Using Machine Learning to Predict Antimicrobial Minimum Inhibitory Concentrations and Associated Genomic Features for Nontyphoidal *Salmonella*

Supplemental Information

Marcus Nguyen, S. Wesley Long, Patrick F. McDermott, Randall J. Olsen, Robert Olson, Rick L. Stevens, Gregory H. Tyson, Shaohua Zhao and James J. Davis

Supplemental Figures

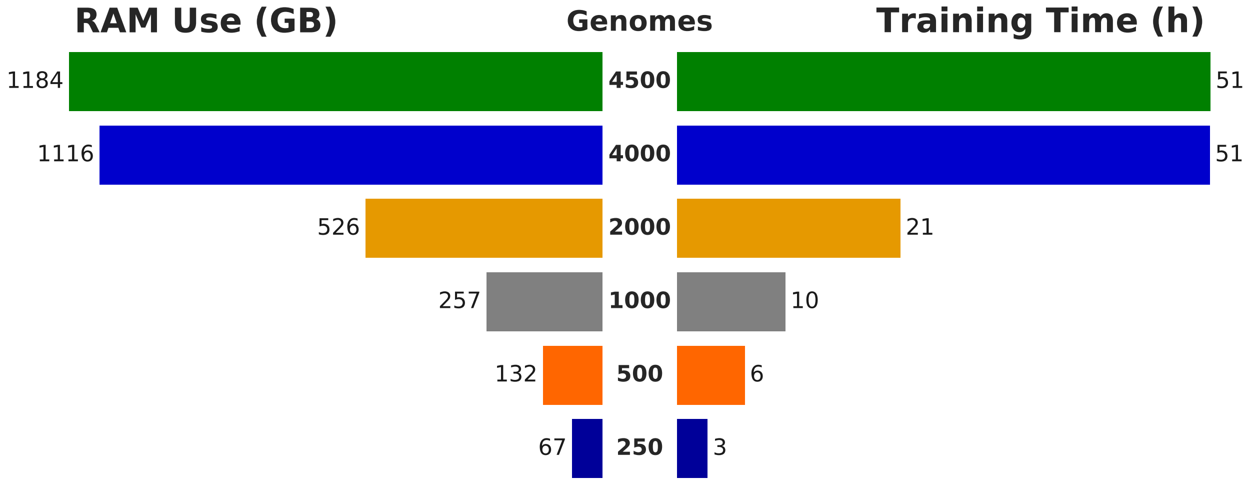

Figure S1. Tornado plot of the peak memory use (GB) and time required to train a 5-fold cross-validated model (hours), using a large server (4 Intel E5-4669v4 CPUs @ 2.20GHz, 1,585,224,876 kB RAM) and the data sets from Figure 1. Training was done using 170 of the 174 available logical cores. From top to bottom, bars in the plot correspond to training sets containing 4500, 4000, 2000, 1000, 500 and 250 genomes respectively. Values on the left depict peak RAM usage, and values on the right depict training time.

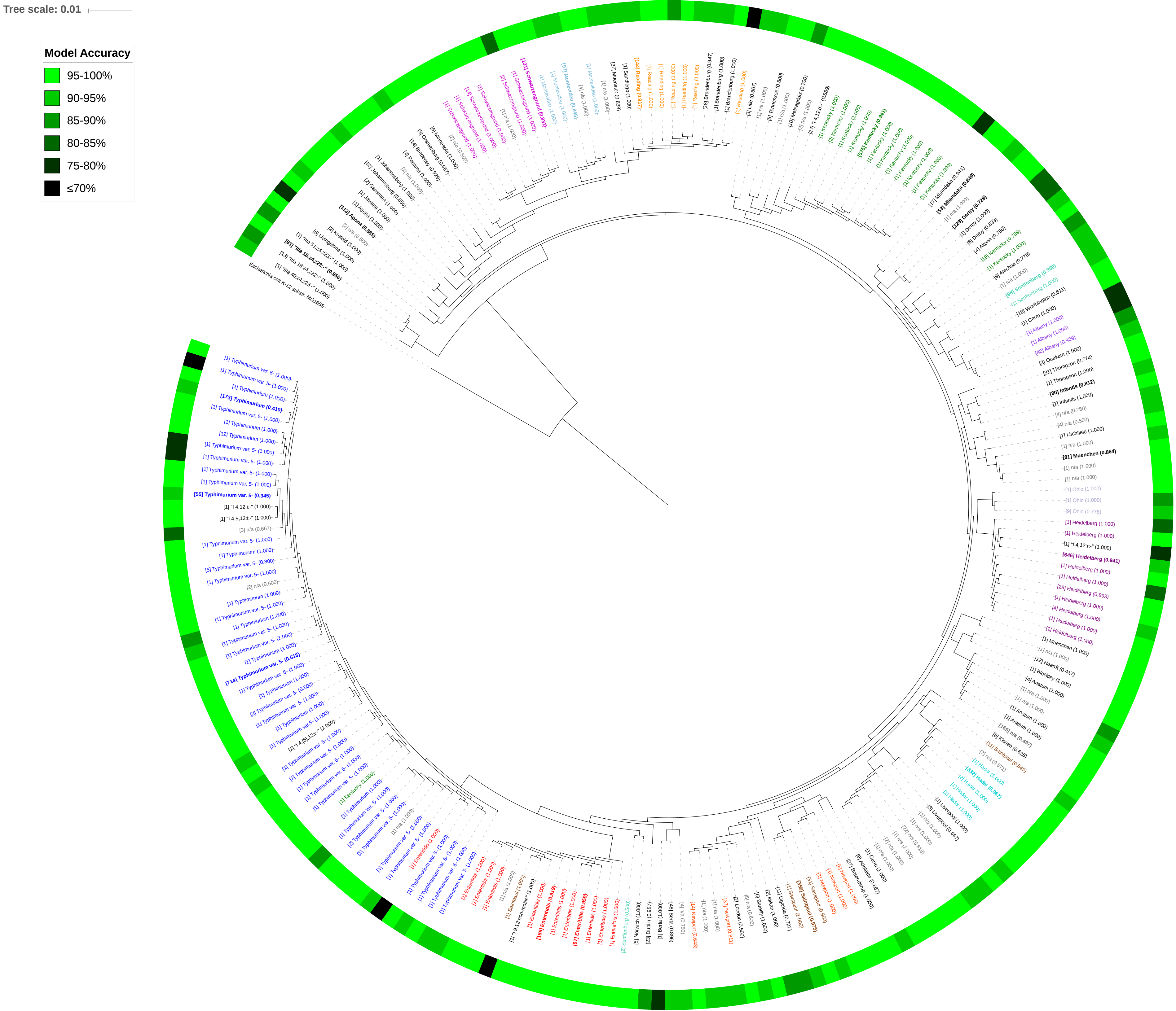
Tree scale: 0.01
Model Accuracy
🟩 95-100%
🟩 90-95%
🟩 85-90%
🟩 80-85%
🟩 75-80%
■ ≤70%
Escherichia coli K-12 substr. MG1655
[1] "IIIa 40:z4,z23:-" (1.000)
[13] "IIIa 18:z4,z32:-" (1.000)
[91] "IIIa 18:z4,z23:-" (0.956)
[1] "IIIa 51:z4,z23:-" (1.000)
[6] Livingstone (1.000)
[2] Krefeld (1.000)
[2] n/a (0.500)
[113] Agona (0.885)
[1] Agona (1.000)
[1] Javiana (1.000)
[2] Gaminara (1.000)
[32] Johannesburg (0.656)
[1] Johannesburg (1.000)
[1] n/a (1.000)
[4] Panama (1.000)
[14] Bredeney (0.929)
[3] Oranienburg (0.667)
[6] Minnesota (1.000)
[2] n/a (0.500)
[1] Schwarzengrund (1.000)
[1] Schwarzengrund (1.000)
[14] Schwarzengrund (1.000)
[1] Schwarzengrund (1.000)
[1] n/a (1.000)
[2] Schwarzengrund (1.000)
[1] Schwarzengrund (1.000)
[131] Schwarzengrund (0.832)
[1] Montevideo (1.000)
[1] Montevideo (1.000)
[97] Montevideo (0.845)
[4] n/a (1.000)
[1] Montevideo (1.000)
[1] n/a (1.000)
[37] Muenster (0.838)
[1] Sandiego (1.000)
[144] Reading (0.917)
[1] Reading (1.000)
[1] Reading (1.000)
[1] Reading (1.000)
[1] Reading (1.000)
[1] Reading (1.000)
[38] Brandenburg (0.947)
[1] Brandenburg (1.000)
[1] Brandenburg (1.000)
[1] Reading (1.000)
[3] Lille (0.667)
[1] n/a (1.000)
[5] Tennessee (0.800)
[1] n/a (1.000)
[10] Meleagridis (0.700)
[2] n/a (1.000)
[27] "I 4,12:d:-" (0.889)
[1] Kentucky (1.000)
[2] Kentucky (1.000)
[1] Kentucky (1.000)
[1] Kentucky (1.000)
[575] Kentucky (0.941)
[1] Kentucky (1.000)
[1] Kentucky (1.000)
[1] Kentucky (1.000)
[1] Kentucky (1.000)
[1] Kentucky (1.000)
[1] Kentucky (1.000)
[17] Mbandaka (0.941)
[53] Mbandaka (0.849)
[1] n/a (1.000)
[129] Derby (0.729)
[1] Derby (1.000)
[6] Derby (0.833)
[4] Altona (0.750)
[19] Kentucky (0.789)
[1] Kentucky (1.000)
[9] Alachua (0.778)
[1] n/a (1.000)
[98] Senftenberg (0.959)
[1] Senftenberg (1.000)
[18] Worthington (0.611)
[1] Cerro (1.000)
[1] Albany (1.000)
[1] Albany (1.000)
[42] Albany (0.929)
[2] Quakam (1.000)
[31] Thompson (0.774)
[1] Thompson (1.000)
[80] Infantis (0.812)
[1] Infantis (1.000)
[4] n/a (0.750)
[4] n/a (0.500)
[7] Litchfield (1.000)
[1] n/a (1.000)
[81] Muenchen (0.864)
[1] n/a (1.000)
[1] n/a (1.000)
[1] Ohio (1.000)
[1] Ohio (1.000)
[9] Ohio (0.778)
[1] Heidelberg (1.000)
[1] Heidelberg (1.000)
[1] "I 4,12:r:-" (1.000)
[646] Heidelberg (0.941)
[1] Heidelberg (1.000)
[1] Heidelberg (1.000)
[28] Heidelberg (0.893)
[1] Heidelberg (1.000)
[4] Heidelberg (1.000)
[1] Heidelberg (1.000)
[1] Heidelberg (1.000)
[1] Muenchen (1.000)
[1] n/a (1.000)
[12] Haardt (0.417)
[1] Blockley (1.000)
[4] Anatum (1.000)
[1] n/a (1.000)
[1] n/a (1.000)
[1] Anatum (1.000)
[1] Anatum (1.000)
[165] n/a (0.497)
[8] Rissen (0.625)
[11] Saintpaul (0.545)
[7] n/a (0.571)
[1] Hadar (1.000)
[332] Hadar (0.967)
[2] Hadar (1.000)
[1] Hadar (1.000)
[1] Hadar (1.000)
[1] Liverpool (1.000)
[3] Liverpool (0.667)
[1] n/a (1.000)
[1] n/a (1.000)
[22] n/a (0.818)
[1] n/a (1.000)
[2] n/a (1.000)
[1] n/a (1.000)
[1] Cerro (1.000)
[9] Adelaide (0.667)
[27] Braenderup (1.000)
[6] Newport (1.000)
[2] Newport (1.000)
[1] Newport (1.000)
[31] Saintpaul (0.903)
[358] Saintpaul (0.975)
[1] Saintpaul (1.000)
[11] Uganda (0.727)
[2] Idikan (1.000)
[6] Bareilly (1.000)
[5] n/a (0.600)
[2] London (0.500)
[37] Newport (0.811)
[1] n/a (1.000)
[1] n/a (1.000)
[14] Newport (0.643)
[4] n/a (0.750)
[48] Berta (0.896)
[1] Berta (1.000)
[23] Dublin (0.957)
[5] Norwich (1.000)
[2] Senftenberg (0.500)
[1] Enteritidis (1.000)
[1] Enteritidis (1.000)
[1] Enteritidis (1.000)
[97] Enteritidis (0.959)
[1] Enteritidis (1.000)
[1] Enteritidis (1.000)
[186] Enteritidis (0.919)
[1] Enteritidis (1.000)
[1] "I 9,12:non-motile" (1.000)
[1] Saintpaul (1.000)
[1] n/a (1.000)
[1] Enteritidis (1.000)
[1] Enteritidis (1.000)
[1] Enteritidis (1.000)
[1] Typhimurium var. 5- (1.000)
[1] Typhimurium var. 5- (1.000)
[1] Typhimurium var. 5- (1.000)
[1] Typhimurium var. 5- (1.000)
[1] Enteritidis (1.000)
[1] Typhimurium var. 5- (1.000)
[1] Typhimurium var. 5- (1.000)
[1] n/a (1.000)
[1] Typhimurium var. 5- (1.000)
[2] Typhimurium var. 5- (1.000)
[1] Typhimurium var. 5- (1.000)
[1] Typhimurium (1.000)
[1] Kentucky (1.000)
[1] Typhimurium var. 5- (1.000)
[1] Typhimurium var. 5- (1.000)
[1] Typhimurium var. 5- (1.000)
[1] Typhimurium var. 5- (1.000)
[1] "I 4,[5],12:i:-" (1.000)
[1] Typhimurium var.5- (1.000)
[1] Typhimurium (1.000)
[1] Typhimurium var. 5- (1.000)
[2] Typhimurium var. 5- (0.500)
[1] Typhimurium (1.000)
[1] Typhimurium var. 5- (1.000)
[714] Typhimurium var. 5- (0.618)
[1] Typhimurium (1.000)
[1] Typhimurium var. 5- (1.000)
[1] Typhimurium var. 5- (1.000)
[1] Typhimurium (1.000)
[1] Typhimurium var. 5- (1.000)
[1] Typhimurium (1.000)
[2] n/a (0.500)
[1] Typhimurium var. 5- (1.000)
[5] Typhimurium var. 5- (0.800)
[1] Typhimurium (1.000)
[1] Typhimurium var. 5- (1.000)
[3] n/a (0.667)
[1] "I 4,5,12:i:-" (1.000)
[1] "I 4,12:i:-" (1.000)
[55] Typhimurium var. 5- (0.345)
[1] Typhimurium var. 5- (1.000)
[1] Typhimurium var. 5- (1.000)
[1] Typhimurium var. 5- (1.000)
[1] Typhimurium var. 5- (1.000)
[12] Typhimurium (1.000)
[1] Typhimurium (1.000)
[1] Typhimurium var. 5- (1.000)
[173] Typhimurium (0.410)
[1] Typhimurium (1.000)
[1] Typhimurium var. 5- (1.000)
[1] Typhimurium var. 5- (1.000)

Figure S2. Phylogenetic tree of the *Salmonella* strains used in this study. The tree is based on a concatenation of alignments for the *rpoB* and *rpoC* genes and is rooted on *E. coli* K-12. Zero-length branches are collapsed and are depicted by a single tip and labeled according to the most common serotype occurring at that tip. The number in square brackets is the number of genomes depicted by a tip, and tips with ≥ 50 taxa are shown in bold. The number in parentheses is the fraction of genomes at a given tip that belong to the listed serotype. Tips representing common serotypes are colored for ease of identification. The color bar surrounding the tree represents the average accuracy (within ± 1 two-fold dilution step) of the 4,500 genome model for the genomes represented by each tip, with black being the least accurate (≤ 70%) and bright green being the most accurate (95-100%).

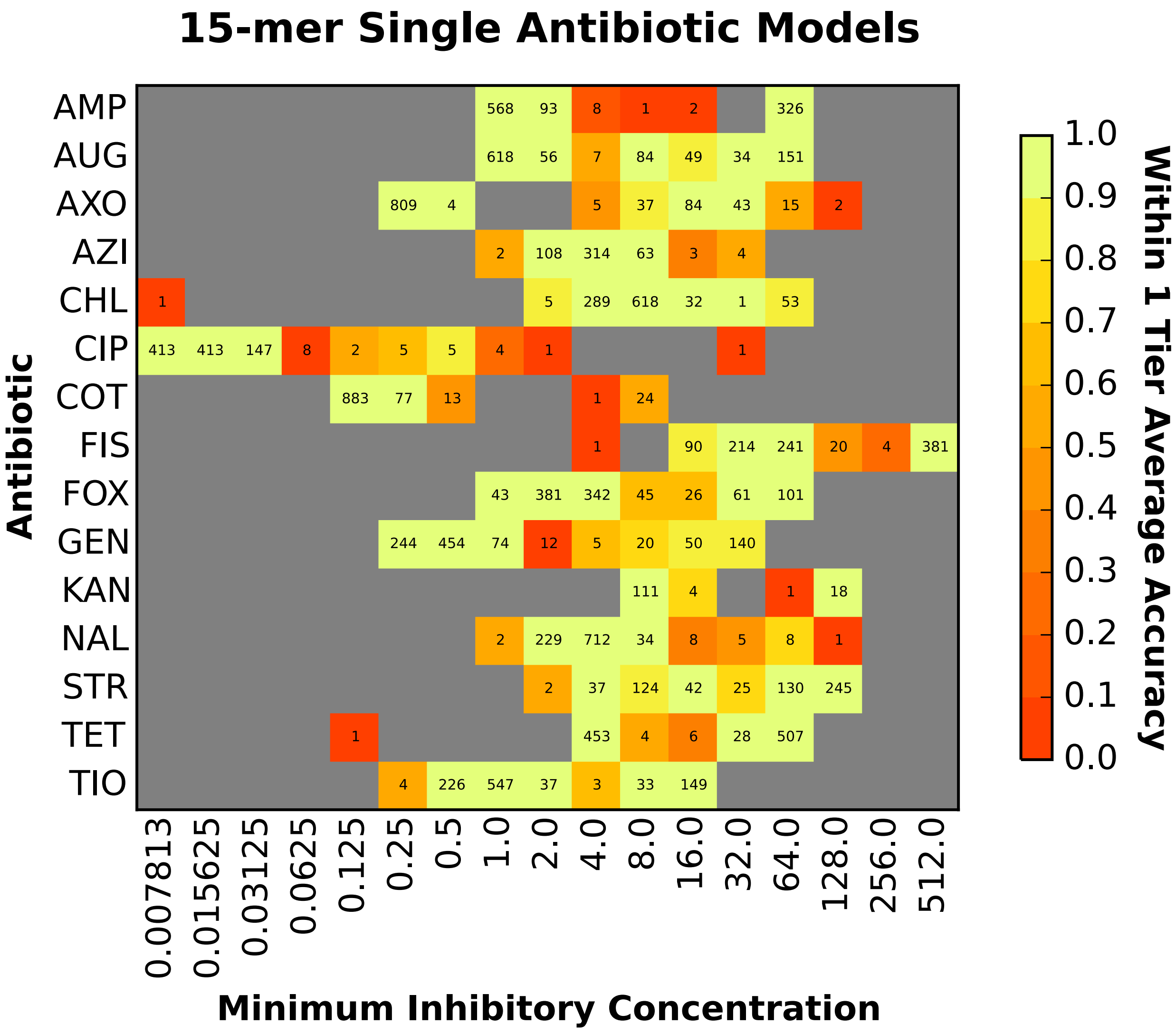

Figure S3. The accuracy of each single antibiotic model based on the set of ≤ 1000 diverse genomes. The heat map depicts the accuracy within ±1 two-fold dilution step of the laboratory-derived MIC. The X-axis shows the MIC (µg/ml) and each antibiotic is shown on the Y-axis. The accuracy for each antibiotic-MIC combination is depicted by color with bright yellow/green being the most accurate and red being the least accurate. The values shown in each cell are the number of genomes with that MIC for a given antibiotic.

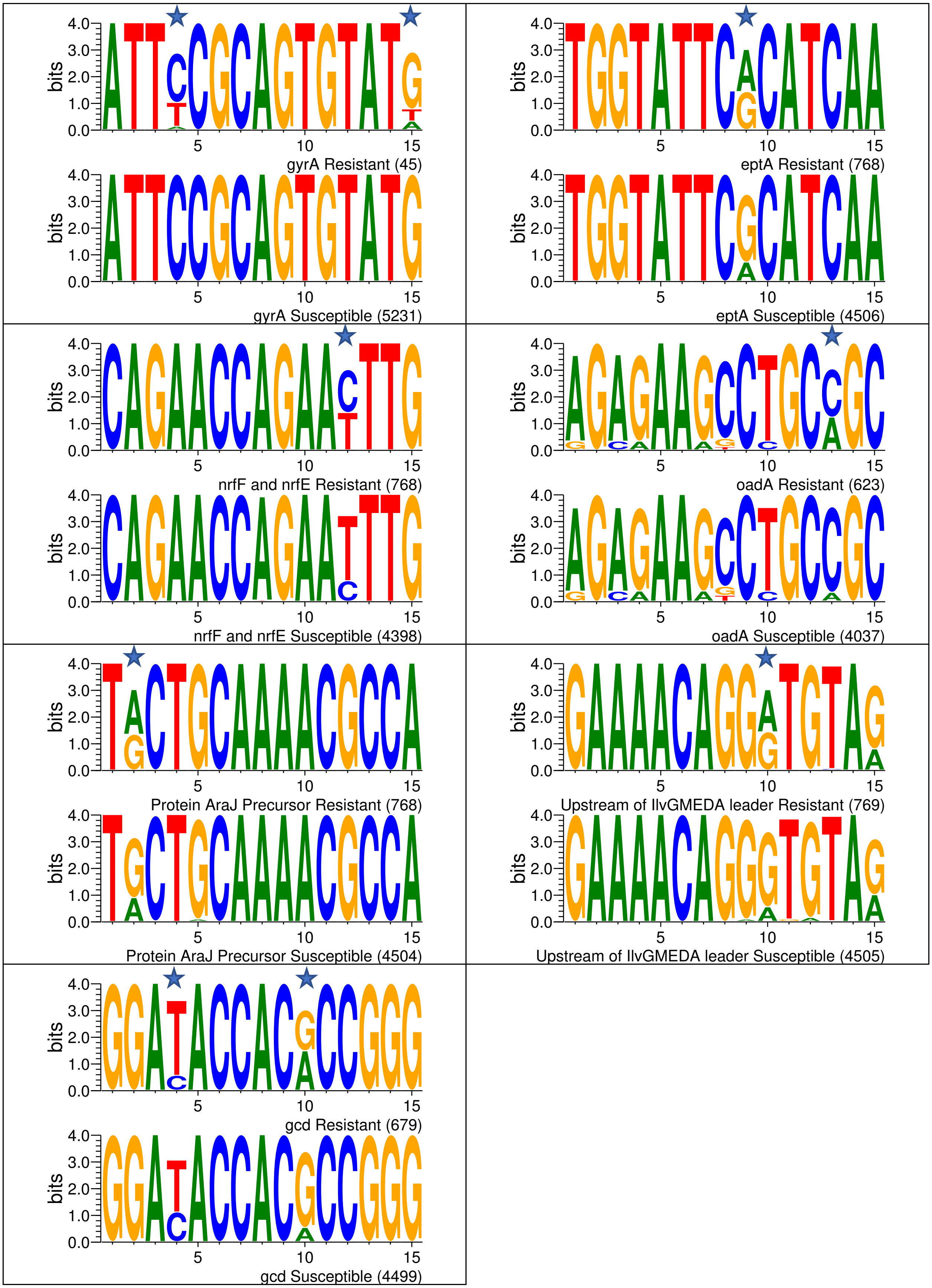
bits
4.0
3.0
2.0
1.0
0.0
5
10
15
gyrA Resistant (45)
gyrA Susceptible (5231)
eptA Resistant (768)
eptA Susceptible (4506)
nrfF and nrfE Resistant (768)
nrfF and nrfE Susceptible (4398)
oadA Resistant (623)
oadA Susceptible (4037)
Protein AraJ Precursor Resistant (768)
Protein AraJ Precursor Susceptible (4504)
Upstream of IlvGMEDA leader Resistant (769)
Upstream of IlvGMEDA leader Susceptible (4505)
gcd Resistant (679)
gcd Susceptible (4499)

Figure S4. WebLogo plots for the susceptibility-related k-mer regions containing significant SNPs from Table 7 (P < 0.001, based on a Chi-square test). SNPs are significant in both the set of 1,000 diverse genomes and in the full set of 5,278 genomes (data are depicted for the full set of 5,278 genomes). Logos for the k-mer region in the resistant genomes is depicted above the logos for the k-mer region from the susceptible genomes. Significant SNPs are depicted with a star. Gene names and short hand descriptions correspond to the following PATRIC annotations: gyrA, DNA gyrase subunit A (EC 5.99.1.3); eptA, Phosphoethanolamine transferase EptA; nrfE and nrfF, Formate-dependent nitrite reductase complex subunit NrfF and Cytochrome c-type heme lyase subunit nrfE, nitrite reductase complex assembly (the k-mer occurs in both genes); oadA, Oxaloacetate decarboxylase alpha chain (EC 4.1.1.3); gcd, Glucose dehydrogenase, PQQ-dependent (EC 1.1.5.2).