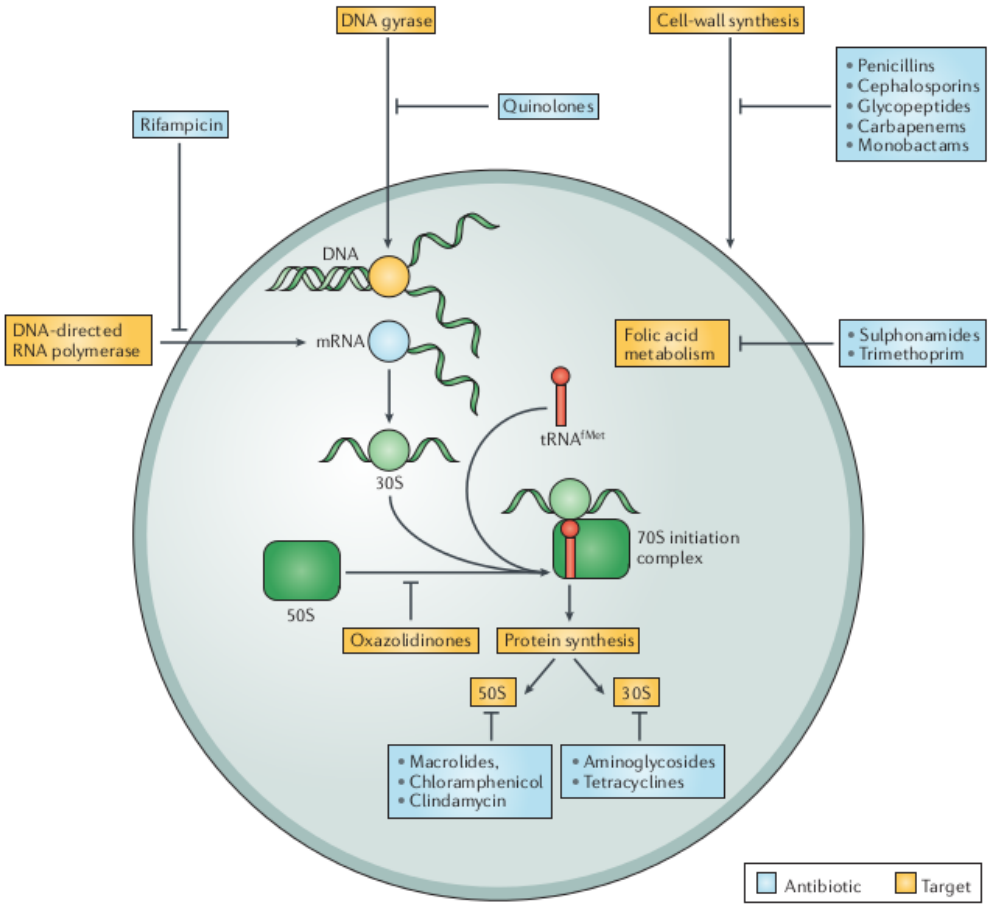
GWAMAR: Genome-wide assessment of mutations associated with drug resistance in bacteria

Michal Wozniak[1,2], Limsoon Wong[2] and Jerzy Tiuryn[1]

[1]University of Warsaw
[2]National University of Singapore

16 December, 2014

Drug action mechanisms



Adopted from: Platforms for antibiotic discovery; Kim Lewis; Nature Reviews; 2013

Timeline of antibiotics

Antibiotic class; example	Discovery	Introduction	Resistance	Mechanism of action	Activity or target species
Sulfadrugs; prontosil	1932	1936	1942	Inhibition of dihydro-pteroate synthetase	Gram-positive bacteria
β-lactams; penicillin	1928	1938	1945	Inhibition of cell wall biosynthesis	Broad-spectrum activity
Aminoglycosides; streptomycin	1943	1946	1946	Binding of 30S ribosomal subunit	Broad-spectrum activity
Chloramphenicols; chloramphenicol	1946	1948	1950	Binding of 50S ribosomal subunit	Broad-spectrum activity
Macrolides; erythromycin	1948	1951	1955	Binding of 50S ribosomal subunit	Broad-spectrum activity
Tetracyclines; chlortetracycline	1944	1952	1950	Binding of 30S ribosomal subunit	Broad-spectrum activity
Rifamycins; rifampicin	1957	1958	1962	Binding of RNA polymerase β-subunit	Gram-positive bacteria
Glycopeptides; vancomycin	1953	1958	1960	Inhibition of cell wall biosynthesis	Gram-positive bacteria
Quinolones; ciprofloxacin	1961	1968	1968	Inhibition of DNA synthesis	Broad-spectrum activity
Streptogramins; streptogramin B	1963	1998	1964	Binding of50S ribosomal subunit	Gram-positive bacteria
Oxazolidinones; linezolid	1955	2000	2001	Binding of 50S ribosomal subunit	Gram-positive bacteria
Lipopetides; daptomycin	1986	2003	1987	Depolarization of cell membrane	Gram-positive bacteria
Fidaxomicin	1948	2011	1977	Inhibition of RNA polymerase	Gram-positive bacteria
Diarylquinolines; bedaquiline	1997	2012	2006	Inhibition of F_1,F_O -ATPase	Narrow-spectrum activity

Timeline of the discovery and introduction of antibiotics (based on Platforms for antibiotic discovery; Kim Lewis; Nature Reviews; 2013).

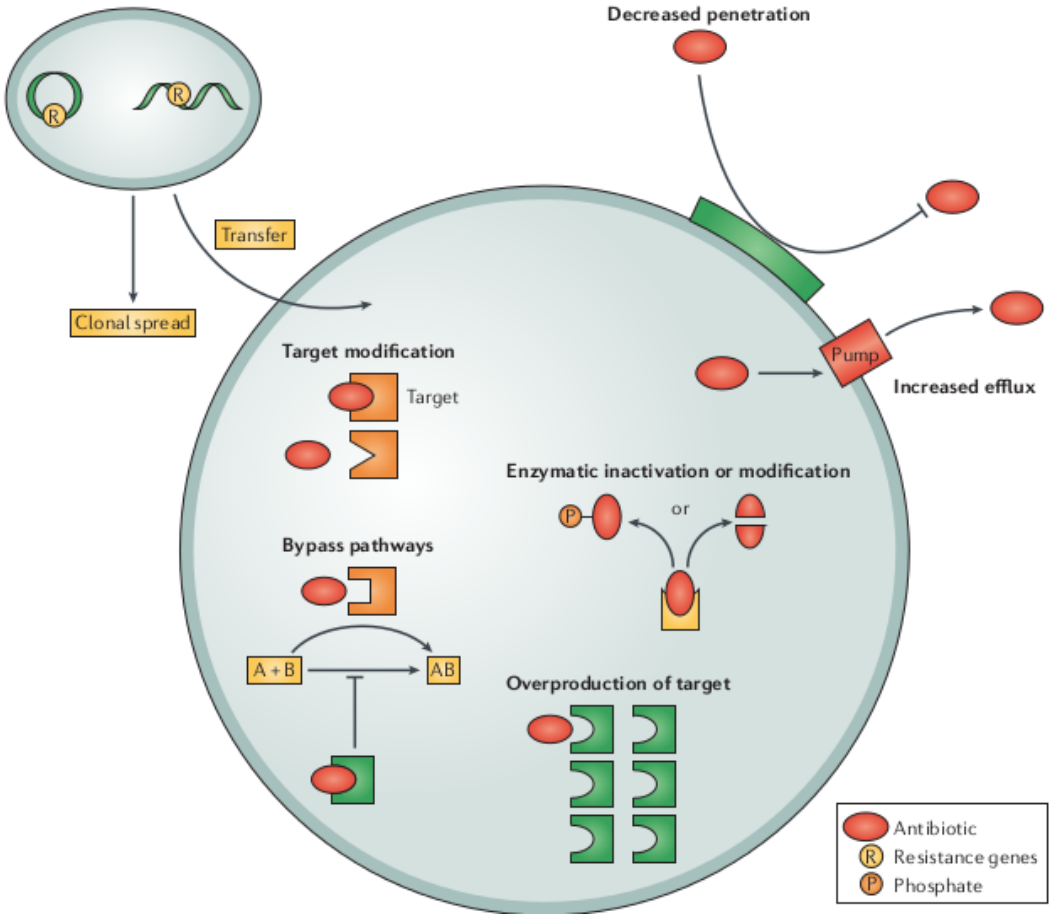
Drug resistance mechanisms I

There are several known drug resistance mechanisms which can be categorized as follows (adopted from: *Wright GD, Chem. Comm., 2011*):

- drug target modification;
- drug molecule modification by specialized enzymes
- reduced accumulation of the drug inside a bacteria cell by decreased cell wall permeability or by pumping out the drug
- alternative metabolic pathways

These drug resistance mechanisms can be acquired either by **chromosomal mutations** or **horizontal gene transfer**.

Drug resistance mechanisms II



Adopted from: Platforms for antibiotic discovery; Kim Lewis; Nature Reviews; 2013

GWAMAR: drug resistance-associated mutations

Goal: identify drug resistance-associated mutations (primary and secondary)

General approach implemented in GWAMAR:

- we use whole-genome comparative approach to identify genetic variations among multiple bacterial strains,
- we retrieve from literature and databases information of the drug resistance phenotypes of the strains,
- we associate the identified mutations with drug resistance-phenotypes based on association scores,
- we propose a new association score, called TGH, which implements scores phylogenetic information.

Genotype and phenotype data

Genotype data

We consider two kinds of genetic variations (determined by eCAMBer based on gene families and their multiple alignments):

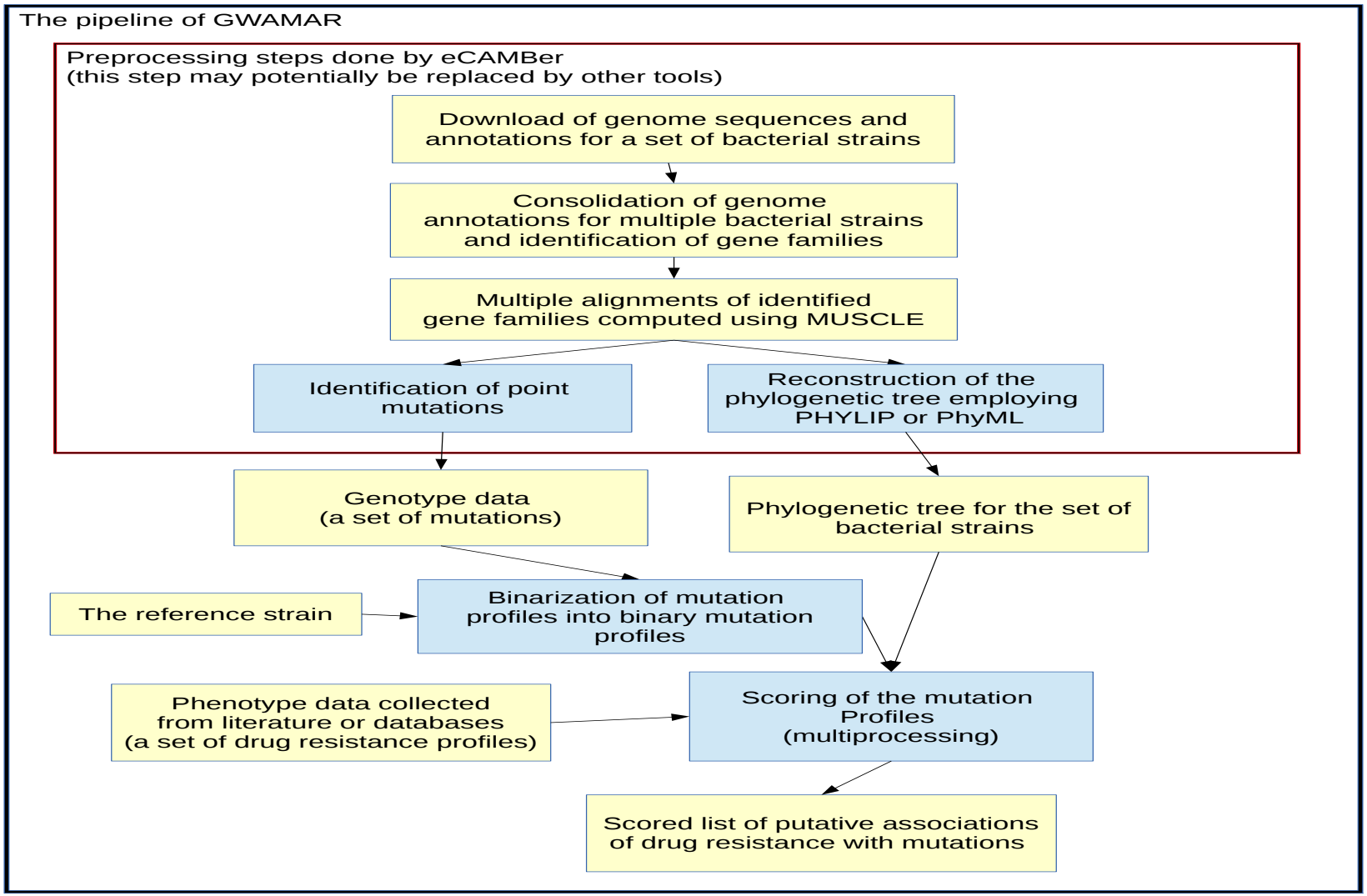
- gene gain/loss,
- amino acid point mutation.

These genetic variations are represented as '0'-'1' vectors (called **mutation profiles**), where '0' denotes the reference state and '1' denotes some change.

Phenotype data (drug susceptibility)

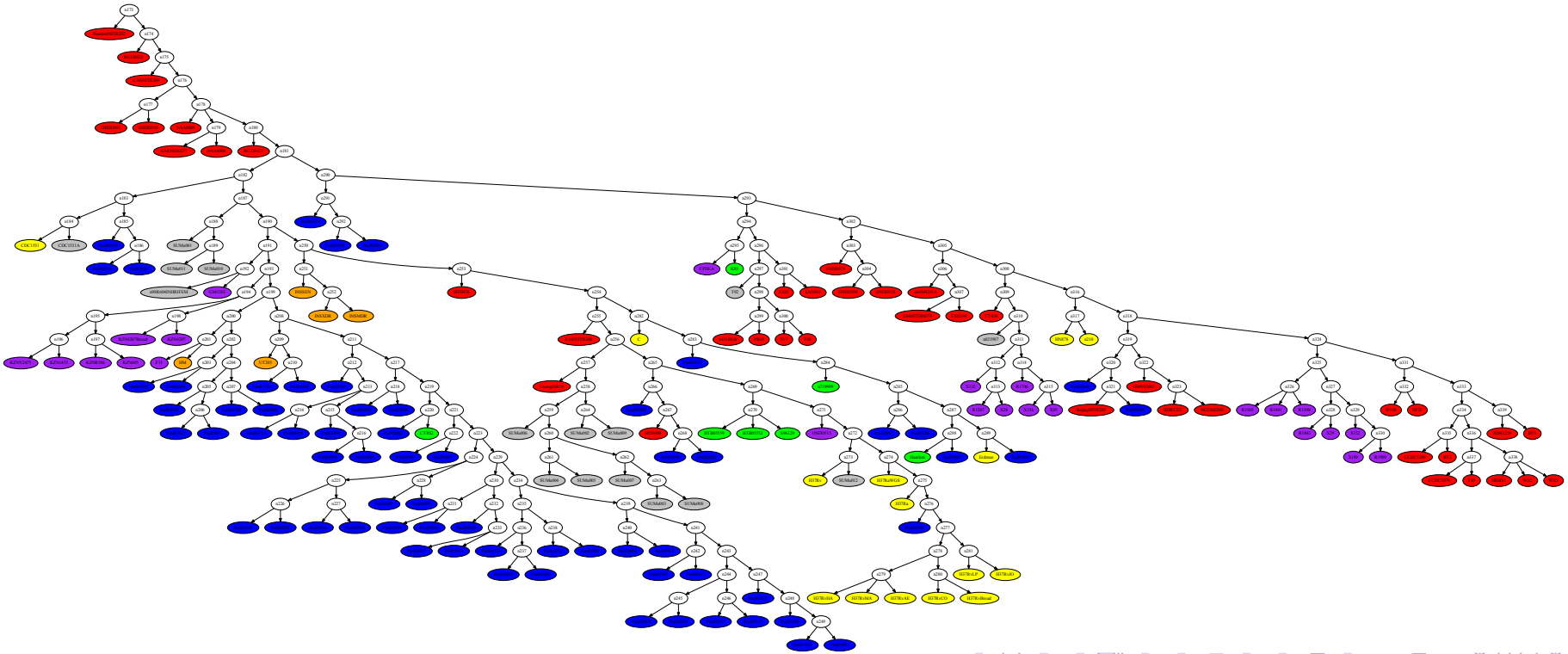
Phenotype data are represented as vectors, called **drug resistance profiles**, with possible states: 'S', 'R', 'I', '?'.

Schema of the framework



Tree-aware scores

We observe that subtrees of the phylogenetic tree very often correspond to geographic locations. Since drug resistance mutations are subject to e volutionary pressure caused by the drug treatment they should be independent of geographic location and therefore be more widely distributed over the tree, as opposed to mutations driven by other environmental factors which tend to rather concentrate in small subtrees.



Classical scores (tree-ignorant) association scores

The classical scores used in genotype-phynotype association studies and co-evolution studies are tree-ignorant.

- odds ratio:

$$\mathrm{OR}(b,r) = \frac{n_1^R \cdot n_0^S}{max(1, n_0^R) \cdot max(1, n_1^S)}$$

- mutual information:

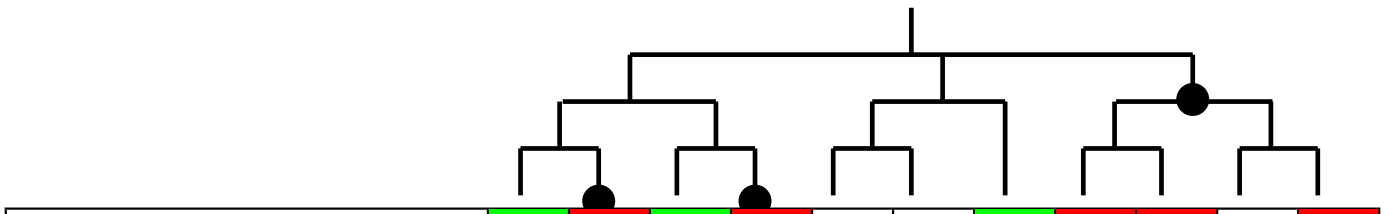
$$\mathrm{MI}(b,r) = \sum_{x \in \{\text{'0','1'}\}} \sum_{y \in \{\text{'S','I','R'}\}} \frac{n_x^y}{n} \cdot log\left(\frac{n_x^y \cdot n}{n_x \cdot n^y}\right)$$

- hypergeometric score

$$\mathrm{H}(b,r) = -log\left(\sum_{i=n^R}^{n} H(n, n^R, n_1, i)\right)$$

Weighted support

Weighted support rewards for drug-resistant strains with the mutation, penalty for drug-susceptible strains with the mutation, where weight $w_T(b, r, i)$ for drug resistant strains is $\frac{1}{k}$, where k denotes the size of the largest subtree with only drug resistant strains.



Drug resistance profile	S	R	S	R	?	?	S	R	R	?	R
Weights	-1.7	1.0	-1.7	1.0	0.0	0.0	-1.7	0.3	0.3	0.0	0.3
mutation_2	0	1	0	1	0	0	0	1	0	0	0
mutation_1	0	0	0	0	0	0	0	1	1	1	1
mutation_3	1	0	0	0	1	0	0	1	1	0	0
...	...										

2.3
1.0
-1.0
...

Weighted support for mutation m is defined as follows:

$$\mathrm{WS}_T(b, r) = \sum_{i \in \mathcal{S}} w_T(b, r, i)[b(i) = \text{'1'}]$$

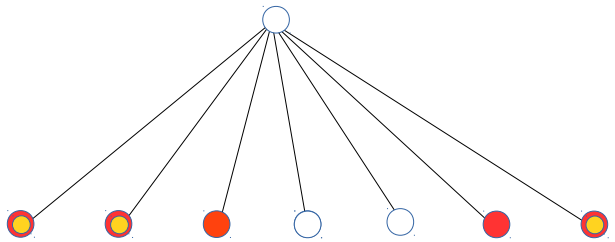
TGH score I

For a given tree T, we call a subset c of its nodes a *coloring*, if it satisfies the following two conditions:

- each path from a leaf to the root contains at most one node from c,
- each internal node in c has a sibling node which does not belong to c.

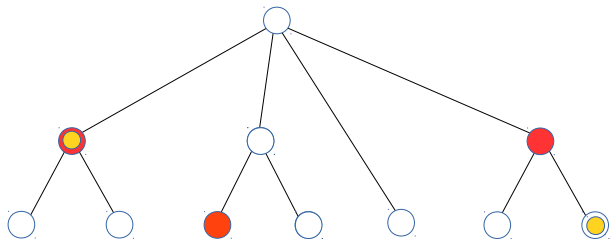
TGH score II

A)



Drug resistance profile	R	R	R	S	S	R	R
Binary mutation profile	1	1	0	0	0	0	1

B)



Drug resistance profile	R	R	R	S	S	R	R
Binary mutation profile	1	1	0	0	0	0	1

(A) an example of coloring $\hat{c}$ induced by a given drug resistance profile (large red nodes) and coloring $\overline{c}$ induced by a given binary mutation profile (small orange nodes) for a flat tree. In this example $|\hat{c}| = 5$, $|\overline{c}| = 3$ and $|L(\hat{c}) \cap \overline{c}| = 3$. (B) another example of colorings $\hat{c}$ and $\overline{c}$ induced by the same pair of profiles but for a different tree. In this example $|\hat{c}| = 3$, $|\overline{c}| = 2$ and $|L(\hat{c}) \cap \overline{c}| = 2$.

TGH score III

We define the TGH score as follows:

$$TGH_T(r, b) = -log(\frac{\sum_{i=k}^{n} B_{T,\hat{c}}(i, n)}{V_T(n)})$$

where:

$$V_T(n) = \#\{c \in C_T : |c| = n\}$$

and:

$$B_{T,\hat{c}}(k, n) = \#\{c \in C_T : |L(\hat{c}) \cap c| = k \text{ and } |c| = n\}$$

GWAMAR implements a dynamic programming approach to calculate the score. The time complexity is $O(D \cdot N^{K-1} \cdot N^2 + D \cdot N \cdot M)$.

Input datasets

We have two datasets of data for *M. tuberculosis*

- 1398 strains with 28 genes sequenced from Broad Institute (mtu_broad)
- 173 fully sequenced strains available in NCBI and PATRIC databases (mtu173)

Genotype data

Point mutation profiles were determined based on gene families identified with *eCAMBer* and their multiple alignments computed with MUSCLE.

Phenotype data (drug susceptibility)

- publications issued together with the fully sequenced genomes;
- other publications found by searching of related literature;
- drug resistance profiles for separate drugs are combined into: Rifampicin, Isoniazid, Fluoroquinolones, Ethambutol, Pyrazinamide, Streptomycin

Collected phenotype data for the mtu173 dataset

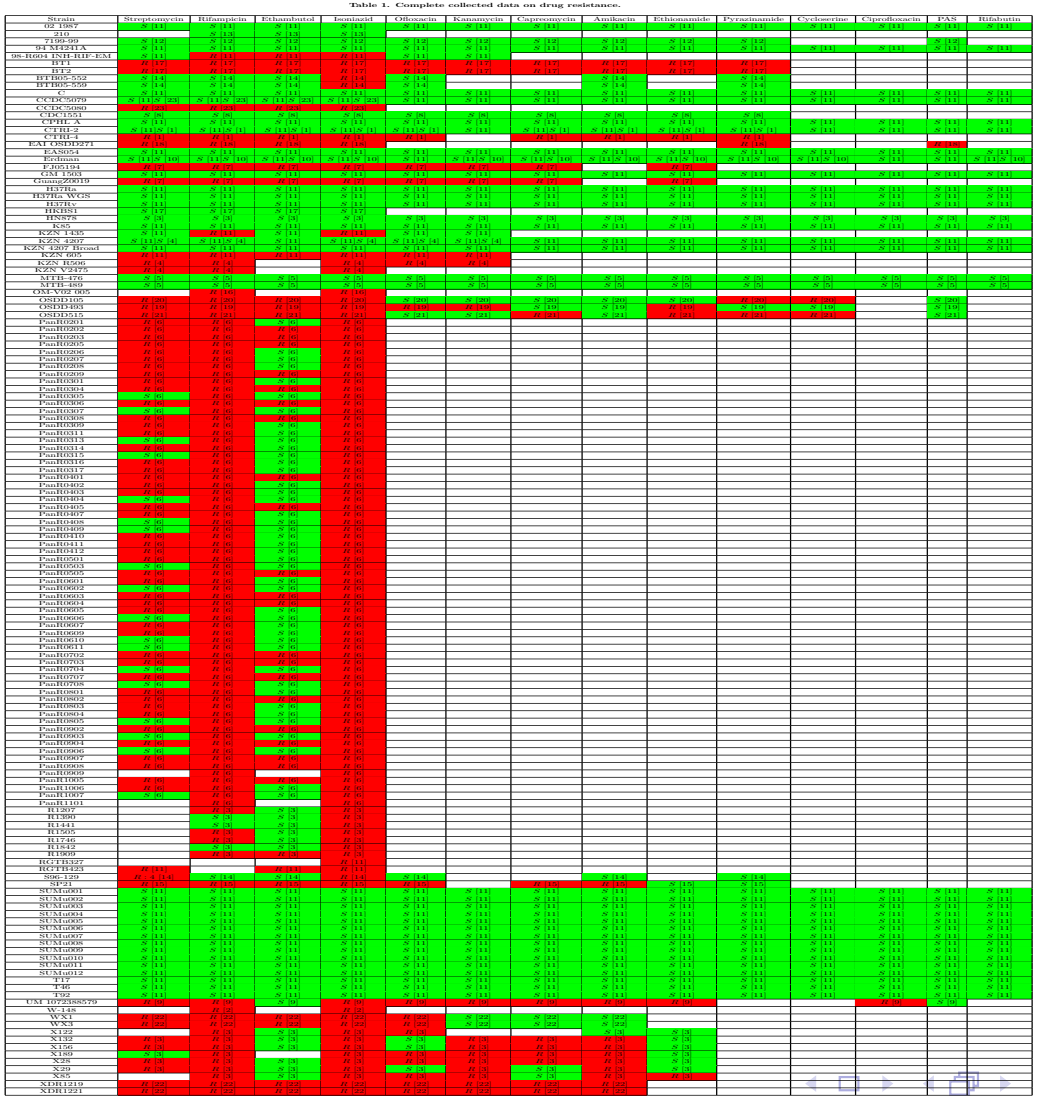
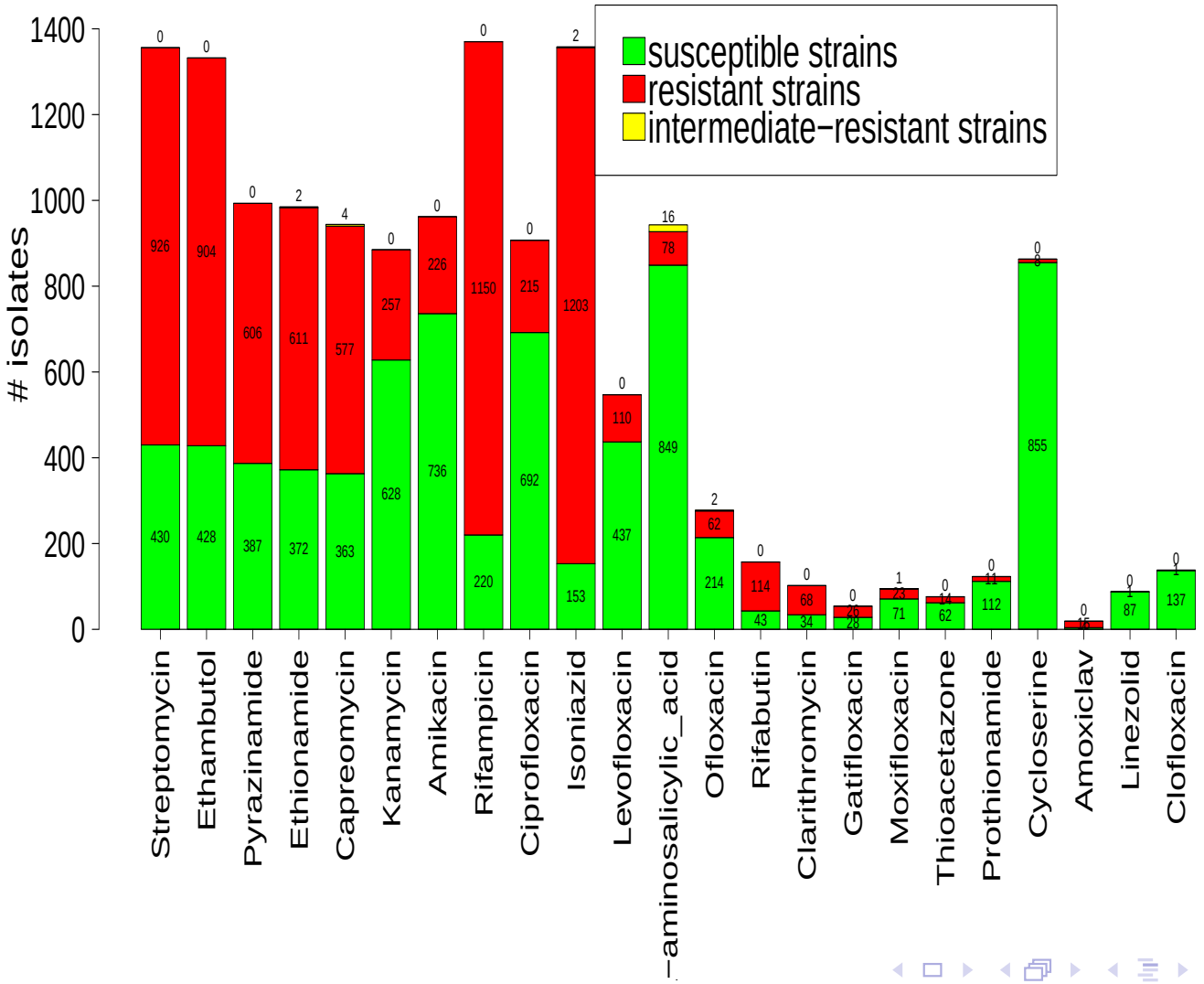


Table 1. Complete collected data on drug resistance.

Strain	Streptomycin	Rifampicin	Ethambutol	Isoniazid	Ofloxacin	Kanamycin	Capreomycin	Amikacin	Ethionamide	Pyrazinamide	Cycloserine	Ciprofloxacin	PAS	Rifabutin
02 1987	S [13]	S [13]	S [13]	S [13]	S [13]	S [13]	S [13]	S [13]	S [13]	S [13]	S [13]	S [13]	S [13]	S [13]
210	S [13]	S [13]	S [13]	S [13]										
7199-99	S [13]	S [13]	S [12]	S [12]	S [12]	S [13]	S [13]	S [13]	S [12]	S [12]			S [12]	
94 M4341A	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]
98-R604 INH-RIF-EMB	S [13]	R [13]	R [13]	R [13]	S [13]	S [13]	S [13]	S [13]	S [13]	S [13]				
BT1	R [17]	R [17]	R [17]	R [17]	R [17]	R [17]	R [17]	R [17]	R [17]	R [17]				
BT2	R [17]	R [17]	R [17]	R [17]	R [17]	R [17]	R [17]	R [17]	R [17]	R [17]				
BT805-552	S [14]	S [14]	S [14]	R [14]	S [14]			S [14]		S [14]				
BT805-559	S [14]	S [14]	S [14]	R [14]	S [14]			S [14]		S [14]				
C	S [13]	S [13]	S [13]	S [13]	S [13]	S [13]	S [13]	S [13]	S [13]	S [13]	S [13]	S [13]	S [13]	S [13]
CCDC5079	S [113 S 63]	S [113 S 23]	S [113 S 63]	S [113 S 23]	S [13]	S [13]	S [13]	S [13]	S [13]	S [13]	S [13]	S [13]	S [13]	S [13]
CCDC5080	R [23]	R [23]	R [23]	R [23]										
CDC1551	S [8]	S [8]	S [8]	S [8]	S [8]	S [8]	S [8]	S [8]	S [8]	S [8]				
CPHL A	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]
CPHL-2	S [11 S 1]	S [11 S 1]	S [11 S 1]	S [11 S 1]	S [11 S 1]	S [11]	S [11 S 1]	S [11 S 1]	S [11 S 1]	S [11 S 1]	S [11]	S [11]	S [11]	S [11]
CPHL-4	R [11]	R [11]	R [11]	R [11]	R [11]		R [11]	R [11]	R [11]	R [11]				
EAI OSD2271	R [18]	R [18]	R [18]	R [18]						R [18]			R [18]	
EA3654	S [11]	S [11]	S [11]		S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]
Erdman	S [11 S 10]	S [11 S 10]	S [11 S 10]	S [11 S 10]	S [11]	S [11 S 10]	S [11 S 10]	S [11 S 10]	S [11 S 10]	S [11 S 10]	S [11 S 10]	S [11]	S [11]	S [11 S 10]
F105194	R [7]	R [7]	R [7]	R [7]	R [7]	R [7]	R [7]	R [7]	R [7]					
GM 1503	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]
Gmag29819	R [7]	R [7]	R [7]	R [7]	R [7]	R [7]	R [7]	R [7]	R [7]	R [7]				
H37Ra	S [1]	S [1]	S [1]	S [1]	S [1]	S [1]	S [1]	S [1]	S [1]	S [1]	S [1]	S [1]	S [1]	S [1]
H37Ra WGS	S [1]	S [1]	S [1]	S [1]	S [1]	S [1]	S [1]	S [1]	S [1]	S [1]	S [1]	S [1]	S [1]	S [1]
H37Rv	S [1]	S [1]	S [1]	S [1]	S [1]	S [1]	S [1]	S [1]	S [1]	S [1]	S [1]	S [1]	S [1]	S [1]
HKBS1	S [17]	S [17]	S [17]	S [17]										
HN878	S [3]	S [3]	S [3]	S [3]	S [3]	S [3]	S [3]	S [3]	S [3]	S [3]	S [3]	S [3]	S [3]	S [3]
K85	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [13]
KZN 1435	S [11]	R [10]	S [11]	R [13]	S [11]	S [11]								
KZN 4207	S [11 S 4]	S [11 S 4]	S [11]	S [11 S 4]	S [11 S 4]	S [11 S 4]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]
KZN 4207 Broad	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]
KZN 605	R [11]	R [11]	S [11]	R [11]	R [11]	R [11]								
KZN 1596	R [4]	R [4]		R [4]	R [4]	R [4]								
KZN V-2475	R [4]	R [4]		R [4]										
MTB-476	S [5]	S [5]	S [5]	S [5]	S [5]	S [5]	S [5]	S [5]	S [5]	S [5]	S [5]	S [5]	S [5]	S [5]
MTB-480	S [5]	S [5]	S [5]	S [5]	S [5]	S [5]	S [5]	S [5]	S [5]	S [5]	S [5]	S [5]	S [5]	S [5]
OSI-V2 985	S [10]	S [10]	S [10]	S [10]										
OSDD105	R [20]	R [20]	R [20]	R [20]	S [20]	S [20]	S [20]	S [20]	S [20]	R [20]	R [20]		S [20]	
OSDD493	R [19]	R [19]	R [19]	R [19]	R [19]		R [19]	S [19]	R [19]	S [19]	S [19]		S [19]	
OSDD515	R [21]	R [21]	R [21]	R [21]	S [21]	S [21]	R [21]	S [21]	R [21]	R [21]	R [21]		S [21]	
PanR0201	R [6]	R [6]	S [6]	R [6]										
PanR0202	R [6]	R [6]	R [6]	R [6]										
PanR0203	R [6]	R [6]	R [6]	R [6]										
PanR0205	R [6]	R [6]	R [6]	R [6]										
PanR0206	R [6]	R [6]	S [6]	R [6]										
PanR0207	R [6]	R [6]	S [6]	R [6]										
PanR0208	R [6]	R [6]	S [6]	R [6]										
PanR0209	R [6]	R [6]	R [6]	R [6]										
PanR0301	R [6]	R [6]	S [6]	R [6]										
PanR0304	R [6]	R [6]	R [6]	R [6]										
PanR0305	S [6]	R [6]	S [6]	R [6]										
PanR0306	R [6]	R [6]	R [6]	R [6]										
PanR0307	S [6]	R [6]	S [6]	R [6]										
PanR0308	R [6]	R [6]	R [6]	R [6]										
PanR0309	R [6]	R [6]	S [6]	R [6]										
PanR0311	R [6]	R [6]	S [6]	R [6]										
PanR0313	S [6]	R [6]	S [6]	R [6]										
PanR0314	R [6]	R [6]	S [6]	R [6]										
PanR0315	S [6]	R [6]	S [6]	R [6]										
PanR0316	R [6]	R [6]	S [6]	R [6]										
PanR0317	R [6]	R [6]	S [6]	R [6]										
PanR0401	R [6]	R [6]	R [6]	R [6]										
PanR0402	R [6]	R [6]	S [6]	R [6]										
PanR0403	R [6]	R [6]	S [6]	R [6]										
PanR0404	S [6]	R [6]	S [6]	R [6]										
PanR0405	R [6]	R [6]	R [6]	R [6]										
PanR0407	R [6]	R [6]	S [6]	R [6]										
PanR0408	S [6]	R [6]	S [6]	R [6]										
PanR0409	S [6]	R [6]	S [6]	R [6]										
PanR0410	R [6]	R [6]	S [6]	R [6]										
PanR0411	R [6]	R [6]	S [6]	R [6]										
PanR0412	R [6]	R [6]	S [6]	R [6]										
PanR0501	R [6]	R [6]	R [6]	R [6]										
PanR0503	S [6]	R [6]	S [6]	R [6]										
PanR0505	S [6]	R [6]	R [6]	R [6]										
PanR0601	R [6]	R [6]	S [6]	R [6]										
PanR0602	S [6]	R [6]	S [6]	R [6]										
PanR0603	R [6]	R [6]	R [6]	R [6]										
PanR0604	R [6]	R [6]	R [6]	R [6]										
PanR0605	R [6]	R [6]	S [6]	R [6]										
PanR0606	S [6]	R [6]	S [6]	R [6]										
PanR0607	R [6]	R [6]	S [6]	R [6]										
PanR0609	R [6]	R [6]	S [6]	R [6]										
PanR0610	S [6]	R [6]	S [6]	R [6]										
PanR0611	S [6]	R [6]	S [6]	R [6]										
PanR0702	R [6]	R [6]	R [6]	R [6]										
PanR0703	R [6]	R [6]	R [6]	R [6]										
PanR0704	S [6]	R [6]	S [6]	R [6]										
PanR0707	R [6]	R [6]	R [6]	R [6]										
PanR0708	S [6]	R [6]	S [6]	R [6]										
PanR0801	R [6]	R [6]	S [6]	R [6]										
PanR0802	R [6]	R [6]	R [6]	R [6]										
PanR0803	R [6]	R [6]	S [6]	R [6]										
PanR0804	R [6]	R [6]	S [6]	R [6]										
PanR0805	S [6]	R [6]	S [6]	R [6]										
PanR0902	R [6]	R [6]	R [6]	R [6]										
PanR0903	S [6]	R [6]	S [6]	R [6]										
PanR0904	S [6]	R [6]	R [6]	R [6]										
PanR0906	S [6]	R [6]	S [6]	R [6]										
PanR0907	R [6]	R [6]	R [6]	R [6]										
PanR0908	R [6]	R [6]	S [6]	R [6]										
PanR0909	R [6]	R [6]		R [6]										
PanR1005	R [6]	R [6]	S [6]	R [6]										
PanR1006	R [6]	R [6]	S [6]	R [6]										
PanR1007	S [6]	R [6]	S [6]	R [6]										
PanR1101		R [6]		R [6]										
RT207		S [3]	S [3]	R [3]										
R1390		S [3]	S [3]	R [3]										
R1441		S [3]	S [3]	R [3]										
R1505		R [3]	S [3]	R [3]										
R1746		R [3]	S [3]	R [3]										
R1842		S [3]	S [3]	R [3]										
R1909		R [13]		R [13]										
RGTB327				R [11]										
RGTB423	R [13]		R [11]	R [11]										
S96-129	R [1 14]	S [14]	S [14]	R [14]	S [14]			S [14]		S [14]				
SP24	R [13]	R [13]	R [13]	R [13]	R [13]		R [13]	R [13]	S [13]	R [13]				
SUMu001	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]
SUMu002	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]
SUMu003	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]
SUMu004	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]
SUMu005	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]
SUMu006	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]
SUMu007	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]
SUMu008	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]
SUMu009	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]
SUMu010	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]
SUMu011	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]
SUMu012	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]
T17	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]
T46	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]
T92	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]	S [11]
UAI 1072388579	R [9]	R [9]	S [9]	R [9]	R [9]	R [9]	R [9]	R [9]	R [9]			R [9]	S [9]	
W-148	R [3]	R [3]		R [3]										
WX1	R [22]	R [22]	R [23]	R [23]	R [23]	S [23]	S [23]	S [23]						
WX3	R [23]	R [23]	R [23]	R [23]	R [23]	S [23]	S [23]	S [23]						
X122	S [3]	S [3]	S [3]	R [3]	S [3]	S [3]		S [3]	S [3]					
X132	R [3]	S [3]	S [3]	R [3]	S [3]	R [3]	R [3]	R [3]	S [3]					
X156	R [3]	S [3]		R [3]	S [3]	R [3]	R [3]	R [3]	S [3]					
X189	S [3]	R [3]		R [3]	R [3]	R [3]	R [3]	R [3]	S [3]					
X28	R [3]	R [3]	S [3]	R [3]	S [3]	R [3]	R [3]	R [3]	S [3]					
X29	R [3]	R [3]	S [3]	R [3]	S [3]	R [3]	S [3]	S [3]	S [3]					
X85		R [3]	S [3]	R [3]	R [31]	R [3]	S [3]		R [31]					
XDR1219	R [23]	R [22]	R [22]	R [22]	R [22]	R [22]	R [23]	R [22]						
XDR1221	R [23]	R [23]	R [22]	R [22]	R [22]	R [23]	R [23]	R [22]						

Phenotype data for the mtu_broad dataset

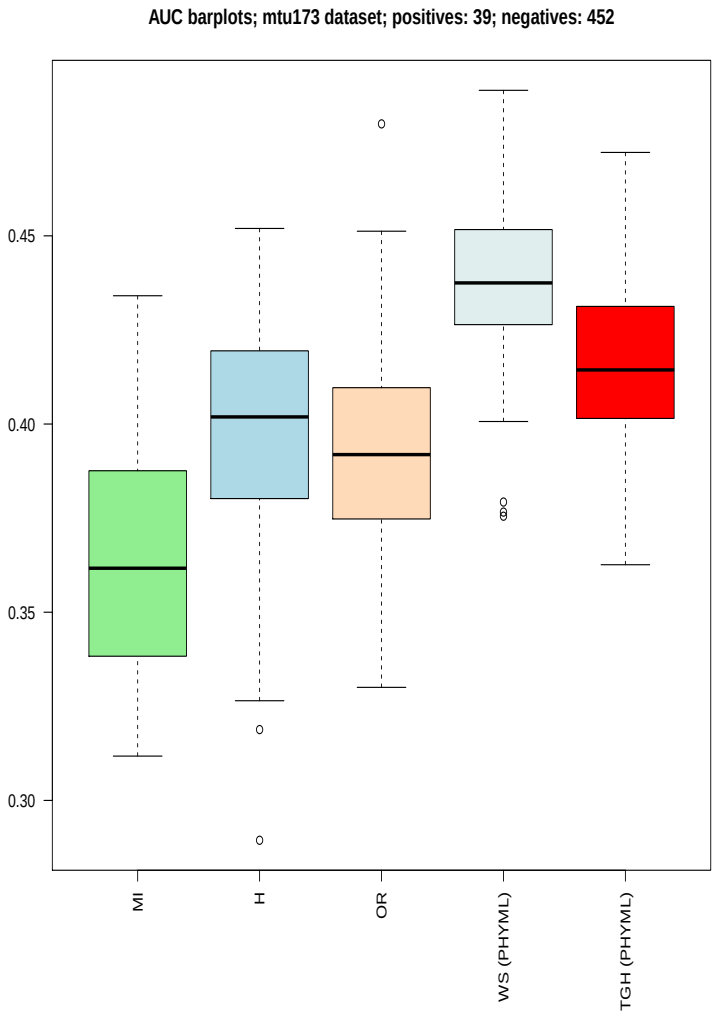


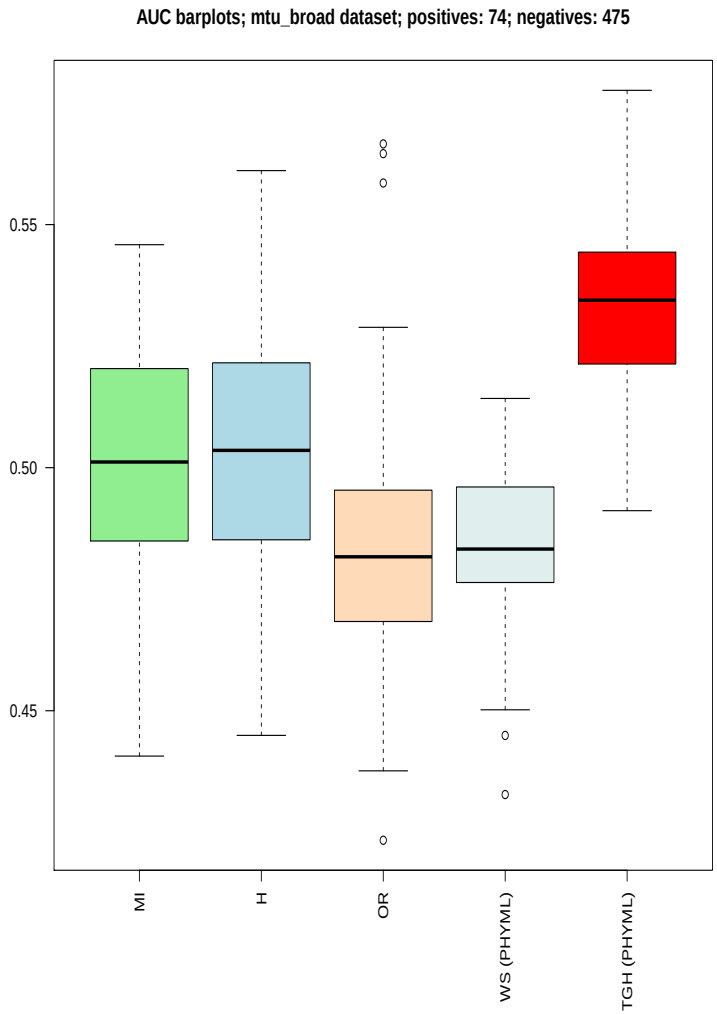
Gold standard associations

We retrieve the gold standard associations from the TBDreamDB database for: Rifampicin, Isoniazid, Fluoroquinolones, Ethambutol, Pyrazinamide, Streptomycin.

drug name	gene name	positions
Fluoroquinolones	gyrA	90,91,94,102,126
	gyrB	538
Ethambutol	embB	306,406,497
Isoniazid	ahpC	-46,-39,21
	fabG1-inhA	-15,-8
	kasA	269
	katG	315
Isoniazid	ahpC	-46,-39,21
Rifampicin	rpoB	432,435,441,445,450,452
Streptomycin	rpsL	43,88
	rrs	492,513,514,517,907
Pyrazinamide	pncA	-11,7,10(60 in total)

Comparison on *mtu173* and *mtu_broad* datasets





Top-scoring mutations on the *mtu173* dataset

drug name	gene id	gene name	mutation	all	h.c.	TGH
Fluoroquinolones	Rv0006	gyrA	D94G/A/H/N/Y	Y	Y	14.1474612056
Rifampicin	Rv0667	rpoB	S450L	Y	Y	13.7718163314
Isoniazid	Rv1908c	katG	S315T/G/N	Y	Y	11.3299511445
Streptomycin	Rv0682	rpsL	K43R	Y	Y	8.48207709024
Rifampicin	Rv0667	rpoB	D435F/V/Y/G/A	Y	Y	6.27518366477
Ethambutol	Rv3795	embB	M306I/V/L	Y	Y	5.86457315376
Isoniazid	Rv1483	fabG1	C-15T	Y	Y	5.54308155286
Streptomycin	Rv0682	rpsL	K88R/M	Y	Y	4.2132021894
Ethambutol	Rv3795	embB	E504G/D	N	N	3.33044011637
Rifampicin	Rv0667	rpoB	H445D/Y/R	Y	Y	2.71070582544
Pyrazinamide	Rv2043c	pncA	H51P	Y	Y	2.7080502011
Pyrazinamide	Rv2043c	pncA	W68L	Y	Y	2.7080502011
Fluoroquinolones	Rv0006	gyrA	A90V/G	Y	Y	2.50721373928
Fluoroquinolones	Rv0005	gyrB	N538T	Y	Y	1.82454929205
Fluoroquinolones	Rv0006	gyrA	S91P	Y	Y	1.82454929205
Fluoroquinolones	Rv0005	gyrB	T539I	N	N	1.82454929205
Ethambutol	Rv3795	embB	D869G	N	N	1.67147330335
Ethambutol	Rv3795	embB	A505T	N	N	1.67147330335
Ethambutol	Rv3795	embB	D1024N	Y	N	1.67147330335
Rifampicin	Rv0667	rpoB	A692T	N	N	1.65151195989
Isoniazid	Rv1483	fabG1	T-8A/C	Y	Y	1.36702851766
Streptomycin	Rvnr01	rrs	A1401G	Y	N	1.31458509629
Rifampicin	Rv0667	rpoB	I1106T	N	N	1.11729579263
Rifampicin	Rv0667	rpoB	E250G	N	N	1.11729579263
Rifampicin	Rv0667	rpoB	L452P	Y	Y	1.11729579263

Top-scoring mutations on the *mtu_broad* dataset

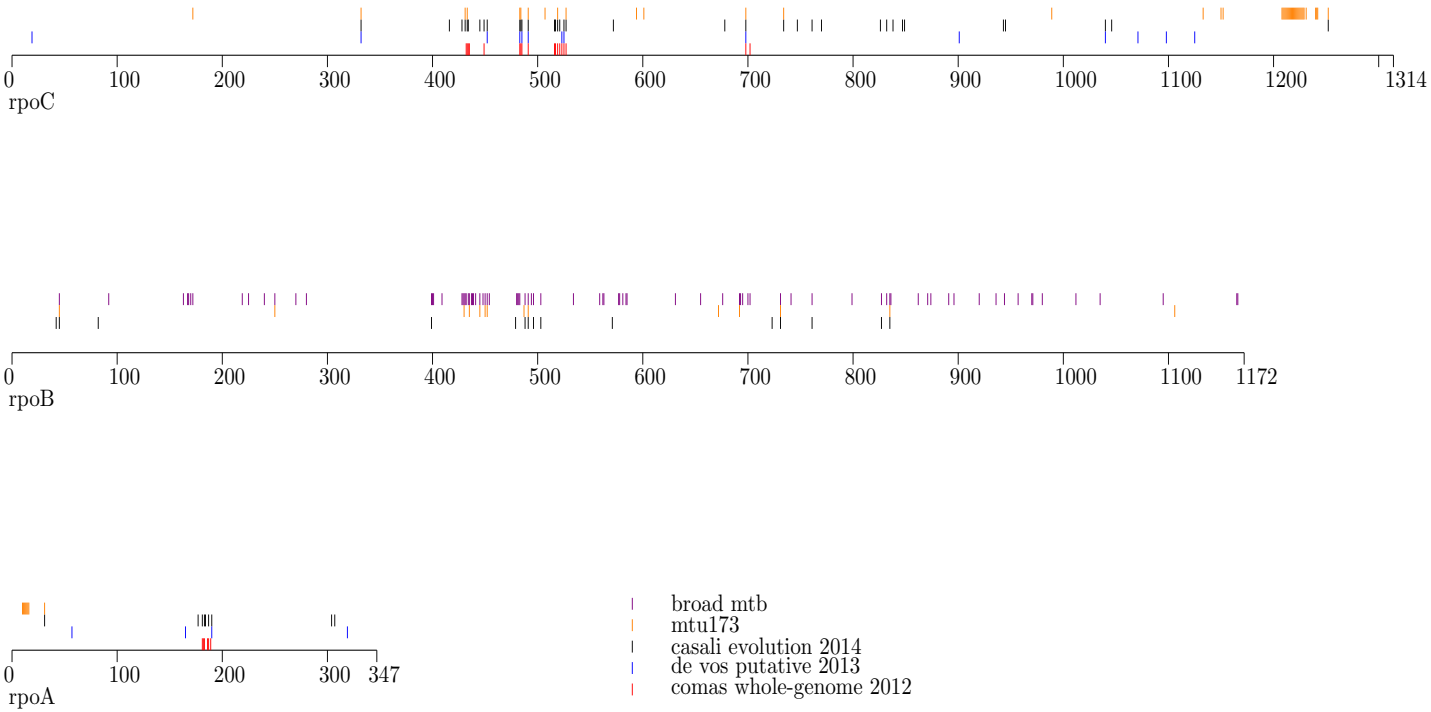
drug name	gene id	gene name	mutation	all	h.c.	TGH
Fluoroquinolones	Rv0006	gyrA	D94A/G/N/Y/H	Y	Y	138.414507992
Ethambutol	Rv3795	embB	M306V/I/L/T	Y	Y	75.677195766
Rifampicin	Rv0667	rpoB	S450L/W	Y	Y	73.8586969047
Fluoroquinolones	Rv0006	gyrA	A90G/V	Y	Y	41.3671629169
Isoniazid	Rv1908c	katG	S315T/G/N/I/R	Y	Y	31.2319396693
Streptomycin	Rv0682	rpsL	K88T/R/Q/M	Y	Y	22.7910550626
Ethambutol	Rv3795	embB	Q497P/R/K/H	Y	Y	18.8920918506
Streptomycin	Rv0682	rpsL	K43R	Y	Y	15.0916682608
Fluoroquinolones	Rv0005	gyrB	N538K/T/D/S	Y	Y	14.6924139234
Rifampicin	Rv0667	rpoB	H445P/D/R/Y/L/N/Q	Y	Y	13.0077710882
Pyrazinamide	Rv2043c	pncA	T135P/A	Y	N	9.80925396944
Pyrazinamide	Rv2043c	pncA	T-11C/G	Y	Y	9.01421480751
Ethambutol	Rv3795	embB	G406S/D/A/C	Y	Y	8.05127512926
Fluoroquinolones	Rv0006	gyrA	S91P	Y	Y	7.62877200647
Streptomycin	Rvnr01	rrs	A1401G	Y	N	7.62190832659
Rifampicin	Rv0667	rpoB	D435Y/V/H/G/A/N	Y	Y	7.54687538111
Streptomycin	Rvnr01	rrs	A514C	Y	Y	7.04416442854
Pyrazinamide	Rv2043c	pncA	Q10R/P/H	Y	Y	6.79992327959
Fluoroquinolones	Rv0006	gyrA	D89G/N	Y	N	6.3023070068
Ethambutol	Rv3795	embB	D328Y/G/H	Y	N	5.93350968238
Pyrazinamide	Rv2043c	pncA	V139G/A/M/L	Y	Y	5.67144832937
Isoniazid	Rv1483	fabG1	C-15T	Y	Y	5.5292894594
Pyrazinamide	Rv2043c	pncA	T76P/I	Y	Y	5.22287436396
Pyrazinamide	Rv2043c	pncA	D12G/A/E/N	Y	Y	5.10653842681
Ethambutol	Rv3795	embB	D354A/G	N	N	5.02061345733

Putative compensatory mutations

Recent publications reporting putative compensatory mutations in *M. tuberculosis*:

- Whole-genome sequencing of rifampicin-resistant Mycobacterium tuberculosis strains identifies compensatory mutations in RNA polymerase genes; Nature Genetics; 2012
- Putative Compensatory Mutations in the rpoC Gene of Rifampin-Resistant Mycobacterium tuberculosis Are Associated with Ongoing Transmission; Antimicrobial Agents and Chemotherapy; 2013
- Evolution and transmission of drug-resistant tuberculosis in a Russian population; Nature Genetics; 2014

Putative compensatory mutations



Interestingly, several mutations identified by GWAMAR that has also been reported in at least one of the papers.

- rpoA: G31S/A
- rpoB: P45S/L, L731P, E761D, R827C, H835P/R
- rpoC: G332R/S, V431M, G433C/S, V483G/A, W484G, I491T/V, L527V, N698K, A734V

Summary

- The fast growing number of fully sequenced bacterial strains enables us to develop and test new methods to identifying drug resistance associated genes and mutations.
- We developed and implemented GWAMAR – a new framework for detection of drug resistance-associated mutations. This software is available at the project website: `http://bioputer.mimuw.edu.pl/gwamar/`.
- We proposed a new association score, called TGH, which employ phylogenetic information. It outperforms the standard tree-ignorant scores, but is more computationally expensive.
- Applying our approach we identified some novel putative drug resistance-associated mutations.
- Future possible direction of research may include: classification of mutations into primary and secondary, grouping of mutations which are close together, incorporation of PPI networks.

Thank you

Thank you!

You are welcome to give comments or ask questions.