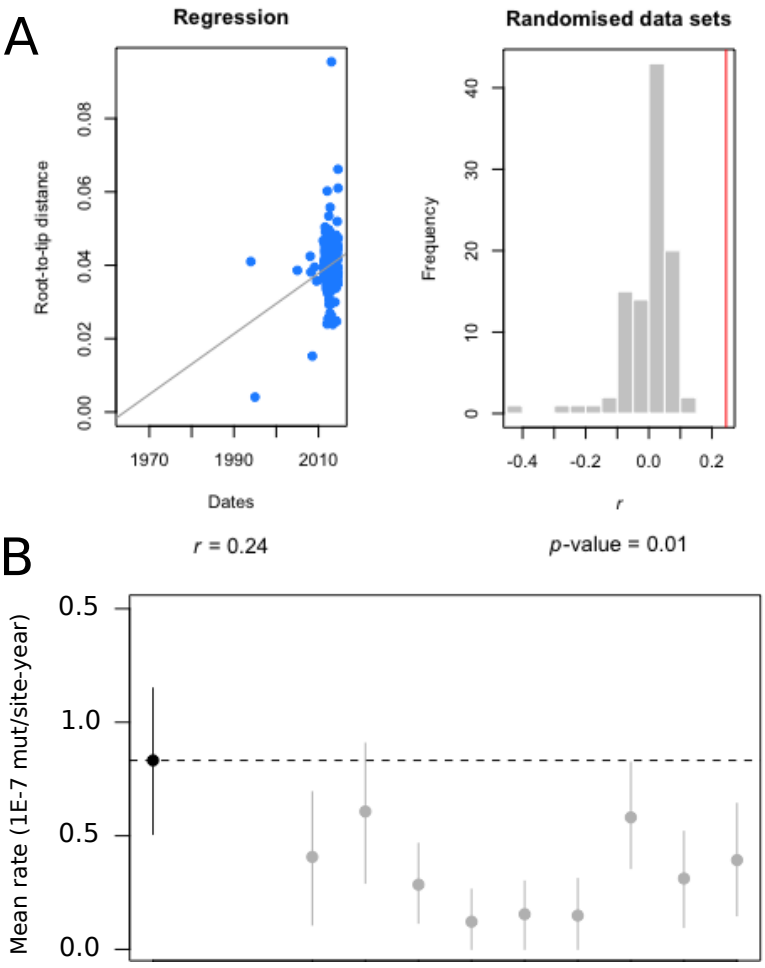


Fig. S13. Testing for temporal signal in tip-dated phylogenies**.** (A) Regression on root-to-tip distance versus tip date with p-value for *r* estimated with 10,000 tip date-randomized data sets. (B) Substitution rate estimated via Bayesian phylogenomic analysis for sequence data with true tip dates (black circle) vs cluster-randomized tip dates (gray). Two data sets with randomized tip dates yielded 95%HPD intervals (whiskers) that overlap with the estimated median value obtained from sequence data with true tip dates, indicating that only weak temporal signal is present in the available data.

Table S1. Comparison between population models in BEAST. MCMC chains were run with 250,000,000 states and 25% burn-in. Log marginal likelihoods (Log ML) estimated via stepping-stone sampling minimally favor the Bayesian Skyline model and logistic growth over constant population size and exponential growth, but the Bayes factors for these comparisons do not indicate significant differences in marginal likelihood between models.

		Population model			
		Constant	**Exponential**	**Logistic**	**Bayesian Skyline**
ESS (likelihood)		2116	1036	929	1010
Mutation rate	**Mean**	1.120E-7	1.300E-7	1.031E-7	1.128E-7
	95%HPD	(8.96E-8, 1.49E-7)	(9.87E-8, 1.61E-7)	(7.26E-8, 1.35E-7)	(8.35E-8, 1.41E-7)
	ESS	8053	8265	6487	6520
Root age	**Mean**	1878	1907	1864	1873
	95%HPD	(1837,1915)	(1877, 1935)	(1810,1912)	(1831, 1913)
	ESS	8872	8429	7470	6845
Log ML		-5951904	-5951884	-5951836	-5951841
Growth rate	**Range**		(0.0256, 0.0941)		

Table S2. Genome-wide values for site frequency spectrum-based neutrality statistics. OG: outgroup. *R2*: Ramos-Onsins and Rozas R_2, *D*: Tajima's *D*, *H*: Fay and Wu's *H*, *E*: Zheng's *E*, $P_{D\text{-}H}$: p-value of the *D*-*H* test. P-values are based on 50,000 coalescent simulations of neutral evolution. Significant negative values for Tajima's *D* indicate a relative abundance of low-frequency alleles, which can result from multiple processes including a selective sweep and population expansion following a bottleneck. Fay and Wu's *H*, which compares high- and intermediate-frequency alleles, is expected to be less influenced by population expansion and thus more sensitive for the detection of selection. The *D*-*H* test, which jointly evaluates *D* and *H*, was developed with goal of detecting selection, and is predicted to be most sensitive for detection of selective sweeps on advantageous alleles prior to fixation (21). Zheng's *E*, which contrasts low- and high-frequency alleles, is expected to be more sensitive to population expansion than selection, and *R2* is a highly sensitive test for population expansion. Despite the predicted behavior of these statistics, all of them are sensitive to both demographics and selection to different degrees (22, 23). *H* and *E* employ an outgroup to determine the mean number of mutations since a most recent common ancestor, and the behavior of these statistics can be strongly influenced by outgroup selection (24). Results across different outgroups show significant departure from neutrality, such that the null hypothesis of an equilibrium population of constant population size can be rejected. Using an isolate from a sister clade (4.3.2) as the outgroup, where more sites are expected to be counted as derived alleles accrued since the (more distant) recent common ancestor, yields neutrality statistics most consistent with positive selection rather than population expansion. Similar results are obtained with a more distant outgroup (2.2.1). With a more phylogenetically proximate outgroup (a pan-susceptible LAM4/KZN isolate ancestral to the XDR LAM4/KZN clade), *H* is non-significant, the *D*-*H* test is non-significant, *E* is significantly negative, indicating that population expansion, rather than selection, is the primary process influencing the site frequency-spectrum over this more recent time period (i.e. since divergence from more recent common ancestor).

OG	*D*	*H*	$P_{D\text{-}H}$	*E*	*R2*
4.3.3	-2.6281 (<0.00001)	-0.7026 (0.1561)	0.0712	-1.7846 (0.0057)	0.0322 (0.0003)
4.3.2	-2.751 (<0.00001)	-5.7400 (0.00002)	<0.00001	2.6176 (0.9970)	0.0322 (0.0003)
2.2.1	-2.751 (<0.00001)	-3.332 (0.00632)	0.0027	0.459 (0.7610)	0.0322 (0.0003)

Table S3. Genetic differentiation between XDR LAM4/KZN isolates collected during different years. The lower triangular matrix shows pairwise F_{ST} values and upper triangular matrix shows corresponding p-values. The lowest p-value (0.027, for the comparison between 2011 and 2013) is non-significant after Bonferroni-correction for multiple testing.

	2011	**2012**	**2013**	**2014**
2011		0.316	0.027	0.098
2012	0.0020		0.598	0.869
2013	0.0854	-0.0086		0.127
2014	0.0670	-0.0479	0.0575	

Table S4. Collection date (year-month) and NCBI BioSample number for *M. tuberculosis* whole genome sequence data used in this study (NCBI BioProject Number PRJNA476470).

Sample ID	BioSample	Date
30569_S7	SAMN09566388	2011-05
30571_S9	SAMN09566389	2011-05
30575_S11	SAMN09566390	2011-07
30577_S1	SAMN09566391	2011-07
30579_S5	SAMN09566392	2011-08
30584_S9	SAMN09566393	2011-09
30585_S13	SAMN09566394	2011-08
30643_S17	SAMN09566395	2011-09
30644_S21	SAMN09566396	2011-07
30646_S29	SAMN09566397	2011-11
30647_S20	SAMN09566398	2011-10
30648_S2	SAMN09566399	2011-11
30994_S14	SAMN09566400	2011-05
30997_S18	SAMN09566403	2012-01
31002_S16	SAMN09566405	2011-10
31006_S30	SAMN09566406	2012-01
31007_S18	SAMN09566407	2012-02
31008_S23	SAMN09566408	2012-02
31010_S7	SAMN09566409	2012-12
31012_S19	SAMN09566410	2012-12
31015_S20	SAMN09566411	2012-02
31023_S22	SAMN09566413	2012-04
31141_S27	SAMN09566414	2011-08
31471_S12	SAMN09566415	2012-11
31737_S1	SAMN09566416	2012-05
31738_S16	SAMN09566417	2012-05
31739_S1	SAMN09566418	2012-03
31740_S2	SAMN09566419	2012-05
31741_S2	SAMN09566420	2012-04
31742_S3	SAMN09566421	2012-05
31743_S4	SAMN09566422	2012-04
31745_S20	SAMN09566423	2012-05
31746_S3	SAMN09566424	2012-06
31747_S24	SAMN09566425	2012-06
31748_S6	SAMN09566426	2012-06
31749_S7	SAMN09566427	2012-05
31750_S26	SAMN09566428	2012-05
31751_S28	SAMN09566429	2013-6
31752_S32	SAMN09566430	2012-06
31753_S1	SAMN09566431	2012-02
31754_S5	SAMN09566432	2012-07
31755_S9	SAMN09566433	2012-02
31756_S13	SAMN09566434	2012-02
31757_S17	SAMN09566435	2012-07
31758_S4	SAMN09566436	2011-11
31759_S21	SAMN09566437	2012-04
31760_S25	SAMN09566438	2012-07
31761_S29	SAMN09566439	2012-07
31766_S27	SAMN09566440	2012-03
31767_S28	SAMN09566441	2012-08
31771_S5	SAMN09566442	2013-6
31772_S29	SAMN09566443	2012-07
31776_S6	SAMN09566444	2013-6
31778_S7	SAMN09566445	2012-09
32060_S8	SAMN09566446	2012-08
32061_S9	SAMN09566447	2012-12
32062_S10	SAMN09566448	2012-12
32063_S11	SAMN09566449	2013-01
32064_S12	SAMN09566450	2012-10
32065_S13	SAMN09566451	2012-10
32204_S8	SAMN09566452	2011-04
32205_S9	SAMN09566453	2011-06
32207_S6	SAMN09566454	2011-10
32208_S10	SAMN09566455	2011-11
32209_S26	SAMN09566456	2013-12
32211_S11	SAMN09566457	2011-10
32212_S12	SAMN09566458	2011-10
32213_S13	SAMN09566459	2011-10
32214_S10	SAMN09566460	2013-01
32215_S27	SAMN09566461	2012-09
32216_S28	SAMN09566462	2013-12
32218_S14	SAMN09566463	2012-10
32219_S15	SAMN09566464	2012-10
32220_S16	SAMN09566465	2012-12
32221_S17	SAMN09566466	2012-07
32222_S14	SAMN09566467	2012-10
32223_S18	SAMN09566468	2012-12
32224_S19	SAMN09566469	2012-11
32225_S20	SAMN09566470	2013-01
32226_S21	SAMN09566471	2012-12
32227_S22	SAMN09566472	2012-12
32228_S24	SAMN09566473	2012-11
32229_S23	SAMN09566474	2012-08
32230_S29	SAMN09566475	2013-02
32231_S30	SAMN09566476	2013-01
32234_S31	SAMN09566477	2013-12
32235_S25	SAMN09566478	2012-04
32236_S30	SAMN09566479	2013-01
32237_S27	SAMN09566480	2013-02
32238_S28	SAMN09566481	2013-12
32240_S18	SAMN09566482	2012-07
32242_S1	SAMN09566483	2013-03
32243_S29	SAMN09566484	2012-10
32244_S30	SAMN09566485	2013-01
32245_S2	SAMN09566486	2013-01
32247_S31	SAMN09566488	2012-09

Table S4 (continued)

Sample ID	BioSample	Date
32248_S32	SAMN09566489	2013-03
32276_S31	SAMN09566490	2013-03
32277_S3	SAMN09566491	2012-09
32278_S22	SAMN09566492	2013-03
32279_S4	SAMN09566493	2012-12
32281_S26	SAMN09566494	2013-01
32283_S30	SAMN09566495	2013-12
32284_S3	SAMN09566496	2012-12
32285_S7	SAMN09566497	2012-08
32286_S11	SAMN09566498	2013-04
32287_S5	SAMN09566499	2013-06
32288_S15	SAMN09566500	2013-05
32289_S6	SAMN09566501	2011-09
32290_S19	SAMN09566502	2011-11
32291_S7	SAMN09566503	2012-05
32292_S8	SAMN09566504	2013-12
32294_S27	SAMN09566505	2013-12
32295_S31	SAMN09566506	2013-06
32296_S4	SAMN09566507	2013-05
32298_S12	SAMN09566508	2013-06
32299_S16	SAMN09566509	2013-03
32301_S10	SAMN09566510	2012-10
32302_S20	SAMN09566511	2013-06
32303_S24	SAMN09566512	2013-05
32304_S28	SAMN09566513	2013-05
32305_S32	SAMN09566514	2013-07
32827_S11	SAMN09566515	2012-07
32828_S12	SAMN09566516	2013-07
32829_S13	SAMN09566517	2013-06
32830_S14	SAMN09566518	2013-05
32832_S16	SAMN09566519	2013-08
32833_S17	SAMN09566520	2013-08
32834_S18	SAMN09566521	2013-05
32835_S19	SAMN09566522	2013-05
32837_S21	SAMN09566523	2012-10
32840_S24	SAMN09566524	2013-09
32841_S25	SAMN09566525	2012-11
32843_S27	SAMN09566526	2013-09
32844_S28	SAMN09566527	2013-11
32845_S29	SAMN09566528	2013-10
32846_S30	SAMN09566529	2013-10
32847_S31	SAMN09566530	2013-11
32848_S1	SAMN09566531	2013-08
32849_S32	SAMN09566532	2013-11
32850_S2	SAMN09566533	2013-11
32851_S3	SAMN09566534	2014-6
32852_S4	SAMN09566535	2013-10
32853_S5	SAMN09566536	2013-10
32854_S6	SAMN09566537	2013-12
32856_S8	SAMN09566538	2013-11
32857_S9	SAMN09566539	2014-01
32858_S10	SAMN09566540	2014-6
32859_S11	SAMN09566541	2013-08
32860_S12	SAMN09566542	2013-07
32861_S13	SAMN09566543	2013-11
32862_S14	SAMN09566544	2013-10
32863_S15	SAMN09566545	2013-12
32864_S16	SAMN09566546	2013-09
32865_S17	SAMN09566547	2013-07
32866_S18	SAMN09566548	2013-07
32867_S19	SAMN09566549	2014-01
32868_S20	SAMN09566550	2013-12
32869_S21	SAMN09566551	2013-08
32870_S22	SAMN09566552	2013-10
32871_S23	SAMN09566553	2013-11
33050_S24	SAMN09566554	2014-8
33051_S25	SAMN09566555	2012-01
33052_S26	SAMN09566556	2011-07
33053_S27	SAMN09566557	2012-04
33054_S28	SAMN09566558	2012-04
33055_S29	SAMN09566559	2012-10
33057_S30	SAMN09566560	2013-11
33058_S31	SAMN09566561	2013-11
33059_S32	SAMN09566562	2013-08
33060_S11	SAMN09566563	2014-02
33061_S12	SAMN09566564	2014-01
33062_S13	SAMN09566565	2013-09
33063_S14	SAMN09566566	2013-08
33064_S15	SAMN09566567	2013-08
33066_S16	SAMN09566568	2013-11
33067_S17	SAMN09566569	2014-02
33068_S18	SAMN09566570	2013-08
33069_S19	SAMN09566571	2014-01
33070_S20	SAMN09566572	2013-12
33071_S21	SAMN09566573	2014-02
33072_S22	SAMN09566574	2014-02
33073_S23	SAMN09566575	2014-02
33075_S24	SAMN09566576	2013-11
33076_S25	SAMN09566577	2014-02
33077_S26	SAMN09566578	2014-01
33078_S27	SAMN09566579	2013-08
33079_S28	SAMN09566580	2014-03
33080_S29	SAMN09566581	2014-02
33081_S30	SAMN09566582	2014-03
33082_S31	SAMN09566583	2014-02
33083_S1	SAMN09566584	2014-02

Table S4 (continued)

Sample ID	BioSample	Date
33084_S2	SAMN09566585	2014-02
33085_S3	SAMN09566586	2014-03
33086_S4	SAMN09566587	2013-12
33087_S5	SAMN09566588	2014-02
33088_S6	SAMN09566589	2014-02
33089_S7	SAMN09566590	2014-01
33090_S8	SAMN09566591	2014-03
33091_S9	SAMN09566592	2014-03
33092_S10	SAMN09566593	2014-03
33093_S3	SAMN09566594	2014-03
33094_S4	SAMN09566595	2014-02
33095_S6	SAMN09566596	2014-8
33096_S5	SAMN09566597	2014-04
33098_S8	SAMN09566598	2014-03
33100_S10	SAMN09566599	2014-04
33101_S11	SAMN09566600	2014-04
33102_S12	SAMN09566601	2014-04
33103_S13	SAMN09566602	2014-03
33104_S14	SAMN09566603	2014-04
33105_S15	SAMN09566604	2014-04
33106_S16	SAMN09566605	2014-03
33107_S11	SAMN09566606	2014-03
33108_S12	SAMN09566607	2014-04
33109_S13	SAMN09566608	2014-04
33110_S14	SAMN09566609	2014-04
33111_S15	SAMN09566610	2014-05
33112_S16	SAMN09566611	2014-05
33113_S17	SAMN09566612	2014-05
33114_S18	SAMN09566613	2014-05
62001_S17	SAMN09566614	2011-02
62009_S27	SAMN09566615	2011-03
62010_S9	SAMN09566616	2011-03
62014_S1	SAMN09566617	2011-05
62015_S28	SAMN09566618	2011-07
62016_S29	SAMN09566619	2011-07
62020_S5	SAMN09566621	2011-07
62021_S2	SAMN09566622	2011-07
62024_S6	SAMN09566623	2011-06
62025_S22	SAMN09566624	2011-08
62026_S8	SAMN09566625	2011-08
62029_S23	SAMN09566626	2011-07
62031_S10	SAMN09566627	2011-09
62032_S24	SAMN09566628	2011-09
62033_S11	SAMN09566629	2011-08
62034_S25	SAMN09566630	2011-10
62037_S12	SAMN09566631	2011-11
62049_S31	SAMN09566632	2011-07
62052_S13	SAMN09566633	2011-11
62059_S13	SAMN09566634	2011-10
62071_S30	SAMN09566635	2011-11
62072_S31	SAMN09566636	2012-02
62074_S29	SAMN09566637	2011-11
62084_S1	SAMN09566638	2011-08
62092_S2	SAMN09566639	2012-01
62095_S4	SAMN09566640	2012-01
62096_S5	SAMN09566641	2012-03
62097_S6	SAMN09566642	2012-03
62098_S7	SAMN09566643	2012-05
62101_S8	SAMN09566644	2011-12
62102_S30	SAMN09566645	2012-04
62105_S10	SAMN09566646	2012-04
62106_S7	SAMN09566647	2012-05
62107_S11	SAMN09566648	2012-03
62134_S14	SAMN09566649	2012-02
62135_S15	SAMN09566650	2012-03
62136_S16	SAMN09566651	2012-02
62137_S17	SAMN09566652	2012-04
62140_S19	SAMN09566653	2012-08
62141_S20	SAMN09566654	2012-06
62142_S12	SAMN09566655	2012-07
62147_S21	SAMN09566656	2012-12
62149_S22	SAMN09566657	2012-07
62152_S23	SAMN09566658	2012-10
62154_S14	SAMN09566659	2012-10
62155_S24	SAMN09566660	2012-10
62156_S15	SAMN09566661	2012-10
62158_S25	SAMN09566662	2012-10
62159_S26	SAMN09566663	2012-10
62164_S4	SAMN09566664	2012-09
62184_S32	SAMN09566665	2012-11
62191_S16	SAMN09566666	2012-11
62211_S27	SAMN09566667	2013-04
62214_S28	SAMN09566668	2013-02
T11_S11	SAMN09566669	2012-10
T18_S17	SAMN09566670	2013-09
T19_S18	SAMN09566671	2014-02
T20_S19	SAMN09566672	2014-02
T21_S20	SAMN09566673	2011-08
T22_S21	SAMN09566674	2011-10
T23_S22	SAMN09566675	2012-09
T24_S23	SAMN09566676	2012-09
T25_S24	SAMN09566677	2014-04
T26_S25	SAMN09566678	2013-09
T27_S26	SAMN09566679	2014-04
T28_S27	SAMN09566680	2013-10
T29_S28	SAMN09566681	2014-05

Table S4 (continued)

Sample ID	BioSample	Date
T30_S29	SAMN09566682	2013-08
T31_S30	SAMN09566683	2014-04
T32_S31	SAMN09566684	2014-05
T33_S32	SAMN09566685	2014-05
T34_S1	SAMN09566686	2013-12
T35_S2	SAMN09566687	2014-06
T38_S5	SAMN09566688	2014-06
T39_S6	SAMN09566689	2014-04
T40_S7	SAMN09566690	2014-06
T41_S8	SAMN09566691	2014-03
T42_S9	SAMN09566692	2014-06
T44_S11	SAMN09566694	2014-05
T45_S12	SAMN09566695	2014-05
T46_S13	SAMN09566696	2014-01
T47_S14	SAMN09566697	2014-06
T48_S15	SAMN09566698	2014-02
T49_S16	SAMN09566699	2014-07
T51_S18	SAMN09566700	2014-06
T52_S19	SAMN09566701	2014-06
T53_S20	SAMN09566702	2014-06
T55_S22	SAMN09566704	2014-05
T56_S23	SAMN09566705	2014-07
T57_S24	SAMN09566706	2014-07
T58_S25	SAMN09566707	2014-05
T59_S26	SAMN09566708	2014-08
T60_S27	SAMN09566709	2014-07
T61_S28	SAMN09566710	2014-07
T62_S29	SAMN09566711	2014-07
T63_S30	SAMN09566712	2014-07
T64_S31	SAMN09566713	2014-07
T65_S32	SAMN09566714	2014-07
T66_S1	SAMN09566715	2014-02

Sample ID	BioSample	Date
T67_S2	SAMN09566716	2014-06
T69_S4	SAMN09566718	2014-06
T70_S5	SAMN09566719	2013-12
T71_S6	SAMN09566720	2014-06
T72_S7	SAMN09566721	2014-08
T74_S9	SAMN09566723	2014-05
T76_S11	SAMN09566725	2014-07
T77_S12	SAMN09566726	2014-07
T80_S15	SAMN09566728	2014-07
T81_S16	SAMN09566729	2014-07
T82_S17	SAMN09566730	2014-08
T83_S18	SAMN09566731	2014-08

Table S5. Rosetta energy value changes for successive *rpoB* mutations. The structural and energetic impact of each mutation was considered by analyzing successive mutations for their overall effect on stability (stability column), the favorability of RNA interaction (RNA binding column), and the favorability of interaction with RNAPβ' (RNAPβ' binding column). Energy values were measured using Rosetta (3.9) as the difference between the mutant and the previous sequence (initially the difference of L452P from wildtype) and are displayed as relative energy units above. Both L452P and D435G are known to be associated with rifampin resistance.

Variant	Stability	RNA binding
L452P	235.78 ± 0.099	0.019 ± 0.011
L452P, D435G	-6.66 ± 0.099	-0.0079 ± 0.010
L452P, D435G, I1106T	1.45 ± 0.099	0.015 ± 0.010

Table S6. VIPUR and Rosetta analysis of single *rpoB* mutations. VIPUR pipeline predictions for the effect of the LAM4/KZN mutations and nine other drug resistance mutations described in Gagneux et al (18) are shown. VIPUR scores greater 0.5 are predicted to disrupt or alter protein function. Values in the *Ess* column represent the difference between structure-based and conservation-based features in VIPUR. High (> 0.2) *Ess* scores indicate mutations that are more conserved than can be explained by their structural disruption, suggesting they may act by altering specific functions or occur at important functional sites. *rpoB* mutations known to rifampin drug resistance consistently obtain VIPUR deleterious scores (>0.5). More destabilizing drug resistance mutations (those with the highest VIPUR scores, including H445P, S441L, and S450W) appear to generally destabilize protein folding, while other less destabilizing mutations disrupt specific side-chain interactions. The *Altered Electrostatics* column indicates whether each mutation alters the charge distribution within the *rpoB* active site.

	VIPUR		Rosetta		
Variant	**Score**	**Ess**	**Stability**	**RNA binding**	**Altered Electrostatics**
L452P	0.387	-0.126	235.78 ± 0.099	0.019 ± 0.011	no
D435G	0.843	0.049	-6.56 ± 0.099	0.027 ± 0.011	yes
I1106T	0.314	0.013	1.40 ± 0.099	-0.015 ± 0.010	no
H445P	0.955	0.17	219.50 ± 0.10	-0.031 ± 0.010	yes
S441L	0.898	-0.019	363.42 ± 0.10	-0.029 ± 0.010	no
S450W	0.882	-0.002	744.39 ± 0.099	0.0077 ± 0.011	no
Q432L	0.865	0.213	5.84 ± 0.099	-0.0055 ± 0.010	no
H445R	0.756	0.184	88.20 ± 0.10	-0.066 ± 0.010	no
H445Y	0.739	-0.025	173.029 ± 0.098	-0.053 ± 0.0097	yes
H445D	0.661	0.282	-6.13 ± 0.10	0.013 ± 0.010	yes
S450L	0.626	0.071	63.64 ± 0.099	-0.0036 ± 0.011	no
R448Q	0.599	0.292	3.01 ± 0.098	0.35 ± 0.010	yes

References

1. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).
2. M. B. O'Neill, T. D. Mortimer, C. S. Pepperell, Diversity of Mycobacterium tuberculosis across Evolutionary Scales. *PLoS Pathog* **11**, e1005257 (2015).
3. H. Li *et al.*, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
4. V. Eldholm *et al.*, Four decades of transmission of a multidrug-resistant Mycobacterium tuberculosis outbreak strain. *Nat Commun* **6**, 7119 (2015).
5. J. T. Simpson *et al.*, ABySS: a parallel assembler for short read sequence data. *Genome Res* **19**, 1117-1123 (2009).
6. G. G. Murray *et al.*, The effect of genetic structure on molecular dating and tests for temporal signal. *Methods Ecol Evol* **7**, 80-89 (2016).
7. A. Roetzer *et al.*, Whole genome sequencing versus traditional genotyping for investigation of a Mycobacterium tuberculosis outbreak: a longitudinal molecular epidemiological study. *PLoS Med* **10**, e1001387 (2013).
8. T. M. Walker *et al.*, Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. *The Lancet. Infectious diseases* **13**, 137-146 (2013).
9. J. Rozas *et al.*, DnaSP 6: DNA Sequence Polymorphism Analysis of Large Data Sets. *Mol Biol Evol* **34**, 3299-3302 (2017).
10. A. Leaver-Fay *et al.*, ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol* **487**, 545-574 (2011).
11. E. H. Baugh *et al.*, Robust classification of protein variation using structural modelling and large-scale data integration. *Nucleic Acids Res* **44**, 2501-2513 (2016).
12. E. H. Kellogg, A. Leaver-Fay, D. Baker, Role of conformational sampling in computing mutation-induced changes in protein structure and stability. *Proteins* **79**, 830-838 (2011).
13. A. Guilhot-Gaudeffroy, C. Froidevaux, J. Aze, J. Bernauer, Protein-RNA complexes and efficient automatic docking: expanding RosettaDock possibilities. *PLoS One* **9**, e108928 (2014).
14. S. Chaudhury *et al.*, Benchmarking and analysis of protein docking performance in Rosetta v3.2. *PLoS One* **6**, e22477 (2011).
15. W. Lin *et al.*, Structural Basis of Mycobacterium tuberculosis Transcription and Transcription Inhibition. *Mol Cell* **66**, 169-179 e168 (2017).
16. R. M. Sivley, X. Dou, J. Meiler, W. S. Bush, J. A. Capra, Comprehensive Analysis of Constraint on the Spatial Distribution of Missense Variants in Human Protein Structures. *Am J Hum Genet* **102**, 415-426 (2018).
17. F. Coll *et al.*, A robust SNP barcode for typing Mycobacterium tuberculosis complex strains. *Nat Commun* **5**, 4812 (2014).
18. S. Gagneux, Fitness cost of drug resistance in Mycobacterium tuberculosis. *Clin Microbiol Infect* **15 Suppl 1**, 66-68 (2009).
19. M. G. Reynolds, Compensatory evolution in rifampin-resistant Escherichia coli. *Genetics* **156**, 1471-1481 (2000).
20. V. Molodtsov, N. T. Scharf, M. A. Stefan, G. A. Garcia, K. S. Murakami, Structural basis for rifamycin resistance of bacterial RNA polymerase by the three most clinically important RpoB mutations found in Mycobacterium tuberculosis. *Mol Microbiol* **103**, 1034-1045 (2017).
21. K. Zeng, Y. X. Fu, S. Shi, C. I. Wu, Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics* **174**, 1431-1439 (2006).
22. J. D. Jensen, Y. Kim, V. B. DuMont, C. F. Aquadro, C. D. Bustamante, Distinguishing between selective sweeps and demography using DNA polymorphism data. *Genetics* **170**, 1401-1410 (2005).
23. H. Li, A new test for detecting recent positive selection that is free from the confounding impacts of demography. *Mol Biol Evol* **28**, 365-375 (2011).

24. E. Baudry, F. Depaulis, Effect of misoriented sites on neutrality tests with outgroup. *Genetics* **165**, 1619-1622 (2003).