Antimicrobial Resistance Prediction in PATRIC and RAST
Supplemental Information

James J. Davis[*1,2], Sébastien Boisvert[3], Thomas Brettin[1,2], Ronald W. Kenyon[4], Chunhong Mao[4], Robert Olson[1,2], Ross Overbeek[2,5], John Santerre[6], Maulik Shukla[1,2], Alice R. Wattam[4], Rebecca Will[4], Fangfang Xia[1,2], Rick Stevens[1,2,6]

[1] Computation Institute, University of Chicago, Chicago, Illinois, 60637, USA
[2] Computing, Environment and Life Sciences, Argonne National Laboratory, Argonne IL, 60439, USA
[3] Gydle Inc., 1332 Avenue du Chanoine Morel, suite 101, Québec, QC, G1S 4B4, Canada
[4] Biocomplexity Institute, Virginia Tech University, Blacksburg, VA 24060, USA.
[5] Fellowship for Interpretation of Genomes, Burr Ridge, IL, 60527, USA
[6] Department of Computer Science, University of Chicago, Chicago, Illinois, 60637, USA

Supplementary Tables

Table S1. The number of *M. tuberculosis* genomes available in PATRIC with distinct AMR phenotypes. Genomes for which the phenotype is unknown or intermediate are depicted by a dash.

Genomes	Ethambutol	Ethionamide	Isoniazid	Kanamycin	Ofloxacin	Rifampicin	Streptomycin
1	—	—	R	—	—	R	R
1	—	—	R	—	R	R	R
1	—	—	R	R	S	R	S
1	—	—	S	S	S	S	—
1	—	R	—	R	R	—	—
1	—	R	—	S	R	—	—
1	—	R	—	S	R	R	S
1	—	R	R	R	R	R	—
1	—	R	R	R	S	R	S
1	—	R	R	S	R	R	—
1	—	R	R	S	R	R	R
1	—	R	R	S	S	R	S
1	—	S	—	S	S	R	S
1	—	S	R	—	R	R	R
1	—	S	R	S	R	R	S
1	—	S	R	S	S	—	—
1	—	S	R	S	S	S	S
1	R	—	R	—	S	R	S
1	R	—	R	S	R	R	S
1	R	R	—	R	R	R	S
1	R	R	—	R	S	R	S
1	R	R	R	—	S	R	R
1	R	R	R	R	—	R	R
1	R	R	R	R	R	S	R
1	R	R	R	R	S	—	S
1	R	R	R	R	S	S	S
1	R	R	R	S	S	S	R
1	R	R	S	S	S	R	R
1	R	S	—	—	R	S	S
1	R	S	R	—	R	R	—
1	R	S	R	—	S	R	—
1	R	S	R	R	—	—	R
1	R	S	R	R	R	R	S
1	R	S	R	R	S	R	R
1	R	S	R	S	R	R	R
1	R	S	R	S	S	R	—
1	R	S	S	R	R	R	R
1	R	S	S	S	S	R	R
1	S	—	R	—	—	S	—
1	S	—	R	—	R	R	R
1	S	—	R	—	S	S	R
1	S	—	R	R	R	R	R
1	S	—	R	R	S	R	S

1	S	—	R	S	R	R	S
1	S	—	R	S	S	S	R
1	S	R	—	S	R	R	S
1	S	R	—	S	S	R	S
1	S	R	R	—	R	R	R
1	S	R	R	—	R	S	—
1	S	R	R	S	S	S	S
1	S	R	S	R	S	S	S
1	S	R	S	S	S	S	R
1	S	S	—	—	R	S	S
1	S	S	—	S	S	S	S
1	S	S	R	—	R	S	S
1	S	S	R	R	R	R	S
1	S	S	R	R	S	R	S
1	S	S	R	S	R	S	R
1	S	S	R	S	S	—	S
1	S	S	S	S	R	R	S
1	S	S	S	S	S	R	—
1	S	S	S	S	S	R	R
2	—	—	R	R	S	S	S
2	—	—	R	S	R	S	R
2	—	R	—	R	R	R	S
2	—	R	R	R	R	R	R
2	—	S	R	R	R	R	S
2	—	S	R	S	S	R	R
2	R	—	R	—	R	R	R
2	R	—	R	R	S	R	S
2	R	R	—	R	R	R	R
2	R	R	R	—	—	R	R
2	R	R	S	S	R	R	R
2	R	S	R	R	S	R	S
2	R	S	S	S	S	S	S
2	S	—	R	R	R	R	S
2	S	—	R	S	S	S	S
2	S	—	S	—	—	R	S
2	S	R	R	—	S	R	—
2	S	R	R	S	S	R	S
2	S	S	R	—	—	R	R
2	S	S	R	—	S	R	R
2	S	S	R	R	S	R	R
2	S	S	R	R	S	S	S
2	S	S	R	S	R	R	R
2	S	S	S	—	—	S	S
2	S	S	S	R	S	R	R
2	S	S	S	R	S	S	S
2	S	S	S	S	R	S	S
3	—	—	R	R	R	R	S
3	—	—	S	S	S	R	S
3	—	R	R	R	R	R	S
3	—	R	R	S	R	R	S
3	—	S	R	S	S	R	—
3	R	—	R	—	—	R	—

3	R	R	R	R	S	R	—
3	R	S	R	R	R	R	R
3	S	—	R	—	—	—	S
3	S	—	S	—	—	—	R
3	S	—	S	—	—	S	—
3	S	R	S	S	S	S	S
3	S	S	R	S	R	R	S
3	S	S	S	S	S	S	—
4	—	—	R	S	S	S	S
4	—	S	R	—	R	R	S
4	R	—	R	—	—	S	R
4	R	—	R	—	—	S	S
4	R	R	R	S	S	R	S
4	R	S	R	R	S	—	R
4	R	S	R	S	R	—	R
4	R	S	R	S	S	—	R
4	S	—	R	—	—	R	—
4	S	—	S	S	S	R	S
4	S	R	—	S	S	S	S
4	S	S	R	—	S	R	—
5	—	—	R	R	S	R	R
5	R	—	R	—	—	R	S
5	R	—	R	R	S	R	R
5	R	—	R	S	R	R	R
5	R	R	R	S	R	R	S
5	S	S	S	S	S	S	R
6	—	—	R	S	S	S	R
6	—	—	S	S	S	S	S
6	—	S	S	S	S	S	S
6	R	—	R	S	S	R	S
6	R	R	R	R	S	R	S
6	S	S	R	—	S	R	S
7	—	—	R	S	R	R	S
7	R	R	R	R	S	R	R
7	R	S	R	S	S	R	S
7	S	S	R	S	S	R	R
7	S	S	R	S	S	S	R
7	S	S	S	—	R	S	S
8	—	S	R	S	S	R	S
8	R	R	R	S	S	R	R
8	S	—	R	—	—	R	R
8	S	S	S	S	S	R	S
9	S	—	R	—	S	R	R
10	—	—	R	S	S	R	S
10	S	—	S	—	—	S	R
12	R	R	R	R	R	—	R
12	R	R	R	R	R	R	S
12	R	R	R	S	R	R	R
12	R	S	R	S	S	R	R
13	R	—	R	R	R	R	R
14	S	S	R	S	S	R	S
16	—	—	R	R	R	R	R

16	S	—	R	—	—	R	S
16	S	—	R	—	—	S	R
16	S	S	R	S	S	S	S
17	—	—	R	S	S	R	R
17	R	—	R	—	S	R	R
17	S	—	S	—	—	—	S
17	S	S	S	—	S	S	S
18	S	—	R	—	—	—	R
18	S	—	R	S	S	R	S
19	R	—	R	S	S	R	R
23	S	—	R	—	—	S	S
26	—	—	R	S	R	R	R
27	R	—	R	—	—	R	R
34	R	—	R	—	—	—	R
47	S	S	S	S	S	S	S
48	S	—	R	S	S	R	R
53	R	R	R	R	R	R	R
68	S	—	S	S	S	S	S
103	—	—	R	—	—	R	—
220	S	—	S	—	—	S	S

Table S2. The correlations between AMR pheotype profiles for *M. tuberculosis* genomes. For each antibiotic the correlations between AMR phenotypes is shown. Columns show correlations for subsets of genomes that were chosen to reduce the overall correlation between AMR profiles.

Antibiotic 1	Antibiotic 2	All available genomes*	<= 250 genomes	<= 200 genomes	<= 150 genomes	<= 100 genomes
Ethambutol						
	Ethambutol	1	1	1	1	1
	Ethionamide	0.356	0.184	0.014	-0.041	-0.237
	Isoniazid	0.570	0.194	0.120	-0.060	-0.091
	Kanamycin	0.289	0.094	-0.004	-0.055	-0.385
	Ofloxacin	0.283	0.056	-0.069	-0.152	-0.388
	Rifampin	0.559	0.242	0.166	0.005	0.081
	Streptomycin	0.516	0.173	0.034	-0.144	-0.141
Ethionamide						
	Ethambutol	0.356	0.618	0.57	0.466	0.216
	Ethionamide	1	1	1	1	1
	Isoniazid	0.191	0.388	0.274	0.113	-0.191
	Kanamycin	0.379	0.508	0.456	0.368	0.113
	Ofloxacin	0.405	0.542	0.497	0.379	0.192
	Rifampin	0.219	0.428	0.333	0.162	-0.100
	Streptomycin	0.213	0.367	0.299	0.139	-0.163
Isoniazid						
	Ethambutol	0.570	0.328	0.347	0.228	0.141
	Ethionamide	0.191	-0.481	-0.532	-0.580	-0.676
	Isoniazid	1	1	1	1	1
	Kanamycin	0.131	-0.659	-0.694	-0.642	-0.755
	Ofloxacin	0.155	-0.703	-0.737	-0.680	-0.757
	Rifampin	0.746	0.611	0.566	0.427	0.429
	Streptomycin	0.590	0.389	0.270	0.113	-0.077
Kanamycin						
	Ethambutol	0.289	0.347	0.305	0	-0.219
	Ethionamide	0.379	0.331	0.272	0.146	-0.173
	Isoniazid	0.131	-0.064	-0.083	-0.088	-0.089
	Kanamycin	1	1	1	1	1
	Ofloxacin	0.514	0.386	0.330	0.129	-0.207
	Rifampin	0.115	-0.058	-0.144	-0.155	-0.135
	Streptomycin	0.147	-0.037	-0.136	-0.161	-0.346
Ofloxacin						
	Ethambutol	0.283	0.194	0.068	-0.328	-0.618
	Ethionamide	0.405	0.268	0.111	-0.053	-0.291
	Isoniazid	0.155	-0.119	-0.178	-0.236	-0.287
	Kanamycin	0.514	0.356	0.232	-0.042	-0.355
	Ofloxacin	1	1	1	1	1
	Rifampin	0.158	-0.066	-0.176	-0.148	-0.242
	Streptomycin	0.185	-0.061	-0.200	-0.207	-0.328
Rifampin						
	Ethambutol	0.559	0.280	0.201	0.207	-0.023

	Ethionamide	0.219	-0.356	-0.427	-0.489	-0.553
	Isoniazid	0.746	0.617	0.524	0.370	0.023
	Kanamycin	0.115	-0.637	-0.664	-0.712	-0.610
	Ofloxacin	0.158	-0.654	-0.694	-0.711	-0.633
	Rifampin	1	1	1	1	1
	Streptomycin	0.506	0.306	0.219	0.022	-0.324
Streptomycin						
	Ethambutol	0.516	0.005	-0.128	-0.410	-0.664
	Ethionamide	0.213	-0.189	-0.293	-0.366	-0.488
	Isoniazid	0.590	0.165	-0.046	-0.193	-0.308
	Kanamycin	0.147	-0.279	-0.376	-0.443	-0.567
	Ofloxacin	0.185	-0.354	-0.405	-0.493	-0.628
	Rifampin	0.506	0.035	-0.108	-0.223	-0.384
	Streptomycin	1	1	1	1	1

*As displayed in Table 1 of the main text.

Table S3. Examples of the top three distinguishing k-mers for rifampicin classifiers built from genome sets ranging from 100 to 300 susceptible and resistant genomes, where the set was chosen to reduce the correlation between rifampin resistance and resistance to other antibiotics (from Supplementary Table S2). Data are shown for *M. tuberculosis* H37Rv and k-mer matches have at least 90% identity.

Number of k-mers with an identical pattern	Corresponding protein-encoding gene	PATRIC/RAST annotation
100 genomes		
25	fig|83332.1.peg.3201	NADH pyrophosphatase (EC 3.6.1.22)
1	fig|83332.1.peg.667	DNA-directed RNA polymerase beta subunit (EC 2.7.7.6)
4	fig|83332.1.peg.1590	hypothetical protein Rv1588c
150 genomes		
1	fig|83332.1.peg.667	DNA-directed RNA polymerase beta subunit (EC 2.7.7.6)
1	fig|83332.1.peg.747	PE-PGRS family protein
1	fig|83332.1.peg.1910	Catalase-peroxidase KatG (EC 1.11.1.21)
200 genomes		
1	fig|83332.1.peg.667	DNA-directed RNA polymerase beta subunit (EC 2.7.7.6)
2	fig|83332.1.peg.2636	PE-PGRS family protein
1	fig|83332.1.peg.1910	Catalase-peroxidase KatG (EC 1.11.1.21)
250 genomes		
1	fig|83332.1.peg.667	DNA-directed RNA polymerase beta subunit (EC 2.7.7.6)
2	fig|83332.1.peg.2636	PE-PGRS family protein
1	fig|83332.1.peg.1910	Catalase-peroxidase KatG (EC 1.11.1.21)
300 genomes		
1	fig|83332.1.peg.1910	Catalase-peroxidase KatG (EC 1.11.1.21)
3	fig|83332.1.peg.746	PE-PGRS family protein
1	fig|83332.1.peg.667	DNA-directed RNA polymerase beta subunit (EC 2.7.7.6)

Table S4. The AMR profiles of resistant genomes used to create the combined multidrug-resistance classifier for *Mycobacterium tuberculosis.* Genomes with intermediate or unknown phenotypes are depicted by a dash.

Genomes	Ethambutol	Ethionamide	Isoniazid	Kanamycin	Ofloxacin	Rifampin	Streptomycin
2	—	R	R	R	R	R	R
13	R	—	R	R	R	R	R
2	R	R	—	R	R	R	R
1	R	R	R	R	—	R	R
12	R	R	R	R	R	—	R
53	R	R	R	R	R	R	R

Table S5. The AMR profiles of susceptible genomes used to create the combined multidrug-resistance classifier for *Mycobacterium tuberculosis.* Genomes with intermediate or unknown phenotypes are depicted by a dash.

Genomes	Ethambutol	Ethionamide	Isoniazid	Kanamycin	Ofloxacin	Rifampin	Streptomycin
6	—	S	S	S	S	S	S
68	S	—	S	S	S	S	S
1	S	S	—	S	S	S	S
17	S	S	S	—	S	S	S
1	S	S	S	S	S	S	—
46	S	S	S	S	S	S	S

Table S6. A description of the top ten k-mers found by AdaBoost for the combined *M. tuberculosis* pan-resistance classifier and their corresponding genomic regions in *M. tuberculosis* TKK_02_0002, TKK_03_0024, TKK-01-0023, H37Rv and KT-0099. Genomes were chosen as examples with exact k-mer matches. The complete list of k-mers is described in the supplementary data file online.

Rank	α-value	k-mers with an identical pattern	corresponding genes	PATRIC annotation
1	1.374	1	fig|1397854.3.peg.2114	Catalase (EC 1.11.1.6) / Peroxidase (EC 1.11.1.7)
2	0.709	31	fig|1397854.3.rna.19	Small Subunit Ribosomal RNA
3	0.800	7	fig|1448395.3.peg.4357	hypothetical protein
4	0.643	31	fig|1397854.3.peg.744	DNA-directed RNA polymerase beta subunit (EC 2.7.7.6)
5	0.630	1	fig|1448395.3.peg.1856	putative cellulose-binding protein
6	0.556	5	fig|1397854.3.peg.1633	Possible regulatory protein Trx
7	0.643	14	fig|1397854.3.peg.9	DNA gyrase subunit A (EC 5.99.1.3)
8	0.531	3	intergenic region	Between fig|1267359.3.peg.43, hypothetical protein and fig|1267359.3.peg.44, hypothetical protein
9	0.532	11	intergenic region	Between fig|83332.12.peg.3135 Type II secretory pathway, component ExeA and fig|83332.12.peg.3136 hypothetical protein
10	0.473	31	fig|1400933.3.peg.3985	Integral membrane indolylacetylinositol arabinosyltransferase EmbB (EC 2.4.2.-)

Supplementary Figures

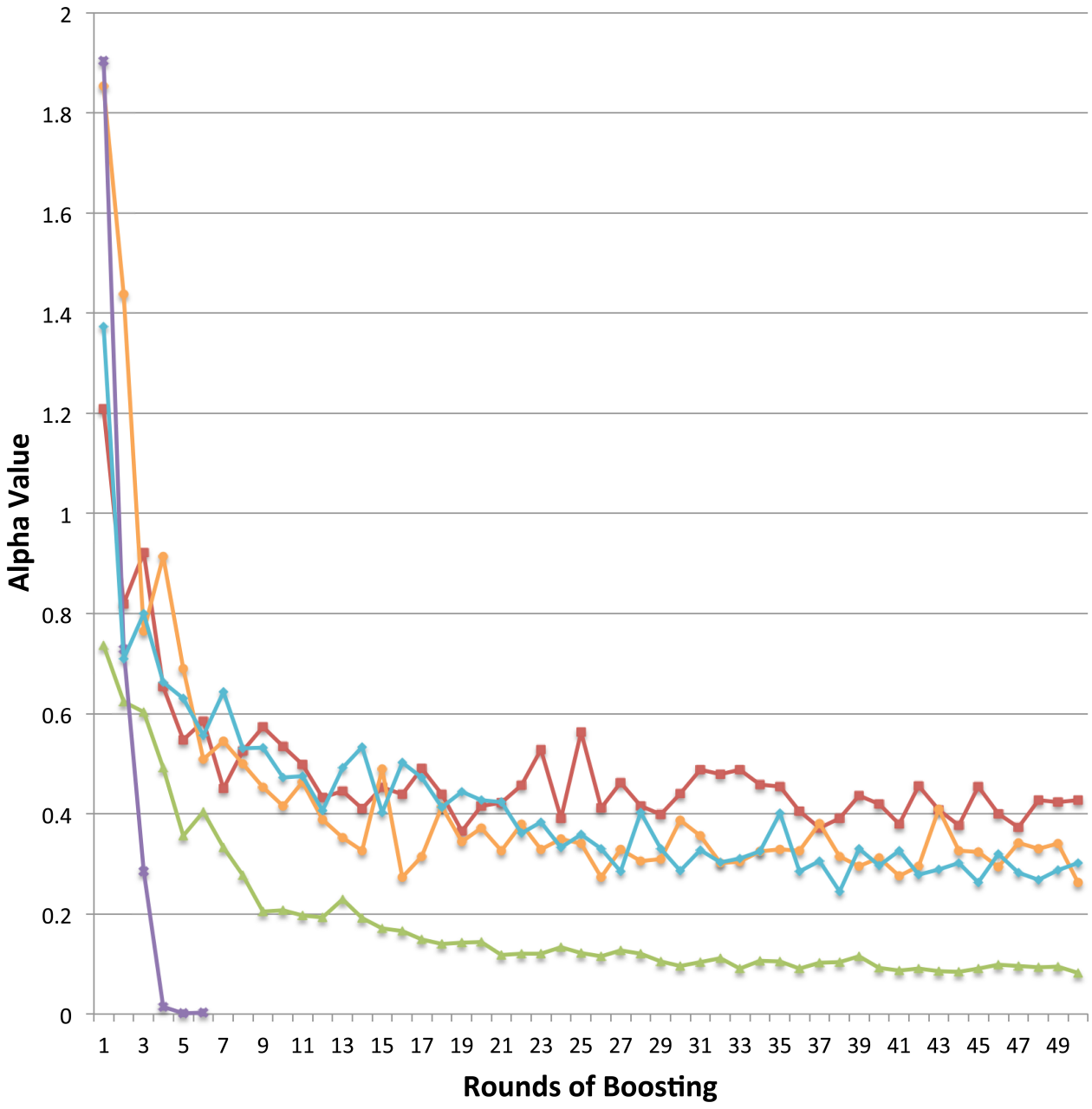

Figure S1. AdaBoost alpha values (Y-axis) are shown for 50 rounds of boosting (X-axis). The *A. baumannii* carbapenem classifier is depicted by the red line with square plot points, the *S. pneumoniae* beta-lactam resistance classifier is depicted by the green line with triangular plot points, the *S. pneumoniae* co-trimoxazole classifier is depicted by the orange line with circular plot points, the combined *M. tuberculosis* classifier is depicted with a teal line and diamond-shaped plot points and the *S. aureus* methicillin classifier is depicted by a purple line with x-shaped plot points. Only the first six plot points for the *S. aureus* classifier are shown because the alpha value goes to zero.

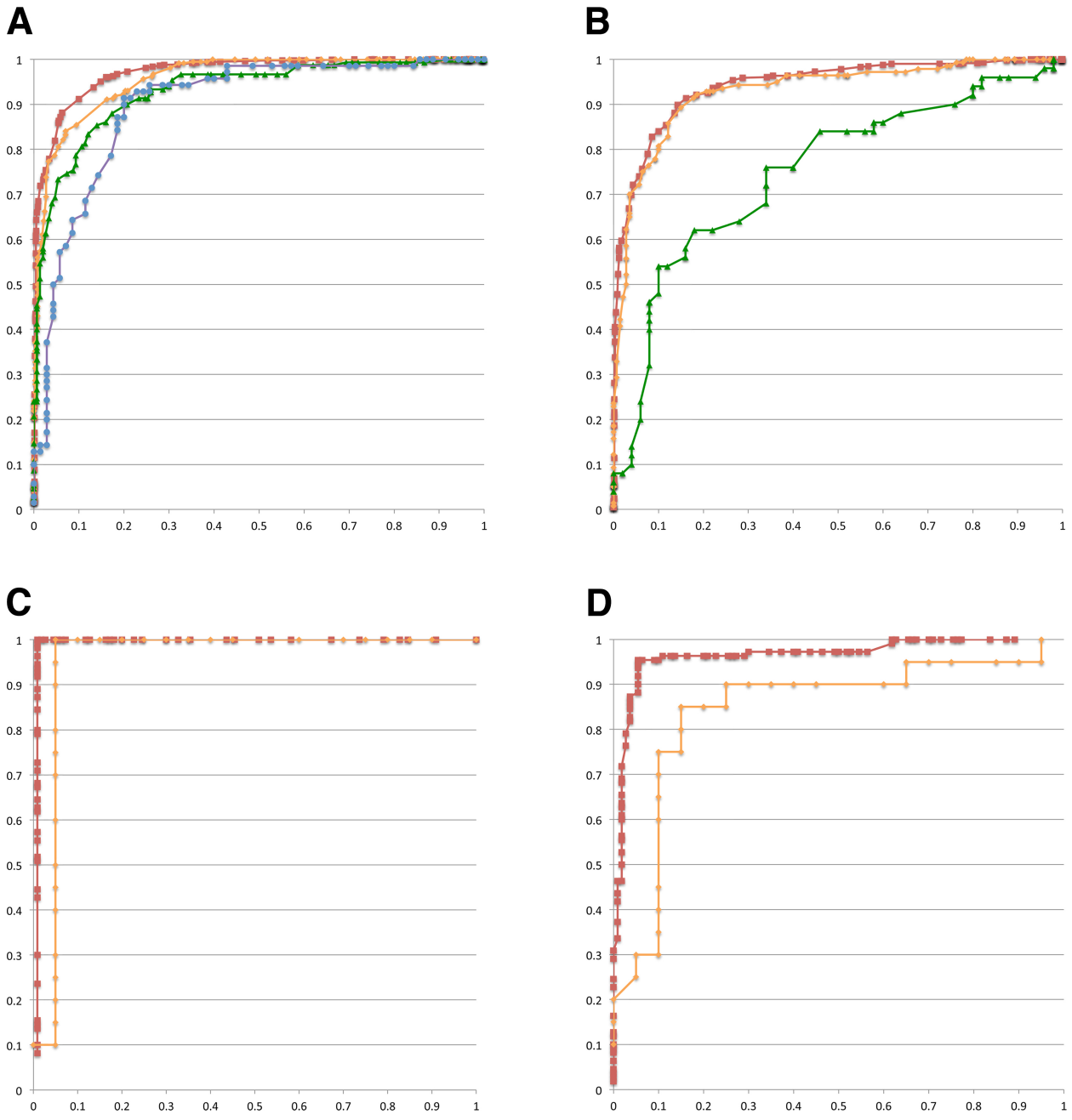

Figure S2. The effect of reducing the number of genomes used to build classifiers. Data are presented as ROC curves for cross validation experiments (see Methods). The X-axis is the false positive rate and the Y-axis is the true positive rate. Data are presented for 100% of the data set presented in Table 1 (red lines with square plot points), 25% of the data set (orange lines with diamond plot points), 10% of the data set (green lines with triangle plot points), and 5% of the data set (blue line with circle plot points) when appropriate. All experiments were balanced to have the same number of resistant and susceptible genomes. A) *S. pneumoniae* beta-lactam resistance, 1504, 376, 150 and 75 resistant and susceptible genomes were used for the 100%, 25%, 10% and 5% sets respectively; B) *S. pneumoniae*

co-trimoxazole resistance, 584, 146 and 58 resistant and susceptible genomes were used for the 100%, 25% and 10% sets respectively; C) *S. aureus* methicillin resistance, 115 and 28 resistant and susceptible genomes were used for the 100% and 25% sets respectively; and D) *A. baumannii* carbapenem resistance 110 and 27 resistant and susceptible genomes were used for the 100% and 25% sets respectively.

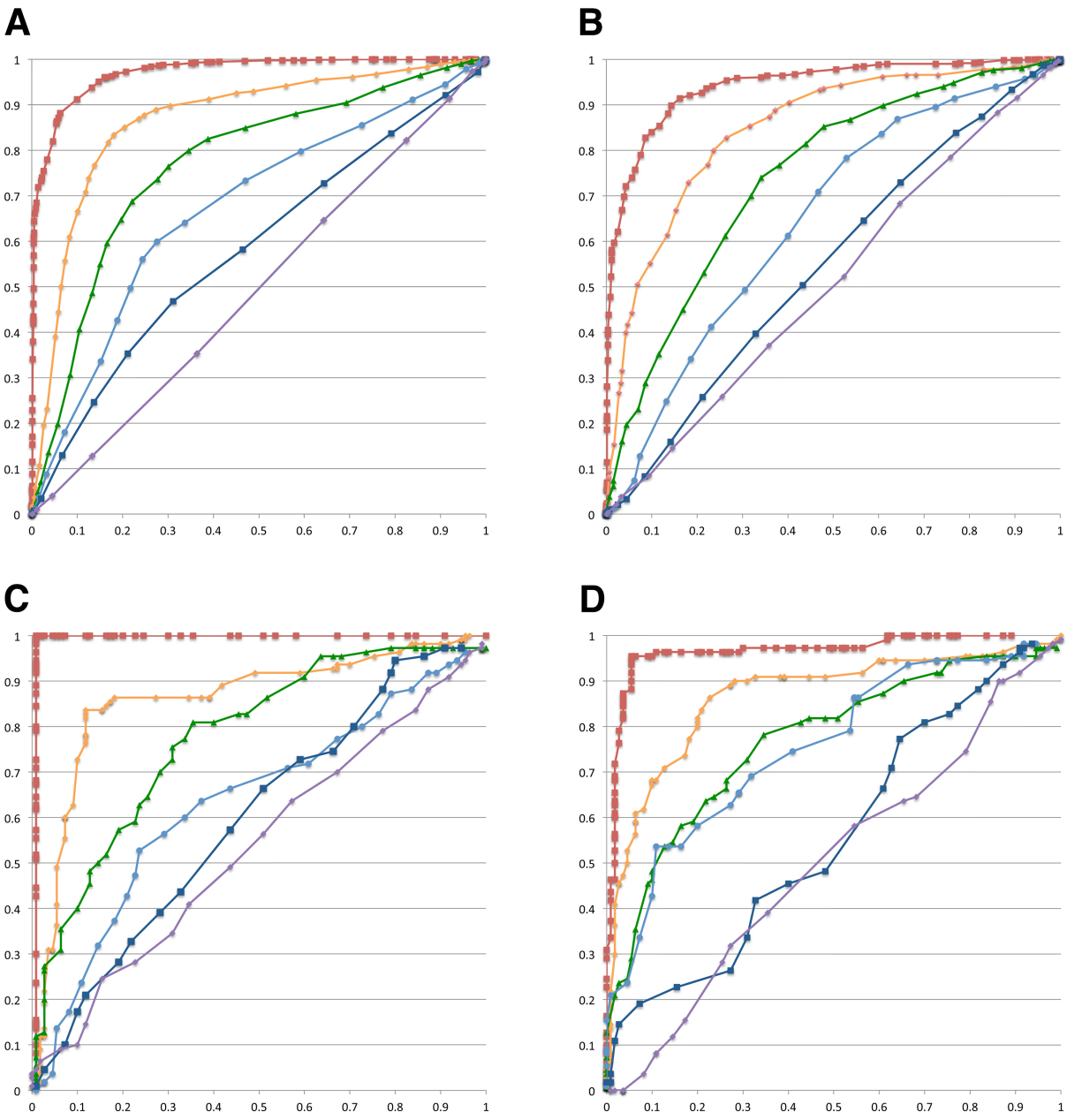

Figure S3. The result of introducing error into the AdaBoost classifiers. In order to determine the effect of unintentionally having misclassified genomes in the training set,

susceptible genomes were mixed with the resistant training set and vice versa prior to building the classifier. The test sets were kept unmixed. Results are displayed as ROC curves for cross validation experiments (see Methods). Experiments were performed for A) *S. pneumoniae* beta-lactam resistance, B) *S. pneumoniae* co-trimoxazole resistance, C) *S. aureus* methicillin resistance, and D) *A. baumannii* carbapenem resistance. The red line with square plot points depicts no mixing, the orange line with diamond plot points depicts 10% mixing, the green line with triangle plot points depicts 20% mixing, the light blue line with circle plot points depicts 30% mixing, the dark blue line with square plot points depicts 40% mixing, and the purple line with diamond plot points depicts 50% mixing. The X-axis is false positive rate and the Y-axis is true positive rate. Each experiment used an equal number of resistant and susceptible genomes (Table 2 main text).

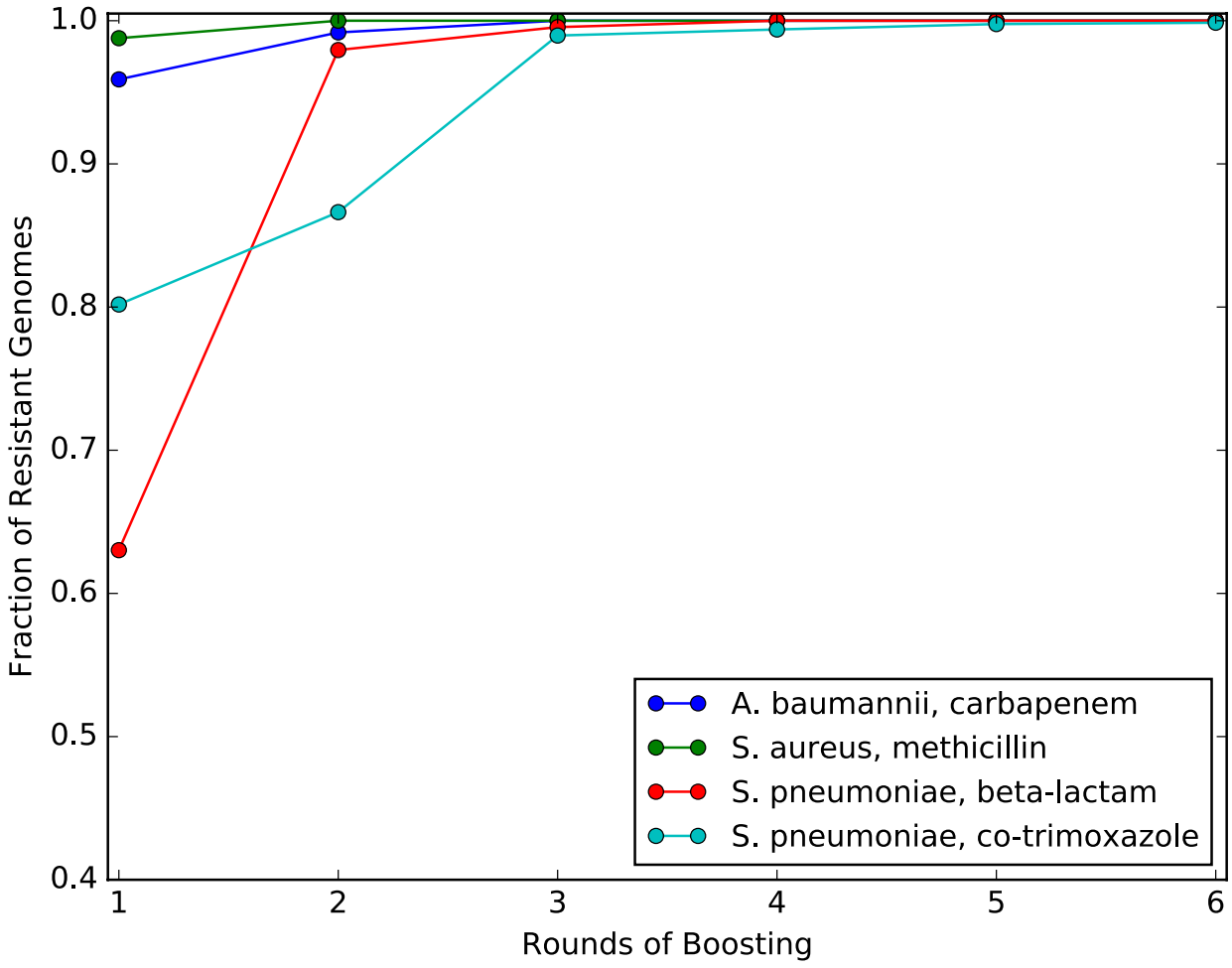

Figure S4. The fraction of *A. baumannii*, *S. aureus*, *and S. pneumoniae* resistant genomes with at least one k-mer match after each successive round of AdaBoost. The number of resistant genomes corresponding to each classifier is shown in Table 2.

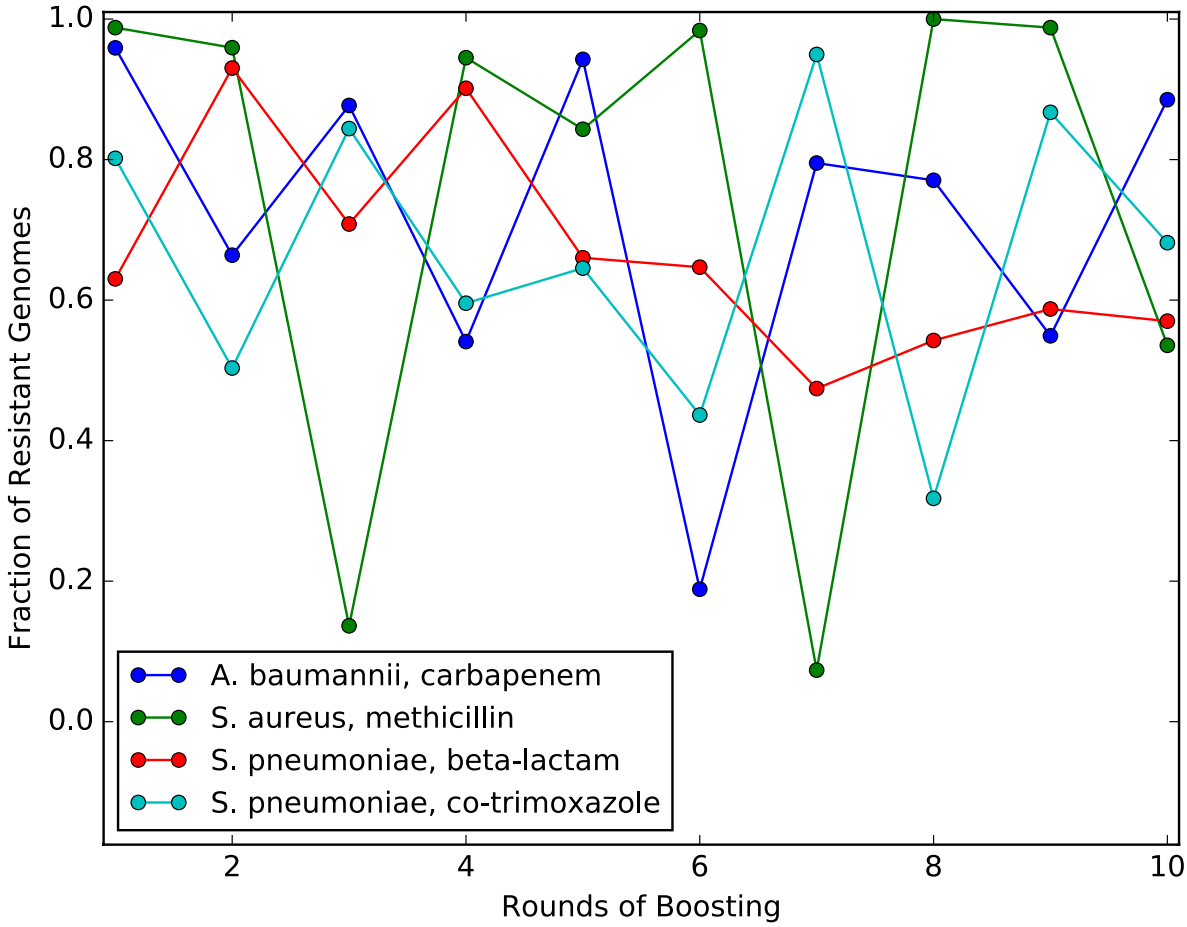

Figure S5. The prevalence of AdaBoost-selected k-mers in *A. baumannii*, *S. aureus*, *and S. pneumoniae* resistant genomes. For each round of AdaBoost, the fraction of *A. baumannii*, *S. aureus*, *and S. pneumoniae* resistant genomes with a matching k-mer is shown. The number of resistant genomes corresponding to each classifier is shown in Table 2.

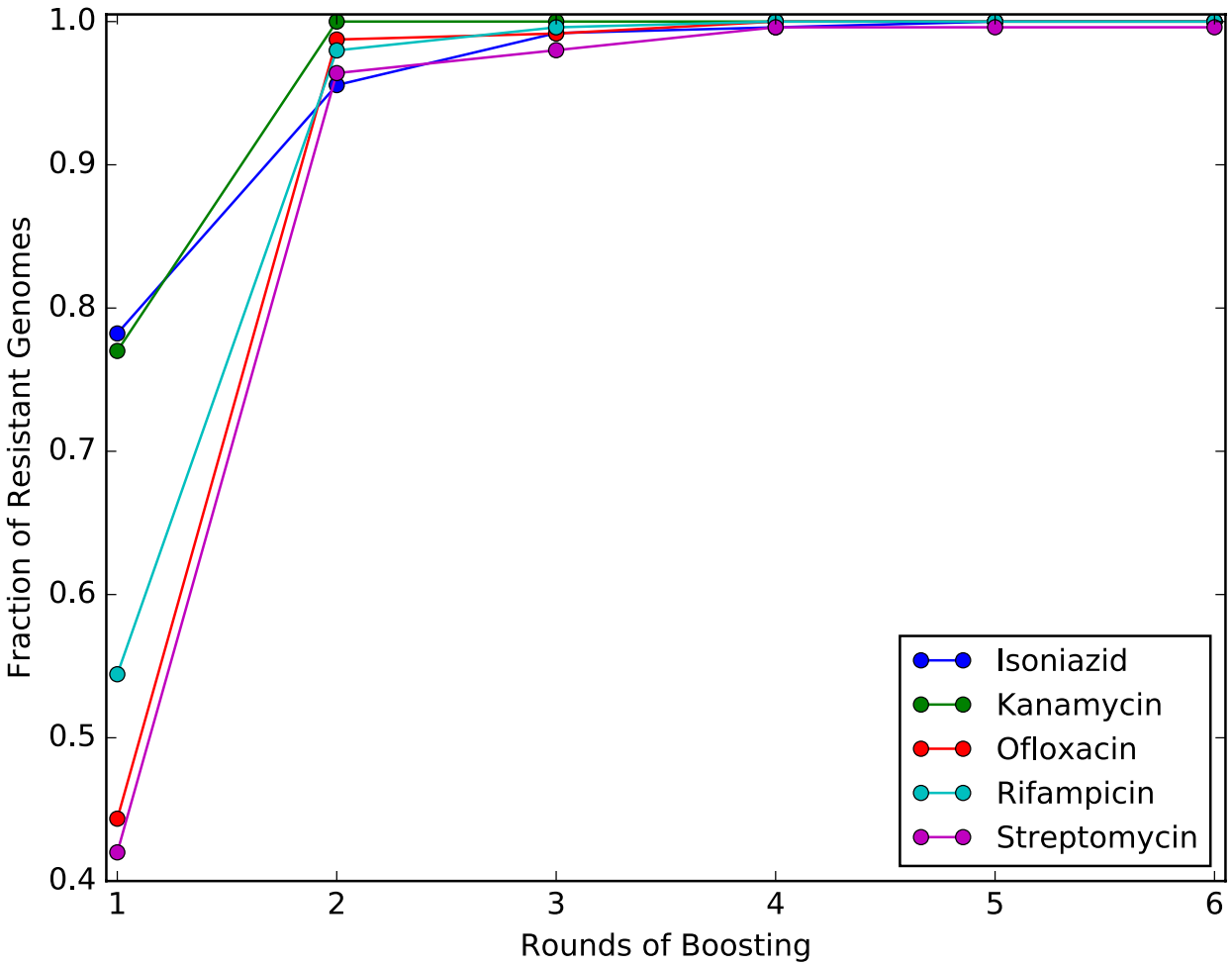

Figure S6. The fraction of *M. tuberculosis* resistant genomes with at least one k-mer match after each successive round of AdaBoost. The number of resistant genomes corresponding to each classifier is shown in Table 4.

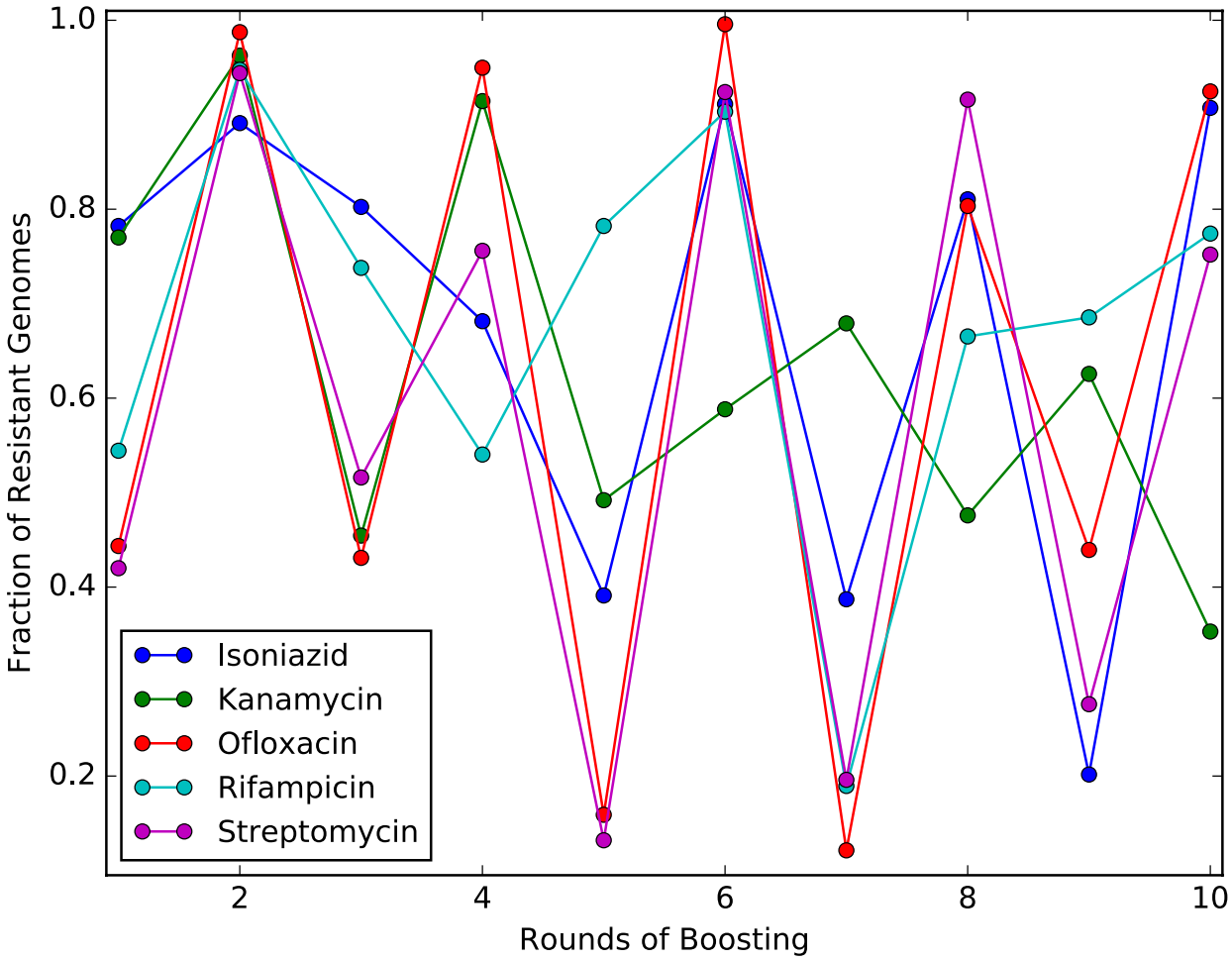

Figure S7. The prevalence of AdaBoost-selected k-mers in *Mycobacterium tuberculosis* resistant genomes. For each round of AdaBoost, the fraction of *M. tuberculosis* resistant genomes with a matching k-mer is shown. The number of resistant genomes corresponding to each classifier is shown in Table 4.