HARVARD MEDICAL SCHOOL | DEPARTMENT OF Biomedical Informatics

MGH 1811 MASSACHUSETTS GENERAL HOSPITAL

Prediction of drug resistance using genotypic data

An argument for machine learning

CPTR Workshop

3/22/2017

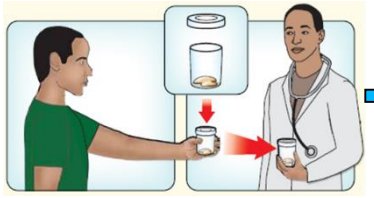
Maha Farhat MD MSc

Department of Biomedical Informatics

Pulmonary and Critical Care Medicine

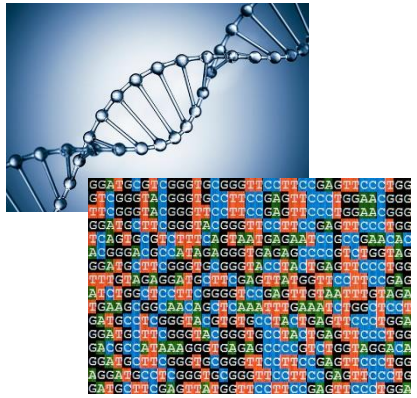
Harvard Medical School

TB & Resistance Diagnosis with Genomic Big Data



Sputum collection

0-1 day



Whole Genome Sequencing

Minutes

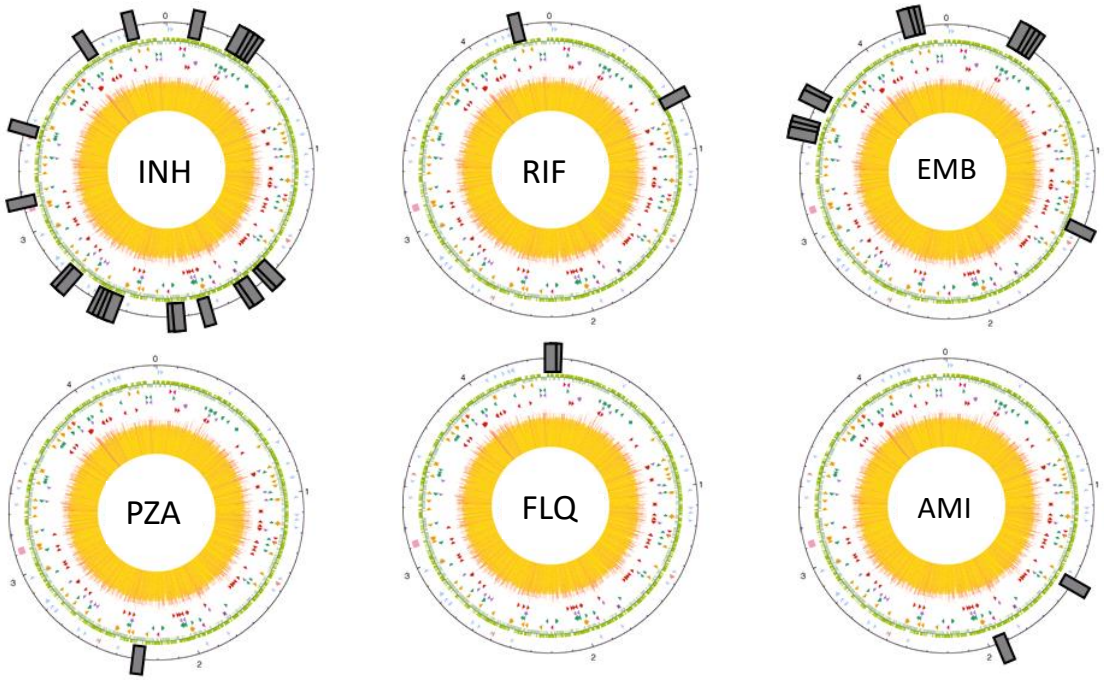
Drug Resistance:

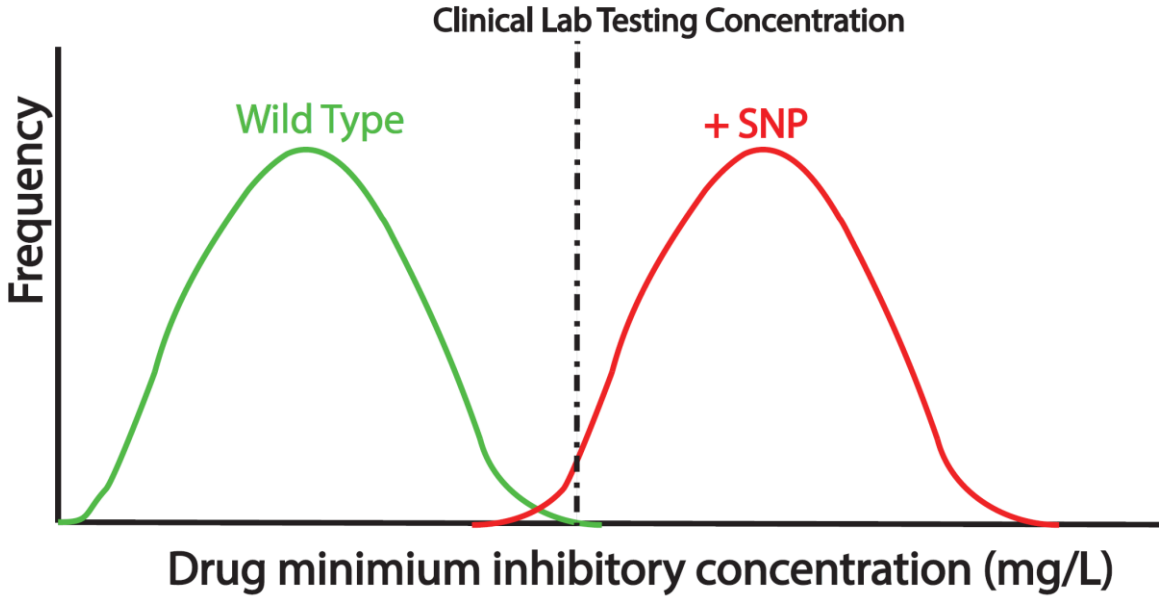
Rifampicin=R
Isoniazid=R
Pyrazinamide=S
Streptomycin=S
Ethambutol=R
Amikacin=S
Capreomycin=S
Kanamycin=S
Ofloxacin=R
Moxifloxacin=S
Ethionamide=R
Bedaquiline=S
Linezolid=S
Cycloserine=S

0-1 day

Treatment started

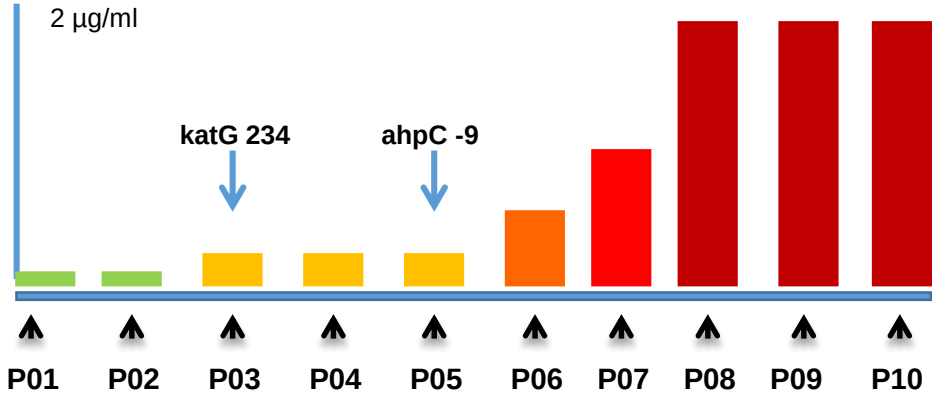
Genetic basis of drug resistance

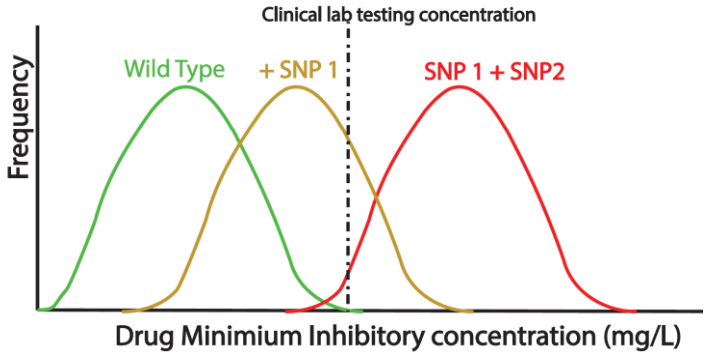


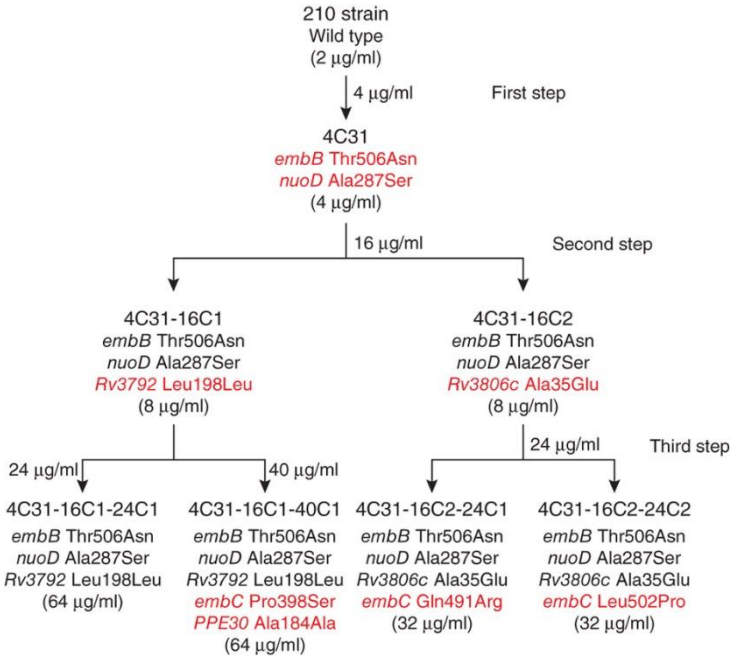


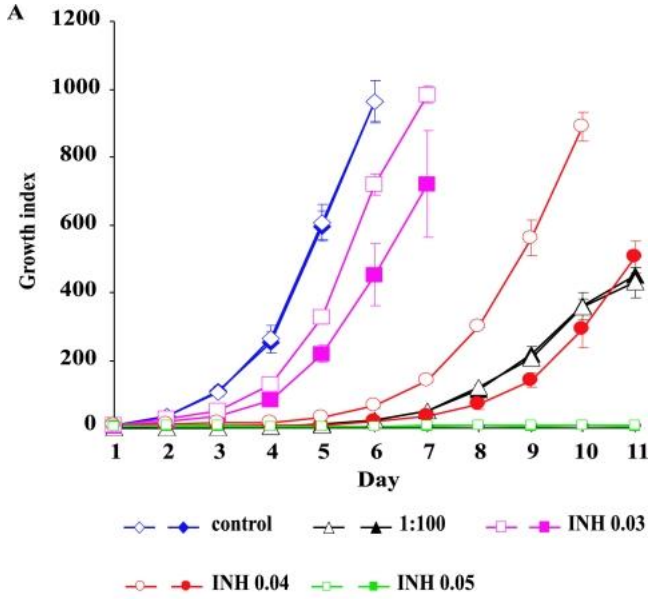
1:1 Mutation : Drug resistance

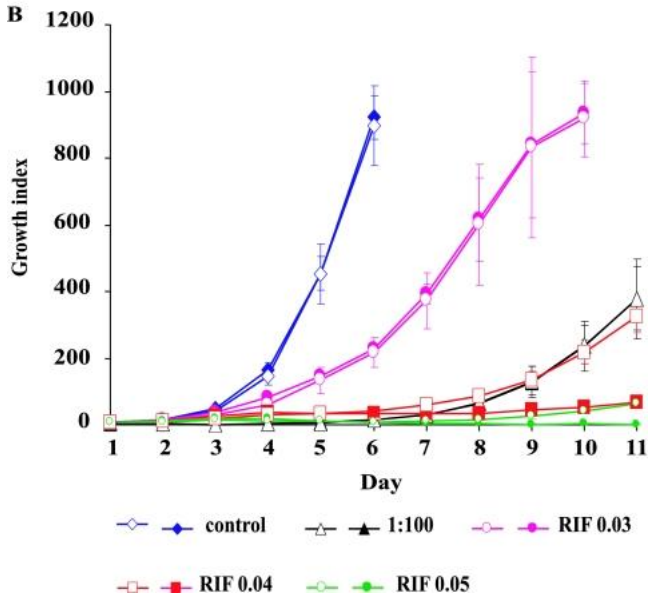
Clinical resistance is more complex





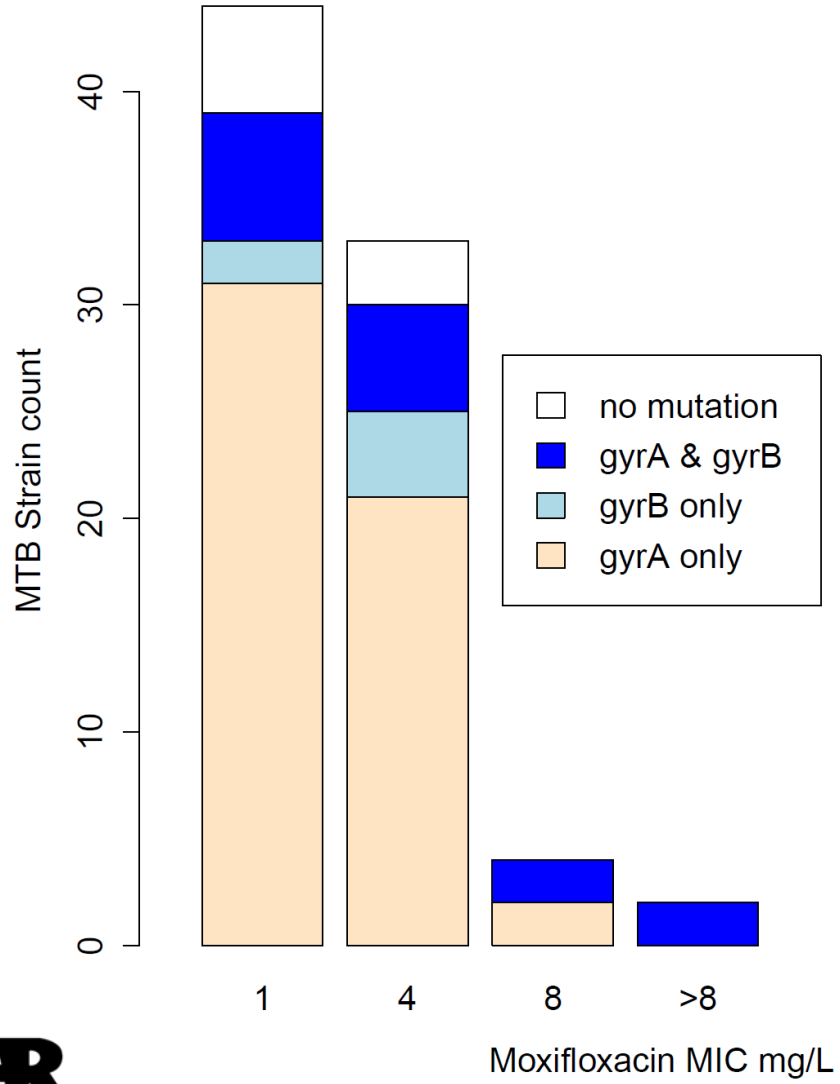


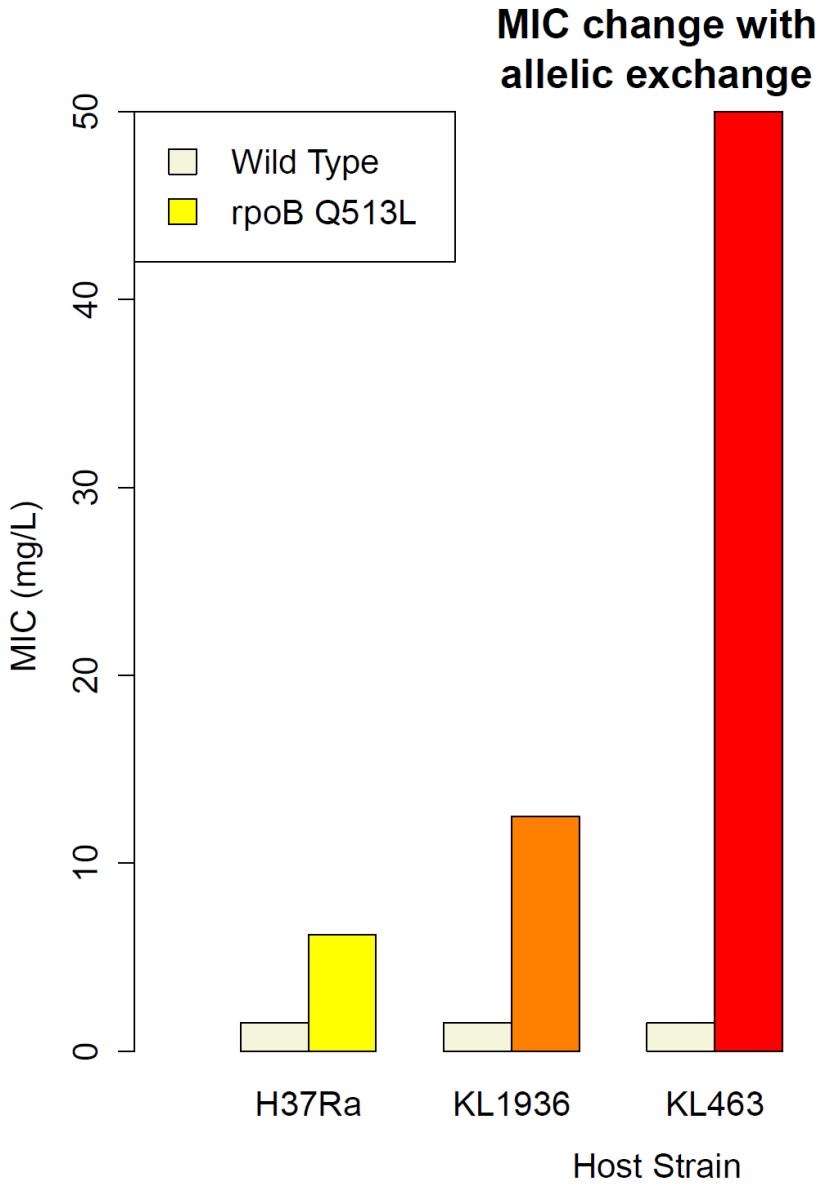




Meacci F et al. J Clin Micro 2005, Safi H et al. Antimicro Agents Chemother, 2008, Safi H et al. Nat Gen 2013

Evidence for epistasis/ gene-gene interaction



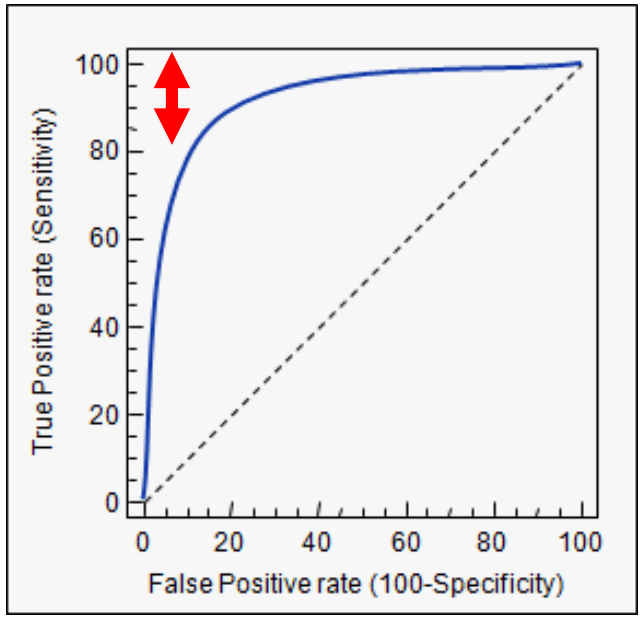


Farhat *et al.* JCM 2016

Nebenzahl-Guimaraes, Jacobson, **Farhat,** Murray. JAC 2014

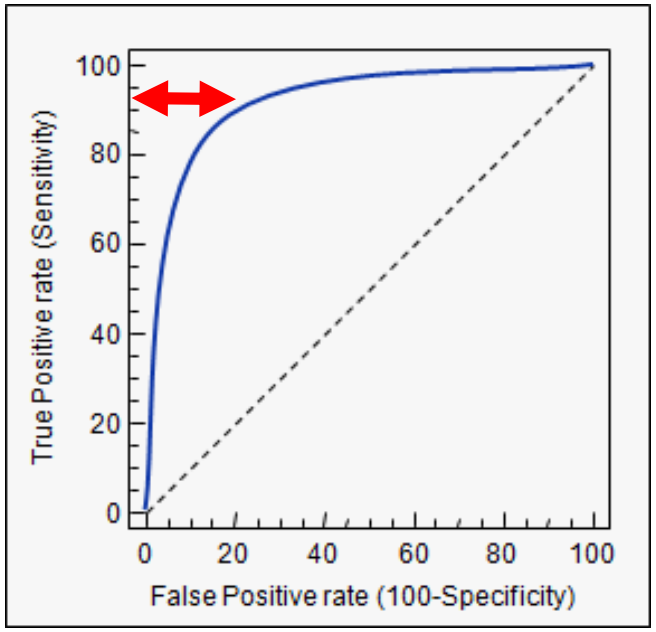
CFAR HARVARD UNIVERSITY CENTER FOR AIDS RESEARCH

The molecular diagnostic gap:



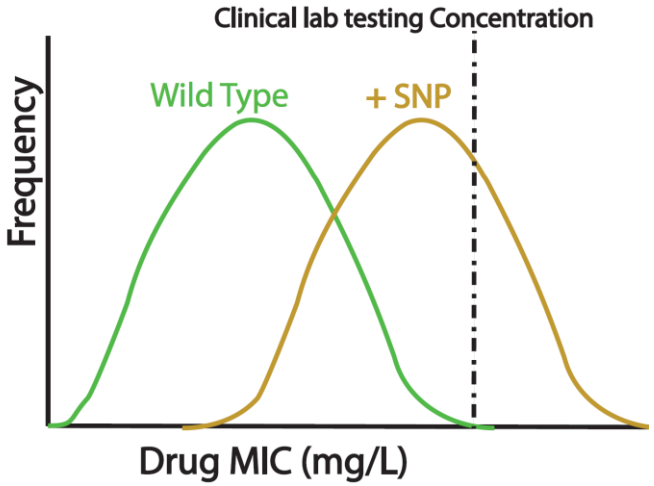
For Sensitivity

- Phenotypic overcall
- Heterogeneity
- Novel mutations or loci
- Epistasis

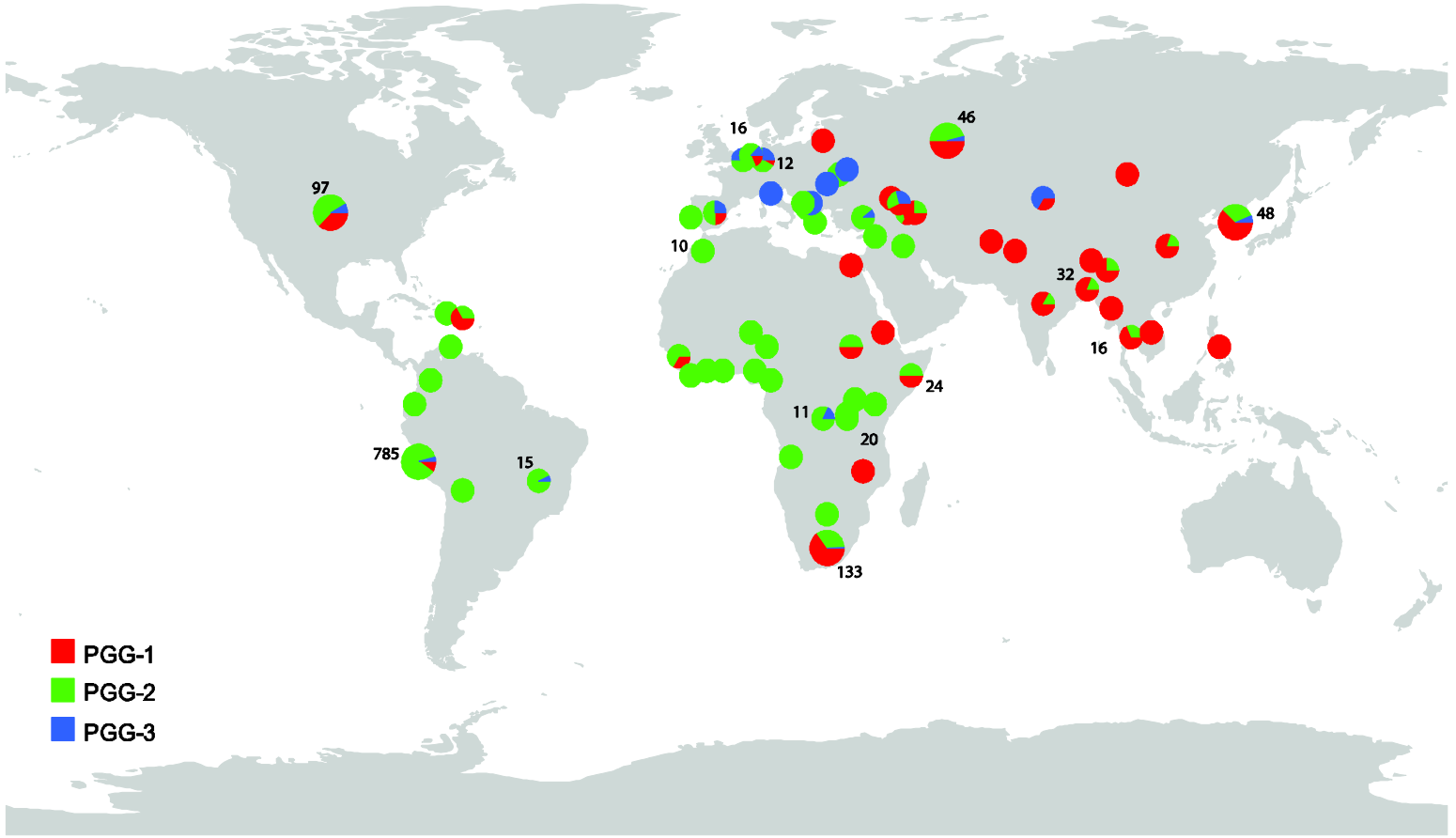


For Specificity

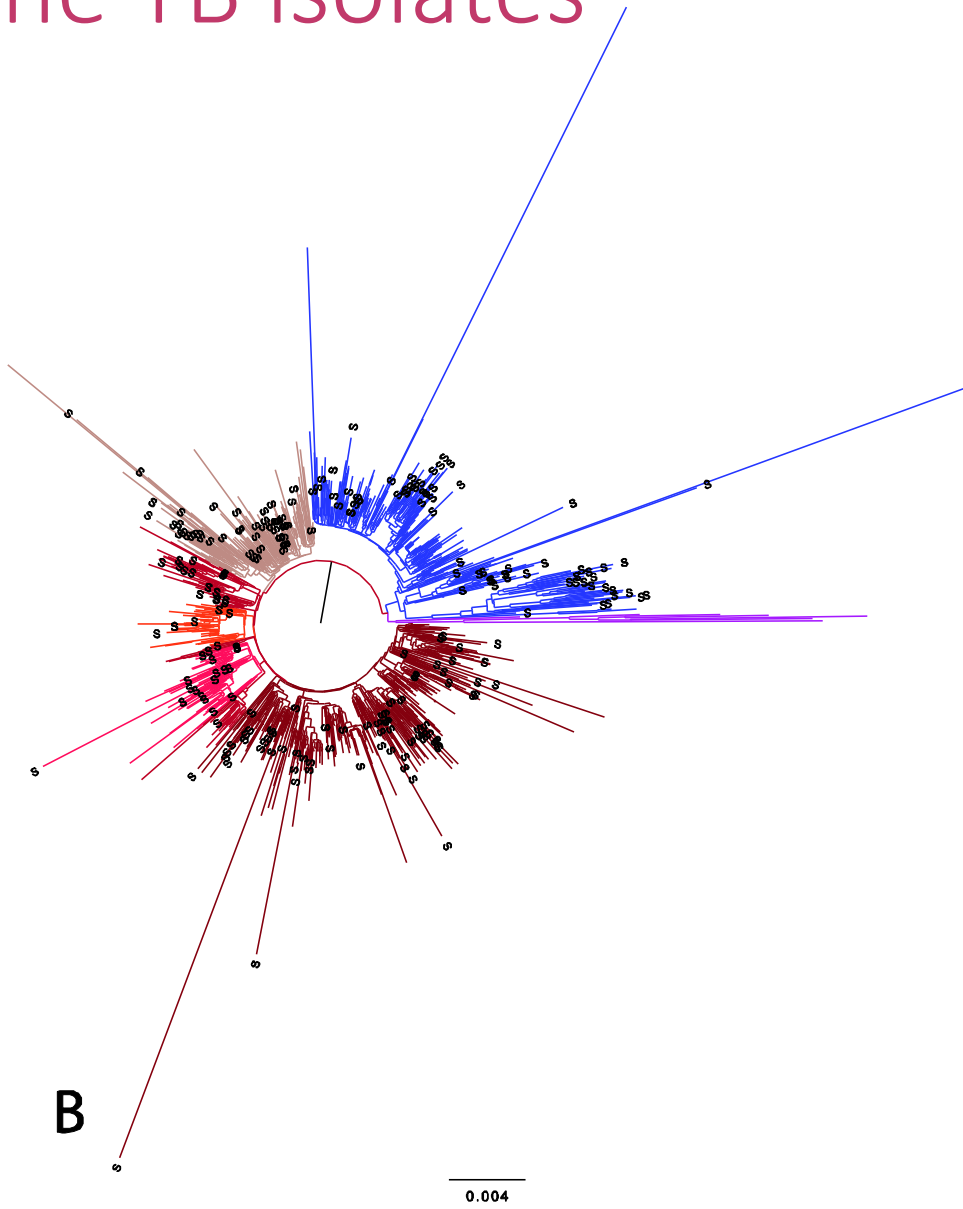
- Phenotypic undercall
- False positive associations
- Intermediate resistance mutations
- Epistasis



Geographic source and diversity of the TB isolates



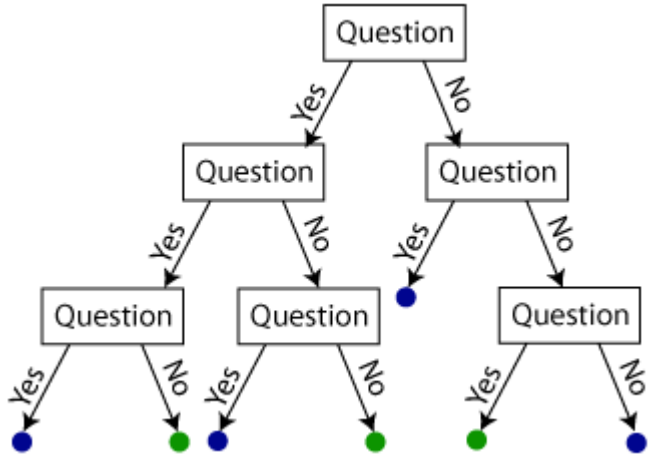
A

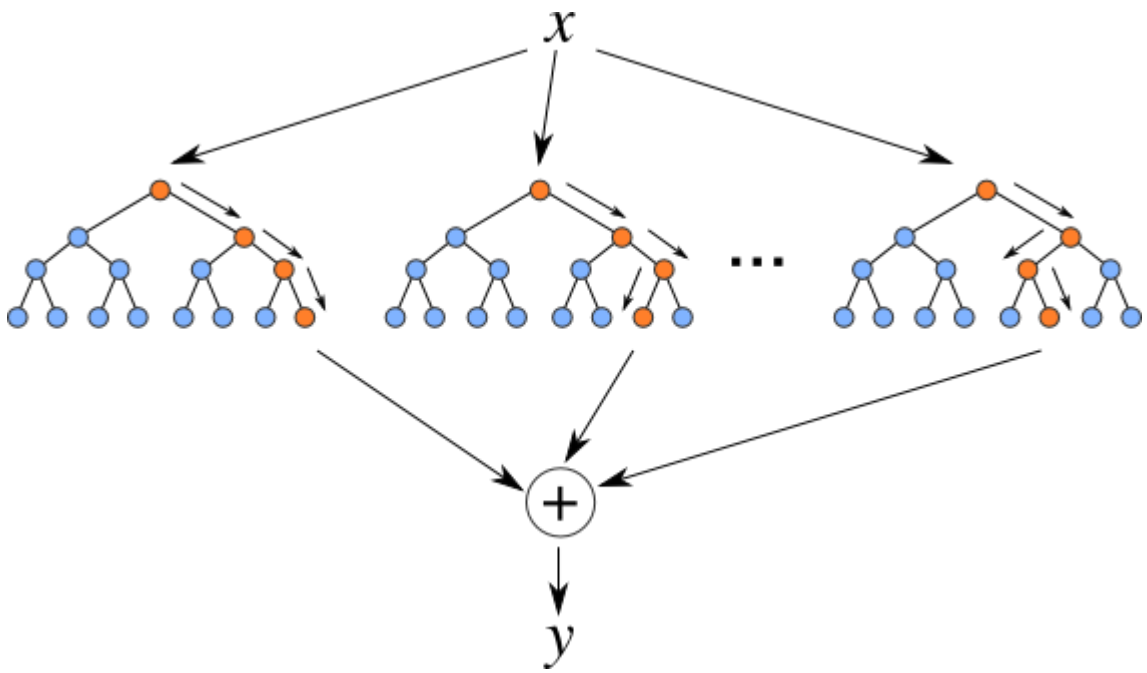


B

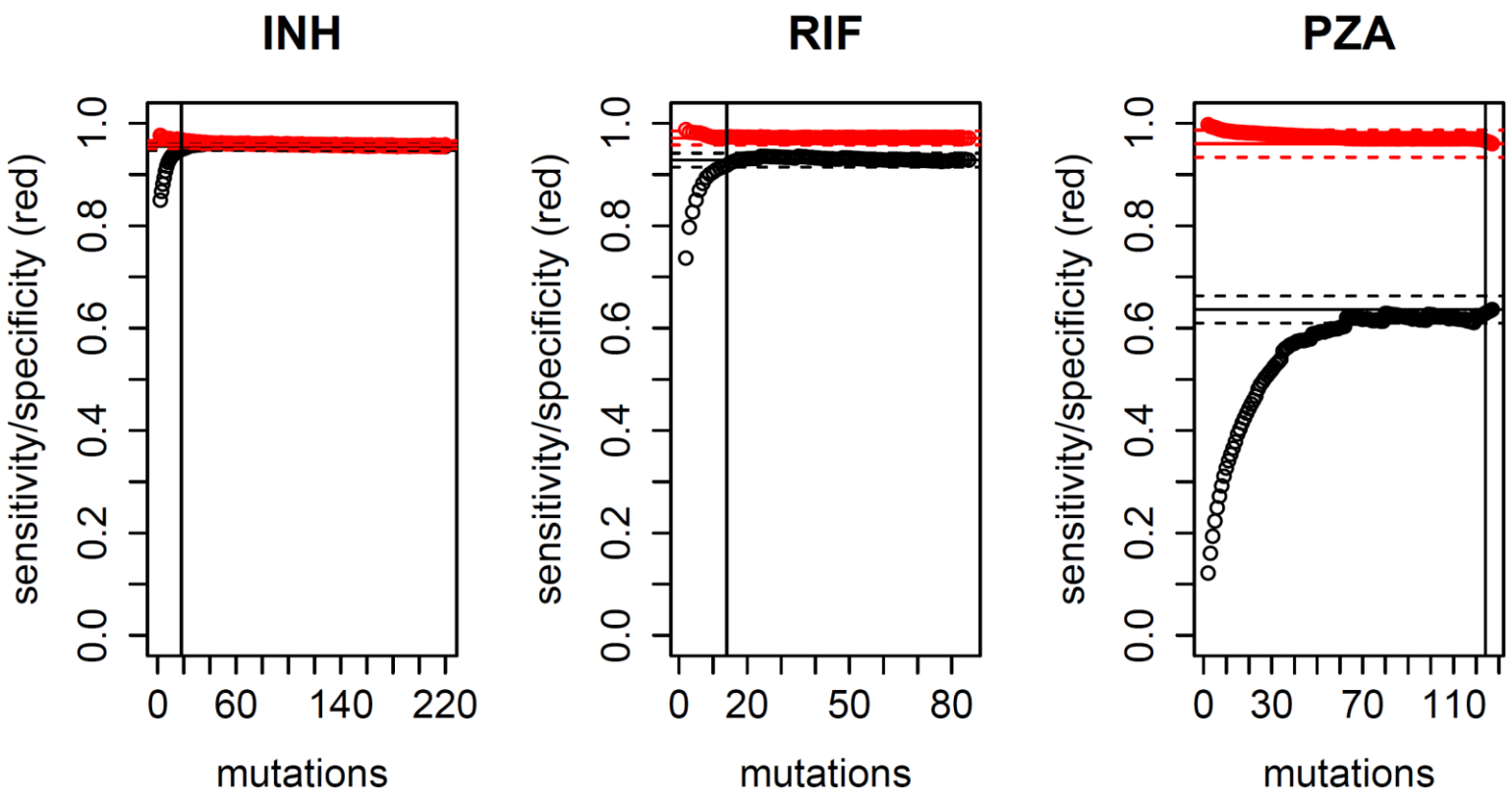
Farhat *et al.* American Journal of Respiratory and Critical Care 2016

Random Forest Classifier



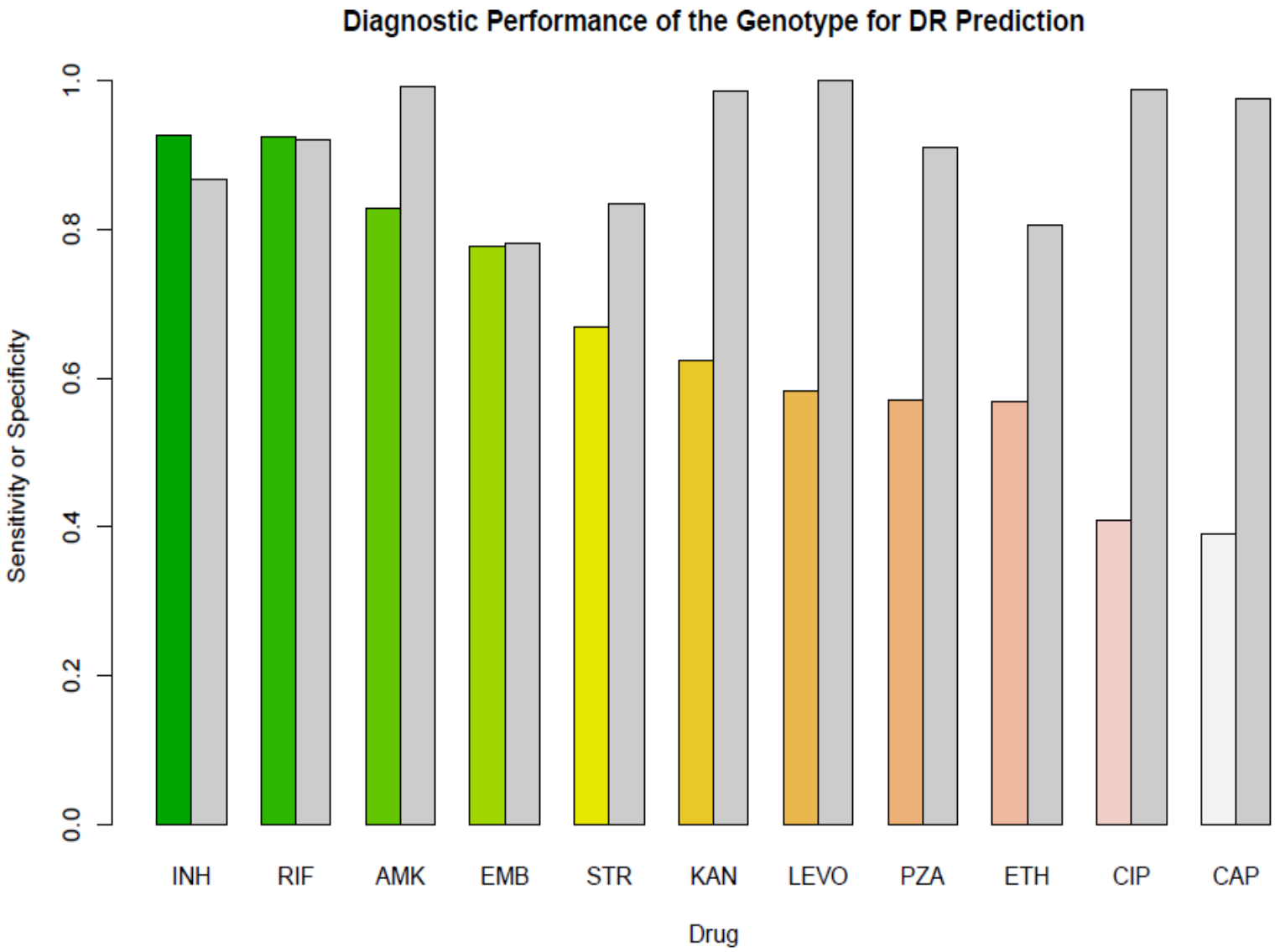


Identifying the minimum set of mutations predictive of drug resistance



Farhat *et al.* AJRCCM 2016

Validated Predictive performance for drug resistance diagnosis



Drug	Selected Mutations
Isonaizid	19
Rifampicin	14
Pyrazinamide	124
Ethambutol	18
Streptomycin	39
Ethionamide	20
Kanamycin	3
Capreomycin	5
Amikacin	2
Ciprofloxacin	7
Levofloxacin	8
Ofloxacin	6
p-aminosalicylic acid	4
Total	***250***

Farhat *et al.* AJRCCM 2016

Predict Upload your genetic data to make drug resistance prediction.

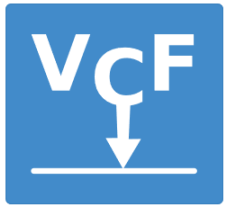
To examine the frequency of individual mutations by drug resistance you can consult the Map function for country specific data, and the Explore function to examine their frequency in all of the data.

New FastQ Pair-Ended Prediction

Create a prediction from a set of pair-ended FastQ genetic sequences. This option involves the largest files and takes more time to process that the vcf or manual options.

New FastQ Single-Ended Prediction

Create a prediction from a single-ended FastQ genetic sequence file. This option involves a large file and takes more time to process that the vcf or manual options.



New Variant Call Format Prediction

Minimal genotypic information for accurate resistance predictions are below. Genetic regions are listed in order of decreasing importance. For more detailed list of genetic variants see reference Farhat MR, Sultana R et al. Genetic Determinants of Drug Resistance in Mycobacterium tuberculosis and Their Diagnostic Value. AJRCCM 2016 and get More information about the vcf format.

New Manually Entered Prediction

Create a prediction from a set of pair-ended FastQ genetic sequences. This option involves the largest files and takes more time to process that the vcf or manual options.

(September 9th, 2016 3:36:19 p.m.)

TwoRavens The Dataverse Network Project IQSS The Institute for Quantitative Social Science

gentb.hms.harvard.edu

RF Validation using Reseq public dataset

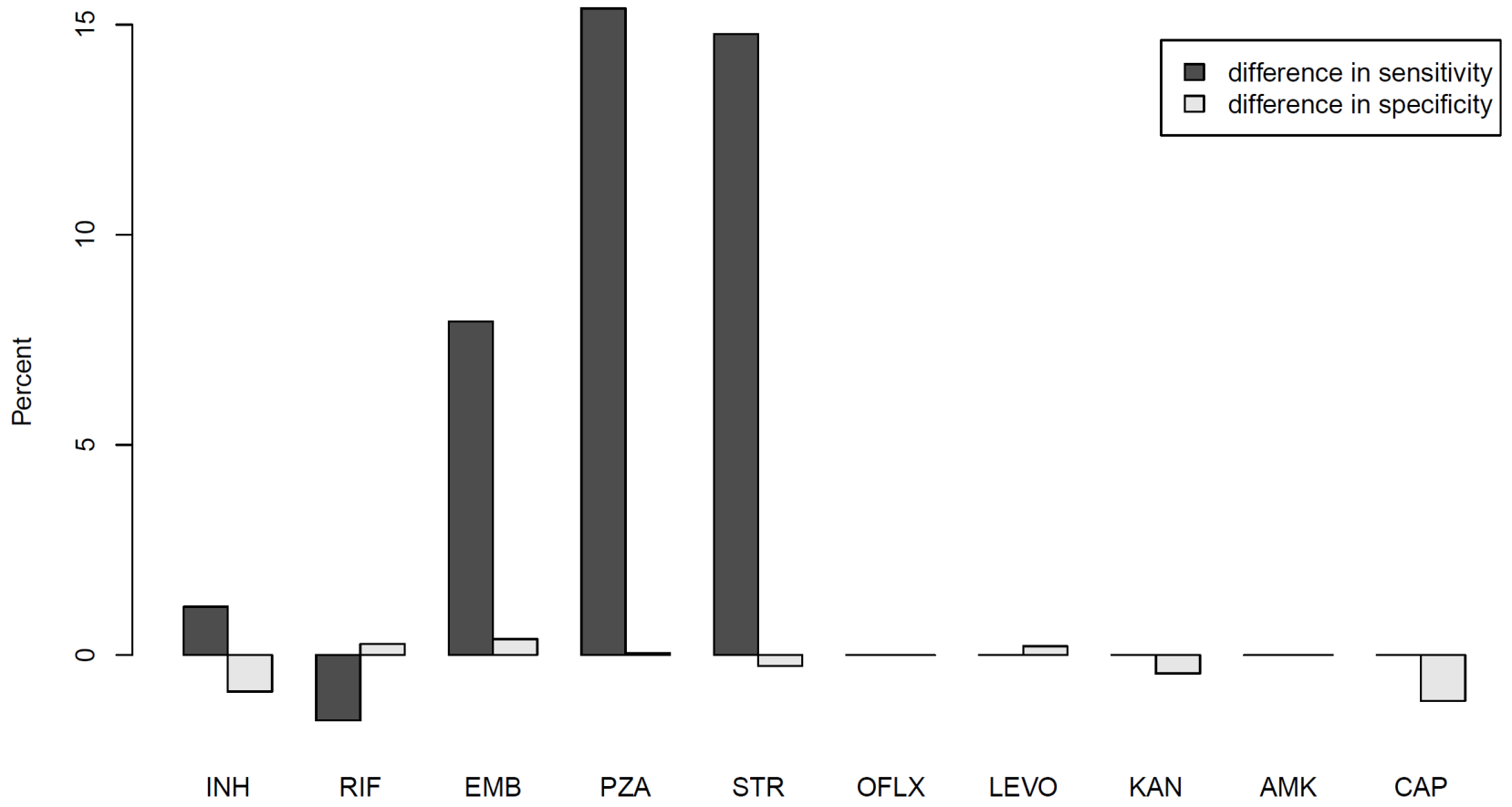
2,429 Isolates:

- 348 INH R
- 129 RIF R
- 63 EMB R
- 91 PZA R
- 88 STR R
- 20 OFLX R

Low INH sensitivity

- 83% (canonical mutations)
- 85% (RF predictor)
- INH-mono resistance ?

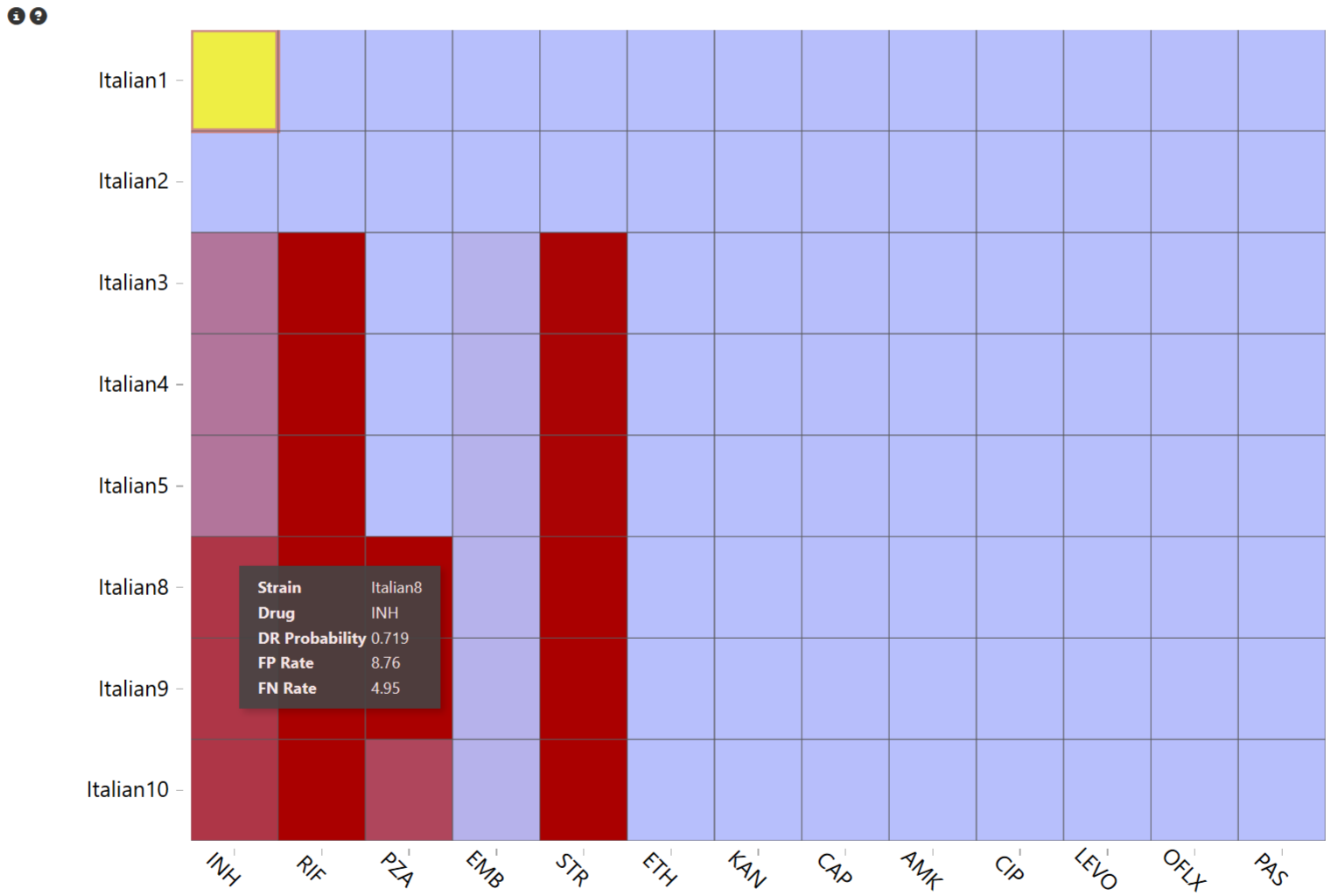
Comparison with canonical mutation +/-



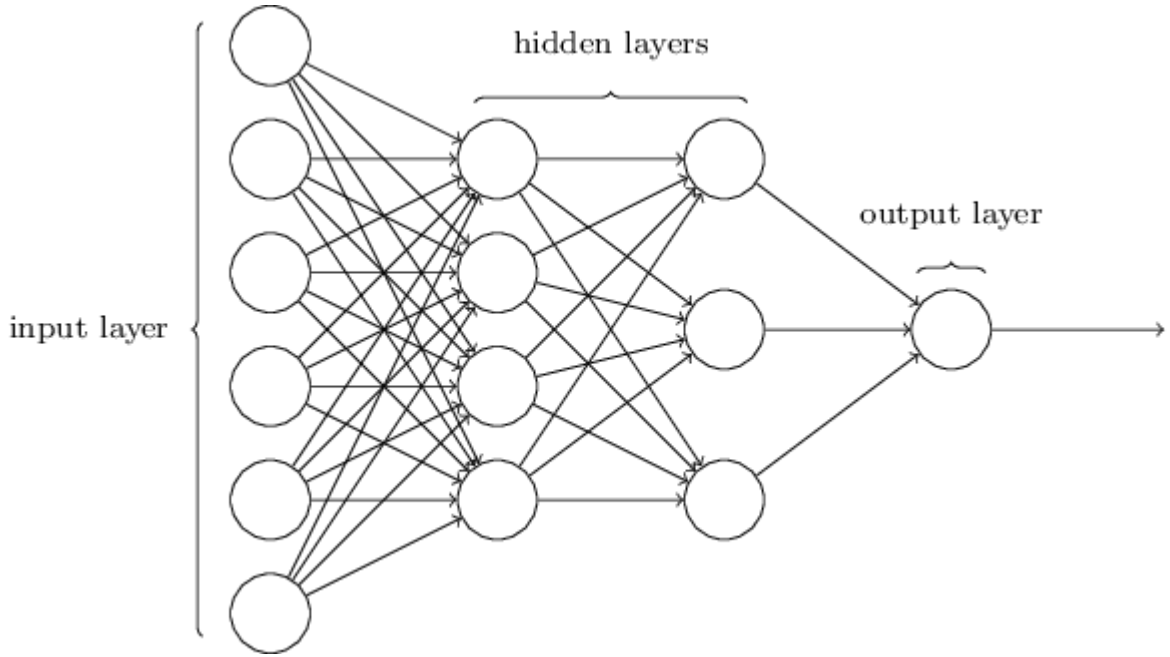
Italian

Serial Strains from the same patient

Processing | Output Files | Resistance Prediction



Neural network model



Pyrazinamide (PZA): Predictive Model Performance

	ROC-AUC	Sensitity [1]	Specificity [2]	[1] + [2]	Optimal Threshold
Original Random Forest		0.640 (0.030)	0.920 (0.030)	1.560	
Random Forest (1)	0.791 (0.019)	0.622 (0.029)	0.929 (0.019)	1.550	0.506 (0.061)
Neural Network (1)	0.856 (0.017)	0.718 (0.027)	0.946 (0.022)	1.664	0.573 (0.050)
Random Forest (2)	0.813 (0.018)	0.649 (0.029)	0.925 (0.020)	1.573	0.505 (0.056)
Neural Network (2)	0.883 (0.016)	0.752 (0.028)	0.943 (0.024)	1.695	0.577 (0.050)

With thanks

Collaborators

- Megan Murray, Department of Global Health and Social Medicine
- PIH/DGHSM (Molly Franke, Carole Mitnick)
- MSLI (Alex Sloutsky and Devinder Kaur)
- Stellenbosch University (Rob Warren, Tommie Victor and Lizma Streicher)
- RIVM (Dick van Soolingen, Hanna Guimaraes)
- US CDC (Bonnie Plikyatis, Jamie Posey)
- PHRI (Barry Kreiswirth, Natalia Kurepina)
- WHO - TDR (Leen Rigouts)
- Boston Medical Center (Karen Jacobson)
- IQSS (Gary King, Merce Crosas, Christine Choirat, Raman Prasad, James Honaker)

Partners In Health

Funders

BILL & MELINDA GATES foundation

AMERICAN LUNG ASSOCIATION®

NIH Big Data to Knowledge(BD2K)

PBF

CFAR HARVARD UNIVERSITY CENTER FOR AIDS RESEARCH