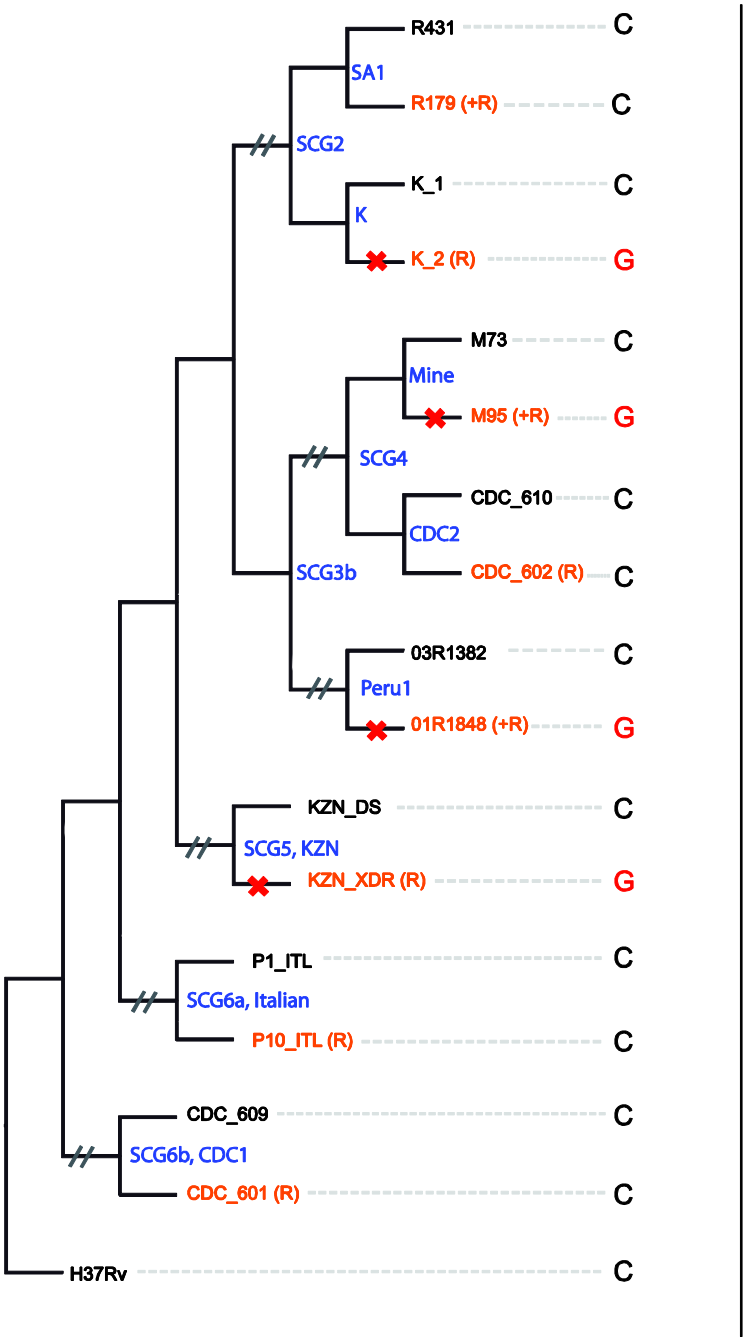
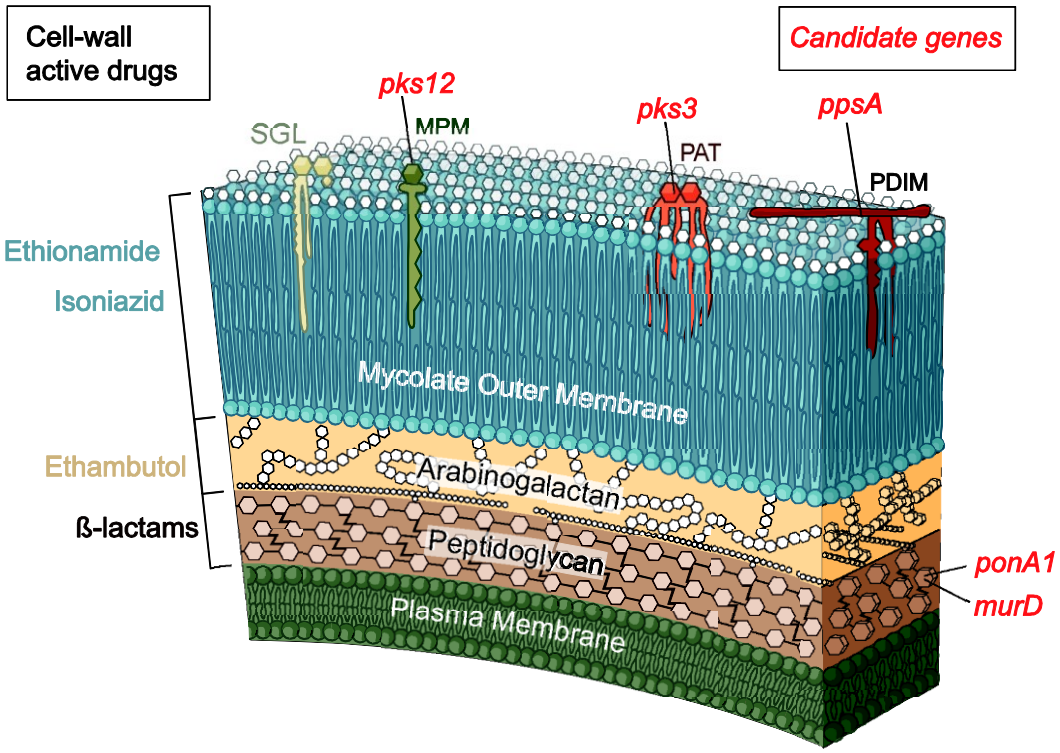
Supplementary Information for *Genomic Analysis Identifies Targets of Convergent Positive Selection in Mycobacterium tuberculosis*:

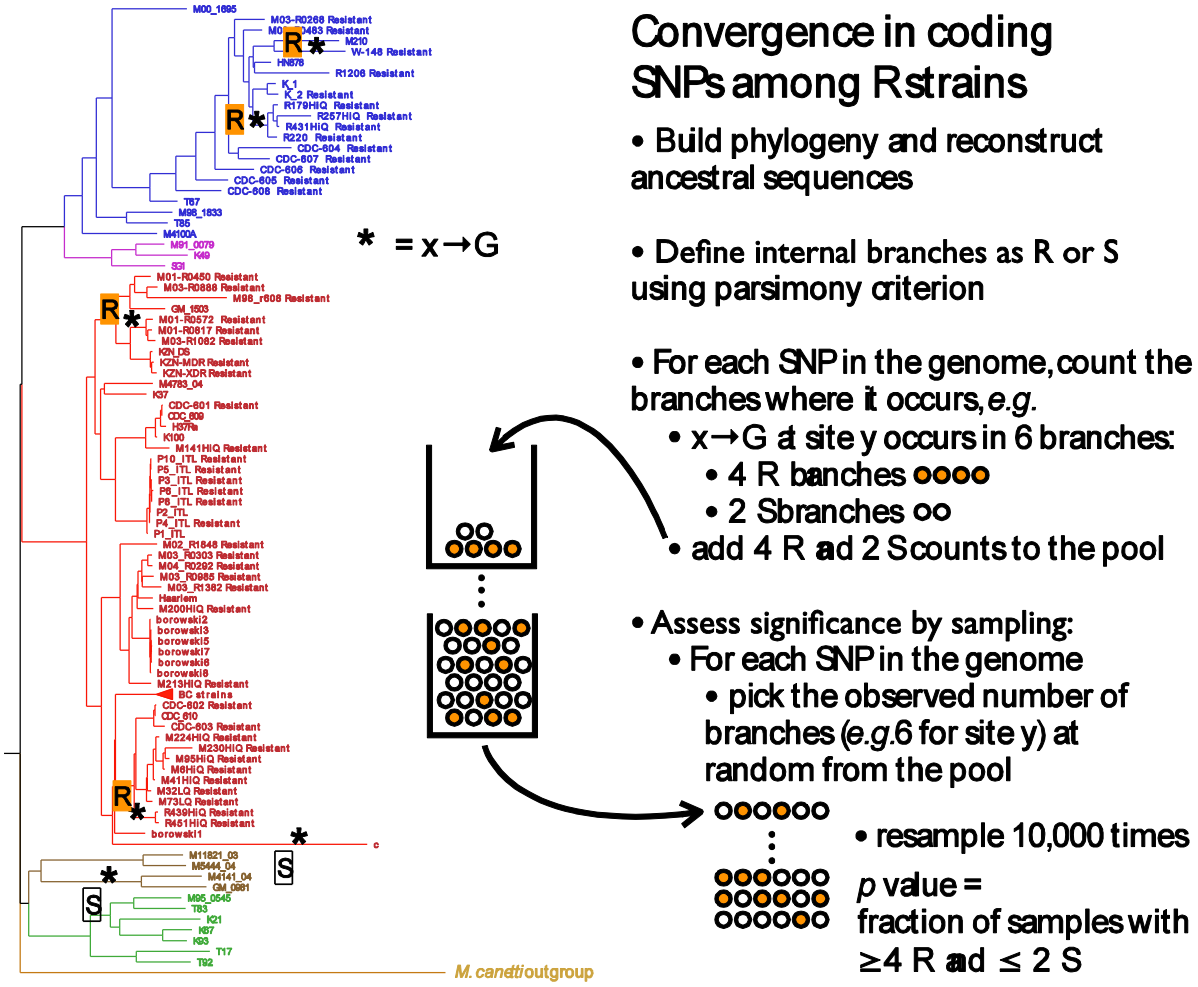
Supplementary Figure 1: Pairwise convergence for a single nucleotide (genomic coordinate 2155168) in gene *katG*. Branch lengths are not to scale. Nodes are labeled by epicluster name and substitution cluster group (SCG), isolates in orange represent the resistant (R) or more resistant (+R) (ie. To additional drugs) member for each epicluster pair. This figure demonstrates the repeated change from base C to G at position 944 of the *katG* with the acquisition of resistance.



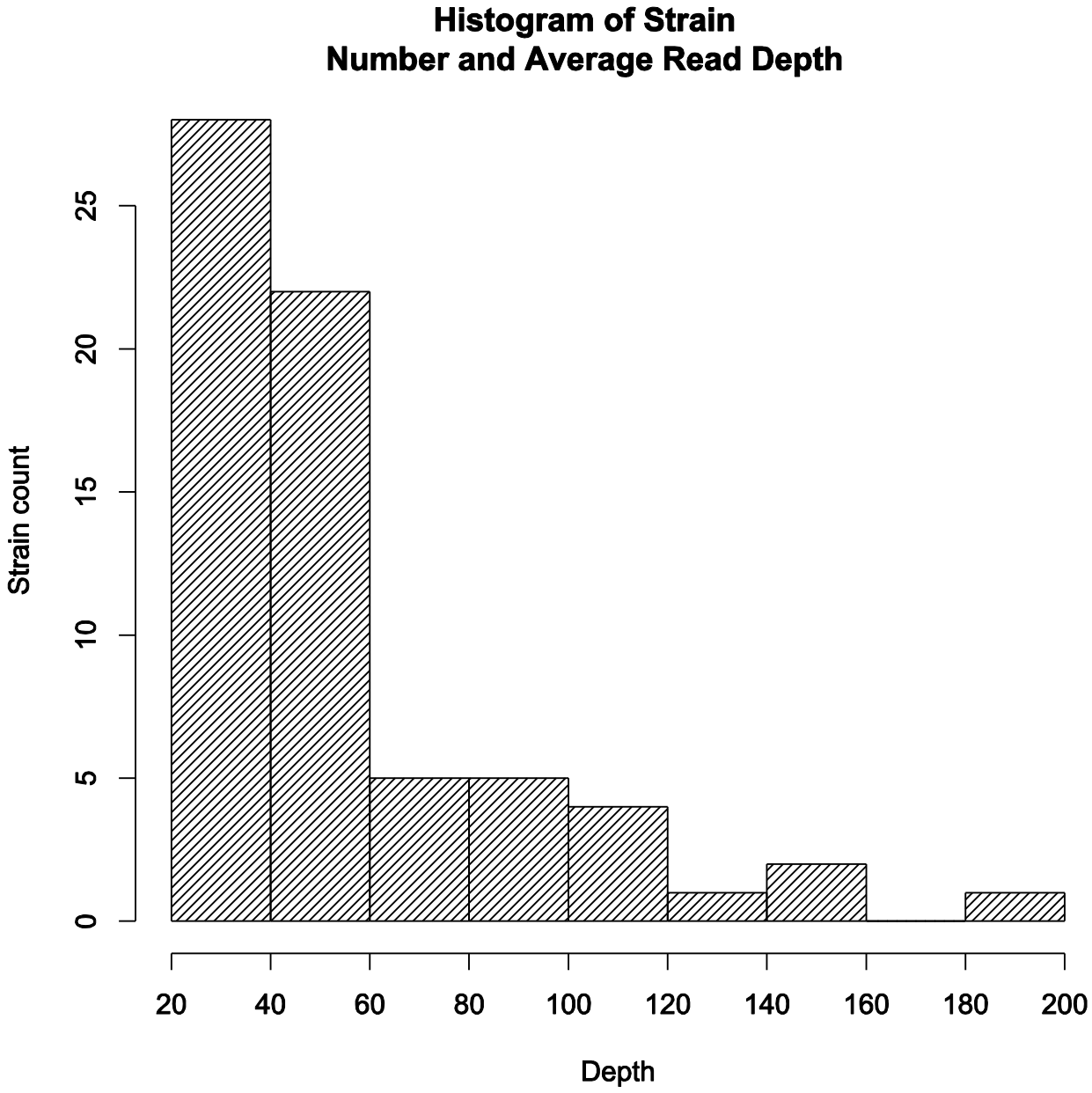
Supplementary Figure 2: MTB cell wall structure indicating drug targets and selected TIM gene functions. Genes are connected to their biosynthetic products with lines. PDIM: phthiocerol dimycocerosate MPM: Mannosyl-beta1-phosphomycoketides, SGL: sulfoglycolipid, PAT: polyacyltrehaloses.



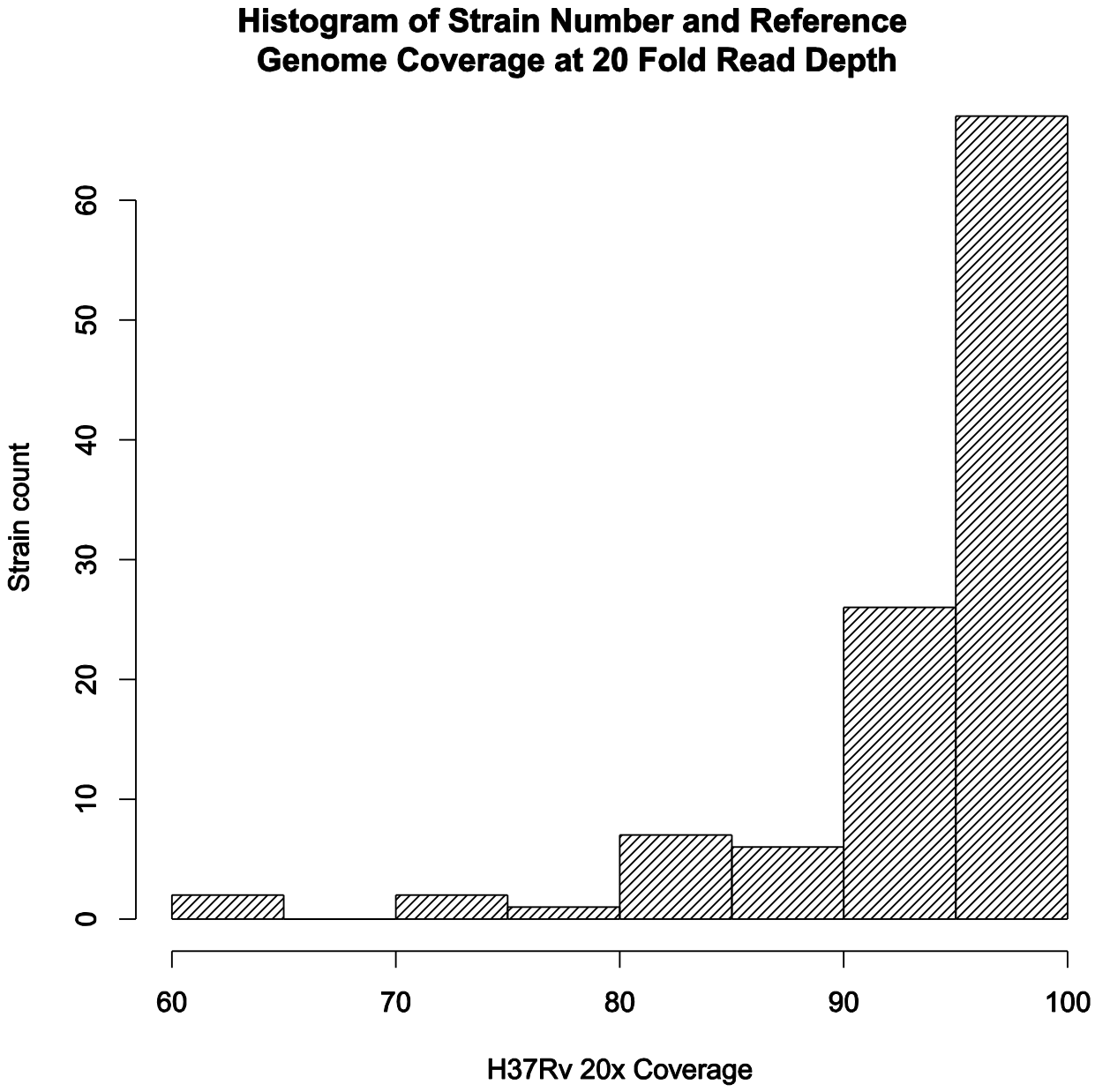
Supplementary Figure 3: Phylogenetic convergence (PhyC): This figure show an example *p*-value calculation for a nucleotide site *y* in the genome that undergoes a nonsynonymous mutation (x->G) in 4 resistance (R) branches and 2 sensitive (S) branches. The *p*-value for this site is obtained by resampling (10,000 times) 6 SNPs from the genomewide distribution of SNPs (depicted as an urn containing balls), including those occurring on R (orange balls) and S (white balls) branches. The *p*-value is equal to the fraction of resamplings (out of 10,000) for which ≥4 R and ≤2 S SNPs are picked. If $p < 0.05$, site *y* is considered to be a significant R-specific target of independent mutation (TIM). Please note that the tree topology here is not accurate and is simply used as an example.



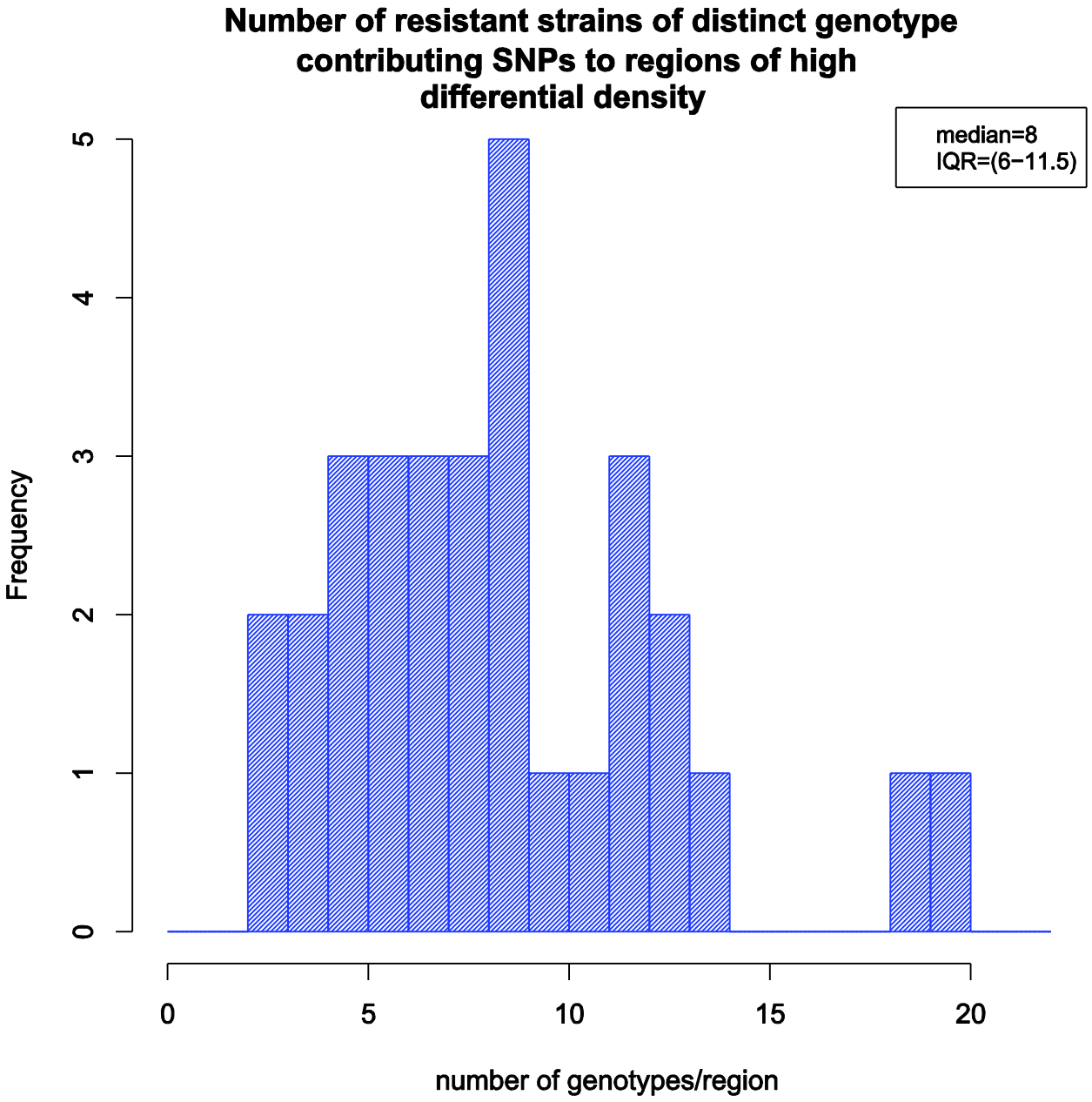
Supplementary Figure 4: Frequency histogram of average read depth per isolate



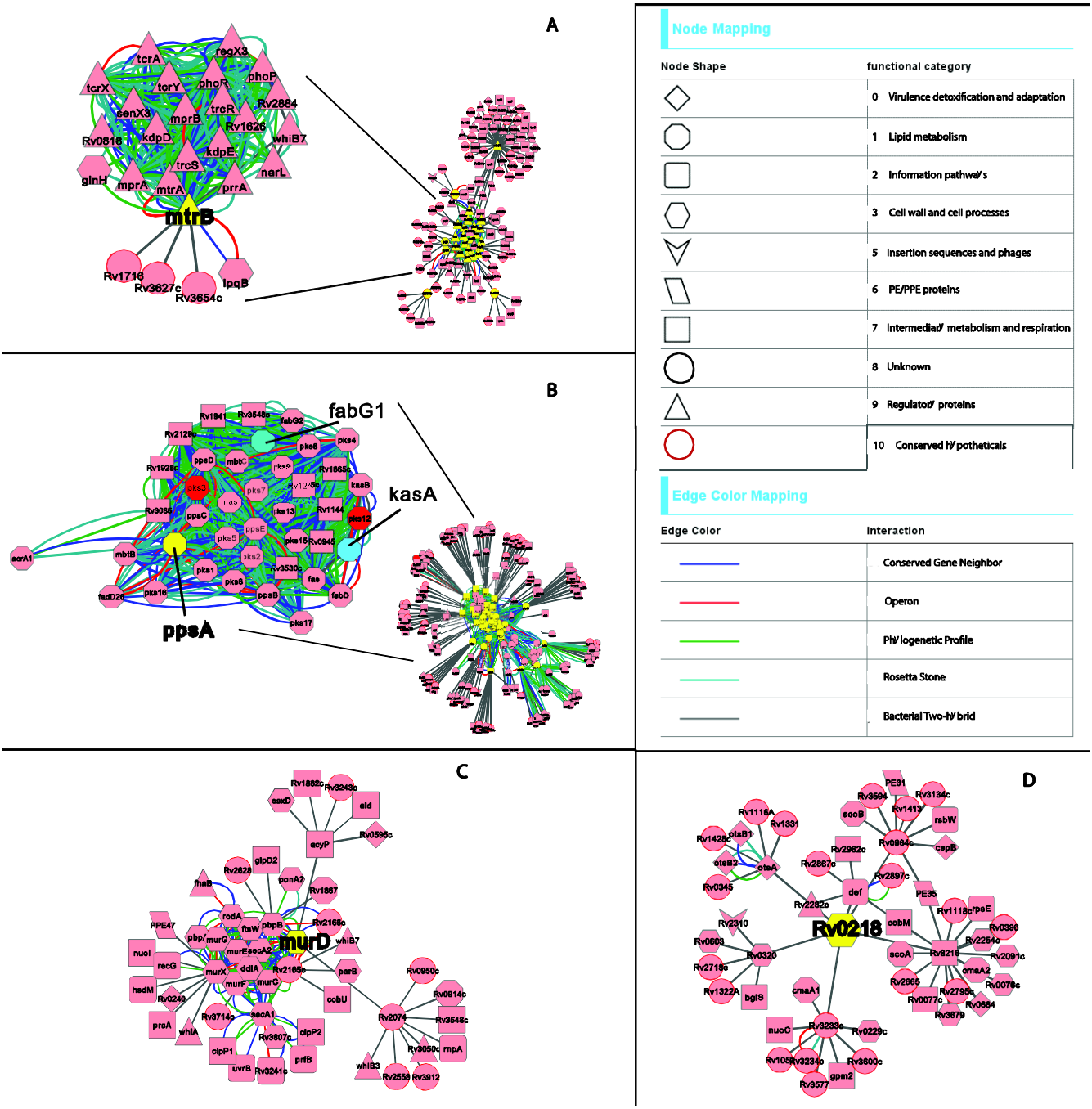
Supplementary Figure 5: Frequency histogram of percentage of H37Rv reference bases that are covered by 20 or more reads.



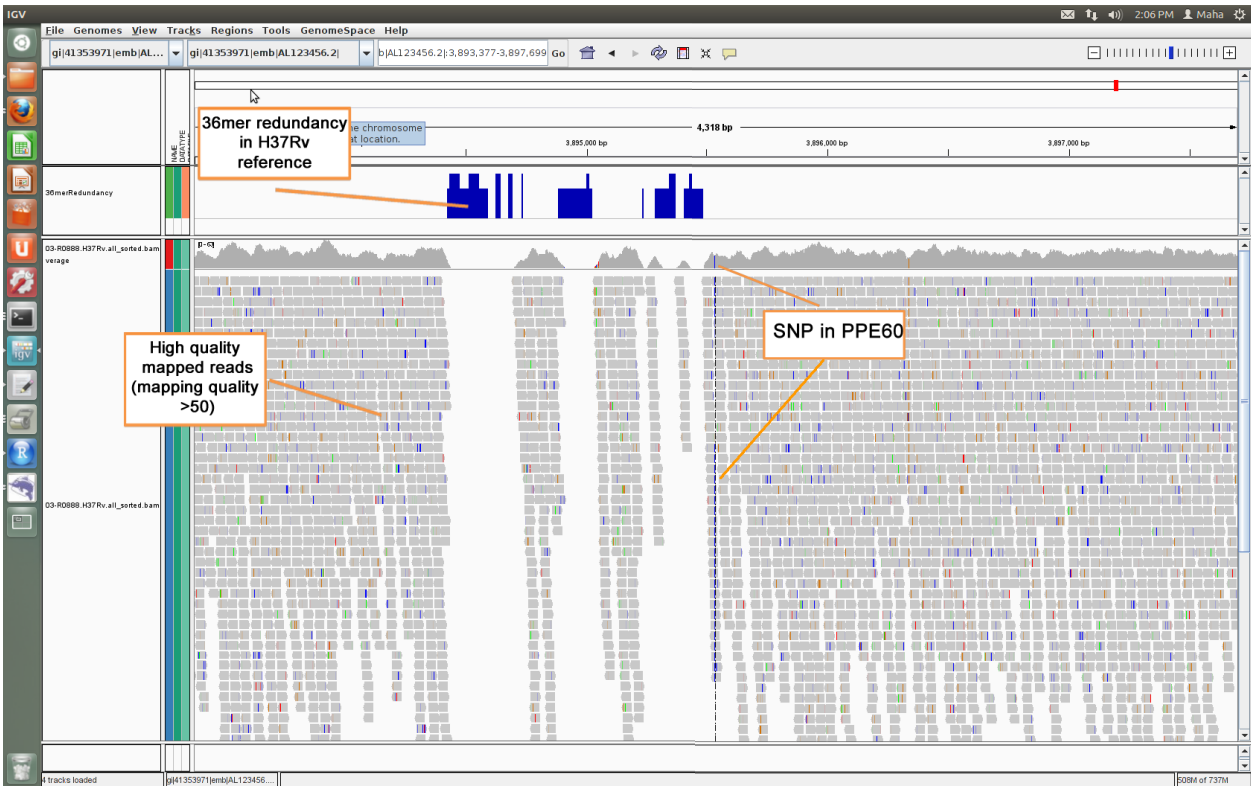
Supplementary Figure 6: Bar chart of the number of resistant isolates showing of distinct "genotype" showing changes in a genomic region with high spatial SNP clustering (Clustering index of ≥5). See supplementary notes for genotype definition.



Supplementary Figure 7: Example protein networks for candidate genes. Please refer to the in figure legend for node shape, and edge key. The gene/protein for which the network was constructed is colored in yellow and labeled in bold larger type font. Any first degree neighbors also found to be candidate resistance genes are colored in red. Any first degree neighbors previously known to be associated with drug resistance from the literature are colored in blue. Panels A and B contain two networks: the magnified network presents only the first degree neighbors of the genes, and the other network presents both the first (colored in yellow) and second degree neighbors. Not displayed in panel C but MurD is connected to PonA1 to the third degree through both of its second degree neighbors PonA2 and PbpA.



Supplementary Figure 8: Diagram demonstrating SNP calling in the PPE60 H37Rv sequence. Displayed are Illumina sequencing reads (represented by grey bars) mapping with high quality (MAQ mapping quality >50) from 03R0888 strain to the 3,893,377-3,897,699 genomic region surrounding and containing PPE60 (coordinates 3,894,425-3,895,607). Colored lines (any color but grey) within the bars represent bases mismatched to the reference (H37Rv) sequence. When the number of high quality reads containing a mismatched base at a particular site is >=20 the mapping algorithm may call a SNP at this position. The height of the blue bars in the '36mer Redundancy' track denotes the number of times a 36mer in this region has been observed elsewhere in the H37Rv genome and is used here as a measure of "repetitiveness". The diagram shows that the SNP present in 03R0888 in PPE60 does not occur in a repetitive region.



Supplementary Table 1: Percentage of resistance not explained by either WGS or ultra-deep sequencing with molecular inversion probes (1 insertion in *pncA* codon 152 detected by molecular inversion probes verification was also included).

Drug[ψ]	Isolates Resistant	Unexplained[‡]	Genes Considered (includes their promoter regions)[†]	Comments
INH	36	1(3%)	*katG*, inhA, fabG1(mabA)*	All isolate with *fabG1* mutations also had either *katG* or *inhA* mutation(s)
RIF	38	1(3%)	*rpoB*	
EMB	24	0(0%)	*embB, embC*	Two isolates only had *embC* mutation(s)
PZA	21	3(15%)	*pncA, rpsA*	All isolates with *rpsA* mutations also had *pncA* mutation(s)
SM	29	0(0%)	*rpsL, rrs, gidB*	
ETH	20	3(15%)	*ethA, inhA, fabG1(mabA)*	
OFLX	6	1(17%)	*gyrA**, gyrB*	
CIP	3	2(67%)	*gyrA**, gyrB*	
LEVO	1	0(0%)	*gyrA**, gyrB*	
AMK	4	0(0%)	*rrs, eis, rrl*	
CAP	6	1(17%)	*rrs, tlyA, eis*	
KAN	18	6(33%)	*rrs, eis, rrl*	
PAS	3	0(0%)	*thyA*	

[ψ]INH: isoniazid, RIF: rifampicin, EMB: ethambutol, STR: streptomycin, PZA: pyrazinamide, ETH: ethionamide, KAN: kanamycin, CAP: capreomycin, AMK: amikacin, OFLX: ofloxacin, LEVO: levofloxacin, CIP: ciprofloxacin, PAS: para-aminosalicylic acid, CYS: cycloserine. [‡] A few strains had unexplained resistance to more than one drug. Supplementary Table 8 details strains and drug type. [†]Excludes the common non-resistance associated polymorphisms ****katG* R463L *****gyrA* E21Q, S95T, G247S, G668D

Supplementary Table 2: Significant genes by the dN/dS method with known drug resistance genes listed first, and other genes second, both in increasing *p*-value.

Rvnumber	Symbol	Drug ‡	Description	*P*-value (<1.25E-5) †	Significant site(s)	Site class dN/dS ψ
Rv0667	*rpoB*	RIF	DNA-directed RNA polymerase beta chain	1.06E-27	435 D 1.000**, 445 H 0.982*, 450 L 1.000**, 452 L 0.983*, 491 I 0.980*, 731 L 0.983*	199
Rv3795	*embB*	EMB	membrane indolylacetylinositol arabinosyltransferase	4.28E-23	306 I 1.000**, 354 D 0.995**, 406 G 0.995**	999
Rv0006	*gyrA*	FLQ	DNA gyrase subunit A	6.93E-15	90 A 0.999**, 94 A 1.000**, 95 T 1.000**, 292 R 0.979*, 376 R 0.980*, 668 D 1.000**	999
Rv1908c	*katG*	INH	catalase-peroxidase-peroxynitritase T	5.98E-06	315 T 0.979*	724
Rv3919c	*gid*	STR	glucose-inhibited division protein B	1.25E-05	92 E 0.973*, 96 R 0.992**, 145 L 0.992**	141
Rv0746	PE_PGRS9		PE-PGRS family protein	1.04E-35	191 G 1.000**, 252 A 0.997**, 280 D 1.000**, 320 A 1.000**, 445 A 1.000**	236
Rv0279c	PE_PGRS4		PE-PGRS family protein	7.36E-28	77 V 1.000**, 325 N 0.995**, 352 G 1.000**, 372 I 0.995**	134
Rv0532	PE_PGRS6		PE-PGRS family protein	1.18E-25	227 D 0.980*, 239 G 1.000**	999
Rv2931	*ppsA*		phenolpthiocerol synthesis type-I polyketide synthase	2.85E-24	624 D 1.000**, 803 A 0.980*, 1194 L 0.990**	826
Rv0747	PE_PGRS10		PE-PGRS family protein	2.51E-23	295 K 1.000**, 300 S 1.000**	146
Rv1753c	PPE24		PPE family protein	5.97E-20	488 T 1.000**	999
Rv2853	PE_PGRS48		PE-PGRS family protein	3.32E-19	180 G 1.000**	999
Rv3512	PE_PGRS56		PE-PGRS family protein	3.67E-19	253 A 1.000**, 306 I 0.988*	999
Rv0050	*ponA1*		bifunctional penicillin-binding protein	6.95E-18	631 P 1.000**	999
Rv0218			conserved membrane protein	3.46E-17	316 R 1.000**	999
Rv2024c			conserved hypothetical protein	5.48E-17	47 R 0.997**, 154 G 1.000**	513
Rv1446c	*opcA*		oxpp cycle protein	1.33E-16	192 R 1.000**	999
Rv2079			conserved hypothetical protein	6.35E-15	47 C 1.000**, 95 L 0.977*	999
Rv2828c			conserved hypothetical protein	9.39E-15	128 S 1.000**	999
Rv0018c	*ppp*		serine/threonine phosphatase	9.94E-15	463 S 1.000**	999

Rv0058	*dnaB*	replicative DNA helicase	1.53E-14	552 R 1.000**	999
Rv0388c	PPE9	PPE family protein	2.39E-14	138 A 1.000**, 139 Q 1.000**	52
Rv2874	*dipZ*	cytochrome C biogenesis protein	1.10E-13	672 D 0.998**	999
Rv2668		exported alanine and valine rich protein	2.09E-13	3 R 1.000**	999
Rv0292		conserved membrane protein	2.43E-13	217 D 1.000**	999
Rv0280	PPE3	PPE family protein	2.43E-13	337 S 1.000**	999
Rv0658c		conserved membrane protein	3.47E-13	75 P 1.000**	999
Rv3077		hydrolase	5.10E-13	310 G 1.000**	197
Rv0236c		conserved membrane protein	7.54E-13	1080 G 1.000**	310
Rv2584c	*apt*	adenine phosphoribosyltransferase	8.07E-13	147 E 1.000**	999
Rv1394c	*cyp132*	cytochrome P450 132	9.51E-13	135 L 1.000**	999
Rv2917		conserved alanine and arginine rich protein	1.12E-12	594 L 1.000**	999
Rv0278c	PE_PGRS3	PE-PGRS family protein	1.68E-12	807 G 1.000**	324
Rv2896c		conserved hypothetical protein	2.02E-12	153 A 1.000**	999
Rv2450c	*rpfE*	resuscitation-promoting factor	2.33E-12	20 T 0.992**, 126 R 1.000**	999
Rv1812c		dehydrogenase	2.38E-12	30 P 0.998**	999
Rv2155c	*murD*	UDP-N-acetylmuramoylalanine-D-glutamate ligase	4.03E-12	247 G 0.999**	999
Rv3711c	*dnaQ*	DNA polymerase III epsilon subunit	5.24E-12	211 L 1.000**	999
Rv3366	*spoU*	tRNA/rRNA methylase	1.11E-11	92 R 1.000**	999
Rv2082		conserved hypothetical protein	1.34E-11	638 R 0.996**	209
Rv0159c	PE3	PE family protein	2.38E-11	14 A 1.000**	999
Rv0064		conserved membrane protein	2.93E-11	457 D 0.982*, 906 R 1.000**	378
Rv3468c		dTDP-glucose-4,6-dehydratase	3.58E-11	62 V 1.000**	999
Rv3479		transmembrane protein	6.12E-11	174 R 0.991**	475
Rv2825c		conserved hypothetical protein	6.33E-11	162 S 0.998**	999
Rv3490	*otsA*	alpha, alpha-trehalose-phosphate synthase	6.85E-11	77 E 0.958*	999
Rv2090		5-3 exonuclease	1.76E-10	358 F 0.999**	999
Rv0964c		hypothetical protein	4.18E-10	124 T 0.999**	999
Rv1093	*glyA1*	serine hydroxymethyltransferase 1	1.02E-09	36 A 0.990*	999

Rv2490c	PE_PGRS43	PE-PGRS family protein	1.40E-09	1399 G 0.996**	999
Rv1971	*mce3F*	MCE-family protein	1.54E-09	396 E 1.000**	999
Rv0082		oxidoreductase	1.89E-09	74 R 0.986*	999
Rv3776		conserved hypothetical protein	1.91E-09	112 S 0.983*, 329 V 0.996**	484
Rv2947c	*pks15*	polyketide synthase	2.69E-09	333 A 0.999**	999
Rv2439c	*proB*	glutamate 5-kinase protein	2.81E-09	226 S 0.993**	999
Rv2458	*mmuM*	homocysteine S-methyltransferase	3.38E-09	125 Y 0.998**	999
Rv3245c	*mtrB*	two component system sensor histidine kinase	3.39E-09	18 P 0.999**, 517 L 0.984*	875
Rv3151	*nuoG*	NADH dehydrogenase I chain G	4.66E-09	474 M 0.989*	999
Rv2488c		transcriptional regulator, luxR-family	6.02E-09	265 T 0.997**	999
Rv3449	*mycP4*	membrane-anchored mycosin	6.57E-09	87 T 0.995**	999
Rv0192		conserved hypothetical protein	6.73E-09	127 P 0.998**	999
Rv0095c		conserved hypothetical protein	8.31E-09	57 D 1.000**	999
Rv3835		conserved membrane protein	1.01E-08	294 L 0.996**	999
Rv1704c	*cycA*	D-serine/alanine/glycine transporter protein	1.01E-08	93 L 0.994**	999
Rv2794c		conserved hypothetical protein	1.08E-08	87 V 0.997**	999
Rv1320c		adenylate cyclase	1.26E-08	531 A 0.996**	999
Rv1463		ABC transporter ATP-binding protein	1.59E-08	198 E 0.992**	999
Rv1895		dehydrogenase	1.66E-08	270 L 0.995**	999
Rv2770c	PPE44	PPE family protein	1.68E-08	194 F 0.995**	999
Rv2059		conserved hypothetical protein	1.81E-08	317 T 0.990**	999
Rv2567		conserved alanine and leucine rich protein	1.85E-08	645 Q 0.992**	999
Rv1378c		conserved hypothetical protein	1.92E-08	37 W 0.994**	999
Rv0194		drugs-transport transmembrane ATP-binding protein ABC transporter	3.06E-08	74 T 0.994**	999
Rv2333c		conserved membrane transport protein	4.79E-08	69 Y 0.994**	999
Rv0338c		iron-sulfur-binding reductase	5.71E-08	506 G 0.992**, 621 V 0.992**	999
Rv0881		rRNA methyltransferase	6.06E-08	115 R 0.999**	634
Rv0417	*thiG*	thiamin biosynthesis protein	6.77E-08	75 C 0.998**	343

Rv0109	PE_PGRS1	PE-PGRS family protein	8.68E-08	346 G 0.995**	999
Rv3425	PPE57	PPE family protein	1.07E-07	63 L 0.972*, 128 T 0.999**	999
Rv3093c		oxidoreductase	1.31E-07	210 C 0.995**	999
Rv1319c		adenylate cyclase	1.46E-07	439 D 0.995**	235
Rv2101	*helZ*	helicase	1.63E-07	462 L 0.986*	330
Rv2017		transcriptional regulator	1.92E-07	262 E 0.993**	999
Rv2769c	PE27	PE family protein	2.19E-07	136 A 0.978*, 270 M 0.999**	999
Rv1186c		conserved hypothetical protein	2.30E-07	207 A 0.988*	999
Rv2741	PE_PGRS47	PE-PGRS family protein	2.79E-07	271 S 0.999**	80
Rv2495c	*pdhC*	dihydrolipoamide S-acetyltransferase E2 component	3.30E-07	107 A 0.997**	696
Rv3777		oxidoreductase	3.50E-07	160 A 0.999**	999
Rv0048c		membrane protein	3.67E-07	248 E 0.992**, 250 V 0.992**	999
Rv3341	*metA*	homoserine O-acetyltransferase	4.12E-07	87 S 0.950	999
Rv3511	PE_PGRS55	PE-PGRS family protein	4.14E-07	396 N 0.987*, 589 G 0.991**	503
Rv1716		conserved hypothetical protein	4.19E-07	178 G 0.977*, 276 A 0.975*	999
Rv0226c		conserved membrane protein	4.34E-07	379 P 0.962*	999
Rv0995	*rimJ*	ribosomal-protein-alanine acetyltransferase	4.34E-07	23 G 0.997**, 105 Y 0.978*	999
Rv0465c		transcriptional regulator	4.49E-07	106 C 0.994**	999
Rv1644	*tsnR*	23S rRNA methyltransferase	5.38E-07	232 P 0.994**	999
Rv3764c		two component system sensor kinase	6.42164E-07	246 R 0.961*	999
Rv1915	*aceAa*	isocitrate lyase	6.78E-07	179 D 0.995**	999
Rv2807		conserved hypothetical protein	8.84E-07	72 E 0.998**, 77 D 0.988*	999
Rv2236c	*cobD*	cobalamin biosynthesis transmembrane protein	9.86E-07	79 C 0.991**	744
Rv0787		hypothetical protein	1.05E-06	267 H 0.991**	802
Rv1618	*tesB1*	acyl-CoA thioesterase II	1.07E-06	121 L 0.995**	999
Rv0425c	*ctpH*	metal cation transporting P-type ATPase	1.13E-06	689 V 0.982*	398
Rv1326c	*glgB*	1,4-alpha-glucan branching enzyme	1.22E-06	470 S 0.908	999
Rv0572c		hypothetical protein	1.28E-06	31 L 0.989*	999
Rv1640c	*lysS*	lysyl-tRNA synthetase 2 lysX	1.36E-06	701 I 0.942	999
Rv3063	*cstA*	carbon starvation protein A	1.45E-06	559 S 0.929	999

Rv3782		L-rhamnosyltransferase	1.66E-06	274 V 0.992**	999
Rv3021c	PPE47	PPE family protein	2.26E-06	222 A 0.952*	64
Rv2433c		hypothetical protein	2.29E-06	26 L 0.998**	999
Rv2316	*uspA*	sugar-transport membrane protein ABC transporter	2.38E-06	67 D 0.996**, 127 L 0.974*	999
Rv1232c		conserved hypothetical protein	2.53E-06	149 G 0.981*	483
Rv0103c	*ctpB*	cation-transporter P-type ATPase B	3.02E-06	22 S 0.981*	999
Rv2611c		acyltransferase	3.22E-06	197 C 0.979*	999
Rv0676c	*mmpL5*	transmembrane transport protein	3.27E-06	948 V 0.934	999
Rv1160	*mutT2*	mutator protein mutT	3.46E-06	58 G 0.969*	999
Rv3879c		hypothetical alanine and proline rich protein	3.68E-06	729 S 0.980*	269
Rv1321		conserved hypothetical protein	3.69E-06	144 R 0.997**	999
Rv3497c	*mce4C*	MCE-family protein	4.09E-06	191 R 0.928	999
Rv1486c		conserved hypothetical protein	4.17E-06	198 N 0.972*	999
Rv3365c		conserved hypothetical protein	4.20E-06	38 P 0.989*, 687 S 0.953*	999
Rv1127c	*ppdK*	pyruvate, phosphate dikinase	4.92E-06	69 E 0.899	999
Rv1400c	*lipI*	lipase lipH	5.04E-06	106 T 0.989*	999
Rv3892c	PPE69	PPE family protein	5.65E-06	19 K 0.962*	999
Rv3655c		conserved hypothetical protein	6.23E-06	100 S 0.979*	119
Rv0557	*pimB*	mannosyltransferase	6.32E-06	none	1
Rv3144c	PPE52	PPE family protein	6.42E-06	226 S 0.969*	999
Rv0758	*phoR*	two component system sensor kinase	7.11E-06	172 L 0.980*	999
Rv1538c	*ansA*	L-aparaginase	7.79E-06	281 G 0.985*	999
Rv2290	*lppO*	lipoprotein	8.61E-06	16 A 0.986*	999
Rv0855	*far*	fatty-acid-CoA racemase	8.78E-06	24 A 0.993**	342
Rv3737		conserved membrane protein	9.15E-06	40 G 0.853	999
Rv2037c		conserved membrane protein	9.16E-06	312 Y 0.914	999
Rv3282	*maf*	conserved hypothetical protein	9.30E-06	80 A 0.988*	999
Rv1300	*hemK*	hypothetical protein	9.49E-06	194 C 0.988*	999
Rv3317	*sdhD*	succinate dehydrogenase hydrophobic membrane anchor subunit	1.00E-05	112 T 0.998**, 114 R 0.990**	999
Rv3347c	PPE55	PPE family protein	1.07E-05	786 V 0.781	1
Rv3591c		hydrolase	1.07E-05	156 F 0.986*	999

Rv0668	*rpoC*	DNA-directed RNA polymerase beta chain	1.13E-05	404 D 0.975*, 484 W 0.975*, 698 N 0.998**, 1040 P 0.976*, 1092 E 0.988*, 1231 R 0.976*	999
Rv3329		aminotransferase	1.16E-05	122 H 0.969*	999
Rv1900c	*lipJ*	lignin peroxidase	1.19E-05	204 M 0.946	999
Rv0538		conserved membrane protein	1.21E-05	228 P 0.960*	999
Rv2482c	*plsB2*	glycerol-3-phosphate acyltransferase	1.22E-05	778 R 0.956*	999

‡ Mutations in the specified gene are previously known to be associated with resistance to the drug listed. Abbreviations detailed in Supplementary Table 1 footnote. †p-value is 0.05 corrected for the multiple testing (3998 tests) to a threshold of 1.25062E-5. Significant sites are determined by PAML Bayesian empiric Bayes method with thresholds of * <0.05, and **<0.01. ψ Site class dN/dS is the dN/dS deduced by the alternative model likelihood model allowing for selection in 2/4 different site categories (*40*).

Supplementary Table 3: Pairwise convergence analysis: SNPs found in two or more resistant isolates relative to the more sensitive ancestor.

SNP	Region	Symbol	Drug*	# pairs	Isolates	p-value
2155168 CG	Rv1908c	*katG*	INH	4	32, 16, 22, 141	2x10-17
781687 AG	Rv0682	*rpsL*	STR	3	32, 141, 13	1 x10-12
761155 CT	Rv0667	*rpoB*	RIF	2	32, 13	4 x 10-8
761161 TC	Rv0667	*rpoB*	RIF	2	16,141	4 x 10-8
4247429 AG	Rv3795	*embB*	EMB	2	16, 37	4 x 10-8
55553 CT	Rv0050	*ponA1*	.	2	37, 13	4 x 10-8
2163375 TC	Rv1917c	PPE34	.	2	37, 13	4 x 10-8
1468208 AC	Rv1313c	.	.	2	37, 13	4 x 10-8
4254290 TG	Rv3798	.	.	2	37, 13	4 x 10-8
921813 CG	Rv0829	.	.	2	37, 13	4 x 10-8
3131473 AC	Rv2823c	.	.	2	16, 37	4 x 10-8
2715346 GA	*eis* promoter	.	.	2	41, 22	4 x 10-8

***Drug abbreviations explained in Supplementary Table 1.**

Supplementary Table 4: Pairwise convergence analysis: Genes with two or more new R specific SNPs

Gene	symbol	Drug*	convergent R isolates #	isolate ids	SNPs	p-value
Rv0667	*rpoB*	RIF	6	32, 13, 16, 37, 22, 141	761161TC (L452P), 761155CT (S450L), 761110AG (D435G), 763123TC (I1106T), 761253CT (P483S)	6 x10-9
Rv3795	*embB*	EMB	5	32, 16, 37, 141, 13	4247431GA (M306I), 4247429AG (M306V), 4248002CA (Q497K), 428003AG (Q497R)	5 x10-7
Rv1908c	*katG*	INH	5	32, 16, 22, 141, 13	2155168CG (S315T), 2155412CT (G234R)	5 x10-7
Rv2043c	*pncA*	PZA	4	32, 22, 141, 13	2289072TC (H57R), 2288730GA (A171V), 2288848CT (G132S), 2288885CT (W119*)	3 x10-5
Rv0682	*rpsL*	STR	3	32, 141, 13	781687AG (K43R)	0.002
Rv3347c	PPE55	-	3	37, 42, 141	3750828AG (V786A), 3752821TC (M122V), 3750421TC (S922G)	0.002
Rv0746	PE_PGRS9	-	3	37, 41, 22	838858AG (T320A), 837033AG (T445A), 836454AG (T252A)	0.002
Rvnr01	*rrs*	SM/AG	2	16, 22	1473246AG, 1472751AG	0.064
Rv0006	*gyrA*	FLQ	2	32, 16	7582AC(D94A), 7570CT(A90V)	0.064
Rv3854c	*ethA*	ETH	2	32, 37	4326630AC(F282V), 4327065AG(C137R)	0.064
eis promoter	-	KAN	2	41, 22	2715346GA	0.064
Rv2813-Rv2814c intergenic	-	-	2	32, 37	3119188TC, 3119957TG	0.064
Rv2048c	*pks12*	-	2	37, 141	2304238AG(S917P), 2295685CA (V3768L)	0.064
Rv0279c	PE_PGRS4	-	2	41, 22	338844AG+338845CT (V77T), 338020AC (C352G)	0.064
Rv3478	PPE60	-	2	37, 42	3894732AG(R103G), 3894784CT(T120M)	0.064
Rv2931	*ppsA*	-	2	42, 13	3247851GA(A803T), 3249411GA(G1323S)	0.064
Rv1917c	PPE34	-	2	37, 13	2163375TC (N1313D)	0.064
Rv0050	*ponA1*	-	2	37, 13	55553CT (P631S)	0.064
Rv0747	PE_PGRS10	-	2	41, 42	839123AG(R225G), 839334AG(K295R), 839348AG(S300G)	0.064

Rv2611c	-	-	2	22, 141	2939373GC(S197C), 2939374AG(C197G), 2939657TC(I102M)	0.064
Rv1313c	-	-	2	37, 13	1468208AC (L433R)	0.064
Rv0064	-	-	2	37, 22	69989GA(G457D), 71336GC(R906P)	0.064
Rv3798	-	-	2	37, 13	4254290TG (L433R)	0.064
Rv3365c	-	-	2	37, 141	3775409TC(Q698R), 3777389GA(P38L)	0.064
Rv0492c	-	-	2	32, 37	583171AC(S70A), 581890CG(V497L)	0.064
Rv3806c	-	-	2	32, 16	4269271AG(V188A), 4269671CG(V55L)	0.064
Rv2823c	-	-	2	16, 37	3131473AC (Y101D)	0.064
Rv0829	-	-	2	37, 13	921813CG (A80G)	0.064

***Drug abbreviations explained in Supplementary Table 1.**

Supplementary Table 5: Significant genomic regions/TIMs by phyC (phylogenetic convergence). Known resistance regions listed first. Regions ordered by increasing p-value.

Gene/ Region	Drug**	Description	Convergence p-value (<0.05)†	Resistant branches***	Sensitive branches ***	Convergent Site(s)
katG	INH	catalase-peroxidase-peroxynitritase	<0.0001	13	5	
pncA	PZA	pyrazinamidase/nicotineamidedase	<0.0001	13	0	
embB	EMB	membrane indolylacetylinositol arabinosyltransferase	0.003, <0.0001*	20	7	916G
ethA	ETH	Monoxygenase	0.0001	10	4	
gyrA	FLQ	DNA gyrase subunit A	0.0003	11	12	
inhA promoter	INH	NADH-dependent enoyl-[acyl-carrier-protein] reductase promoter	0.011	6	1	
eis promoter	KAN	Enhanced intracellular survival protein promoter	0.0021	6	0	
rpoB	RIF	DNA-direction RNA polymerase beta chain	<0.0001, 0.0001, <0.0001*	26	3	1304T, 1349T
gid	SM	glucose-inhibited division protein B	0.0002	12	7	
rpsL	SM	30S ribosomal protein S12	0.0016, 0.0008*	11	0	128G
rrs	SM/AG	16S ribosomal RNA	0.0044, <0.0001*	12	3	1401G
Rv0218		conserved membrane protein	<0.0001, 0.0001*	11	5	946T
PE_PGRS4		PE-PGRS family protein	<0.0001	20	23	
PE_PGRS6		PE-PGRS family protein	<0.0001	18	16	
PE_PGRS9		PE-PGRS family protein	<0.0001	28	23	
PPE9		PPE family protein	<0.0001	22	21	
ppsA		phenolpthiocerol synthesis type-I polyketide synthase	<0.0001	23	21	
Rv0064		conserved membrane protein	<0.0001	19	18	
Rv2082		conserved hypothetical protein	<0.0001	20	18	
PE_PGRS10		PE-PGRS family protein	<0.0001	19	17	
PPE55		PPE family protein	<0.0001	16	24	
PE_PGRS3		PE-PGRS family protein	0.0003	15	18	
ponA1		bifunctional penicillin-binding protein	0.0005	12	10	
Rv3680-*whib4* intergenic region		Intergenic area between anion transporter ATPase (Rv3680) and transcriptional regulator whib-like	0.0005	12	9	
pks12		hypothetical protein	0.0006	13	17	
PE_PGRS1		PE-PGRS family protein	<0.0001, 0.0006*	11	13	1036C
Rv3093c		Oxidoreductase	0.0009	9	5	
rpoC		DNA-directed RNA polymerase beta chain	0.0011	10	6	
pks3		polyketide beta-ketoacyl synthase	0.0011	7	10	
opcA		oxpp cycle protein	0.0015	10	8	

rbsK	Ribokinase	0.0018	10	9	
PPE54	PPE family protein	0.0022	16	28	
Rv1319c	Possible adenylate cyclase	0.0032	9	7	
Rv3446c	hypothetical alanine and valine rich protein	0.0039	9	5	
PPE3	PPE family protein	0.0039, 0.0028*	13	12	1009T
Rv2896c	conserved hypothetical protein	<0.0001, 0.0053*	9	8	457T
Rv2024c	conserved hypothetical protein	0.0066	12	13	
PE_PGRS48	PE-PGRS family protein	0.0068, 0.0006*	14	6	538C
mtrB	two component system sensor histidine kinase	0.0069	8	4	
Rv0658c	conserved membrane protein	0.0013, 0.0072*	10	11	224T
murD	UDP-N-acetylmuramoylalanine-D-glutamate ligase	0.0075, 0.0001*	15	8	739C
Rv1147-Rv1148c intergenic region	intergenic area or possible promoter between two conserved hypothetical protein	0.0075, 0.0092*	8	7	408C
PE_PGRS50	PE-PGRS family protein	0.0077	9	4	
PPE47	PPE family protein	0.0101	10	15	
dnaQ	DNA polymerase III epsilon subunit	0.0126	7	9	
PE_PGRS47	PE-PGRS family protein	0.0143	8	8	
coaE-Rv1632c intergenic region	intergenic area, ? Promoter of hypothetical protein, preceeded by dephospho-CoA kinase	0.0152	8	5	
PPE60	PPE family protein	0.0184	9	5	
PE_PGRS53	PE-PGRS family protein	0.0224	8	13	
murA-rrs intergenic region	Probable rrs, rrl, rrf promoter	0.0328, 0.0112*	6	1	83C

† Average p-value over all replicates of trees (ml, Bayesian, parsimony, and ml and parsimony reconstructions, site had to be significant by all replicate trees/reconstructions to be considered significant) * Significant by both nucleotide site and gene analysis, p-values for all significant sites/ and gene analysis listed, in order site1, (site2), gene. ** Mutations in the specified gene are previously known to be associated with resistance to the drug listed. Abbreviations detailed in Supplementary Table 1 footnote. AG aminoglycosides (kanamycin, capreomycin, amikacin) * these counts represent the number of parsimony phylogenetic branches with one or more changes anywhere along the gene or intergenic region, all changes were treated equal. The tree contained a total of 87 resistant branches and 158 sensitive branches.**

Supplementary Table 6: Significant genomic regions by the differential density method

Gene/Region	Drug*	Description	p-value** (<0.05)/(CI>5)
pncA	PZA	pyrazinamidase/nicotineamidedase	<0.0001
embB	EMB	membrane indolylacetylinositol arabinosyltransferase	<0.0001
rrs	SM/AG	16S ribosomal RNA	<0.0001
rpoB	RIF	DNA-direction RNA polymerase beta chain	<0.0001
ethA	ETH	monoxygenase	0.036
rpsL	SM	30S ribosomal protein S12	0.036
katG	INH	catalase-peroxidase-peroxynitritase	0.0441 (CI=8)
gyrA	FLQ	DNA gyrase subunit A	0.0675 (CI=5)
eis promoter	KAN	Enhanced intracellular survival protein promoter	0.2567 (CI=6)
gid	SM	glucose-inhibited division protein B	0.4901 (CI=5)
PE_PGRS9		PE-PGRS family protein	0.0026
ppsA		phenolpthiocerol synthesis type-I polyketide synthase	0.01
Rv0218		conserved membrane protein	0.1229 (CI=7)
rpoC		DNA-directed RNA polymerase beta chain	0.4901 (CI=5)
Rv2024c		conserved hypothetical protein	0.1085 (CI=6)
PE_PGRS48		PE-PGRS family protein	0.0133
murD		UDP-N-acetylmuramoylalanine-D-glutamate ligase	0.2143 (CI=6)
PE_PGRS16		PE-PGRS family protein	0.0338
Rv0236c		Conserved membrane protein	0.051 (CI=7)
PE_PGRS56		PE-PGRS family protein	0.051 (CI=7)
PE_PGRS24		PE-PGRS family protein	0.3041 (CI=6)
pks1		Polyketide synthase	0.4901 (CI=5)
Rv0192		Conserved hypothetical protein	0.7048 (CI=5)
rimJ		Ribosomal-protein-alanine acetyltransferase	0.7048 (CI=5)
Rv2828c		Conserved hypothetical protein	0.7436 (CI=5)
Rv3468c		dTDP-glucose-4,6-dehydratase	0.7436 (CI=5)
Rv0749A-Rv0750 intergenic region		Flanking genes are conserved hypothetical proteins	0.7378 (CI=5)
fadD36		Fatty-acid-CoA ligase	0.7378 (CI=5)
PE_PGRS26		PE-PGRS family protein	0.7378 (CI=5)

*** Mutations in the specified gene are previously known to be associated with resistance to the drug listed. Abbreviations detailed in Supplementary Table 1 footnote. AG aminoglycosides (kanamycin, capreomycin, amikacin) **Regions and p-values highlighted in orange or yellow are also detected by phyC. Regions are displayed if they have an empirical p-value of <0.05 (orange highlight) or a clustering index (CI, see text) ≥5 (yellow highlight).**

Supplementary Table 7: Genes significant by two or more of the three following methods: phyC, dN/dS and differential density.

Rvnumber	Symbol	Description	Method
Known genes			
Rv0667	*rpoB*	DNA-direction RNA polymerase beta chain	All
Rv1908c	*katG*	catalase-peroxidase-peroxynitritase	All
Rv0006	*gyrA*	DNA gyrase subunit A	All
Rv3919	*gid*	glucose-inhibited division protein B	All
Rv3795	*embB*	membrane indolylacetylinositol arabinosyltransferase	All
Rv3854c	*ethA*	monoxygenase	phyC + differential density
Rv0682	*rpsL*	30S ribosomal protein S12	phyC + differential density
Rv2043c	*pncA*	pyrazinamidase/nicotineamidedase	phyC + differential density
Rvnr01	*rrs*	16S ribosomal RNA	phyC + differential density
Rv2416c	*eis*	Enhanced intracellular survival protein	phyC + differential density (promoter)
Rv1484	*inhA*	NADH-dependent enoyl-[acyl-carrier-protein] reductase	phyC + differential density (promoter)
Other genes			
Rv0218		conserved membrane protein	All
Rv0668	*rpoC*	DNA-directed RNA polymerase beta chain	All
Rv0746	PE_PGRS9	PE-PGRS family protein	All
Rv2155c	*murD*	UDP-N-acetylmuramoylalanine-D-glutamate ligase	All
Rv2931	*ppsA*	phenolpthiocerol synthesis type-I polyketide synthase	All
Rv2024c		conserved hypothetical protein	All
Rv2853	PE_PGRS48	PE-PGRS family protein	All
Rv0236c		conserved membrane protein	dN/dS + differential density
Rv3512	PE_PGRS56	PE-PGRS family protein	dN/dS + differential density
Rv0192		conserved hypothetical protein	dN/dS + differential density
Rv0995	*rimJ*	ribosomal-protein-alanine acetyltransferase	dN/dS + differential density
Rv2828c		conserved hypothetical protein	dN/dS + differential density
Rv3468c		dTDP-glucose-4,6-dehydratase	dN/dS + differential density
Rv0050	*ponA1*	bifunctional penicillin-binding protein	dN/dS + phyC

Rv0064		conserved membrane protein	dN/dS + phyC
Rv0109	PE_PGRS1	PE-PGRS family protein	dN/dS + phyC
Rv0278c	PE_PGRS3	PE-PGRS family protein	dN/dS + phyC
Rv0279c	PE_PGRS4	PE-PGRS family protein	dN/dS + phyC
Rv0280	PPE3	PPE family protein	dN/dS + phyC
Rv0388c	PPE9	PPE family protein	dN/dS + phyC
Rv0532	PE_PGRS6	PE-PGRS family protein	dN/dS + phyC
Rv0658c		conserved membrane protein	dN/dS + phyC
Rv0747	PE_PGRS10	PE-PGRS family protein	dN/dS + phyC
Rv1319c		adenylate cyclase	dN/dS + phyC
Rv1446c	*opcA*	oxpp cycle protein	dN/dS + phyC
Rv2082		conserved hypothetical protein	dN/dS + phyC
Rv2741	PE_PGRS47	PE-PGRS family protein	dN/dS + phyC
Rv2896c		conserved hypothetical protein	dN/dS + phyC
Rv3021c	PPE47	PPE family protein	dN/dS + phyC
Rv3093c		Oxidoreductase	dN/dS + phyC
Rv3245c	*mtrB*	two component system sensor histidine kinase	dN/dS + phyC
Rv3347c	PPE55	PPE family protein	dN/dS + phyC
Rv3711c	*dnaQ*	DNA polymerase III epsilon subunit	dN/dS + phyC

Supplementary Table 8: Amino acid (AA) substitutions in TIMs for strains with unexplained resistance. Substitutions that also occurred in isolates sensitive to the respective drug were excluded. Substitutions are named relative to the reference H37Rv amino-acid sequence.

Drug*	Isolate	Epicluster	Nonsynonymous or noncoding SNPs
INH	M213	-	*pks12* Q3283R
RIF	M213	-	*pks12* Q3283R
PZA	R439	SA2	*rpoC* E750D, *PPE54* A2181V & F2182L, *PPE55* Q1766R, Rv1319c G457R
	R451	SA2	*rpoC* E750D, PPE54 A2181V & F2182L, PPE55 Q1766R, PPE60 G146A, Rv2082 A185P
	CDC607	-	PE_PGRS9 V118A
ETH	M213	-	*pks12* Q3283R, *rpoC* P1040R
	M200	-	-
	02R1848	Peru1	-
OFLX	M230	-	-
CIP	03R1082	Russia2	-
	04R0292	Peru1	-
KAN	M213	-	*pks12* Q3283R
	CDC602	CDC2	*ppsA* H955P, PPE60 T120M
KAN CAP	03R0303	Peru1	*mtrB* G174A
	M41	Mine	-
	03R0888	Russia1	PPE60 S371R
	03R1082	Russia2	-
	03R1082	Russia2	-

* Drug abbreviations expanded in Supplementary Table 1 footnote.

Supplementary Table 9: Synonymous mutations significant by a secondary application of PhyC. Permutation based significance levels were determined using the dataset of all mutations synonymous, nonsynonymous and non-coding.

Gene/ Region	Description	Convergence p-value (<0.05)**	Resistant branches***	Sensitive branches***	Convergent Site(s)
pks12	Polyketide synthase 12	0.02	8	4	6750T
Rv2205c	Conserved hypothetical protein	0.001	8	5	315A
Rv0236c	Conserved membrane protein	0.001	8	1	981T
Rv3228	Conserved hypothetical protein	0.01	6	2	96G
PE_PGRS4*	PE_PGRS family protein	0.0009, 0.02	6,6	0,2	171G, 621T
PE_PGRS9	PE_PGRS family protein	0.002	8	4	591G
PPE32	PPE family protein	0.004	6	1	993A
PE_PGRS7	PE_PGRS family protein	0.007	6	1	3426G

*** Two sites within PE_PGRS4 were significant, the p-values, and number of branches for the two significant sites are listed in order of sites listed in 6th column. ** average p-value over all replicates of trees (ml, Bayesian, parsimony, and ml and parsimony reconstructions, site had to be significant by all replicate trees/reconstructions to be considered significant) ***these counts represent the number of parsimony phylogenetic branches with one or more changes anywhere along the gene or intergenic region, all changes were treated equal. The tree contained a total of 87 resistant branches and 158 sensitive branches.**

Supplementary Table 10: Genes significant by each of phyC, dN/dS, and differential density for rifampicin resistance in increasing order of p-value. In bold and underlined are genes previously associated with resistance to rifampicin.

Phylogenetic Convergence

Number	Rv-number	Symbol	Description	P-value (<0.05) *Site & Gene convergence. ** Site convergence only (p-values ordered by site1, (site2), gene)	Convergent site(s)
1	**Rv0667**	***rpoB***	**DNA-directed RNA polymerase beta chain**	**0, 0.0017, 0.0075***	**1349T, 1304T**
2	Rv0682	*rpsL*	30S ribosomal protein S12	<0.0001, 00017*	128G
3	Rv3795	*embB*	membrane indolylacetylinositol arabinosyltransferase	<0.0001, 0.0386*	916G
4	Rv2043c	*pncA*	pyrazinamidase/nicotinamidas	<0.0001	
5	Rv0746	PE_PGRS9	PE-PGRS family protein	0.0004	
6	Rv2896c		conserved hypothetical protein	0.0075**	457T
7	Rv2853	PE_PGRS48	PE-PGRS family protein	0.0089	
8	Rvnr01	*rrs*	ribosomal RNA 16S	0.0115	
9	Rv0064		conserved membrane protein	0.0122	
10	Rv0236c		conserved membrane protein	0.014	
11	Rv2931	*ppsA*	phenolpthiocerol synthesis type-I polyketide synthase	0.0164	
12	Rv0388c	PPE9	PPE family protein	0.0164	
13	Rv1908c	*katG*	catalase-peroxidase-peroxynitritase T	0.0244	
14	Rv3919c	*gid*	glucose-inhibited division protein B	0.0244	
15	Rv2082		conserved hypothetical protein	0.0269	
16	Rv2155c	*murD*	UDP-N-acetylmuramoylalanine-D-glutamate ligase	0.0279	
17	Rv0218		conserved membrane protein	0.0326**	946T
18	Rv0050	*ponA1*	bifunctional penicillin-binding protein	0.0386**	1891C
19	Rv3854c	*ethA*	monooxygenase	0.042	

Differential Density

Numbe	Rv-	Symbol	Description	P-value (<0.05)	Clustering Index

r	number					
1	Rv2043c	*pncA*	pyrazinamidase/nicotinamidas	<0.0001	15	
2	**Rv0667**	***rpoB***	**DNA-directed RNA polymerase beta chain**	**<0.0001**	**32**	
3	Rv3795	*embB*	membrane indolylacetylinositol arabinosyltransferase	<0.0001	14	
4	Rvnr01	*rrs*	ribosomal RNA 16S	<0.0001	12	
5	Rv0977	PE_PGRS16	PE-PGRS family protein	0.0207	8	
6	Rv3854c	*ethA*	monooxygenase	0.0219	8	
7	Rv0236c		conserved membrane protein	0.0291	7	
8	Rv0006	*gyrA*	DNA gyrase subunit A	0.0345	5	
9	Rv2931	*ppsA*	phenolpthiocerol synthesis type-I polyketide synthase	0.046	4	
10	Rv2853	PE_PGRS48	PE-PGRS family protein	0.0497	7	
11	Rv2024c		conserved hypothetical protein	0.0666	6	
12	Rv0682	*rpsL*	30S ribosomal protein S12	0.0837	7	
13	Rv1908c	*katG*	catalase-peroxidase-peroxynitritase T	0.1842	6	
14	Rv1325c	PE_PGRS24	PE-PGRS family protein	0.2183	6	
15	Rv0668	*rpoC*	DNA-directed RNA polymerase beta chain	0.3638	5	
16	Rv2946c	*pks1*	polyketide synthase	0.3638	5	
17	Rv3919c	*gid*	glucose-inhibited division protein B	0.3638	5	
18	Rv0218		conserved membrane protein	0.4529	5	
19	Rv0192		conserved hypothetical protein	0.5719	5	
20	Rv2828c		conserved hypothetical protein	0.6208	5	
21	Rv3468c		dTDP-glucose-4,6-dehydratase	0.6208	5	
22	Rv0988		hypothetical exported protein	0.6262	5	
23	Rv1193	*fadD36*	fatty-acid-CoA ligase	0.6262	5	

dN/dS

Number	Rv-number	Symbol	Description	P-value (<1.25E-5)	Significant site(s) by Bayesian empiric bayes method * <0.05, **<0.01	Class dN/dS ψ (Supplementary Table 7)
1	Rv0746	PE_PGRS9	PE-PGRS family protein	1.04E-35	191 G 1.000**, 252 A 0.997**, 280 D	236

					1.000**, 320 A 1.000**, 445 A 1.000**	
2	Rv0279c	PE_PGRS4	PE-PGRS family protein	7.36E-28	77 V 1.000**, 325 N 0.995**, 352 G 1.000**, 372 I 0.995**	134
3	**Rv0667**	***rpoB***	**DNA-directed RNA polymerase beta chain**	**1.06E-27**	**435 D 1.000**, 445 H 0.982*, 450 L 1.000**, 452 L 0.983*, 491 I 0.980*, 731 L 0.983***	**199**
4	Rv0532	PE_PGRS6	PE-PGRS family protein	1.18E-25	227 D 0.980*, 239 G 1.000**	999
5	Rv2931	*ppsA*	phenolpthiocerol synthesis type-I polyketide synthase	2.85E-24	624 D 1.000**, 803 A 0.980*, 1194 L 0.990**	826
6	Rv0747	PE_PGRS10	PE-PGRS family protein	2.51E-23	295 K 1.000**, 300 S 1.000**	146
7	Rv3795	*embB*	membrane indolylacetylinositol arabinosyltransferase	4.28E-23	306 I 1.000**, 354 D 0.995**, 406 G 0.995**	999
8	Rv1753c	PPE24	PPE family protein	5.97E-20	488 T 1.000**	999
9	Rv2853	PE_PGRS48	PE-PGRS family protein	3.32E-19	180 G 1.000**	999
10	Rv3512	PE_PGRS56	PE-PGRS family protein	3.67E-19	253 A 1.000**, 306 I 0.988*	999
11	Rv0050	*ponA1*	bifunctional penicillin-binding protein	6.95E-18	631 P 1.000**	999
12	Rv0218		conserved membrane protein	3.46E-17	316 R 1.000**	999
13	Rv2024c		conserved hypothetical protein	5.48E-17	47 R 0.997**, 154 G 1.000**	513
14	Rv1446c	*opcA*	oxpp cycle protein	1.33E-16	192 R 1.000**	999
15	Rv2079		conserved hypothetical protein	6.35E-15	47 C 1.000**, 95 L 0.977*	999
16	Rv0006	*gyrA*	DNA gyrase subunit A	6.93E-15	90 A 0.999**, 94 A 1.000**, 95 T 1.000**, 292 R 0.979*, 376 R 0.980*, 668 D 1.000**	999
17	Rv2828c		conserved hypothetical protein	9.39E-15	128 S 1.000**	999
18	Rv0018c	*ppp*	serine/threonine phosphatase	9.94E-15	463 S 1.000**	999
19	Rv0058	*dnaB*	replicative DNA helicase	1.53E-14	552 R 1.000**	999
20	Rv0388c	PPE9	PPE family protein	2.39E-14	138 A 1.000**, 139 Q 1.000**	52
21	Rv2874	*dipZ*	cytochrome C biogenesis protein	1.10E-13	672 D 0.998**	999
22	Rv2668		exported alanine and valine rich protein	2.09E-13	3 R 1.000**	999
23	Rv0292		conserved membrane protein	2.43E-13	217 D 1.000**	999
24	Rv0280	PPE3	PPE family protein	2.43E-13	337 S 1.000**	999
25	Rv0658c		conserved membrane protein	3.47E-13	75 P 1.000**	999

26	Rv3077		hydrolase	5.10E-13	310 G 1.000**	197
27	Rv0236c		conserved membrane protein	7.54E-13	1080 G 1.000**	310
28	Rv2584c	*apt*	adenine phosphoribosyltransferase	8.07E-13	147 E 1.000**	999
29	Rv1394c	*cyp132*	cytochrome P450 132	9.51E-13	135 L 1.000**	999
30	Rv2917		conserved alanine and arginine rich protein	1.12E-12	594 L 1.000**	999
31	Rv0278c	PE_PGRS3	PE-PGRS family protein	1.68E-12	807 G 1.000**	324
32	Rv2896c		conserved hypothetical protein	2.02E-12	153 A 1.000**	999
33	Rv2450c	*rpfE*	resuscitation-promoting factor	2.33E-12	20 T 0.992**, 126 R 1.000**	999
34	Rv1812c		dehydrogenase	2.38E-12	30 P 0.998**	999
35	Rv2155c	*murD*	UDP-N-acetylmuramoylalanine-D-glutamate ligase	4.03E-12	247 G 0.999**	999
36	Rv3711c	*dnaQ*	DNA polymerase III epsilon subunit	5.24E-12	211 L 1.000**	999
37	Rv3366	*spoU*	tRNA/rRNA methylase	1.11E-11	92 R 1.000**	999
38	Rv2082		conserved hypothetical protein	1.34E-11	638 R 0.996**	209
39	Rv0159c	PE3	PE family protein	2.38E-11	14 A 1.000**	999
40	Rv0064		conserved membrane protein	2.93E-11	457 D 0.982*, 906 R 1.000**	378
41	Rv3468c		dTDP-glucose-4,6-dehydratase	3.58E-11	62 V 1.000**	999
42	Rv3479		transmembrane protein	6.12E-11	174 R 0.991**	475
43	Rv2825c		conserved hypothetical protein	6.33E-11	162 S 0.998**	999
44	Rv3490	*otsA*	alpha, alpha-trehalose-phosphate synthase	6.85E-11	77 E 0.958*	999
45	Rv2090		5-3 exonuclease	1.76E-10	358 F 0.999**	999
46	Rv0964c		hypothetical protein	4.18E-10	124 T 0.999**	999
48	Rv1093	*glyA1*	serine hydroxymethyltransferase 1	1.02E-09	36 A 0.990*	999
50	Rv2490c	PE_PGRS43	PE-PGRS family protein	1.40E-09	1399 G 0.996**	999
51	Rv1971	*mce3F*	MCE-family protein	1.54E-09	396 E 1.000**	999
52	Rv0082		oxidoreductase	1.89E-09	74 R 0.986*	999
53	Rv3776		conserved hypothetical protein	1.91E-09	112 S 0.983*, 329 V 0.996**	484
54	Rv2947c	*pks15*	polyketide synthase	2.69E-09	333 A 0.999**	999
55	Rv2439c	*proB*	glutamate 5-kinase protein	2.81E-09	226 S 0.993**	999
56	Rv2458	*mmuM*	homocysteine S-methyltransferase	3.38E-09	125 Y 0.998**	999
57	Rv3245c	*mtrB*	two component system sensor histidine kinase	3.39E-09	18 P 0.999**, 517 L 0.984*	875
58	Rv3151	*nuoG*	NADH dehydrogenase I	4.66E-09	474 M 0.989*	999

			chain G			
59	Rv2488c		transcriptional regulator, luxR-family	6.02E-09	265 T 0.997**	999
60	Rv3449	*mycP4*	membrane-anchored mycosin	6.57E-09	87 T 0.995**	999
61	Rv0192		conserved hypothetical protein	6.73E-09	127 P 0.998**	999
62	Rv0095c		conserved hypothetical protein	8.31E-09	57 D 1.000**	999
63	Rv3835		conserved membrane protein	1.01E-08	294 L 0.996**	999
64	Rv1704c	*cycA*	D-serine/alanine/glycine transporter protein	1.01E-08	93 L 0.994**	999
65	Rv2794c		conserved hypothetical protein	1.08E-08	87 V 0.997**	999
66	Rv1320c		adenylate cyclase	1.26E-08	531 A 0.996**	999
68	Rv1463		ABC transporter ATP-binding protein	1.59E-08	198 E 0.992**	999
69	Rv1895		dehydrogenase	1.66E-08	270 L 0.995**	999
70	Rv2770c	PPE44	PPE family protein	1.68E-08	194 F 0.995**	999
71	Rv2059		conserved hypothetical protein	1.81E-08	317 T 0.990**	999
72	Rv2567		conserved alanine and leucine rich protein	1.85E-08	645 Q 0.992**	999
73	Rv1378c		conserved hypothetical protein	1.92E-08	37 W 0.994**	999
74	Rv0194		drugs-transport transmembrane ATP-binding protein ABC transporter	3.06E-08	74 T 0.994**	999
75	Rv2333c		conserved membrane transport protein	4.79E-08	69 Y 0.994**	999
76	Rv0338c		iron-sulfur-binding reductase	5.71E-08	506 G 0.992**, 621 V 0.992**	999
77	Rv0881		rRNA methyltransferase	6.06E-08	115 R 0.999**	634
78	Rv0417	*thiG*	thiamin biosynthesis protein	6.77E-08	75 C 0.998**	343
79	Rv0109	PE_PGRS1	PE-PGRS family protein	8.68E-08	346 G 0.995**	999
80	Rv3425	PPE57	PPE family protein	1.07E-07	63 L 0.972*, 128 T 0.999**	999
82	Rv3093c		oxidoreductase	1.31E-07	210 C 0.995**	999
83	Rv1319c		adenylate cyclase	1.46E-07	439 D 0.995**	235
84	Rv2101	*helZ*	helicase	1.63E-07	462 L 0.986*	330
85	Rv2017		transcriptional regulator	1.92E-07	262 E 0.993**	999
86	Rv2769c	PE27	PE family protein	2.19E-07	136 A 0.978*, 270 M 0.999**	999
87	Rv1186c		conserved hypothetical protein	2.30E-07	207 A 0.988*	999
88	Rv2741	PE_PGRS47	PE-PGRS family protein	2.79E-07	271 S 0.999**	80

89	Rv2495c	*pdhC*	dihydrolipoamide S-acetyltransferase E2 component	3.30E-07	107 A 0.997**	696
90	Rv3777		oxidoreductase	3. E-07	160 A 0.999**	999
91	Rv0048c		membrane protein	3.67E-07	248 E 0.992**, 250 V 0.992**	999
92	Rv3341	*metA*	homoserine O-acetyltransferase	4.12E-07	87 S 0.950	999
93	Rv3511	PE_PGRS55	PE-PGRS family protein	4.14E-07	396 N 0.987*, 589 G 0.991**	503
94	Rv1716		conserved hypothetical protein	4.19E-07	178 G 0.977*, 276 A 0.975*	999
95	Rv0226c		conserved membrane protein	4.34E-07	379 P 0.962*	999
96	Rv0995	*rimJ*	ribosomal-protein-alanine acetyltransferase	4.34E-07	23 G 0.997**, 105 Y 0.978*	999
97	Rv0465c		transcriptional regulator	4.49E-07	106 C 0.994**	999
98	Rv1644	*tsnR*	23S rRNA methyltransferase	5.38E-07	232 P 0.994**	999
99	Rv3764c		two component system sensor kinase	6.42164E-07	246 R 0.961*	999
100	Rv1915	*aceAa*	isocitrate lyase	6.78E-07	179 D 0.995**	999
101	Rv2807		conserved hypothetical protein	8.84E-07	72 E 0.998**, 77 D 0.988*	999
102	Rv2236c	*cobD*	cobalamin biosynthesis transmembrane protein	9.86E-07	79 C 0.991**	744
103	Rv0787		hypothetical protein	1.05E-06	267 H 0.991**	802
104	Rv1618	*tesB1*	acyl-CoA thioesterase II	1.07E-06	121 L 0.995**	999
105	Rv0425c	*ctpH*	metal cation transporting P-type ATPase	1.13E-06	689 V 0.982*	398
106	Rv1326c	*glgB*	1,4-alpha-glucan branching enzyme	1.22E-06	470 S 0.908	999
107	Rv0572c		hypothetical protein	1.28E-06	31 L 0.989*	999
108	Rv1640c	*lysS*	lysyl-tRNA synthetase 2 lysX	1.36E-06	701 I 0.942	999
109	Rv3063	*cstA*	carbon starvation protein A	1.45E-06	559 S 0.929	999
110	Rv3782		L-rhamnosyltransferase	1.66E-06	274 V 0.992**	999
111	Rv3021c	PPE47	PPE family protein	2.26E-06	222 A 0.952*	64
112	Rv2433c		hypothetical protein	2.29E-06	26 L 0.998**	999
113	Rv2316	*uspA*	sugar-transport membrane protein ABC transporter	2.38E-06	67 D 0.996**, 127 L 0.974*	999
114	Rv1232c		conserved hypothetical protein	2.53E-06	149 G 0.981*	483
115	Rv0103c	*ctpB*	cation-transporter P-type ATPase B	3.02E-06	22 S 0.981*	999
116	Rv2611c		acyltransferase	3.22E-06	197 C 0.979*	999
117	Rv0676c	*mmpL5*	transmembrane transport protein	3.27E-06	948 V 0.934	999
118	Rv1160	*mutT2*	mutator protein mutT	3.46E-06	58 G 0.969*	999
119	Rv3879c		hypothetical alanine and proline rich protein	3.68E-06	729 S 0.980*	269

120	Rv1321		conserved hypothetical protein	3.69E-06	144 R 0.997**	999
121	Rv3497c	*mce4C*	MCE-family protein	4.09E-06	191 R 0.928	999
122	Rv1486c		conserved hypothetical protein	4.17E-06	198 N 0.972*	999
123	Rv3365c		conserved hypothetical protein	4.20E-06	38 P 0.989*, 687 S 0.953*	999
124	Rv1127c	*ppdK*	pyruvate, phosphate dikinase	4.92E-06	69 E 0.899	999
125	Rv1400c	*lipI*	lipase lipH	5.04E-06	106 T 0.989*	999
126	Rv3892c	PPE69	PPE family protein	5.65E-06	19 K 0.962*	999
127	Rv1908c	*katG*	catalase-peroxidase-peroxynitritase T	5.98E-06	315 T 0.979*	724
128	Rv3655c		conserved hypothetical protein	6.23E-06	100 S 0.979*	119
129	Rv0557	*pimB*	mannosyltransferase	6.32E-06	none	1
130	Rv3144c	PPE52	PPE family protein	6.42E-06	226 S 0.969*	999
131	Rv0758	*phoR*	two component system sensor kinase	7.11E-06	172 L 0.980*	999
132	Rv1538c	*ansA*	L-aparaginase	7.79E-06	281 G 0.985*	999
133	Rv2290	*lppO*	lipoprotein	8.61E-06	16 A 0.986*	999
134	Rv0855	*far*	fatty-acid-CoA racemase	8.78E-06	24 A 0.993**	342
135	Rv3737		conserved membrane protein	9.15E-06	40 G 0.853	999
136	Rv2037c		conserved membrane protein	9.16E-06	312 Y 0.914	999
137	Rv3282	*maf*	conserved hypothetical protein	9.30E-06	80 A 0.988*	999
138	Rv1300	*hemK*	hypothetical protein	9.49E-06	194 C 0.988*	999
139	Rv3317	*sdhD*	succinate dehydrogenase hydrophobic membrane anchor subunit	1.00E-05	112 T 0.998**, 114 R 0.990**	999
140	Rv3347c	PPE55	PPE family protein	1.07E-05	786 V 0.781	1
141	Rv3591c		hydrolase	1.07E-05	156 F 0.986*	999
142	Rv0668	*rpoC*	DNA-directed RNA polymerase beta chain	1.13E-05	404 D 0.975*, 484 W 0.975*, 698 N 0.998**, 1040 P 0.976*, 1092 E 0.988*, 1231 R 0.976*	999
143	Rv3329		aminotransferase	1.16E-05	122 H 0.969*	999
144	Rv1900c	*lipJ*	lignin peroxidase	1.19E-05	204 M 0.946	999
145	Rv0538		conserved membrane protein	1.21E-05	228 P 0.960*	999
146	Rv2482c	*plsB2*	glycerol-3-phosphate acyltransferase	1.22E-05	778 R 0.956*	999
147	Rv3919c	*gid*	glucose-inhibited division protein B	1.25E-05	92 E 0.973*, 96 R 0.992**, 145 L 0.992**	141

Supplementary Table 12: Genes significant by each of phyC, dN/dS, and differential density for isoniazid resistance in increasing order of p-value. In bold and underlined are genes previously associated with resistance to isoniazid.

Phylogenetic Convergence

Number	Rvnumber	Symbol	Description	P-value (<0.05) *Site & Gene convergence. ** Site convergence only (p-values ordered by site1, (site2), gene)	Convergent site(s)
1	Rv3795	*embB*	membrane indolylacetylinositol arabinosyltransferase	<0.0001	
2	Rv2043c	*pncA*	pyrazinamidase/nicotinamidas	<0.0001	
3	Rv0682	*rpsL*	30S ribosomal protein S12	<0.0001, 0.0005*	128G
4	Rv0667	*rpoB*	DNA-directed RNA polymerase beta chain	<0.0001, 0.0022, 0.0038*	1349T, 1304T
5	Rv2931	*ppsA*	phenolpthiocerol synthesis type-I polyketide synthase	0.0018	
6	Rv2896c		conserved hypothetical protein	0.0038**	457T
7	Rvnr01	*rrs*	ribosomal RNA 16S	0.0053	
8	Rv2155c	*murD*	UDP-N-acetylmuramoylalanine-D-glutamate ligase	0.0062	
9	Rv0746	PE_PGRS9	PE-PGRS family protein	0.0062	
10	Rv2853	PE_PGRS48	PE-PGRS family protein	0.0085	
11	**Rv1908c**	***katG***	**catalase-peroxidase-peroxynitritase T**	**0.0118**	
12	Rv0388c	PPE9	PPE family protein	0.0179	
13	Rv3854c	*ethA*	monooxygenase	0.0223	
14	Rv0064		conserved membrane protein	0.0287	
15	Rv2082		conserved hypothetical protein	0.0346	

Differential Density

Number	Rv-number	*Symbol*	Description	P-value (<0.05)	Clustering Index
1	Rv0667	*rpoB*	DNA-directed RNA polymerase beta chain	<0.0001	26
2	Rv3795	*embB*	membrane indolylacetylinositol arabinosyltransferase	<0.0001	12
3	Rv2043c	*pncA*	pyrazinamidase/nicotinamidas	<0.0001	15
4	Rvnr01	*rrs*	ribosomal RNA 16S	0.0001	12
5	Rv2853	PE_PGRS48	PE-PGRS family protein	0.0052	9
6	Rv2931	*ppsA*	phenolpthiocerol synthesis type-I polyketide synthase	0.0096	6
7	Rv3854c	*ethA*	monooxygenase	0.0164	8
8	Rv0006	*gyrA*	DNA gyrase subunit A	0.024	5

9	Rv2024c		conserved hypothetical protein	0.0515	6	
10	Rv0682	*rpsL*	30S ribosomal protein S12	0.0674	7	
11	Rv0977	PE_PGRS16	PE-PGRS family protein	0.0822	6	
12	**Rv1908c**	***katG***	**catalase-peroxidase-peroxynitritase T**	**0.1445**	**6**	
13	Rv1325c	PE_PGRS24	PE-PGRS family protein	0.1778	6	
14	Rv2946c	*pks1*	polyketide synthase	0.301	5	
15	Rv0995	*rimJ*	ribosomal-protein-alanine acetyltransferase	0.5038	5	
16	Rv2828c		conserved hypothetical protein	0.5564	5	
17	Rv3468c		dTDP-glucose-4,6-dehydratase	0.5564	5	
18	Rv1193	*fadD36*	fatty-acid-CoA ligase	0.5693	5	
19	Rv0988		hypothetical exported protein	0.5693	5	
20	Rv1441c	PE_PGRS26	PE-PGRS family protein	0.5693	5	
21	593			0.5693	5	

dN/dS

Number	Rv-number	Symbol	Description	P-value (<1.25E-5)	Significant site(s) by Bayesian empiric bayes method * <0.05, **<0.01	Class dN/dS ψ (Supplementary Table 7)
1	Rv0746	PE_PGRS9	PE-PGRS family protein	1.03E-47	191 G 1.000**, 252 A 0.998**, 280 D 1.000**, 320 A 1.000**, 445 A 1.000**	354
2	Rv0532	PE_PGRS6	PE-PGRS family protein	6.00E-28	227 D 0.965*, 239 G 1.000**	999
3	Rv0279c	PE_PGRS4	PE-PGRS family protein	1.19E-26	77 V 1.000**, 325 N 0.995**, 352 G 1.000**, 372 I 0.986*	130
4	Rv0667	*rpoB*	DNA-directed RNA polymerase beta chain	1.30E-25	435 D 1.000**, 445 H 0.974*, 450 L 1.000**, 452 L 0.976*	228
5	Rv0747	PE_PGRS10	PE-PGRS family protein	4.23E-25	295 K 1.000**, 300 S 1.000**	321
6	Rv2931	*ppsA*	phenolpthiocerol synthesis type-I polyketide synthase	6.77E-24	624 D 1.000**, 803 A 0.980*, 1194 L 0.990**	824
7	Rv2853	PE_PGRS48	PE-PGRS family protein	1.90E-22	180 G 1.000**	999
8	Rv3795	*embB*	membrane indolylacetylinositol arabinosyltransferase	5.28E-21	306 I 1.000**, 354 D 0.997**, 406 G 0.997**, 497 Q 0.974*	999
9	Rv1753c	PPE24	PPE family protein	6.34E-17	488 T 1.000**	999
10	Rv2828c		conserved hypothetical protein	5.71E-16	128 S 1.000**	999
11	Rv0278c	PE_PGRS3	PE-PGRS family protein	2.09E-15	807 G 1.000**	312
12	Rv2024c		conserved hypothetical protein	2.88E-15	47 R 0.996**, 154 G 1.000**	472
13	Rv0058	*dnaB*	replicative DNA helicase	6.33E-15	552 R 1.000**	999
14	Rv2079		conserved hypothetical protein	6.35E-15	47 C 1.000**, 95 L 0.977*	999
15	Rv0006	*gyrA*	DNA gyrase subunit A	6.94E-15	90 A 0.999**, 94 A 1.000**, 95 T 1.000**,	999

					292 R 0.979*, 376 R 0.980*, 668 D 1.000**	
16	Rv0658c		conserved membrane protein	1.22E-14	75 P 1.000**	999
17	1		conserved membrane protein	1.52E-14	316 R 1.000**	999
18	Rv2825c		conserved hypothetical protein	1.90E-14	162 S 1.000**	999
19	Rv1446c	*opcA*	oxpp cycle protein	3.49E-14	192 R 1.000**	999
20	Rv0018c	*ppp*	serine/threonine phosphatase	1.24E-13	463 S 1.000**	999
21	Rv3512	PE_PGRS56	PE-PGRS family protein	2.06E-13	253 A 0.957*, 306 I 0.975*	999
22	Rv2917		conserved alanine and arginine rich protein	1.18E-12	594 L 1.000**	999
23	Rv0280	PPE3	PPE family protein	1.28E-12	337 S 1.000**	999
24	Rv0236c		conserved membrane protein	1.55E-12	1080 G 0.999**	323
25	Rv0050	*ponA1*	bifunctional penicillin-binding protein	1.61E-12	631 P 1.000**	999
26	Rv0800	*pepC*	aminopeptidase	1.80E-12	139 L 1.000**	999
27	Rv2896c		conserved hypothetical protein	2.02E-12	153 A 1.000**	999
28	Rv2450c	*rpfE*	resuscitation-promoting factor	2.33E-12	20 T 0.992**, 126 R 1.000**	999
29	Rv1812c		dehydrogenase	2.41E-12	30 P 0.998**	999
30	Rv0388c	PPE9	PPE family protein	2.71E-12	138 A 1.000**, 139 Q 1.000**	47
31	Rv2155c	*murD*	UDP-N-acetylmuramoylalanine-D-glutamate ligase	3.64E-12	247 G 1.000**	999
32	Rv3711c	*dnaQ*	DNA polymerase III epsilon subunit	5.25E-12	211 L 1.000**	999
33	Rv3366	*spoU*	tRNA/rRNA methylase	1.11E-11	92 R 1.000**	999
34	Rv3077		hydrolase	2.62E-11	310 G 1.000**	197
35	Rv1394c	*cyp132*	cytochrome P450 132	4.94E-11	135 L 0.999**	999
36	Rv2490c	PE_PGRS43	PE-PGRS family protein	5.13E-11	1399 G 0.996**	999
37	Rv0964c		hypothetical protein	4.18E-10	124 T 0.999**	999
39	Rv3245c	*mtrB*	two component system sensor histidine kinase	1.43E-09	18 P 0.999**, 517 L 0.984*	875
41	Rv3776		conserved hypothetical protein	1.91E-09	112 S 0.983*, 329 V 0.996**	484
42	Rv0082		oxidoreductase	1.95E-09	74 R 0.986*	999
43	Rv2584c	*apt*	adenine phosphoribosyltransferase	2.15E-09	147 E 0.999**	999
44	Rv0064		conserved membrane protein	2.17E-09	906 R 1.000**	347
45	Rv2823c		conserved hypothetical protein	2.39E-09	101 Y 0.999**	999
46	Rv2458	*mmuM*	homocysteine S-methyltransferase	2.64E-09	125 Y 0.998**	999
47	Rv2439c	*proB*	glutamate 5-kinase protein	2.81E-09	226 S 0.993**	999
48	Rv3511	PE_PGRS55	PE-PGRS family protein	2.86E-09	396 N 0.988*, 589 G 0.991**	503
49	Rv2947c	*pks15*	polyketide synthase	3.13E-09	333 A 0.999**	999

50	Rv3468c		dTDP-glucose-4,6-dehydratase	4.46E-09	62 V 0.999**	999
51	**Rv1908c**	***katG***	**catalase-peroxidase-peroxynitritase T**	**5.77E-09**	**315 T 0.998****	**939**
52	Rv2488c		transcriptional regulator, luxR-family	6.03E-09	265 T 0.997**	999
53	Rv2567		conserved alanine and leucine rich protein	8.13E-09	645 Q 0.988*	999
54	Rv3835		conserved membrane protein	1.01E-08	294 L 0.996**	999
55	Rv1704c	*cycA*	D-serine/alanine/glycine transporter protein	1.01E-08	93 L 0.994**	999
56	Rv2874	*dipZ*	cytochrome C biogenesis protein	1.16E-08	672 D 0.986*	999
57	Rv2059		conserved hypothetical protein	1.81E-08	317 T 0.990**	999
58	Rv1378c		conserved hypothetical protein	1.92E-08	37 W 0.994**	999
59	Rv1463		ABC transporter ATP-binding protein	2.73E-08	198 E 0.992**	999
60	Rv1319c		adenylate cyclase	2.80E-08	439 D 0.993**	206
61	Rv0417	*thiG*	thiamin biosynthesis protein	4.03E-08	75 C 0.998**	343
62	Rv0758	*phoR*	two component system sensor kinase	4.09E-08	172 L 0.996**	999
63	Rv2090		5-3 exonuclease	4.33E-08	358 F 0.994**	999
64	Rv0338c		iron-sulfur-binding reductase	5.72E-08	506 G 0.992**, 621 V 0.992**	999
65	Rv0881		rRNA methyltransferase	6.06E-08	115 R 0.999**	633
66	Rv2905	*lppW*	alanine rich lipoprotein	6.25E-08	81 Q 0.992**	999
67	Rv0159c	PE3	PE family protein	7.45E-08	14 A 0.994**	999
68	Rv0980c	PE_PGRS18	PE-PGRS family protein	8.30E-08	None	1
69	Rv1193	*fadD36*	fatty-acid-CoA ligase	1.32E-07	124 P 0.996**	999
70	Rv1186c		conserved hypothetical protein	1.57E-07	207 A 0.988*	999
71	Rv1971	*mce3F*	MCE-family protein	2.03E-07	396 E 0.997**	999
72	Rv2017		transcriptional regulator	2.15E-07	262 E 0.993**	999
73	Rv3630		conserved membrane protein	2.70E-07	40 T 0.995**	999
74	Rv1027c	*kdpE*	transcriptional regulator	2.79E-07	60 G 0.970*	999
76	Rv2495c	*pdhC*	dihydrolipoamide S-acetyltransferase E2 component	3.30E-07	107 A 0.997**	696
77	Rv0834c	PE_PGRS14	PE-PGRS family protein	3.54E-07	804 N 0.990*	999
78	Rv0048c		membrane protein	3.67E-07	248 E 0.992**, 250 V 0.992**	999
79	Rv1716		conserved hypothetical protein	4.19E-07	178 G 0.977*, 276 A 0.975*	999
80	Rv0995	*rimJ*	ribosomal-protein-alanine acetyltransferase	4.34E-07	23 G 0.997**, 105 Y 0.978*	999
81	Rv0109	PE_PGRS1	PE-PGRS family protein	4.46E-07	346 G 0.995**	999
82	Rv0465c		transcriptional regulator	4.49E-07	106 C 0.994**	999
83	Rv0259c		conserved hypothetical protein	4.67E-07	182 V 0.997**	999
84	Rv3389c		dehydrogenase	4.73E-07	165 P 0.974*	999

85	Rv0226c		conserved membrane protein	4.87E-07	379 P 0.968*	999
86	Rv1449c	*tkt*	transketolase	6.34E-07	18 D 0.902	719
87	Rv3764c		two component system sensor kinase	6.44E-07	246 R 0.961*	999
88	Rv1915	*aceAa*	isocitrate lyase	6.78E-07	179 D 0.995**	999
89	Rv3490	*otsA*	alpha, alpha-trehalose-phosphate synthase	7.26E-07	77 E 0.958*	999
90	Rv0192		conserved hypothetical protein	7.53E-07	127 P 0.986*	999
91	Rv2436	*rbsK*	ribokinase	7.90E-07	282 A 0.994**	252
92	Rv0538		conserved membrane protein	9.02E-07	228 P 0.992**	999
93	Rv3449	*mycP4*	membrane-anchored mycosin	9.92E-07	87 T 0.976*	999
94	Rv0787		hypothetical protein	1.05E-06	267 H 0.991**	802
95	Rv1326c	*glgB*	1,4-alpha-glucan branching enzyme	1.22E-06	470 S 0.908	999
96	Rv3341	*metA*	homoserine O-acetyltransferase	1.23E-06	87 S 0.950	999
97	Rv2769c	PE27	PE family protein	1.31E-06	136 A 0.978*, 270 M 0.999**	999
98	Rv1640c	*lysS*	lysyl-tRNA synthetase 2 lysX	1.37E-06	701 I 0.942	999
99	Rv2048c	*pks12*	hypothetical protein	1.50E-06	917 S 0.863	477
100	Rv3199c	*nudC*	NADH pyrophosphatase	1.55E-06	239 P 0.987*	999
101	Rv0323c		conserved hypothetical protein	1.64E-06	142 G 0.991**	999
102	Rv3782		L-rhamnosyltransferase	1.66E-06	274 V 0.992**	999
103	Rv3425	PPE57	PPE family protein	2.21E-06	128 T 0.994**	725
104	Rv2433c		hypothetical protein	2.35E-06	26 L 0.998**	999
105	Rv2316	*uspA*	sugar-transport membrane protein ABC transporter	2.38E-06	67 D 0.996**, 127 L 0.974*	999
106	Rv1538c	*ansA*	L-aparaginase	2.40E-06	281 G 0.985*	999
108	Rv1300	*hemK*	hypothetical protein	2.48E-06	194 C 0.969*	999
109	Rv1232c		conserved hypothetical protein	2.53E-06	149 G 0.981*	483
110	Rv1895		dehydrogenase	2.74E-06	270 L 0.969*	999
111	Rv0103c	*ctpB*	cation-transporter P-type ATPase B	2.91E-06	22 S 0.981*	999
112	Rv3479		transmembrane protein	2.99E-06	174 R 0.937	355
113	Rv3590c	PE_PGRS58	PE-PGRS family protein	3.26E-06	314 A 0.856	999
114	Rv1160	*mutT2*	mutator protein mutT	3.46E-06	58 G 0.969*	999
115	Rv0425c	*ctpH*	metal cation transporting P-type ATPase	3.53E-06	689 V 0.982*	398
116	Rv3879c		hypothetical alanine and proline rich protein	3.68E-06	729 S 0.980*	269
117	Rv1321		conserved hypothetical protein	3.68E-06	144 R 0.997**	999
118	Rv3347c	PPE55	PPE family protein	4.01E-06	786 V 0.894	103
119	Rv3497c	*mce4C*	MCE-family protein	4.09E-06	191 R 0.928	999
120	Rv1486c		conserved hypothetical protein	4.18E-06	198 N 0.972*	999
121	Rv3365c		conserved hypothetical protein	4.19E-06	38 P 0.989*, 687 S	999

					0.953*	
122	Rv0118c	*oxcA*	oxalyl-CoA decarboxylase	4.82E-06	253 A 0.961*	503
123	Rv3837c		phosphoglycerate mutase	4.89E-06	15 G 0.989*	999
124	Rv1400c	*lipI*	lipase lipH	5.04E-06	106 T 0.989*	999
125	Rv1489		conserved hypothetical protein	5.42E-06	52 K 0.990*	999
126	Rv3892c	PPE69	PPE family protein	5.65E-06	19 K 0.962*	999
127	Rv3144c	PPE52	PPE family protein	6.42E-06	226 S 0.969*	999
128	Rv0376c		conserved hypothetical protein	6.93E-06	14 T 0.826	102
129	Rv2982c	*gpsA*	glycerol-3-phosphate dehydrogenase gpdA2	7.16E-06	133 A 0.985*	374
130	Rv3805c		conserved membrane protein	8.25E-06	397 D 0.967*	999
131	Rv2290	*lppO*	lipoprotein	8.61E-06	16 A 0.986*	999
132	Rv0095c		conserved hypothetical protein	8.90E-06	57 D 0.998**	999
133	Rv3737		conserved membrane protein	9.18E-06	40 G 0.853	999
134	Rv1320c		adenylate cyclase	9.89E-06	531 A 0.977*	999
135	Rv2015c		conserved hypothetical protein	1.13E-05	None	1
136	Rv3329		aminotransferase	1.16E-05	122 H 0.969*	999
137	Rv2741	PE_PGRS47	PE-PGRS family protein	1.18E-05	271 S 0.994**	60
138	Rv2482c	*plsB2*	glycerol-3-phosphate acyltransferase	1.22E-05	778 R 0.956*	999

Supplementary Table 13: Genes significant by each of phyC, dN/dS, and differential density for ethambutol resistance in increasing order of p-value. In bold and underlined are genes previously associated with resistance to ethambutol.

Phylogenetic Convergence

Number	Rv-number	Symbol	Description	P-value (<0.05) *Site & Gene convergence. ** Site convergence only (p-values ordered by site1, (site2), gene)	Convergent site(s)
1	Rv0667	*rpoB*	DNA-directed RNA polymerase beta chain	<0.0001	
2	**Rv3795**	***embB***	**membrane indolylacetylinositol arabinosyltransferase**	**0.0008**	
3	Rv0050	*ponA1*	bifunctional penicillin-binding protein	0.0024**	1891C
4	Rv2043c	*pncA*	pyrazinamidase/nicotinamidas	0.0043	
5	Rv0746	PE_PGRS9	PE-PGRS family protein	0.0086	
6	Rv3919c	*gid*	glucose-inhibited division protein B	0.0198	
7	Rv0747	PE_PGRS10	PE-PGRS family protein	0.0466	

Differential Density

Number	Rv-number	Symbol	Description	P-value (<0.05)	Clustering Index
1	Rv0667	rpoB	DNA-directed RNA polymerase beta chain	<0.0001	18
2	**Rv3795**	**embB**	**membrane indolylacetylinositol arabinosyltransferase**	**0.0082**	**4**
3	Rv2853	PE_PGRS48	PE-PGRS family protein	0.0173	5

dN/dS

Number	Rv-number	Symbol	Description	P-value (<1.25E-5)	Significant site(s) by Bayesian empiric bayes method * <0.05, **<0.01	Class dN/dS ψ (Supplementary Table 7)
1	Rv0746	PE_PGRS9	PE-PGRS family protein	8.24E-33	191 G 0.997**, 252 A 0.973*, 280 D 1.000**, 320 A 1.000**, 445 A 1.000**	263
2	Rv0747	PE_PGRS10	PE-PGRS family protein	7.12E-26	295 K 1.000**, 300 S 1.000**	213
3	Rv2931	*ppsA*	phenolpthiocerol synthesis type-I polyketide synthase	9.66E-22	624 D 1.000**, 1194 L 0.989*	999

4	Rv0279c	PE_PGRS4	PE-PGRS family protein	8.13E-21	77 V 1.000**, 325 N 0.995**, 352 G 1.000**, 372 I 0.995**	139
5	Rv0667	*rpoB*	DNA-directed RNA polymerase beta chain	6.99E-20	435 D 1.000**, 445 H 0.974*, 450 L 1.000**, 452 L 0.976*, 491 I 0.973*	174
6	Rv0278c	PE_PGRS3	PE-PGRS family protein	3.39E-19	807 G 1.000**	999
7	Rv0050	*ponA1*	bifunctional penicillin-binding protein	6.98E-18	631 P 1.000**	999
8	Rv0388c	PPE9	PPE family protein	5.43E-16	138 A 1.000**, 139 Q 1.000**	92
9	Rv2828c		conserved hypothetical protein	4.93E-15	128 S 1.000**	999
10	Rv2853	PE_PGRS48	PE-PGRS family protein	7.66E-15	180 G 1.000**	999
11	Rv1753c	PPE24	PPE family protein	1.37E-14	488 T 1.000**	999
12	Rv1446c	*opcA*	oxpp cycle protein	3.53E-14	192 R 1.000**	999
13	Rv2090		5-3 exonuclease	4.59E-14	358 F 0.999**	999
14	Rv2024c		conserved hypothetical protein	3.00E-13	47 R 0.977*, 154 G 1.000**	999
15	Rv3512	PE_PGRS56	PE-PGRS family protein	3.41E-13	253 A 0.935	999
16	Rv0532	PE_PGRS6	PE-PGRS family protein	3.77E-13	227 D 0.979*, 239 G 1.000**	472
17	Rv3077		hydrolase	4.82E-13	310 G 1.000**	224
18	Rv0082		oxidoreductase	1.64E-12	74 R 0.998**	999
19	Rv0058	*dnaB*	replicative DNA helicase	3.81E-12	552 R 1.000**	999
20	Rv0323c		conserved hypothetical protein	1.28E-11	142 G 0.991**	999
21	Rv0292		conserved membrane protein	1.37E-11	217 D 0.998**	999
22	Rv0006	*gyrA*	DNA gyrase subunit A	2.52E-11	90 A 0.999**, 94 A 1.000**, 95 T 0.982*, 376 R 0.981*, 668 D 1.000**	999
23	Rv0658c		conserved membrane protein	3.00E-11	75 P 1.000**	999
24	Rv2874	*dipZ*	cytochrome C biogenesis protein	3.08E-11	672 D 0.998**	999
25	Rv2450c	*rpfE*	resuscitation-promoting factor	4.39E-11	126 R 1.000**	999
26	Rv3711c	*dnaQ*	DNA polymerase III epsilon subunit	2.31E-10	211 L 0.999**	999
27	Rv2917		conserved alanine and arginine rich protein	3.35E-10	594 L 0.999**	999
28	Rv0800	*pepC*	aminopeptidase	4.33E-10	139 L 0.999**	999
30	Rv3835		conserved membrane protein	6.17E-10	294 L 0.972*	999
31	Rv1093	*glyA1*	serine hydroxymethyltransferase 1	1.05E-09	36 A 0.990*	999

32	**Rv3795**	***embB***	**membrane indolylacetylinositol arabinosyltransferase**	**1.15E-09**	**306 I 0.999**, 406 G 0.991****	**481**
33	Rv1812c		dehydrogenase	1.35E-09	30 P 0.984*	999
34	Rv2584c	*apt*	adenine phosphoribosyltransferase	1.38E-09	147 E 0.999**	999
35	Rv1971	*mce3F*	MCE-family protein	1.60E-09	396 E 0.999**	999
36	Rv2458	*mmuM*	homocysteine S-methyltransferase	2.07E-09	125 Y 0.998**	999
37	Rv2947c	*pks15*	polyketide synthase	3.09E-09	333 A 0.999**	999
38	Rv3347c	PPE55	PPE family protein	3.95E-09	786 V 0.976*, 2259 P 0.965*	149
39	Rv0280	PPE3	PPE family protein	5.47E-09	337 S 0.999**	999
40	Rv0833	PE_PGRS13	PE-PGRS family protein	6.24E-09	584 S 0.982*	554
41	Rv0236c		conserved membrane protein	8.98E-09	1080 G 0.968*	224
42	Rv3479		transmembrane protein	9.85E-09	174 R 0.991**	474
43	Rv3468c		dTDP-glucose-4,6-dehydratase	1.01E-08	62 V 0.999**	999
44	Rv1704c	*cycA*	D-serine/alanine/glycine transporter protein	1.02E-08	93 L 0.994**	999
45	Rv2048c	*pks12*	hypothetical protein	1.14E-08	917 S 0.931	701
46	Rv1895		dehydrogenase	1.67E-08	270 L 0.995**	999
47	Rv1378c		conserved hypothetical protein	1.93E-08	37 W 0.994**	999
48	Rv0417	*thiG*	thiamin biosynthesis protein	2.39E-08	75 C 0.999**	708
49	Rv1463		ABC transporter ATP-binding protein	2.76E-08	198 E 0.992**	999
50	Rv3343c	PPE54	PPE family protein	3.21E-08	103 A 0.899	705
51	Rv2333c		conserved membrane transport protein	4.80E-08	69 Y 0.994**	999
52	Rv0064		conserved membrane protein	4.92E-08	457 D 0.984*, 906 R 0.984*	478
53	Rv0964c		hypothetical protein	7.27E-08	124 T 0.993**	999
54	Rv1394c	*cyp132*	cytochrome P450 132	7.53E-08	135 L 0.989*	999
55	Rv2896c		conserved hypothetical protein	8.43E-08	153 A 0.993**	999
56	Rv1716		conserved hypothetical protein	9.04E-08	178 G 0.982*	999
57	Rv3093c		oxidoreductase	1.31E-07	210 C 0.995**	999
58	Rv0787		hypothetical protein	1.43E-07	267 H 0.994**	999
59	Rv1570	*bioD*	dethiobiotin synthetase	2.02E-07	191 T 0.999**	999
60	Rv2155c	*murD*	UDP-N-acetylmuramoylalanine-D-glutamate ligase	2.43E-07	247 G 0.995**	999
61	Rv2079		conserved hypothetical protein	2.90E-07	47 C 0.989*	850

62	Rv0194		drugs-transport transmembrane ATP-binding protein ABC transporter	2.91E-07	74 T 0.994**	999
63	Rv2490c	PE_PGRS43	PE-PGRS family protein	2.98E-07	1399 G 0.990*	999
64	Rv1319c		adenylate cyclase	3.09E-07	439 D 0.956*	248
65	Rv0095c		conserved hypothetical protein	4.28E-07	57 D 1.000**	999
66	Rv0302		transcriptional regulator, tetR/acrR-family	4.86E-07	84 H 0.995**	999
67	Rv2769c	PE27	PE family protein	5.32E-07	136 A 0.978*, 270 M 0.999**	999
68	Rv2825c		conserved hypothetical protein	6.23E-07	162 S 0.951*	999
69	Rv0425c	*ctpH*	metal cation transporting P-type ATPase	6.48E-07	689 V 0.987*	773
70	Rv0218		conserved membrane protein	7.20E-07	316 R 0.901	999
71	Rv0538		conserved membrane protein	8.99E-07	228 P 0.992**	999
72	Rv1326c	*glgB*	1,4-alpha-glucan branching enzyme	1.25E-06	470 S 0.909	999
73	Rv3144c	PPE52	PPE family protein	1.73E-06	226 S 0.969*	999
74	Rv3511	PE_PGRS55	PE-PGRS family protein	1.93E-06	396 N 0.960*, 589 G 0.995**	999
75	Rv0109	PE_PGRS1	PE-PGRS family protein	2.30E-06	346 G 0.995**	999
76	Rv0284		conserved membrane protein	2.59E-06	214 R 0.975*	999
77	Rv2037c		conserved membrane protein	3.03E-06	312 Y 0.933	999
78	Rv3590c	PE_PGRS58	PE-PGRS family protein	3.41E-06	314 A 0.877	999
79	Rv2101	*helZ*	helicase	3.46E-06	462 L 0.941	336
80	Rv2488c		transcriptional regulator, luxR-family	3.66E-06	265 T 0.964*	999
81	Rv0103c	*ctpB*	cation-transporter P-type ATPase B	3.67E-06	22 S 0.949	999
82	Rv1321		conserved hypothetical protein	3.68E-06	144 R 0.997**	999
83	Rv2741	PE_PGRS47	PE-PGRS family protein	3.84E-06	271 S 0.999**	135
84	Rv1486c		conserved hypothetical protein	4.21E-06	198 N 0.972*	999
85	Rv2611c		acyltransferase	4.21E-06	197 C 0.910	999
86	Rv1320c		adenylate cyclase	4.49E-06	531 A 0.977*	999
87	Rv0018c	*ppp*	serine/threonine phosphatase	4.82E-06	463 S 0.980*	999
88	Rv2770c	PPE44	PPE family protein	5.01E-06	194 F 0.974*	999
89	Rv0881		rRNA methyltransferase	6.64E-06	115 R 0.990*	477

90	Rv2495c	*pdhC*	dihydrolipoamide S-acetyltransferase E2 component	6.98E-06	107 A 0.987*	389
91	Rv2982c	*gpsA*	glycerol-3-phosphate dehydrogenase gpdA2	7.14E-06	133 A 0.985*	374
92	Rv0355c	PPE8	PPE family protein	7.39E-06	2571 A 0.665	999
93	Rv2059		conserved hypothetical protein	7.39E-06	317 T 0.940	999
95	Rv1127c	*ppdK*	pyruvate, phosphate dikinase	8.04E-06	69 E 0.900	999
96	Rv1300	*hemK*	hypothetical protein	9.54E-06	194 C 0.988*	999
97	Rv2290	*lppO*	lipoprotein	1.01E-05	16 A 0.986*	999
98	Rv3329		aminotransferase	1.17E-05	122 H 0.969*	999
99	Rv3919c	*gid*	glucose-inhibited division protein B	1.17E-05	92 E 0.968*, 96 R 0.989*, 145 L 0.989*	170
100	Rv1900c	*lipJ*	lignin peroxidase	1.20E-05	204 M 0.947	999
101	Rv2482c	*plsB2*	glycerol-3-phosphate acyltransferase	1.23E-05	778 R 0.956*	999

Supplementary Table 14: Genes significant by each of phyC, dN/dS, and differential density for streptomycin resistance in increasing order of p-value. In bold and underlined are genes previously associated with resistance to streptomycin.

Phylogenetic Convergence

Number	Rv-number	Symbol	Description	P-value (<0.05) *Site & Gene convergence. ** Site convergence only (p-values ordered by site1, (site2), gene)	Convergent site(s)
1	**Rv0682**	***rpsL***	**30S ribosomal protein S12**	**0.0001, 0.0012***	**128G**
2	Rv0667	*rpoB*	DNA-directed RNA polymerase beta chain	0.0002	
3	Rv2896c		conserved hypothetical protein	0.0021**	457T
4	Rv2931	*ppsA*	phenolpthiocerol synthesis type-I polyketide synthase	0.008	
5	Rv0746	PE_PGRS9	PE-PGRS family protein	0.0327	
6	Rv0218		conserved membrane protein	0.0367**	946T
7	Rv0658c		conserved membrane protein	0.0408**	224T

Differential density

Number	Rv-number	Symbol	Description	P-value (<0.05)	Clustering Index
1	Rv0667	*rpoB*	DNA-directed RNA polymerase beta chain	<0.0001	6
2	**Rv0682**	***rpsL***	**30S ribosomal protein S12**	**0.0597**	**5**
3	Rv3468c		dTDP-glucose-4,6-dehydratase	0.0851	5

dN/dS

Number	Rv-number	Symbol	Description	P-value (<1.25E-5)	Significant site(s) by Bayesian empiric bayes method * <0.05, **<0.01	Class dN/dS ψ (Supplementary Table 7)
1	Rv2931	*ppsA*	phenolpthiocerol synthesis type-I polyketide synthase	3.65E-25	624 D 1.000**, 803 A 0.978*, 1194 L 0.988*	999
2	Rv0746	PE_PGRS9	PE-PGRS family protein	9.79E-23	191 G 0.997**, 280 D 1.000**, 320 A 1.000**, 445 A 1.000**	303
3	Rv0279c	PE_PGRS4	PE-PGRS family protein	2.31E-20	77 V 1.000**	799
4	Rv0278c	PE_PGRS3	PE-PGRS family protein	3.27E-18	807 G 1.000**	999
5	Rv0218		conserved membrane protein	7.26E-17	316 R 1.000**	999
6	Rv0667	*rpoB*	DNA-directed RNA	2.70E-15	435 D 0.999**, 450 L	185

			polymerase beta chain		1.000**, 452 L 0.967*	
7	Rv2828c		conserved hypothetical protein	4.92E-15	128 S 1.000**	999
8	Rv0658c		conserved membrane protein	8.23E-15	75 P 1.000**	999
9	Rv3795	*embB*	membrane indolylacetylinositol arabinosyltransferase	1.30E-14	306 I 1.000**, 354 D 0.995**, 497 Q 0.964*	999
10	Rv3512	PE_PGRS56	PE-PGRS family protein	2.41E-14	253 A 0.952*, 306 I 0.952*	999
11	Rv0747	PE_PGRS10	PE-PGRS family protein	7.42E-14	295 K 0.999**, 300 S 1.000**	261
12	Rv2874	*dipZ*	cytochrome C biogenesis protein	2.23E-13	672 D 0.986*	999
13	Rv0018c	*ppp*	serine/threonine phosphatase	3.36E-13	463 S 1.000**	999
14	Rv2079		conserved hypothetical protein	5.99E-13	47 C 1.000**	999
15	Rv1753c	PPE24	PPE family protein	1.96E-12	488 T 1.000**	999
16	Rv2896c		conserved hypothetical protein	2.01E-12	153 A 1.000**	999
17	Rv0532	PE_PGRS6	PE-PGRS family protein	2.76E-12	239 G 1.000**	999
18	Rv0006	*gyrA*	DNA gyrase subunit A	9.55E-12	90 A 0.999**, 94 A 1.000**, 95 T 0.984*, 292 R 0.983*, 376 R 0.984*, 668 D 1.000**	999
19	Rv3366	*spoU*	tRNA/rRNA methylase	1.11E-11	92 R 1.000**	999
20	Rv1446c	*opcA*	oxpp cycle protein	1.32E-11	192 R 0.998**	999
21	Rv2024c		conserved hypothetical protein	4.18E-11	47 R 0.969*, 154 G 1.000**	790
22	Rv3343c	PPE54	PPE family protein	4.32E-11	103 A 0.928	645
23	Rv1394c	*cyp132*	cytochrome P450 132	9.74E-11	135 L 0.999**	999
24	Rv0050	*ponA1*	bifunctional penicillin-binding protein	1.60E-10	631 P 0.999**	999
25	Rv3468c		dTDP-glucose-4,6-dehydratase	1.85E-10	62 V 1.000**	999
26	Rv0064		conserved membrane protein	2.59E-10	906 R 1.000**	539
27	Rv0758	*phoR*	two component system sensor kinase	2.64E-10	172 L 0.999**	999
28	Rv2917		conserved alanine and arginine rich protein	3.35E-10	594 L 0.999**	999
29	Rv0800	*pepC*	aminopeptidase	9.12E-10	139 L 0.999**	999
30	Rv0058	*dnaB*	replicative DNA helicase	1.10E-09	552 R 0.998**	999
32	Rv3711c	*dnaQ*	DNA polymerase III epsilon subunit	1.20E-09	211 L 0.999**	999
33	Rv1812c		dehydrogenase	1.34E-09	30 P 0.984*	999
34	Rv3835		conserved membrane protein	1.41E-09	294 L 0.977*	999

35	Rv2458	*mmuM*	homocysteine S-methyltransferase	2.07E-09	125 Y 0.998**	999
36	Rv0236c		conserved membrane protein	2.07E-09	1080 G 0.993**	317
37	Rv2947c	*pks15*	polyketide synthase	2.66E-09	333 A 0.999**	999
38	Rv0082		oxidoreductase	2.71E-09	74 R 0.986*	999
39	Rv2450c	*rpfE*	resuscitation-promoting factor	2.82E-09	20 T 0.991**, 126 R 0.997**	999
40	Rv0095c		conserved hypothetical protein	3.34E-09	57 D 1.000**	999
41	Rv3425	PPE57	PPE family protein	4.66E-09	128 T 0.999**	999
42	Rv1704c	*cycA*	D-serine/alanine/glycine transporter protein	1.02E-08	93 L 0.994**	999
43	Rv0109	PE_PGRS1	PE-PGRS family protein	1.14E-08	346 G 0.999**	999
44	Rv2770c	PPE44	PPE family protein	1.69E-08	194 F 0.995**	999
45	Rv0292		conserved membrane protein	1.76E-08	217 D 0.990*	999
46	Rv1378c		conserved hypothetical protein	1.93E-08	37 W 0.994**	999
47	Rv2584c	*apt*	adenine phosphoribosyltransferase	2.24E-08	147 E 0.997**	999
48	Rv1463		ABC transporter ATP-binding protein	2.57E-08	198 E 0.992**	999
49	Rv2090		5-3 exonuclease	4.06E-08	358 F 0.994**	999
50	Rv0881		rRNA methyltransferase	6.05E-08	115 R 0.999**	634
51	Rv3776		conserved hypothetical protein	6.16E-08	329 V 0.997**	611
52	Rv0964c		hypothetical protein	7.26E-08	124 T 0.993**	999
53	Rv1186c		conserved hypothetical protein	1.56E-07	207 A 0.988*	999
54	Rv2439c	*proB*	glutamate 5-kinase protein	1.95E-07	226 S 0.993**	999
55	Rv3151	*nuoG*	NADH dehydrogenase I chain G	2.31E-07	474 M 0.983*	999
56	Rv3630		conserved membrane protein	2.69E-07	40 T 0.995**	999
57	Rv3077		hydrolase	2.83E-07	310 G 0.999**	174
58	Rv0388c	PPE9	PPE family protein	2.92E-07	138 A 1.000**, 139 Q 0.986*	48
59	Rv0280	PPE3	PPE family protein	3.21E-07	337 S 0.999**	999
60	Rv0048c		membrane protein	3.68E-07	248 E 0.992**, 250 V 0.992**	999
61	Rv3341	*metA*	homoserine O-acetyltransferase	4.18E-07	87 S 0.950*	999
62	Rv1232c		conserved hypothetical protein	4.47E-07	149 G 0.989*	999
63	Rv0465c		transcriptional	4.49E-07	106 C 0.994**	999

			regulator			
64	Rv0259c		conserved hypothetical protein	4.66E-07	182 V 0.997**	999
65	Rv2436	*rbsK*	ribokinase	4.88E-07	174 G 0.966*, 282 A 0.999**	999
66	Rv0226c		conserved membrane protein	4.91E-07	379 P 0.968*	999
67	Rv2155c	*murD*	UDP-N-acetylmuramoylalanine -D-glutamate ligase	5.32E-07	247 G 0.995**	999
68	Rv2769c	PE27	PE family protein	5.33E-07	136 A 0.978*, 270 M 0.999**	999
69	Rv2825c		conserved hypothetical protein	6.22E-07	162 S 0.951*	999
70	Rv1915	*aceAa*	isocitrate lyase	6.34E-07	179 D 0.995**	999
71	Rv3764c		two component system sensor kinase	6.51E-07	246 R 0.961*	999
72	Rv1093	*glyA1*	serine hydroxymethyltransfer ase 1	6.71E-07	36 A 0.903	999
73	Rv3084	*lipR*	acetyl-hydrolase/esterase	7.63E-07	none	1
74	Rv3879c		hypothetical alanine and proline rich protein	9.05E-07	729 S 0.983*	472
75	Rv0787		hypothetical protein	1.05E-06	267 H 0.991**	802
76	Rv3777		oxidoreductase	1.20E-06	160 A 0.999**	999
77	Rv1326c	*glgB*	1,4-alpha-glucan branching enzyme	1.24E-06	470 S 0.908	999
78	Rv2495c	*pdhC*	dihydrolipoamide S-acetyltransferase E2 component	1.27E-06	107 A 0.995**	999
79	**Rv0682**	***rpsL***	**30S ribosomal protein S12**	**1.50E-06**	**43 R 0.993**, 88 K 0.985***	**68**
80	Rv3144c	PPE52	PPE family protein	1.73E-06	226 S 0.969*	999
81	Rv3590c	PE_PGRS58	PE-PGRS family protein	1.79E-06	314 A 0.877	999
82	Rv0425c	*ctpH*	metal cation transporting P-type ATPase	2.17E-06	689 V 0.979*	571
83	Rv2433c		hypothetical protein	2.41E-06	26 L 0.998**	999
84	Rv2037c		conserved membrane protein	2.67E-06	312 Y 0.933	999
85	Rv1895		dehydrogenase	2.75E-06	270 L 0.969*	999
86	Rv2741	PE_PGRS47	PE-PGRS family protein	3.43E-06	271 S 0.944	76
87	Rv0355c	PPE8	PPE family protein	3.63E-06	2571 A 0.648	999
88	Rv1321		conserved hypothetical protein	3.68E-06	144 R 0.997**	999
89	Rv0103c	*ctpB*	cation-transporter P-type ATPase B	3.96E-06	22 S 0.949	999
90	Rv3666c	*dppA*	periplasmic dipeptide-	4.13E-06	4 Q 0.801	999

			binding lipoprotein			
91	Rv3365c		conserved hypothetical protein	4.23E-06	38 P 0.989*, 687 S 0.953*	999
92	Rv3514	PE_PGRS57	PE-PGRS family protein	4.24E-06	1462 T 0.780	999
93	Rv1320c		adenylate cyclase	4.50E-06	531 A 0.977*	999
94	Rv2333c		conserved membrane transport protein	4.94E-06	69 Y 0.964*	999
95	Rv1644	*tsnR*	23S rRNA methyltransferase	5.10E-06	232 P 0.968*	999
96	Rv0417	*thiG*	thiamin biosynthesis protein	5.15E-06	75 C 0.991**	537
97	Rv1300	*hemK*	hypothetical protein	5.55E-06	194 C 0.969*	999
98	Rv1127c	*ppdK*	pyruvate, phosphate dikinase	6.39E-06	69 E 0.906	999
99	Rv0159c	PE3	PE family protein	6.98E-06	14 A 0.967*	999
100	Rv0284		conserved membrane protein	7.57E-06	214 R 0.992**	999
101	Rv2290	*lppO*	lipoprotein	8.59E-06	16 A 0.986*	999
102	Rv2160A		conserved hypothetical protein	8.64E-06	155 C 0.956*	605
103	Rv3303c	*lpdA*	dihydrolipoamide dehydrogenase	8.97E-06	308 L 0.976*	999
104	Rv3511	PE_PGRS55	PE-PGRS family protein	9.15E-06	589 G 0.985*	496
105	Rv3347c	PPE55	PPE family protein	1.06E-05	786 V 0.969*	133
106	Rv1193	*fadD36*	fatty-acid-CoA ligase	1.13E-05	124 P 0.963*	999
107	Rv3329		aminotransferase	1.16E-05	122 H 0.969*	999
108	Rv2482c	*plsB2*	glycerol-3-phosphate acyltransferase	1.23E-05	778 R 0.956*	999
109	Rv3497c	*mce4C*	MCE-family protein	1.24E-05	none	2

Supplementary Table 15: Genes significant by each of phyC, dN/dS and differential density for pyrazinamide resistance in increasing order of p-value. In bold and underlined are genes previously associated with resistance to pyrazinamide.

Phylogenetic Convergence

Number	Rv-number	Symbol	Description	P-value (<0.05) *Site & Gene convergence. ** Site convergence only (p-values ordered by site1, (site2), gene)	Convergent site(s)
1	Rv0667	*rpoB*	DNA-directed RNA polymerase beta chain	<0.0001	
2	**Rv2043c**	***pncA***	**pyrazinamidase/nicotinamidas**	**0.0015**	
3	Rv3795	*embB*	membrane indolylacetylinositol arabinosyltransferase	0.0131	
4	Rv1971	*mce3F*	MCE-family protein	0.0345	1187C

Differential density

Number	Rvnumber	Symbol	Description	P-value (<0.05)	Clustering Index
1	**Rv2043c**	***pncA***	**pyrazinamidase/nicotinamidas**	**0.0001**	**5**

dN/dS

Number	Rvnumber	Symbol	Description	P-value (<1.25E-5)	Significant site(s) by Bayesian empiric bayes method * <0.05, **<0.01	Class dN/dS ψ (Supplementary Table 7)
1	Rv0747	PE_PGRS10	PE-PGRS family protein	1.19E-21	295 K 1.000**, 300 S 1.000**	505
2	Rv2931	*ppsA*	phenolpthiocerol synthesis type-I polyketide synthase	3.31E-12	624 D 1.000**	768
3	Rv0746	PE_PGRS9	PE-PGRS family protein	1.24E-11	252 A 0.965*, 280 D 1.000**, 320 A 0.963*	329
4	Rv2584c	*apt*	adenine phosphoribosyltransferase	1.55E-11	147 E 1.000**	999
5	Rv0532	PE_PGRS6	PE-PGRS family protein	1.77E-11	239 G 1.000**	999
6	Rv0278c	PE_PGRS3	PE-PGRS family protein	1.31E-10	807 G 0.998**	999
8	Rv2917		conserved alanine and arginine rich protein	3.55E-10	594 L 0.999**	999
9	Rv1753c	PPE24	PPE family protein	7.46E-10	488 T 0.996**	999
10	Rv0280	PPE3	PPE family protein	1.45E-09	337 S 0.991**	999
11	Rv0388c	PPE9	PPE family protein	3.77E-09	138 A 0.997**, 139 Q 1.000**	101
12	Rv3711c	*dnaQ*	DNA polymerase III epsilon subunit	9.22E-09	211 L 0.993**	999
13	Rv2741	PE_PGRS47	PE-PGRS family protein	1.24E-08	271 S 0.995**	164
14	Rv1378c		conserved hypothetical protein	1.94E-08	37 W 0.994**	999

15	Rv2024c		conserved hypothetical protein	2.16E-08	154 G 1.000**	999
16	Rv0279c	PE_PGRS4	PE-PGRS family protein	5.38E-08	77 V 0.995**, 352 G 0.994**	88
17	Rv2155c	*murD*	UDP-N-acetylmuramoylalanine-D-glutamate ligase	5.70E-08	247 G 0.995**	999
18	Rv0658c		conserved membrane protein	1.33E-07	75 P 0.999**	999
19	Rv0050	*ponA1*	bifunctional penicillin-binding protein	1.37E-07	631 P 0.991**	999
20	Rv3795	*embB*	membrane indolylacetylinositol arabinosyltransferase	1.39E-07	306 I 0.989*, 354 D 0.996**	420
21	Rv3343c	PPE54	PPE family protein	1.43E-07	103 A 0.854	877
22	Rv0058	*dnaB*	replicative DNA helicase	1.68E-07	552 R 0.988*	999
23	Rv0236c		conserved membrane protein	2.44E-07	1080 G 0.960*	334
24	Rv1971	*mce3F*	MCE-family protein	2.65E-07	396 E 0.994**	999
25	Rv0417	*thiG*	thiamin biosynthesis protein	3.62E-07	75 C 0.996**	999
26	Rv0881		rRNA methyltransferase	1.20E-06	115 R 0.996**	999
27	Rv0667	*rpoB*	DNA-directed RNA polymerase beta chain	1.50E-06	435 D 0.994**, 452 L 0.983*	73
28	Rv2450c	*rpfE*	resuscitation-promoting factor	1.81E-06	126 R 0.991**	645
29	Rv0292		conserved membrane protein	2.15E-06	217 D 0.940	999
30	Rv2825c		conserved hypothetical protein	3.30E-06	162 S 0.951*	999
31	Rv0159c	PE3	PE family protein	3.93E-06	14 A 0.994**	999
32	Rv0323c		conserved hypothetical protein	3.96E-06	142 G 0.948	999
33	Rv1704c	*cycA*	D-serine/alanine/glycine transporter protein	4.23E-06	93 L 0.956*	999
34	Rv1895		dehydrogenase	5.38E-06	270 L 0.969*	999
35	Rv1396c	PE_PGRS25	PE-PGRS family protein	8.15E-06	66 S 0.988*	999
36	Rv0193c		hypothetical protein	9.92E-06	417 E 0.960*	999
37	Rv1394c	*cyp132*	cytochrome P450 132	1.13E-05	135 L 0.949	999
38	Rv1900c	*lipJ*	lignin peroxidase	1.20E-05	204 M 0.947	999
39	Rv0758	*phoR*	two component system sensor kinase	1.23E-05	172 L 0.963*	999
40	Rv3458c	*rpsD*	30S ribosomal protein S4	1.23E-05	none	4

Supplementary Table 16: Non-synonmous or non-coding SNPs relative to the phylogenetic ancestor in TIMs present in resistant isolates.

Rvnumber / intergenic region no	gene symbol*	SNP position and base	number of unique sites
Rv0667	*rpoB*	1226-G 1268-C 1289-C 1303-T 1304-G 1304-T 1333-T 1334-T 1349-C 1349-T 1355-C 1447-T 1471-T 1684-A 2192-C 509-C	14
Rv0006	*gyrA*	1127-T 2003-G 269-T 281-C 281-G 284-G 61-G 874-G	7
Rv1908c	*katG*	1388-G 440-G 610-C 944-C 944-G 945-A	5
Rv3919c	*gid*	142-T 149-G 236-C 276-A 276-C 287-T 299-C 435-G 435-T 47-T	8
Rv3854c	*ethA*	1141-C 127-A 149-G 227-A 409-C 452-C 616-T 750-G 760-T 844-G 904-G	11
Rv0682	*rpsL*	128-A 128-G 263-C	2
1183	*inhA* promoter	125-T 127-T 134-A 1183-G 1183-A 1183-G	4
Rvnr01	*rrs*	1401-A 1401-G 499-T 514-C 517-T 906-G	5
Rv3795	*embB*	1040-T 1061-C 1217-A 1217-C 1490-G 1807-A 916-A 916-G 918-A 918-G 983-G	8
Rv2043c	*pncA*	152-G 170-G 172-C 188-C 30-C 309-G 357-A 388-A 394-A 395-C 403-C 502-C 512-T	13
1899	*eis* promoter	11-G 11-T 15-A 38-A	3
Rv0050	*ponA1*	1095-T 123-G 1891-C 1891-T	3
Rv0218		946-C 946-T	1
Rv2650c		302-C 302-T 646-G 8-A	3
Rv3093c		630-C 630-G	1
Rv1446c	*opcA*	112-A 575-C 575-G	2
Rv3245c	*mtrB*	1107-C 1549-A 1549-C 52-C 52-T 521-C	4
Rv0658c		125-A 224-C 224-T	2
Rv2048c	*pks12*	11302-G 2749-C 2919-T 4954-C 4954-T 6441-A 7791-C 9011-C 9011-T 9848-G	8
Rv2436	*rbsK*	520-A 845-C 845-T	2
Rv2931	*ppsA*	1872-C 1872-G 2407-A 2407-G 2630-A 2746-A 2864-C 3581-T 3967-A	7
922	pre-*Rv1148c*	922-G 408-C 408-T 683-A	3
Rv1180	*pks3*	1467-A	1
Rv3711c	*dnaQ*	227-A 227-G 631-G	2
1058	pre-Rvnr01*(rrs)*	107-A 83-C	2
1292	pre-Rv1632c	51-C 51-G 1292-G	2
2867	pre-Rv3681c*(whiB4)*	158-A 230-T 29-A 29-G	3
Rv2082		103-A 103-G 1127-C 1396-C 1913-C 1913-G 286-A 286-G 547-A 553-C 65-A 65-G 817-G 817-T	9
Rv2155c	*murD*	239-C 739-C 739-G 892-A 892-G	3
Rv0064		1252-T 1370-G 1648-G 1681-G 2197-A 2305-T 2717-C 2717-G 719-T	8
Rv1319c		1165-T 1317-G 1317-T 1369-C 317-A	4

Rv0668	*rpoC*	1211-C 1450-G 2093-G 2094-A 2250-C 3098-C 3119-G 3276-C 3691-T	9
Rv2896c		457-G 457-T	1
Rv3446c		1184-G 718-G 718-T 851-C 851-G	3
Rv2024c		139-C 139-T 461-A 461-G	2
Rv0280	PPE3	1009-C 1009-T 769-G	2
Rv3478	PPE60	1111-C 1160-C 1160-T 307-G 359-T 437-C	5
Rv3345c	PE_PGRS50	1417-A 2399-C 380-C 4090-A 4090-C 4099-A 4223-C 638-G	7
Rv3021c	PPE47	1021-A 1021-G 371-G 665-C 665-G	3
Rv0532	PE_PGRS6	229-T 680-A 680-G 716-C 716-G	3
Rv0279c	PE_PGRS4	1054-G 1054-T 1115-G 1115-T 229-A 229-G 230-GT 355-G 974-A 974-G	6
Rv0746	PE_PGRS9	1333-A 1333-G 2228-G 353-C 572-A 572-G 754-A 754-G 838-A 838-G 958-A 958-G	7
Rv0388c	PPE9	17-C 364-A 364-C 412-A 412-G 417-G 438-G 456-G	6
Rv0278c	PE_PGRS3	1012-C 120-G 1328-A 2419-C 2419-G	4
Rv2741	PE_PGRS47	752-A 811-A 811-G	2
Rv0747	PE_PGRS10	1234-A 1729-G 673-A 673-G 679-C 679-G 884-A 884-G 898-A 898-G	6
Rv3507	PE_PGRS53	1010-C 2530-G 3908-A 3908-C 400-T 416-T	5
Rv2853	PE_PGRS48	1816-A 365-A 538-C 538-G	3
Rv3343c	PPE54	1123-C 1396-G 1855-G 2747-A 308-A 308-C 4075-G 4592-T 6320-C 6320-T 6542-T	9
Rv3347c	PPE55	1928-G 2357-C 2764-G 2768-A 2768-T 2975-C 364-A 364-G 5297-G 6776-T 9090-G	9
Rv0109	PE_PGRS1	1036-C 1036-G	1
		known dr=	80
		PPE=	78
		other=	85
		Total=	**243**

*** genes highlighted in yellow represent PE/PPE family genes, in green are regions known to be associated with drug resistance in TB, not highlighted are other TIMs.**

Supplementary Table 18: Culture, Drug Sensitivity Testing, Fingerprinting.

Isolate Source	Culture	Drug Sensitivity Testing*	Fingerprinting (*1, 2*)
Stellenbosch University, South Africa	BACTEC MGIT 960 system (BD Diagnostics Systems, Sparks, MD)	Indirect proportion method on Middlebrook 7H11 agar slants supplemented individually with: RIF (1.0 µg/ml), INH (0.2 µg/ml), EMB (7.5 µg/ml), OFLX (2.0 µg/ml), KAN (5.0 µg/ml), STR (2.0 µg/ml), AMK (5.0 µg/ml), CAP (10 µg/ml). PZA sensitivity was tested using the MGIT system (100 µg/ml).	Spoligotyping/ RFLP
Center for Disease Control, Atlanta, USA	Middlebrook 7H9 broth supplemented with 10% (vol/vol) albumin-dextrose-catalase enrichment (Difco Laboratories) and 0.05% (vol/vol) Tween 80 (Sigma-Aldrich) at 37°C until they reached an approximate optical density at 600 nm of 1.0 (corresponding to 5 x 10^8 CFU/ml)	Indirect proportion method on Middlebrook 7H10 agar slants supplemented individually with: RIF (1 µg/ml), INH (0.2, 1, and 5µg/ml), EMB (5 µg/ml), OFLX (2 µg/ml), CIP (2 µg/ml), KAN (5 µg/ml), CAP (10µg/ml), and AMI (4 µg/ml). PZA was tested using the BACTEC 460 (100 µg/ml), MGIT (100 µg/ml), or agar proportion (25 µg/ml) method.	Spoligotyping
University of San Francisco, USA	Lowenstein-Jensen slant culture, 7H11, 7H11 selective and BACTEC 460	BACTEC MGIT 960 AST system	RFLP IS6110 and PGRS
British Columbia Center for Disease Control, Vancouver, Canada	Lowenstein-Jensen slant culture and BACTEC MGIT 960 system (BD Diagnostics Systems, Sparks, MD)	BACTEC MGIT 960 AST system. INH (0.1 ug/ml) STR (1.0 ug/ml) EMB (5.0 ug/ml) RIF (1.0ug/ml)	MIRU-VNTR, RFLP
Massachusetts State Laboratory (source country Peru, Russia)	Radiometric BACTEC 460 TB system (Becton-Dickinson)	Indirect proportion method on Middlebrook 7H10 agar plates supplemented with : INH (0.2, 1, and 5 µg/mL), RIF (1 µg/mL), EMB (5 µg/mL), STR (2 and 10 µg/mL), KAN (5 µg/mL), CAP (10 µg/mL), ETH (5 µg/mL), CYS (30 µg/mL), PAS (1 µg/mL), AMK (6 µg/mL), LEVO (1 µg/mL), OFLX (2 µg/mL), and CIP (2 µg/mL). PZA was tested using the the BACTEC (100 µg/mL).	Spoligotype
Universita di Siena, Italy	Lowenstein-Jensen medium and by the radiometric BACTEC 460 TB system (Becton-Dickinson).	MICs were determined on Middlebrook 7H11 agar (Difco). Plates containing different drug concentrations were inoculated with approximately 2 x10^2 and 2 x10^3 CFU by a semiautomated inoculator (Multipoint Inoculator A400; Denley) and incubated at	RFLP

		37°C for 21 days. The following drugs were tested: INH, RIF, EMB, STR, RFB, CIP, OFLX, sparfloxacin, AMI, KAN, CAP, CS, PAS, thiacetazone, viomycin, and ETH. The MIC was defined as the lowest drug concentration inhibiting >99% of the inoculum. Susceptibility to PZA was determined by the BACTEC 460 TB method.	
Public Health Reseach Institute, UMDNJ, Newark, NJ	Lowenstein-Jensen slant culture	Indirect agar proportion method using Middlebrook 7H10 agar plates containing the following drugs: RIF (1 μg/ml), INH (0.2, 1, and 5μg/ml), EMB (5 μg/ml), CIP (2 μg/ml), KAN (5 μg/ml), CAP (10μg/ml)	RFLP

***Drug abbreviations explained in Supplementary Table 1.**

Supplementary Table 19: F_{ST} values for the 14 epiclusters. Distance measured by pairwise difference. Epicluster names and F_{ST} values underlined are those selected for pairwise convergence analysis.

	Italian	K	Peru1	BC	SF	CDC1	CDC2	KZN	Mine	Peru2	SA1	SA2	Russia1	Russia2
Italian	0	0.96	0.84	0.93	0.87	0.92	0.94	0.94	0.87	0.93	0.93	0.95	0.90	0.92
K	0.96	0	0.82	0.95	0.86	0.94	0.91	0.93	0.87	0.41	0.54	0.93	0.84	0.91
Peru1	0.84	0.82	0	0.87	0.40	0.72	0.69	0.77	0.67	0.75	0.80	0.69	0.66	0.73
BC	0.93	0.95	0.87	0	0.87	0.93	0.88	0.93	0.86	0.94	0.94	0.89	0.91	0.92
SF	0.87	0.86	0.40	0.87	0	0.81	0.74	0.82	0.71	0.81	0.84	0.76	0.75	0.80
CDC1	0.92	0.94	0.72	0.93	0.81	0	0.91	0.93	0.81	0.82	0.87	0.96	0.78	0.88
CDC2	0.94	0.91	0.69	0.88	0.74	0.91	0	0.90	0.52	0.83	0.87	0.76	0.80	0.87
KZN	0.94	0.93	0.77	0.93	0.82	0.93	0.90	0	0.82	0.85	0.89	0.93	0.71	0.72
Mine	0.87	0.87	0.67	0.86	0.71	0.81	0.52	0.82	0	0.82	0.84	0.58	0.76	0.80
Peru2	0.93	0.41	0.75	0.94	0.81	0.82	0.83	0.85	0.82	0	0.36	0.82	0.71	0.82
SA1	0.93	0.54	0.80	0.94	0.84	0.87	0.87	0.89	0.84	0.36	0	0.87	0.80	0.86
SA2	0.95	0.93	0.69	0.89	0.76	0.96	0.76	0.93	0.58	0.82	0.87	0	0.81	0.89
Russia1	0.90	0.84	0.66	0.91	0.75	0.78	0.80	0.71	0.76	0.71	0.80	0.81	0	0.62
Russia2	0.92	0.91	0.73	0.92	0.80	0.88	0.87	0.72	0.80	0.82	0.86	0.89	0.62	0

Supplementary Table 20: Pairwise convergence analysis: Epicluster pairs showing progressive resistance.

Sensitive isolate (Id)	Resistant isolate (Id)	SNP cluster group ***(41)***	New resistance to*	SNPs total	SNPs non coding	Variant Genes	Average # SNP/Gene
P1_ITL (6)	P10_ITL (13)	6a	INH RIF RFB PZA STR	39	4	32	1.1
KZN_DS (14)	KZN_XDR (16)	5	INH RIF EMB STR CAP KAN OFLX	23	6	15	1.1
M73 (29)	M95 (32)	4	ETH EMB ?PZA ?RIF STR ?OFLX THA	43	7	36	1.0
03R1382 (24)	02R1848 (22)	3b	INH ETH PZA STR KAN PAS	89	7	79	1.0
R431 (36)	R179 (37)	2	ETH EMB THA	90	9	63	1.3
K-1 (140)	K-2 (141)	2	INH RIF EMB PZA STR	76	9	66	1.0
CDC609 (49)	CDC601 (41)	6b	KAN	15	0	14	1.1
CDC610 (50)	CDC602 (42)	4	KAN	4	1	3	1.0

***Drug abbreviations explained in Supplementary Table 1. ? =possible because one of the two pairs of isolates was not tested for resistance to the drug listed.**

Supplementary Table 21: K-S p-values for the differential density test.

Resistance to*	K-S p-value
Any drug	4.9×10^{-46}
INH	3.6×10^{-32}
RIF	2.0×10^{-32}
EMB	1.7×10^{-9}
STR	3.6×10^{-17}
AG	0.51
PZA	1
FLQ	1

***Drug abbreviations explained in Supplementary Table 1.**

Supplementary Note:

Culture, Drug Sensitivity Testing, Fingerprinting:

Isolates culture, drug sensitivity testing and molecular fingerprint methods are detailed in Supplementary Table 18. All isolates were grown in Middlebrook 7H9 medium (Difco) and stored at -80°C before DNA extraction. Restriction fragment length polymorphism analysis was performed based on the insertion sequence *IS6110*. Spoligotyping and MIRU-VNTR -24 was performed using established methods[1,2].

Sequencing:

DNA was extracted by Qiagen method with minor modifications.

As the isolate collection and sequencing spanned a duration of 3-4 years, the sequencing methodology varied slightly for each isolate set. Isolates were sequenced by the Illumina GAIIx. Most isolates received one lane of 35 basepair (bp) single end reads. The Italian isolates (P2_ITL, P4_ITL, P5_ITL, and P10_ITL) and isolates with identification numbers 51-73 each received one lane of 50 bp single end reads. Isolates from Vancouver were sequenced using one lane of 50 bp paired reads. Isolate CDC_603 received one lane of 75 bp reads. Isolates P3_ITL, P6_ITL, P8_ITL, R179, R257, R451, and R439 were sequenced with one lane of 35 bp reads and resequenced on another lane of 75 bp reads to improve coverage.

For isolates Haarlem, C, W_148 and 98_r604 two whole genome shotgun plasmid libraries with 4kb and 10kb inserts were constructed for each isolate from genomic DNA. For W_148 an additional fosmid library was created. Sequences were produced with Sanger technology and assembled using Arachne[3]. A draft assembly was generated for Mycobacterium tuberculosis C, W_148, and 98_r604 and a finished assembly for Mycobacterium tuberculosis Haarlem. The total coverage was 6.8 fold for isolate C, 15.1 fold for Haarlem, 6.5 fold for 98_r604 and 10.5 fold for W_148.

The raw sequence or genomic sequence data was not publically available for isolates K1 and K2 (ID 140,141), instead the substitutions (single nucleotide polymorphisms, SNPs) published by Niemann *et al.*[4] were used directly.

Alignment and SNP calling:

Sequence reads were aligned to the reference genome sequence for H37Rv using MAQ version 0.6.6[5]. Where relevant each read of a read pair was aligned independently. Reads that aligned with more than 3 mismatches in the first 24bp or that aligned to multiple locations were discarded. Sequence depth at each base was calculated as the number of unique reads aligned over each position. Supplementary Figure 4 displays the frequency histogram of average depth per isolate. Reference coverage was calculated as the percent of H37Rv bases covered by 20 reads or more, and displayed per isolate as a frequency histogram in Supplementary Figure 5. SNPs were called relative to H37Rv with the stringent minimum depth of 20 fold, and consensus quality of 20. The required maximum mapping quality of reads covering the SNP was set at 50.

SNPs that were within 5bp of an indel (insertion or deletion) or did not have an adjacent consensus quality of 20 were also discarded.

Whole-genome alignments of the completed genomes or contigs were generated using MUMMER version 3.22[6] and the H37Rv reference sequence to identify maximal stretches of perfectly matching regions and selecting optimal order-preserving assignments of matches between the genomes based on the longest increasing subsequence algorithm. SNPs were called relative to H37Rv using this alignment. SNPs occurring within 10bp of each other were excluded to minimize false positive calls.

F_{ST} analysis:

Fixation index (F_{ST}) was calculated as a measure of genetic differentiation between the 14 epiclusters using the standard formula in Arlequin v 3.5.1.2[7]. Pairwise difference was used as the distance method, and results were permuted 100 times to obtain significance values. Overall the level of differentiation was high between epiclusters (Figure 1C, Supplementary Table 19). Of the 8 epiclusters used in the pairwise convergence analysis only 2 pairs of epiclusters Mine-CDC2 and K-SA1 showed borderline differentiation with F_{ST} of 0.520 (still significantly high differentiation; Permutation test, $p < 0.05$), and 0.535 (not significant) respectively; otherwise F_{ST} values between these epiclusters were all greater than 0.667.

Intergenic regions definition:

Intergenic regions were defined as a contiguous set of 1 or more bases not annotated to code for protein on either strand according to the TubercuList annotation of the reference isolate H37Rv[8]. They were numbered sequentially along the circular MTB genome starting from the intergenic region between genes Rv0001 and Rv0002.

Unexplained drug resistance:

To determine whether any of the candidate selected genes underlie previously unexplained drug resistance, we first performed deep targeted resequencing of the known resistance determinant regions using molecular inversion probes (MIP) to confirm the absence of known drug resistance mutations. There were 16 isolates (of the 44 resistant isolates WGS sequenced as part of this project) that appeared to have resistance unexplained by mutations in known resistance genes by whole genome sequencing (WGS). Thirty five resistant isolates (15 of the above 16 isolates and an additional 19 control resistant isolates) underwent MIP targeted resequencing. One isolate appeared to have unexplained resistance (03R0988 to kanamycin) from WGS but could not undergo MIP resequencing because of technical reasons, this isolate was excluded from this analysis. The average depth of MIP sequencing of the resistance regions was 2500 fold (range 100-38000). Among the 28 individual drug resistance phenotypes unexplained by WGS mutations, 8 were further explained by MIP sequencing through the capture of a total of 7 SNPs and 1 insertion in the known resistance genes. Overall of a total of 649 SNPs captured by MIP sequencing 64 were missed by WGS. A subsequent examination of the whole genome sequence results revealed that all missed SNPs had been detected but filtered out due to low read depth (<20 fold) at their respective sites. This is an expected advantage of deep sequencing over standard WGS. However there were no SNPs captured by WGS and not by MIPs, consistent with a very low false positive rate using the rigorous depth threshold of ≥ 20 fold. The genomic regions assessed and the number of isolates with verified unexplained resistance are both detailed in Supplementary Table 1. Of note, a relatively high proportion of fluoroquinolone

resistant strains (2/3 to ciprofloxacin, and 1/6 to ofloxacin) had phenotypic resistance but did not carry any *gyrA/B* resistance mutations. This may have at least partly resulted from our strain selection that over-represented strains without known mutations to explain drug resistance.

Overall, the conservative criteria for SNP calling likely led to some missed mutations that is likely to differentially affect strains sequenced with lower depth, because an important criterion for SNP calling at a site was for there to be >20 reads carrying a variant base at that site. The average depth of sequencing reads per strain averaged over all resistant strains was 57 fold, and the depth averaged over all sensitive strains was 68 fold (the difference is borderline statistically significant with p-value of 0.05 by the t-test). Thus we expect on average to miss slightly more SNPs in the resistant strains vs. the sensitive strains or at least miss SNPs at random from both pools of strains. Coupled with the fact that we had less WGS sequenced resistant (42 strains) than sensitive strains (74), our power for detecting SNPs relevant to drug resistance may have been weakened by this non-zero false negative SNP rate. However despite this we were able to find significant genes and SNP as associated with drug resistance. Furthermore our permutation test for significance was a conservative test that identified outlier SNPs and genes by generating the null distribution of SNP/genes using permutation of the resistance labels, and should have controlled for random sequencing error.

Supplementary Tests for Selection:

1. dN/dS:

To identify genes with a significantly high dN/dS in the resistant branches of the phylogeny, we fit and compared three maximum likelihood models using the software package PAML v4.4[9]. We fit three likelihood models in parallel for each of MTB's 3998 protein coding genes. Two genes that contained an in-frame stop codon as part of the H37Rv sequence (Rv1792 and Rv3128c) were analyzed excluding this codon. We excluded the pseudogene Rv2427Ac from the analysis. The first model calculated the likelihood of the sequence data without positive selection. All tree branches were treated the same, with codons classified into four classes of purifying selection (dN/dS < 1) or neutral evolution (dN/dS = 1). The second model allowed for relaxed but not fully positive selection to varying levels on four codon classes in the resistant branches of the tree, and four codon classes of purifying selection in the sensitive branches (branch-site models). The last allowed for full positive selection (dN/dS > 1) on one codon class in the resistant branches of the tree only. This paralleled the recommended models by Zhang et al. 2005[9], the parameters used to fit the branch-site models paralleled those in the lysozyme example of the PAML software unless otherwise described above. A gene was considered positively selected if the model allowing for full positive selection was significantly more likely than the other two models by the likelihood ratio test with significance thresholds as recommended in Zhang et al. 2005[9] adjusted by the Bonferroni correction. Supplementary Table 2 lists all genes found to be under positive selection using the dN/dS method.

2. Pairwise convergence:

Two types of convergence tests were performed: the first was based on resistant/sensitive isolate pairs chosen from isolates groups that shared a molecular fingerprint (epiclusters) and were isolated from the same epidemiological outbreak setting or patient. This test was independent of the isolate phylogeny. Eight epiclusters had a maximum within cluster SNP distance of 220 SNPs (98% percentile for within epicluster SNP distance). These 8 epiclusters each had a single

common ancestor in the SNP based phylogeny that was distinct from the ancestor for any of the other 7 epiclusters included. Within each such epicluster we chose the single most extensively resistant isolate and the most sensitive isolate as a resistant/sensitive (R/S) pair (Supplementary Table 20). We then identified R-specific SNPs, each defined as a nucleotide carried by the R isolate that differed from the nucleotide at the same site in the S isolate and in the H37Rv reference sequence (Supplementary Figure 1). As a measure of convergence among resistant isolates from different lineages, we counted R-specific SNPs that occurred in two or more epiclusters (Supplementary Table 2). Twelve such convergent SNPs were observed across the genome, whereas 0.02 would have been expected by chance (significant excess convergence; p =4e-8), assuming a Poisson distribution for the number of convergent mutations per site in R isolates, with rate parameter set as the total number of R-specific SNPs (summed over all R isolates) per site. We also identified genes that had two or more R specific SNP in any location along the gene length. Twenty six such genes were found whereas 10 would have been expected by chance assuming a Poisson distribution of R-specific SNP-containing genes across the genome (for genes containing 2 R-specific SNPs, $p = 0.064$) (Supplementary Table 3).

The SNP cluster group (SCG) was determined for each epicluster based on the presence or absence of 9 previously described SNPs for lineage classification[10]. The 8 epiclusters belonged to 6 different SCGs. The K and SA epiclusters belonged to the same SCG. This was also the case for the Mine-CDC2 epiclusters (Supplementary Table 20). This observation is in congruence with the borderline F_{ST} values of about 0.5 described above for these two pairs of epiclusters. However each of these epiclusters had a distinct ancestor in the phylogeny, an inclusion requirement for the pairwise convergence analysis. Additionally R-specific SNPs shared by epiclusters belonging to the same SCG were not overrepresented (Supplementary Tables 2 & 3).

3. The density of resistance-specific SNPs:

We assembled a data set of all non-synonymous or intergenic SNPs inferred in leaf branches (extant isolates) of the phylogeny, and divided these substitutions into separate pools for resistant and sensitive branches. We compared the distribution of distances between the SNPs in each pool using the Kolmogorov-Smirnov test, and found that resistance-associated SNPs tended to cluster significantly closer together in the genome than sensitive-associated SNPs (Supplementary Table 21). To identify which regions of the genome were contributing to the difference in SNP distribution between the sensitive and resistance pools we defined the difference in the number of resistant and sensitive substitutions as a resistant SNP 'clustering index' for each gene or intergenic region. We calculated significance values by resampling (10,000 times) to determine the empirical null distribution for a region having ≥x resistant SNPs and ≤y sensitive SNPs, similar to our approach in the convergence methodology above. The empirical significance values were likely conservative, as two benchmark resistance genes, *gyrA* and *gid*, did not meet the significance threshold of $p<0.05$ (Supplementary Table 6). Those two genes nevertheless showed a clustering index of 5 (99.5th percentile for clustering index). We therefore reasoned that a cutoff of clustering index ≥5 would provide more power to identify genes that were of high relevance to resistance.

As the pool of resistant isolates was of different size than the pool of sensitive isolates (47 vs 76 isolates), we performed a sensitivity analysis to confirm that this difference did not affect our analysis results. We sampled the most resistant isolate from each epicluster and non-clustered isolates into a pool of 20 resistant isolates. We similarly sampled sensitive isolates from each epicluster and at random from non-clustered isolates to make up a pool of 20 sensitive isolates.

We compared the distributions of distances between resistance- and sensitive-associated SNPs, using the Kolmogorov-Smirnov test as described above, and repeatedly obtained a highly significant *p*-value (order of magnitude 10^{-45}-10^{-50}). This supports that the tight clustering of resistance-associated SNPs in the genome is not merely due to oversampling relative to sensitive isolates.

We only used the SNPs inferred in the terminal/leaf branches of the phylogeny to control for lineage specific SNPs that reflect neutral evolution or evolution unrelated to drug resistance. Most of the resistance seen in our phylogeny arises in the terminal branches, whereas most of lineage-determining SNPs are inferred deeper in the phylogeny. Some resistance does arise in the 2-3 branches prior to the leaf, and thus excluding these does decrease our method's sensitivity to some extent. We accepted this loss in sensitivity in order to keep our method conservative. Supplementary Figure 6 shows the number of distinct resistant isolate 'genotypes' that SNPs in regions with a clustering index of ≥5. Each epicluster was considered to have a distinct genotype, additionally each isolate isolated from non-outbreak settings or one that had a unique molecular fingerprint was considered to have a distinct genotype. Supplementary Figure 6 demonstrates that isolates accruing these resistance-associated SNPs were diverse (belonged to 3 or more different genotypes) for each region of high clustering index. If lineage-determining SNPs biased this analysis we would have expected that some regions of high clustering index derived all or most of their SNPs from isolates with the same genotype, and this was not the case. We repeated this analysis defining genotype as SNP cluster categories (9 possible categories). This showed similar results with isolates contributing SNPs belonging to a median of 5 SCGs with interquartile range of 4-6 SCGs.

4. PhyC analysis including synonymous sites.

A secondary PhyC analysis was performed, this time including synonymous sites as well as nonsynonymous and intergenic sites. Significance levels were again obtained using a permutation based test as described above but including all mutations including at synonymous sites. In total, 8 of the TIMs associated with drug resistance ($p < 0.05$), were synonymous (Supplementary Table 9).

Functional annotation:

Enrichment testing was performed using the gene functional category definitions for H37Rv[8], and using COGs[11]. Standard Fisher exact tests and EASE scores were calculated as defined by DAVID bioinformatics resources[12]. All three candidate gene sets (identified by convergence, differential density, or dN/dS) were significantly enriched in the PE/PPE gene family of H37Rv genes with EASE scores: convergence 1.64×10^{-10}, dN/dS 2.25×10^{-10}, and differential density 0.00491. The PE/PPE gene/protein group share sequence motifs and amino-acid composition but are not technically grouped by protein function. In fact many members of this gene family are of unknown function[8]. The differential density regions were borderline enriched for the cell membrane, cell wall/envelope COG category (Fisher exact test, $p = 0.046$, *EASE* $p = 0.1822$), otherwise the gene sets were not significantly enriched in other functional or COGs.

We assessed the potential significance of each of the TIMs through a systematic search for an association with known drug resistance genes, drug efflux pumps, genes involved in cell wall biogenesis or remodeling, genes that affect intrinsic drug resistance in MTB or non-tuberculous mycobacteria, and genes involved in DNA repair, replication or recombination.

To identify genes closely associated with DR genes, we first determined whether selected regions were known promoters of known DR genes/loci. We searched TB Database[13] to determine whether the expression of any TIMs was associated with exposure to rifampin, INH, ethambutol, pyrazinamide, streptomycin, amikacin, kanamycin, or a fluoroquinolone. We considered an association significant if the z-score was greater or equal to 3. Next, we used the String 9.0 database of known and predicted protein-protein interactions[14] to determine interactors of each gene, using a confidence score of .4 and assessing no more than 50 interactors. Lastly, we looked to see if known DR genes were co-expressed with each selected candidate, using an arbitrary cut-off of .6 as evidence of moderate co-expression. We classified selected genes as strongly associated with drug resistance genes if they met at least two of these criteria. Only two loci met these criteria: the *rrs* promoter and *rpoC*.

Next, we cross-matched our convergence results against a list of known and putative drug efflux pumps[15]. Although no efflux pumps were identified among genes that met our statistical criteria for TIMs, we found that several efflux pumps, including the ABC transporters Rv0194 and Rv1463, have a larger number of independent mutations in resistant strains relative to sensitive strain. In addition, 2 unrelated XDR isolates acquired a mutation in the putative efflux pump, Rv3806, in the 'pairwise' convergence analysis.

We then conducted a systematic search to determine whether our TIMs played a role in cell wall biogenesis, remodeling or intrinsic drug resistance. Using the following search terms: tuberculosis, cell wall, permeability and intrinsic drug resistance, we searched for each gene and each search term. Publication titles and abstracts were scanned to identify papers that assessed the role of each gene in either MTB or non-tuberculous mycobacteria and experiments that document loss of cell wall permeability to antibiotics upon deletion of each gene were recorded. We identified 5 genes that contribute to cell wall structure: *ppsA*, *pks3*, *pks12*, *ponA1* and *murD*. Three of these were also found to be associated with loss of intrinsic antibiotic resistance in either MTB or a non-tuberculous mycobacteria. In addition, *mtrB* was shown to affect intrinsic antibiotic resistance in MAI.

Next, we matched our genes against a list of known and putative genes involved in DNA repair, replication and recombination in MTB[16]. We also matched the list against genes identified in screens for *lexA* promoter binding motifs[17] and for RecA independent binding motifs[18]. One gene, *dnaQ*, was identified as a known DNA repair gene, one was identified in the screen for *recA* independent motifs (Rv2024) and another was a "soft" hit for a *lexA* binding site (*rbsK*).

Network construction:

In order to further identify interactions between TIMs and other genes known to be related to drug resistance we constructed protein-protein interaction networks for each of the TIMs. Two global protein-protein interaction networks of TB have been published. The first deduced interactions using bacterial two hybrid analysis and was verified to have more than 60% success rate[19]. The second deduced them based on agreement of two or more of the Rosetta stone, Phylogenetic profile, Conserved Gene Neighbor, or Operon methods[20]. We used linkages from either network to construct local networks for each gene TIM (example networks displayed in Supplementary Figure 7) using Cytoscape v2.8.2[21]. We examined first and second neighbors for each gene of interest. In addition to protein-protein interactions shows in Supplementary Figure 7, proteins PonA1 and MurD were third neighbors connected though both PbpA and PonA2.

PE/PPE/diversity analysis:

The PE/PPE gene family, characterized by unique sequence motifs and a specific amino-acid composition, was significantly overrepresented among TIMs (enrichment EASE score for convergence hits 1.64 $x10^{-10}$). This family has not been previously associated with drug resistance or compensation. Consistent with previous studies[22], we observe extremely high diversity in this gene family. To obtain a measure of diversity in PE/PPE genes, and compare it to other genes, we counted the number of SNPs inferred in the terminal branches of the phylogeny, per gene or intergenic region. We pooled this count for all 123 *M. tuberculosis* isolates, and excluded the root isolate *M. canetti*. Genome-wide, the diversity count had a median and interquartile range of 0 (0-2 SNPs/region). The PE/PPE family of genes as whole was more diverse with a SNP count median of 3 (IQR 1-8/gene). The Wilcoxon rank sum test showed this higher SNP density in PE/PPE genes to be significant, with a *p* value of $2x10^{-16}$. The subset of the PE/PPE genes that were TIMs showed more extreme levels of diversity with a median of 43 (IQR 16-52.5). The known drug resistance genes also showed high levels of diversity with a median of 14.0 (IQR 10.5-19.5), but not to the level of the PE/PPE genes. Furthermore the non-PE/PPE subset of candidate TIMs showed similar levels of diversity as the known drug resistant genes with a median of 13.5 (IQR 11-15.56 SNPs/gene; not significantly different from the known resistance genes by the Wilcoxon rank sum test).

Such high diversity makes it difficult to distinguish between positive selection or a combination of relaxed purifying selection on PE/PPE genes and resistant isolate-specific population bottlenecks. If only a small fraction of PE/PPE diversity is fixed after population bottlenecks in independent resistant lineages, this would cause spurious signals of selection in these genes (high dN/dS, frequent convergence, and a high density of mutations) in resistant isolates. At some point in the last ~70 years, since the advent of antibiotics for TB, MTB isolates have likely been subjected to severe bottlenecks due to drug treatment. Resistant isolates would thus be descended from the survivors of these bottlenecks, 'scarred' with many substitutions in PE/PPE genes, even in the absence of positive selection. Distinguishing between positive selection and bottlenecks as causes of resistance-associated substitutions in PE/PPE genes provides a challenge for future research.

SNP calling in repetitive regions

The stringent MAQ mapping qualities required to call SNPs were designed to prevent false-positive SNP calling in repetitive genomic regions, prone to mismatching. To further guard against mis-mapping, we defined the "36merRedundancy" score as the number of times a unique 36 bp sequence is observed in the H37Rv reference genome. We picked a 36bp length because this was the minimum sequence read length for any strain in our data set. Supplementary Figure 8 shows the 36merRedundancy score overlaid with a known repetitive region of the genome (PPE60). In regions with high 36merRedundancy scores, no reads are mapped and therefore no SNPs are called. Yet in the non-repetitive region within the PPE60 gene, there was low 36merRedundancy and good quality read mapping, providing confidence in the SNP call. This example is representative of other repetitive regions of the genome.

The example described above suggests that mis-mapping is not likely to affect the quality of our SNP calls. Nevertheless, we repeated our analyses after excluding SNPs called at the borders of repetitive regions. At the borders of repetitive sequences, uniquely mapped reads (to the

neighboring unique sequence) can provide data on the sequence within repetitive regions. Of the 924 SNP sites in all 50 convergent regions (11 known, 39 candidate), only 39 SNP sites occurred in these border regions (as measured by the 36mer-redundancy mentioned above). We repeated the PhyC test excluding these 39 SNPs, and our results were minimally different: all the genes and regions of known function are equally significant. The only two regions that no longer achieve significance with exclusion of these sites are intergenic region 2867 (Intergenic area between Rv3680 and transcriptional regulator whib-like) and PPE47. All other SNPs (924-39=885 SNPs) map to unique sequences in the reference genome and thus remain the same as our original results.

References:

1. Kamerbeek, J. *et al.* Simultaneous detection and strain differentiation of Mycobacterium tuberculosis for diagnosis and epidemiology. *J. Clin. Microbiol.* **35,** 907–914 (1997).
2. Supply, P. *et al.* Proposal for Standardization of Optimized Mycobacterial Interspersed Repetitive Unit-Variable-Number Tandem Repeat Typing of Mycobacterium Tuberculosis. *J. Clin. Microbiol.* **44,** 4498–4510 (2006).
3. Jaffe, D. B. *et al.* Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res.* **13,** 91–96 (2003).
4. Niemann, S. *et al.* Genomic diversity among drug sensitive and multidrug resistant isolates of Mycobacterium tuberculosis with identical DNA fingerprints. *PLoS ONE* **4,** e7407 (2009).
5. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18,** 1851–1858 (2008).
6. Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biol.* **5,** R12 (2004).
7. Excoffier, L. & Lischer, H. E. L. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resour.* **10,** 564–567 (2010).
8. Lew, J. M., Kapopoulou, A., Jones, L. M. & Cole, S. T. TubercuList--10 years after. *Tuberc. Edinb. Scotl.* **91,** 1–7 (2011).
9. Zhang, J., Nielsen, R. & Yang, Z. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol. Biol. Evol.* **22,** 2472–2479 (2005).
10. Alland, D. *et al.* Role of large sequence polymorphisms (LSPs) in generating genomic diversity among clinical isolates of Mycobacterium tuberculosis and the utility of LSPs in phylogenetic analysis. *J. Clin. Microbiol.* **45,** 39–46 (2007).
11. Tatusov, R. L. *et al.* The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4,** 41 (2003).
12. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4,** 44–57 (2009).
13. Galagan, J. E. *et al.* TB database 2010: overview and update. *Tuberc. Edinb. Scotl.* **90,** 225–235 (2010).
14. Szklarczyk, D. *et al.* The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.* **39,** D561–568 (2011).
15. Louw, G. E. *et al.* A balancing act: efflux/influx in mycobacterial drug resistance. *Antimicrob. Agents Chemother.* **53,** 3181–3189 (2009).

16. Dos Vultos, T., Mestre, O., Tonjum, T. & Gicquel, B. DNA repair in Mycobacterium tuberculosis revisited. *FEMS Microbiol. Rev.* **33,** 471–487 (2009).
17. Smollett, K. L. *et al.* Global analysis of the regulon of the transcriptional repressor LexA, a key component of SOS response in Mycobacterium tuberculosis. *J. Biol. Chem.* **287,** 22004–22014 (2012).
18. Gamulin, V., Cetkovic, H. & Ahel, I. Identification of a promoter motif regulating the major DNA damage response mechanism of Mycobacterium tuberculosis. *FEMS Microbiol. Lett.* **238,** 57–63 (2004).
19. Wang, Y. *et al.* Global protein-protein interaction network in the human pathogen Mycobacterium tuberculosis H37Rv. *J. Proteome Res.* **9,** 6665–6677 (2010).
20. Strong, M. *et al.* Visualization and interpretation of protein networks in Mycobacterium tuberculosis based on hierarchical clustering of genome-wide functional linkage maps. *Nucleic Acids Res.* **31,** 7099–7109 (2003).
21. Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P.-L. & Ideker, T. Cytoscape 2.8: new features for data integration and network visualization. *Bioinforma. Oxf. Engl.* **27,** 431–432 (2011).
22. Machowski, E. E., Barichievy, S., Springer, B., Durbach, S. I. & Mizrahi, V. In vitro analysis of rates and spectra of mutations in a polymorphic region of the Rv0746 PE_PGRS gene of Mycobacterium tuberculosis. *J. Bacteriol.* **189,** 2190–2195 (2007).
23. Maus, C. E., Plikaytis, B. B. & Shinnick, T. M. Molecular Analysis of Cross-Resistance to Capreomycin, Kanamycin, Amikacin, and Viomycin in Mycobacterium tuberculosis. *Antimicrob. Agents Chemother.* **49,** 3192–3197 (2005).
24. Agerton, T. *et al.* Transmission of a highly drug-resistant strain (strain W1) of Mycobacterium tuberculosis. Community outbreak and nosocomial transmission via a contaminated bronchoscope. *JAMA J. Am. Med. Assoc.* **278,** 1073–1077 (1997).
25. Morlock, G. P. *et al.* Phenotypic characterization of pncA mutants of Mycobacterium tuberculosis. *Antimicrob. Agents Chemother.* **44,** 2291–2295 (2000).
26. Ioerger, T. R. *et al.* Genome Analysis of Multi- and Extensively-Drug-Resistant Tuberculosis from KwaZulu-Natal, South Africa. *PLoS One* **4,** (2009).
27. Gardy, J. L. *et al.* Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. *N. Engl. J. Med.* **364,** 730–739 (2011).
28. Calver, A. D. *et al.* Emergence of increased resistance and extensively drug-resistant tuberculosis despite treatment adherence, South Africa. *Emerg. Infect. Dis.* **16,** 264–271 (2010).
29. Meacci, F. *et al.* Drug resistance evolution of a Mycobacterium tuberculosis strain from a noncompliant patient. *J. Clin. Microbiol.* **43,** 3114–3120 (2005).
30. Johnson, R. *et al.* An outbreak of drug-resistant tuberculosis caused by a Beijing strain in the western Cape, South Africa. *Int. J. Tuberc. Lung Dis. Off. J. Int. Union Tuberc. Lung Dis.* **10,** 1412–1414 (2006).
31. Kato-Maeda, M. *et al.* Use of whole genome sequencing to determine the microevolution of Mycobacterium tuberculosis during an outbreak. *PloS One* **8,** e58235 (2013).
32. Comas, I. *et al.* Human T cell epitopes of Mycobacterium tuberculosis are evolutionarily hyperconserved. *Nat. Genet.* **42,** 498–503 (2010).
33. Gagneux, S. *et al.* Variable host-pathogen compatibility in Mycobacterium tuberculosis. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 2869–2873 (2006).

34. Hirsh, A. E., Tsolaki, A. G., DeRiemer, K., Feldman, M. W. & Small, P. M. Stable association between strains of Mycobacterium tuberculosis and their human host populations. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 4871–4876 (2004).
35. Ioerger, T. R. *et al.* The non-clonality of drug resistance in Beijing-genotype isolates of Mycobacterium tuberculosis from the Western Cape of South Africa. *BMC Genomics* **11,** 670 (2010).