Genomic analysis identifies targets of convergent positive selection in drug-resistant *Mycobacterium tuberculosis.*

Farhat et al., *Nature Genetics,* 2013

Journal Club, 06.12.2013

Triinu Kõressaar

Introduction (1/2)

Rapid evolution of antibiotic-resistant bacteria

Resistance is encoded in the bacterial genome

Resistance associated mutations - biomarkers that can be rapidly identified by PCR/sequencing

Article about method to identify biomarkers of drug resistance in a rapid and unbiased manner

The method can be applied to different microbes with different phenotypes

Introduction (2/2)

In this work, the method is applied to identify biomarkers of *Mycobacterium tuberculosis*

Multidrug-resistant (MDR) tuberculosis - tuberculosis that is resistant to isoniazid and rifampicin, the two most effective drugs

Resistance is thought to arise through the serial acquisitions of point mutations in genes encoding drug-activating enzymes or drug targets

Causative mutations have not been identified in 10-40% of clinically resistant isolates

Mutations in four classes of genes may confer selective advantage in the presence of drugs:

Classical drug resistance genes – encoding protein target of the drug or drug-metabolizing enzyme

Mutations that reduce cell wall permeability or increase the activity of drug efflux pumps

Mutations that ameliorate the fitness costs of other resistance conferring mutations

Mutator phenotypes can increase the rate at which rare beneficial mutations occur

Genomes

NGS whole genome sequencing of 116 *M.tuberculosis* isolates
+ 7 publicly available genomes

-among them 47 isolates resistant to at least one tuberculosis drug
(including 9 XDR tuberculosis)

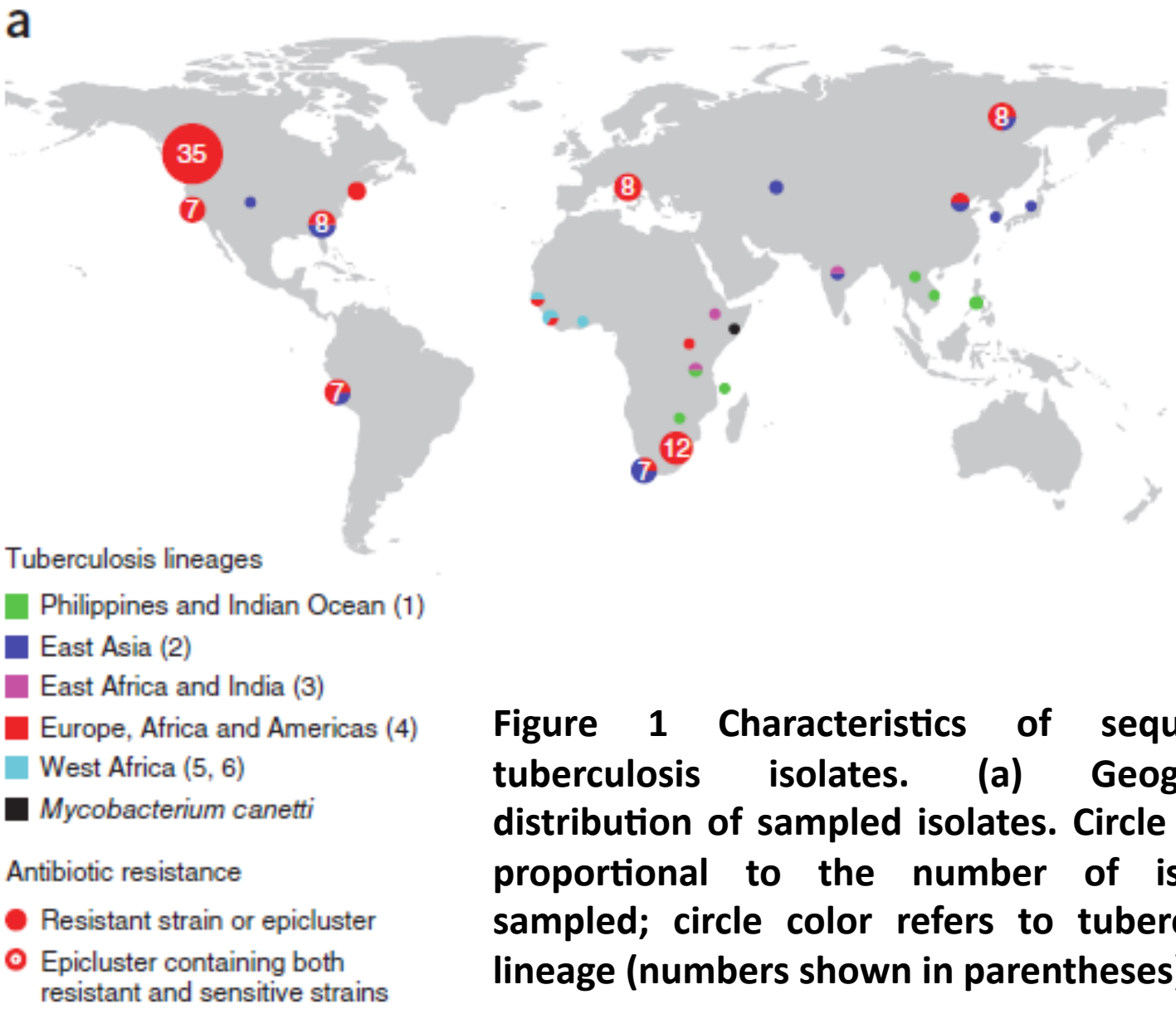


Figure 1 Characteristics of sequenced tuberculosis isolates. (a) Geographic distribution of sampled isolates. Circle size is proportional to the number of isolates sampled; circle color refers to tuberculosis lineage (numbers shown in parentheses).

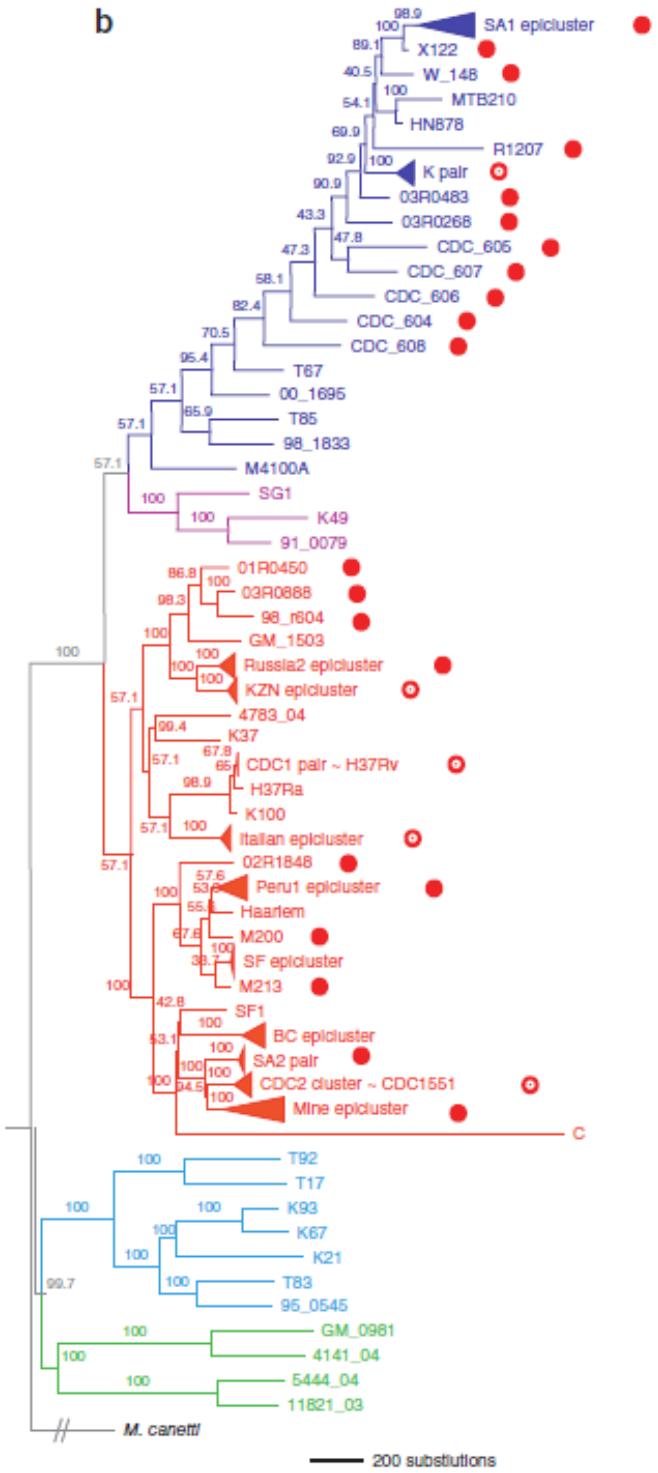


Figure 1. (b) Parsimony-based phylogenetic tree with node bootstrap support. Root length is not to scale. Epiclusters are merged into triangles for clarity, with the exception of two paraphyletic epiclusters, Peru2 and Russia1.

Methods (1/2)

Sequencing - Illumina Genome Analyzer IIx, reads 36bp or more

Mapping – MAQ (Mapping and Assembly with Qualities), reference genome H37Rv

Reads that aligned with more than three mismatches in the first 24bp or that aligned to multiple locations were discarded

SNP calling – min depth 20x and consensus quality score 20

Whole genome alignements - MUMMER

Phylogeny construcion – from superset of SNPs relative to reference genome (SNPs in repeats were discarded), ~23k SNPs were concatenated and were used to build a tree

Tree building – with three methods, PHYLIP, MrBayes v3.2 (Bayesian Markov chain Monte Carlo), PhyML v3.0 (maximum likelihood tree)

Methods (2/2)

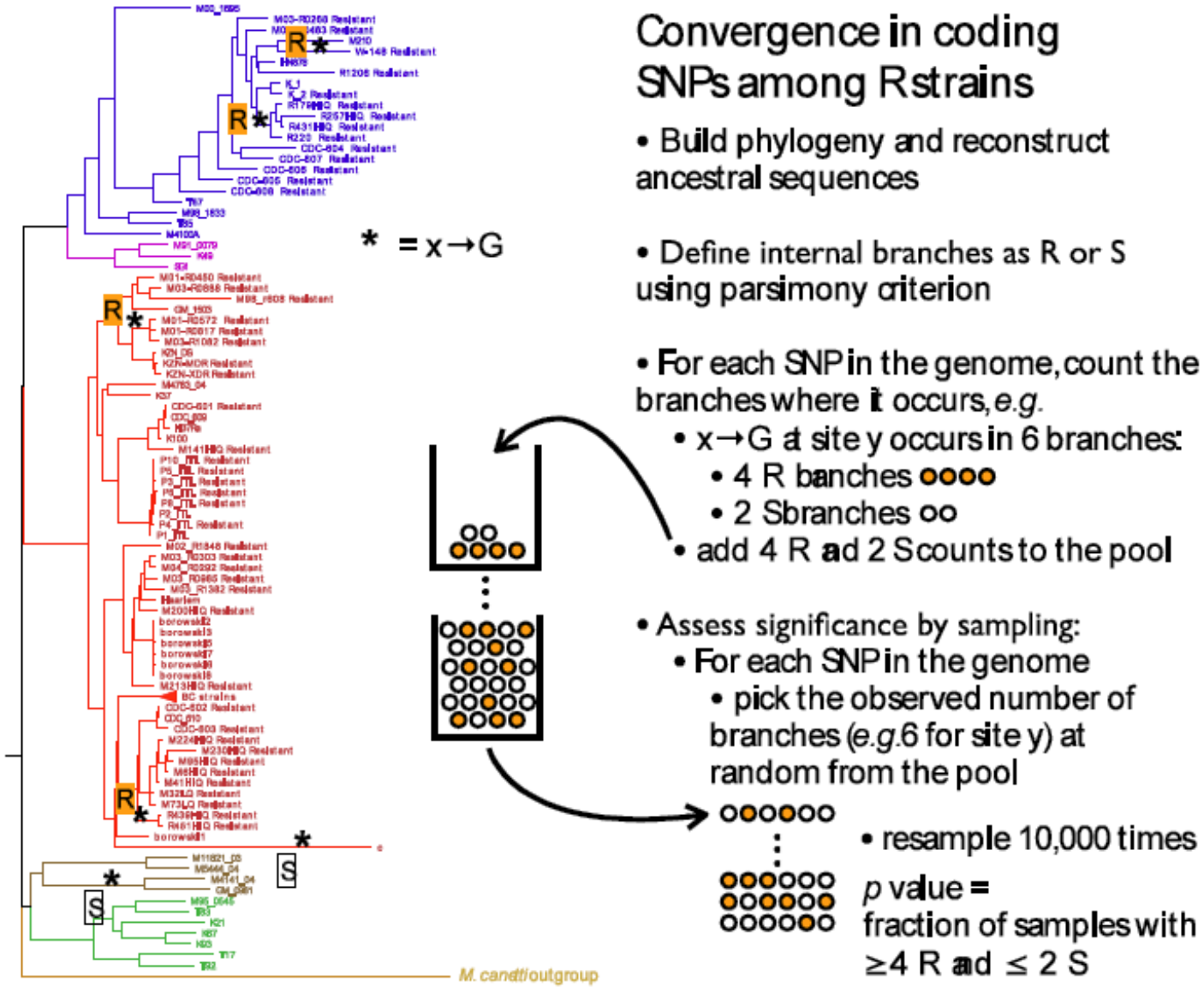
Phylogenetic convergence test for selection (PhyC)

Three trees, *Mycobacterium canetti* was used as outgroup

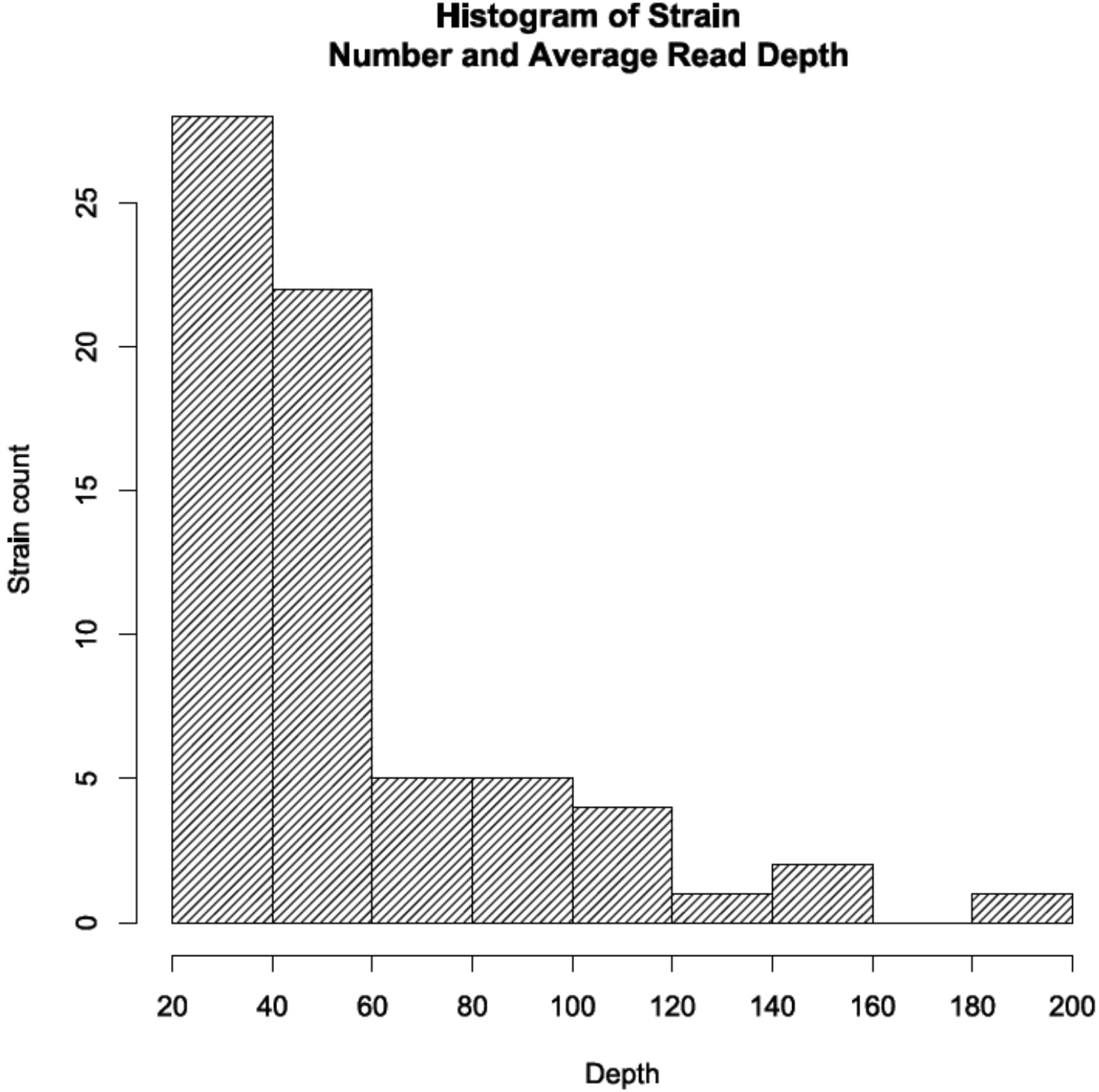
All ambiguously constructed states were excluded

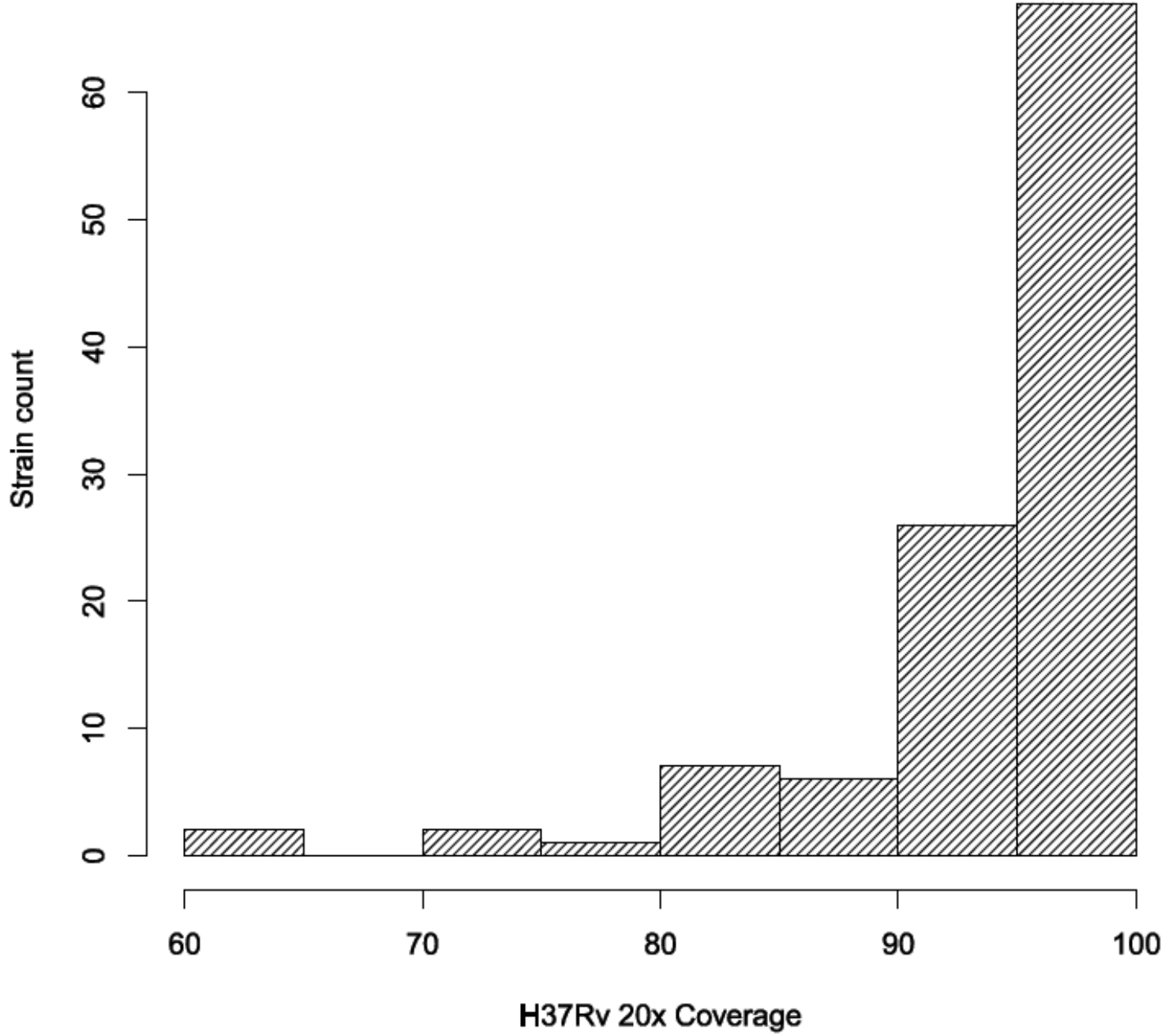
For each nucleotide position, they counted the number of convergent SNPs in resistant and sensitive branches

For a SNP that converges in x resistant and y sensitive branches, we sampled x + y branches from the distribution of all SNPs in all branches across the genome, repeated this 10,000 times and recorded the proportion of times substitutions were observed in ≥x resistant and ≤y sensitive branches. This proportion serves as an empirical P value for an unexpectedly high level of convergence among resistant branches, suggesting the action of selection.



Supplementary Figure 7: 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.


Histogram of Strain
Number and Average Read Depth
Strain count
0
5
10
15
20
25
20
40
60
80
100
120
140
160
180
200
Depth


Histogram of Strain Number and Reference Genome Coverage at 20 Fold Read Depth
Strain count
0
10
20
30
40
50
60
60
70
80
90
100
H37Rv 20x Coverage

Results

PhyC detected all 11 known resistance determinants

They identified 39 new targets of independent mutation not previously associated with resistance (7 nonsyn, 2 noncoding, 28 genes and 2 intergenic regions)

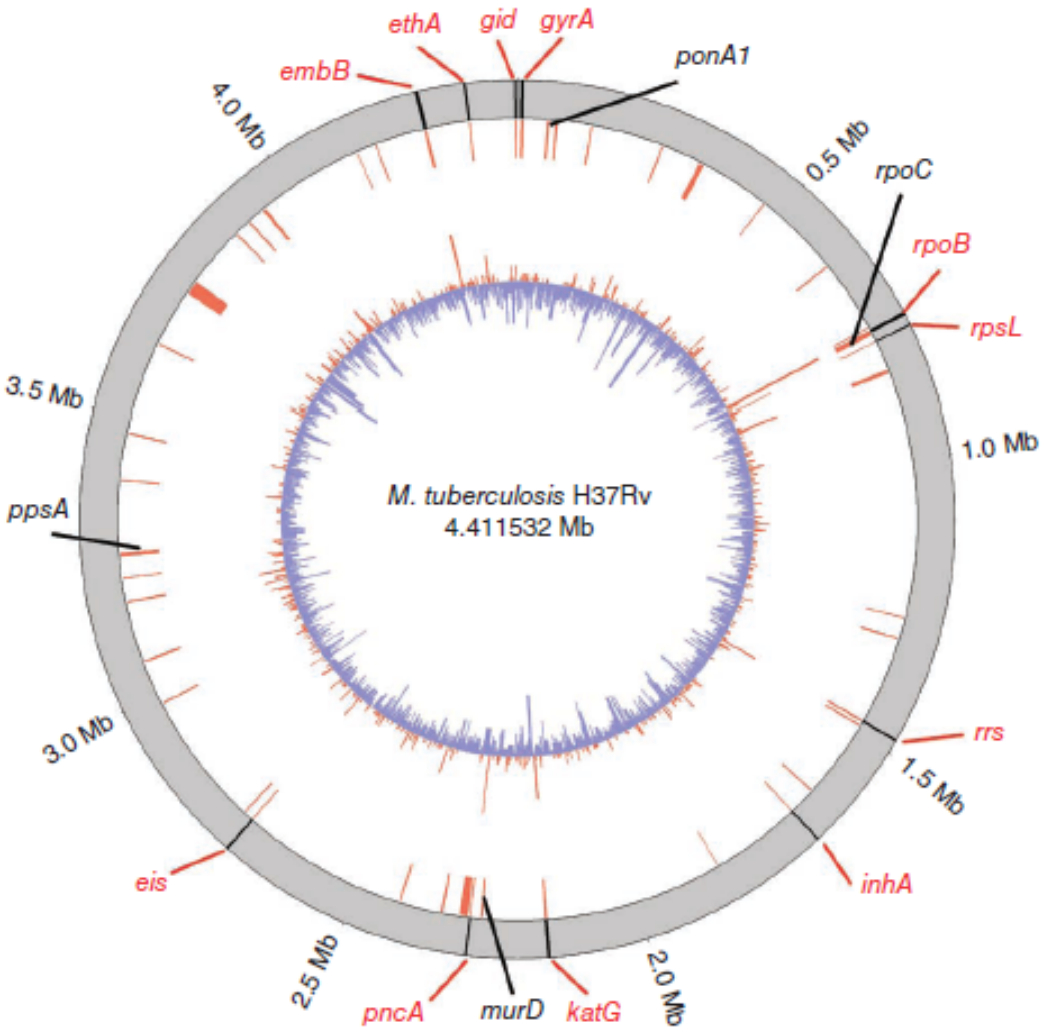


Figure 2 Candidate genes under selection in resistant *M. tuberculosis*. Circular plot of gene locations. Outer black lines represent the 11 benchmark drug resistance–associated genes in the H37Rv reference genome (red text). Inner red lines represent the locations of targets of independent mutation. Four new targets of independent mutation of interest are shown in black text. The innermost bar plot shows the number of mutations per gene or intergenic region in resistant (red) and sensitive (blue) isolates. Plotted using Circos[29].

Functions of candidate selected loci

Table 1 Targets of independent mutation with annotated function

		Cellular function					Resistance association		
Gene	Rv number	Synthesis or regulation of surface-exposed lipids	Peptidoglycan homeostasis	Transcriptional regulation	DNA replication and repair	Glucose metabolism and antioxidation	Associated with resistance in *M. tuberculosis*	Associated with resistance in NTM	Associated with resistance in other bacteria
ppsA	*Rv2931*	19					27	25	
pks3	*Rv1180*	21							
pks12	*Rv2048c*	20					24	24,26	
ponA1	*Rv0050*		22					23	30
murD	*Rv2155c*		22						
mtrB	*Rv3245c*			31				32,33	34
rpoC	*Rv0668*				12		11,12		13
dnaQ	*Rv3711c*				35				15
opcA	*Rv1446c*					35	36		
rbsK	*Rv2436*				37	35			
rrs promoter (pre-*Rvnr01*)							38	39	

Numbers refer to literature references. Genes involved in cell wall biosynthesis include *ppsA*, *pks3*, *pks12*, *ponA1* and *murD*. A complete list of targets of independent mutation is given in **Supplementary Table 5**. NTM, non-tuberculous mycobacteria.

Out of 39 newly associated genomic regions, only 11 of them have known functions, 16 belong to family of genes (PE/PPE) of principally unknown function, 12 unknown function

Conclusions

The method for comprehensive genome-wide screen for genes under selection is provided

39 new genes and inergenic regions associated with resistance have identified as promising new targets for molecular diagnostics