

45 FOX: cefoxitin
46 FSIS: USDA Food Safety and Inspection Service
47 GEN: gentamicin
48 KAN: kanamycin
49 ME: major error
50 MIC: minimum inhibitory concentration
51 NAL: nalidixic acid
52 NARMS: National Antimicrobial Resistance Monitoring System
53 SNP: single nucleotide polymorphism
54 STR: streptomycin
55 TET: tetracycline
56 TIO: ceftiofur
57 USDA: United States Department of Agriculture
58 VME: very major error
59 WGS: whole genome sequencing
60 XGBoost: Extreme Gradient Boosting

61
62
63
64
65
66
67

68 **Keywords**

69 Machine learning
70 Deep learning
71 Antimicrobial susceptibility testing
72 Genome sequencing
73 Diagnostics

74
75
76

77 Nontyphoidal *Salmonella* species are the leading bacterial cause of food-borne disease in the
78 United States. Whole genome sequences and paired antimicrobial susceptibility data are
79 available for *Salmonella* strains because of surveillance efforts from public health agencies. In
80 this study, a collection of 5,278 nontyphoidal *Salmonella* genomes, collected over 15 years in
81 the United States, were used to generate XGBoost-based machine learning models for
82 predicting minimum inhibitory concentrations (MICs) for 15 antibiotics. The MIC prediction
83 models have average accuracies between 95-96% within ± 1 two-fold dilution factor and can
84 predict MICs with no *a priori* information about the underlying gene content or resistance
85 phenotypes of the strains. By selecting diverse genomes for training sets, we show that highly
86 accurate MIC prediction models can be generated with fewer than 500 genomes. We also show
87 that our approach for predicting MICs is stable over time despite annual fluctuations in
88 antimicrobial resistance gene content in the sampled genomes. Finally, using feature selection,
89 we explore the important genomic regions identified by the models for predicting MICs. To
90 date, this is one of the largest MIC modeling studies to be published. Our strategy for
91 developing whole genome sequence-based models for surveillance and clinical diagnostics can
92 be readily applied to other important human pathogens.
93
94
95

96 Introduction

97 Nontyphoidal *Salmonella* species are the leading bacterial cause of food-borne disease in the
98 United States[1, 2], causing over one million illnesses per year[3] and an estimated 80 million
99 illnesses annually world-wide[4]. Nontyphoidal *Salmonella* causes acute gastroenteritis and is
100 usually contracted via fecal contamination of food sources[5]. Although these infections are
101 usually self-limiting and typically do not require antibiotic treatment[6], severe infections can
102 occur[7]. Antimicrobial resistance (AMR) is prevalent in *Salmonella* isolates and infections
103 caused by highly antimicrobial resistant *Salmonella* strains result in worse outcomes than
104 infections caused by susceptible strains[8-11].

105
106 In 1996, the National Antimicrobial Resistance Monitoring System (NARMS) was established as
107 a collaboration between the United States Centers for Disease Control and Prevention (CDC),
108 U.S. Food and Drug Administration (FDA), U.S. Department of Agriculture (USDA), and state and
109 local health departments. A primary goal of NARMS is to monitor antimicrobial resistance in
110 *Salmonella* and other food-borne bacteria, including *Campylobacter*, *Escherichia* and
111 *Enterococcus*[12]. The data collected by NARMS is used to inform public health decisions aimed
112 at identifying contaminated food sources and reducing the spread of AMR through enhanced
113 stewardship. In recent years, NARMS has adopted whole genome sequencing (WGS) as a
114 routine monitoring tool. The WGS data are used to determine the source of outbreak strains,
115 the virulence factor and AMR genes carried by each strain. As a result, a large collection of
116 bacterial whole genome sequences with extensive metadata is available for downstream
117 research efforts[13].

118
119 Whole genome sequencing is now routinely used for public health surveillance and to guide
120 diagnostic and patient care decisions[14-18]. For routine surveillance, WGS provides the
121 highest possible resolution for individuating traits in bacteria, assessing phylogenetic
122 relationships, conducting outbreak investigations, and predicting virulence and epidemicity.
123 From the clinical perspective, rapid diagnostics are key to improving patient care. For a
124 conventional microbiology laboratory diagnosis, the total time for organism growth, isolation,
125 taxonomic identification, and antimicrobial minimum inhibitory concentration (MIC)
126 determination may exceed 36 hours for relatively fast-growing bacteria and several days for
127 slower growing organisms[19-21]. Since reducing the time to optimal antimicrobial therapy
128 significantly improves patient outcomes[22-24], rapid sequencing-based approaches that
129 predict MICs may have clinical utility. The extensive WGS datasets generated by health
130 agencies and the scientific community, such as nontyphoidal *Salmonella* strains, provides the
131 necessary training sets required for building predictive models.

132
133 Several investigations have recently built models for predicting AMR phenotypes from WGS
134 data. To date, the most common approach has relied on using a curated reference set of genes
135 and polymorphisms that are implicated in AMR[25-33]. This reference-guided approach best
136 predicts susceptibility and resistance when organisms are well studied and the AMR
137 mechanisms are known. As larger collections of genomes have become available, several
138 studies have used machine learning algorithms to predict susceptible and resistant
139 phenotypes[27, 29, 31, 34-38]. By using WGS and AMR phenotype data to train a machine

140 learning model, predictions without *a priori* information about the underlying gene content of
141 the genome or molecular mechanism for resistance to each agent are possible. Although this
142 reference-free approach requires many genomes, it is unbiased and can potentially be used to
143 enable the discovery of new genomic features that are involved in AMR[36, 37]. These two
144 general approaches have also been used to predict MICs for *Streptococcus*, *Neisseria*, and
145 *Klebsiella*[35, 38-40]. When a curated reference collection of genes and SNPs is used for
146 predicting MICs, a rules-based or machine learning algorithm is required for determining how
147 much a given feature contributes to the MIC. Thus, for MIC prediction, both reference-guided
148 and reference-free approaches are expected to have similar advantages and disadvantages if
149 the collection of genes and SNPs used by the reference-guided method is sufficient for
150 predicting all MICs, including those that are in the susceptible range. For example, in previous
151 work, we built a machine learning model to predict MICs for a comprehensive population-based
152 collection of 1,668 *Klebsiella pneumoniae* clinical isolates[38]. For each genome, we used
153 nucleotide 10-mers and the MICs for each antibiotic as features to train the model. Extreme
154 gradient boosting (XGBoost) was chosen as the machine learning algorithm[41]. The model
155 could rapidly predict the MICs for 20 antibiotics with an average accuracy of 92%. This
156 demonstrated that it is possible to successfully predict MICs without using a precompiled set of
157 AMR genes or polymorphisms.

158
159 In this study, we build models that use whole genome sequence data to predict MICs for
160 nontyphoidal *Salmonella* based on strains collected and sequenced by NARMS from 2002-2016.
161 Our strategy can be used to guide responses to outbreaks and inform antibiotic stewardship
162 decisions.

163
164

165 **Materials and Methods**

166

167 **Genomes and Metadata**

168 A total of 5,278 nontyphoidal *Salmonella* genome sequences were used in this study. All strains
169 were collected and sequenced as part of the NARMS program. The strains were recovered
170 from either raw retail meat and poultry or directly from livestock animals at slaughter.
171 Antimicrobial susceptibility testing was performed using broth microdilution on the Sensititre®
172 system (Thermo Scientific) for 15 antibiotics: ampicillin (AMP), amoxicillin/clavulanic acid
173 (AUG), ceftriaxone (AXO), azithromycin (AZI), chloramphenicol (CHL), ciprofloxacin (CIP),
174 trimethoprim/sulfamethoxazole (COT), sulfisoxazole (FIS), cefoxitin (FOX), gentamicin (GEN),
175 kanamycin (KAN), nalidixic acid (NAL), streptomycin (STR), tetracycline (TET), and ceftiofur (TIO)
176 at FDA and USDA NARMS laboratories[13]. Clinical breakpoints are based on CLSI and FDA
177 guidelines[42]. Whole genome sequencing was performed using the Illumina HiSeq and MiSeq
178 platforms using standard methods[25]. Accession numbers and MICs for each isolate are listed
179 in Table S1. All non-AMR metadata including serotypes, host, geographic location of isolation
180 and isolation year were taken from the metadata associated with each NCBI SRA entry.

181

182 **Genomic Analyses**

183 The short read sequence data for each strain was assembled with the PATRIC genome assembly
184 service[43], using the “Full SPAdes” pipeline which uses BayesHammer[44] for read correction
185 and SPAdes for assembly[45]. All genomes were annotated using the PATRIC annotation
186 service[43], which uses a variation of the RAST tool kit annotation pipeline[46]. Annotated
187 genomes are available on the PATRIC website (<https://patricbrc.org>). PATRIC genome
188 identifiers are displayed in Table S1. Protein annotations, including those specifically asserted
189 to be involved in AMR[47] were downloaded from the PATRIC workspace and used for
190 subsequent analyses. A phylogenetic tree was generated for the strains in the analysis by
191 aligning the genes for the beta and beta prime subunits of the RNA polymerase using
192 MAFFT[48], concatenating the alignments, and computing a tree with FastTree[49]. The tree
193 was rendered using iTOL[50].

194

195 **MIC Prediction**

196 *Model Generation.* A model for predicting minimum inhibitory concentrations for the 15
197 antibiotics was built following the methods previously described by Nguyen and colleagues[38].
198 Briefly, each genome was divided into the set of nonredundant overlapping nucleotide 10-mers
199 using the k-mer counting program KMC[51]. A matrix was built where the k-mers, antibiotics,
200 and MICs are treated as features for each genome. Each row in the matrix contains the k-mers
201 for a genome as well as the MIC for a single antibiotic. The MIC prediction model was built
202 using an XGBoost[41] regressor predicting linearized MICs. All model parameters were identical
203 to those used by Nguyen et al[38]. Ten-fold cross validations were used to assess the overall
204 accuracy and sensitivity of every model used in this study. A non-overlapping training set (80%
205 of the data), validation set (10% of the data), and test set (10% of the data) were generated for
206 each fold. The validation set was used to monitor each model to prevent overfitting. Unless
207 otherwise stated, the accuracy is reported as the ability to predict the correct MIC within ± 1
208 two-fold dilution step of the laboratory-derived MIC. Defining an accuracy to be within one
209 two-fold dilution step is consistent with FDA requirements for automated MIC measuring
210 device standards and is consistent with established clinical microbiology practices[20, 52, 53].
211 A comparison of raw accuracies and accuracies within ± 1 two-fold dilution step is shown in
212 Table S2. To assess the accuracy of a model over various metadata categories including date,
213 serotype source, and location, the training set genomes are used to make the model. The test
214 set genomes are used to assess the model accuracy for a given fold. For models based on date
215 ranges, all parameters are identical and the accuracy is reported over the genomes from the
216 held-out dates.

217

218 *Subsampling.* In order to perform the model building on a machine with 1.5 TB of RAM
219 (machines with more memory are currently somewhat uncommon), we reduced the matrix size
220 to sets of size n , where $n \leq 250, 500, 1000, 2000, 3000, 4000$, and 4500 genomes respectively.
221 To create a diverse subset of size n , a hierarchical clustering method[54] was used to create n
222 clusters by using the 10-mer distribution of each genome as input features. To avoid the curse
223 of dimensionality[55, 56], the taxicab/Manhattan distance (L_1 norm) was used, rather than the
224 Euclidean distance (L_2 norm), since previous research has shown it to be both computationally
225 fast and more accurate for high dimensional data[57]. From the resulting n clusters, one

226 genome from each cluster was randomly selected from a uniform distribution to create the
227 subset containing n genomes. For each subset of genomes, a matrix was generated, and
228 models were generated as described above.

229

230 *Feature identification.* In order to unambiguously identify k-mers that are important to MIC
231 prediction, we built separate models for each individual antibiotic using the method described
232 above, except that we increased the k-mer length to 15 nucleotides in order to reduce the
233 number of redundant k-mers within each genome and to enable analyses with BLAST[58]. We
234 also measured k-mer hits as presence versus absence, rather than counts, in order to simplify
235 the analysis. Each model was built using the set of 1,000 diverse genomes from the
236 subsampling experiment described above and 10-fold cross validations were performed on
237 each model.

238

239 XGBoost's internal feature importance was computed for each fold within the 10-fold cross
240 validation. This results in an importance score per feature (15-mer) from each fold. In order to
241 generate an overall importance score for the top features, we summed the feature importance
242 scores from each fold for the top ten features. This overall importance score captures both the
243 importance of the 15-mer to a given fold and the number of times that 15-mer was chosen as a
244 top feature within each of the ten folds.

245

246 XGBoost's internal feature importance is unable to provide correlations between features and
247 label values, and thus does not provide an indication of whether a k-mer is related to antibiotic
248 resistance or susceptibility. This is partially due to the fact that many non-linear correlations
249 exist that may use multiple features. In order to see if the high scoring k-mers correlate with
250 resistance or susceptibility, we computed the distribution of MICs for the genomes containing
251 each high scoring k-mer. For example, a k-mer conferring susceptibility should be found in
252 more genomes with lower MICs, while a k-mer conferring resistance should exist in genomes
253 with higher MICs. Each high scoring k-mer was also compared to the set of protein encoding
254 genes within each *Salmonella* genome. If a k-mer was found within a known AMR gene, that
255 gene was reported. Otherwise, all protein-encoding genes within 3kb of the k-mer were
256 reported in order to assess the neighborhood of the k-mer.

257

258 To find k-mers that are being used by the individual antibiotic models to predict susceptible
259 MICs, we computed the presence or absence of each k-mers with high XGBoost feature
260 importance values (described above) for the entire data set of 5278 genomes. The k-mers with
261 the largest difference in occurrence between the susceptible and resistant genomes are the
262 ones that are being chosen by the models for predicting susceptible MICs. To demine if there
263 were significant SNPs in these k-mers, we found the genomic features containing the k-mer—
264 protein encoding gene, RNA gene, or intergenic region—using BLASTn[58]. The corresponding
265 feature or region was then found for all genomes in the collection. The features were aligned
266 using MAFFT[48] and manually curated using Jalview[59]. Poor quality sequence was removed,
267 all duplicates and paralogs were removed, and the subalignment covering the k-mer was
268 extracted. To prevent possible biases due to clonality that may exist in the full set of genomes,
269 the analysis was repeated on the diverse subset of 1000 genomes (described above). We

270 report a SNP in a k-mer region as being significant if the susceptible and resistant sets are
271 significantly different (P-value < 0.001) for a given nucleotide position based on a Chi-square
272 test for both the full set of 5278 genomes and the set of 1000 diverse genomes. Sequence
273 logos for k-mers containing significant SNPs were generated using WebLogo[60]. K-mers from
274 the Azithromycin and Ciprofloxacin models were excluded from this analysis because they each
275 had seven resistant genomes. Comparisons of codon usage were computed versus the genome
276 average, genome mode, and high expression gene sets as described previously[61, 62].

277

278 **Software availability**

279 The *Salmonella* MIC prediction model based on 4,500 genomes—including the software and
280 documentation for running the model—is available at our GitHub page,

281 https://github.com/PATRIC3/mic_prediction.

282

283

284 **Results**

285

286 **Model Construction**

287 For this study, we used a publicly available collection of 5,278 *Salmonella* whole genome
288 sequences generated by the NARMS project between 2002 and 2016. The strains were isolated
289 from retail meat and food animal sources in the United States. The collection includes 98
290 different serotypes, including Heidelberg (678 genomes), Kentucky (618 genomes), and
291 Typhimurium var. 5- (588 genomes) from 41 states (Table S1). Isolates were tested for
292 resistance to up to 15 antimicrobial agents using the broth microdilution method. Many of the
293 strains were randomly selected for sequencing as part of a compulsory nation-wide collection
294 program (Table 1).

295

296 The nontyphoidal *Salmonella* MIC prediction model was built similar to our previously
297 described strategy used to predict MICs for *K. pneumoniae* clinical isolates[38]. Since the
298 *Salmonella* data set has many more genomes and greater sampling in the range of susceptible
299 MICs, it provides a critical test case for determining if the approach remains robust for other
300 pathogens. In the *Klebsiella* study, we built individual models for each antibiotic, as well as a
301 single large integrated model by combining the data from all antibiotics. We found that the
302 combined model achieved slightly higher overall accuracies (by ~1-2%), however the matrix
303 that was necessary to train this model had a large memory footprint. Indeed, if we were to
304 build a similar matrix for the current *Salmonella* data set using all 5,278 genomes, the model
305 training would exceed 1.5 TB of RAM. Therefore, we first built models for all antibiotics using
306 subsets of the genomes ranging in size from 250-4,500 genomes that were rationally selected
307 to maximize genetic diversity (Figure 1). A matrix built from 4,500 genomes is the largest we
308 can train on a 1.5 TB machine using this protocol. As the training set size increases from 250 to
309 1000 genomes, the accuracy increases from 88.5% to 91.4%. Then as the training set increases
310 beyond 1000 genomes, the accuracy continues to improve more modestly, with the 4,500-
311 genome model having an average accuracy of 95.2%. Results indicate that the overall MIC
312 prediction approach, which was developed previously for *Klebsiella pneumoniae*, also works for
313 *Salmonella* despite the differences in sampling, genetic diversity and MICs. Also, we discovered

314 that a smaller number of well-chosen diverse genomes can serve as a useful proxy for
315 representing the entire set, since models built from ≥ 500 genomes have accuracies exceeding
316 90%.

317

318 **Model Accuracy**

319 We computed the overall accuracy for each antibiotic using the model that is based on 4,500
320 genomes. For this model, all 15 antibiotics have average accuracies $\geq 90\%$, with their Q_1 quartile
321 bound $\geq 89\%$ (Figure 2). Chloramphenicol and ceftiofur had the highest accuracies (99%), and
322 gentamicin and tetracycline had the lowest accuracies (91% and 90%, respectively) (Table S2).
323 Since the model is robust to the various mechanisms of resistance for the 15 antibiotics, it is
324 possible that the slightly lower accuracies for gentamicin and tetracycline could be due to the
325 distribution of multiple AMR genes/mechanisms across the population of strains with resistant
326 genomes (which will be analyzed in more detail below). Figure 3 depicts the accuracy of the
327 4,500-genome model for each MIC. Overall, the model is robust for both the resistant and
328 susceptible MICs, and it tends to be more accurate when a MIC is represented by many
329 genomes. The model tends to have lower accuracies for the highest and lowest MICs, perhaps
330 because of underlying genetic differences between strains that have been reported with \geq or \leq
331 values, which represents a range of MICs rather than a discrete value.

332

333 The utility of AMR diagnostic devices is often described in terms of error rate. Major errors
334 (MEs) are defined as susceptible genomes that have been incorrectly assigned resistant MICs by
335 the model. Very major errors (VMEs) are defined as resistant genomes that have been
336 incorrectly assigned susceptible MICs by the model. FDA standards for automated systems
337 recommend a major error rate $\leq 3\%$ [53]. All antibiotics used in the model have ME rates within
338 this range (Table 2). The FDA standards for VME rates indicate that the lower 95% confidence
339 limit should be $\leq 1.5\%$ and upper limit should be $\leq 7.5\%$ [53]. Seven of the 15 antibiotics—
340 amoxicillin/clavulanic acid, ceftriaxone, chloramphenicol, cefoxitin, streptomycin, tetracycline
341 and ceftiofur—have acceptable VME rates by this measure. Ampicillin and sulfisoxazole have
342 VME rates with 95% confidence intervals approaching this range: [0.022, 0.033] and [0.026,
343 0.053] respectively. The VME rates are higher for some of the remaining antibiotics because
344 there are fewer resistant genomes. As more resistant genomes are collected, and the data set
345 becomes more balanced, we expect VME rates to be reduced.

346

347 In addition to the extensive MIC data, NARMS reports rich metadata including isolation date,
348 food or animal source, collection year, geographic location and serotype. We computed the
349 accuracy of the model over each available metadata category to determine if the model is
350 robust to these differences and to ensure that no subset is biasing the model. The genomes
351 span a 15-year collection period, with all the years except 2002 (the oldest) and 2016 (the most
352 recent) having over 100 isolates. The model accuracy ranges from 94-97% over each collection
353 year (Table 3). That is, the genetic factors that contribute to the MICs have either remained
354 stable over the 15-year period or have been learned as the model was trained. Although the
355 data set is mostly comprised of poultry meat or live animal isolates, the accuracy ranges
356 between 94-96% over the four contamination sources: turkey, beef, pork, and chicken (Table
357 4). No obvious biases were detected in the accuracies based on the state of isolation (an

358 average of 95% accuracy over 41 states with a 95% CI equal to [0.95-0.96]) (Figure 4) or the
359 serovars of each isolate (94% accuracy over 97 serovars with a 95% CI equal to [0.94-0.96])
360 (Table S3). Since the traditional *Salmonella* serotyping scheme is based the lipopolysaccharide
361 O and flagellar H antigens, which are encoded by genes that influence the cell surface[63], we
362 also constructed a phylogenetic tree for *Salmonella* genomes to observe the model accuracy
363 over the various clades. Overall, no phylogenetic bias in the model accuracy was detected
364 (Figure S1).

365
366 One concern of using a model that is trained on the data from previous years, in some cases
367 over 15 years old, is that the training set is not representative of currently circulating strains.
368 That is, the model may be inaccurate for predicting MICs for genomes of strains that are
369 currently circulating or will emerge in the future. For example, shifts in clonal groups, evolution
370 of AMR-associated genes, or introduction of AMR genes by horizontal gene transfer is
371 possible[64, 65]. We evaluated this possibility by building models from subsets of the whole
372 genome sequence data using strains collected in earlier years and measuring the accuracy of
373 the models on genomes collected in later years. Models were built for years prior to 2009
374 through 2014 and tested on the remaining genomes (Table 5). These models have accuracies
375 ranging from 86-92%. As the number of years used for building the models increases, the
376 number of genomes available for testing decreases, so we also tested each model on only the
377 462 genomes from 2015 and 2016. Similarly, the accuracy of each model on the 2015 and 2016
378 genomes ranges from 87-90% (Table S4). The results indicate that within this data set, models
379 generated from genomes collected at earlier dates yield stable MIC predictions for genomes
380 collected at later dates. This finding is consistent with the pattern of AMR genes that is
381 observed within the data set. Although AMR gene content may vary from year to year, we do
382 not observe any major sweeps or fixation events that drastically alter the AMR gene content of
383 the collection between years, which would cause the MIC predictions to fail for a large fraction
384 of the genomes (Table S5). Taken together, these data suggest that the MIC prediction models
385 generated in this study are likely to be sustainable over time.

386 387 **Genomic regions important for MIC prediction**

388 The 4,500-genome model described above contains data from all antibiotics and MICs, making
389 feature extraction to determine which k-mers contribute to the MIC predictions for each
390 antibiotic difficult. To address this limitation, we modified the protocol by building separate
391 models for each antibiotic. Instead of using 10-mers, we increased the k-mer length to 15
392 nucleotides to reduce redundancy and make them identifiable using BLAST[58]. We also
393 searched for presence or absence of k-mers, rather than using k-mer counts, to simplify the
394 analysis of the XGBoost decision trees. Since a 15-mer matrix can be 4⁵ times larger than a 10-
395 mer matrix, we used <= 1000 diverse genomes to reduce the memory footprint during training.
396 Overall, the average accuracy for the individual models is nearly identical to the average
397 accuracy for the combined 4,500-genome model (96% vs. 95%, respectively), and in nearly all
398 cases, the 95% confidence intervals overlap between the combined and single antibiotic models
399 (Table S6). Thus, for this data set, single antibiotic models with fewer genomes and larger k-
400 mers perform as well as a combined model (Figure S2).

401

402 During model training, XGBoost assigns an importance value to each k-mer used in a decision
403 tree. When the model is used to predict the MICs for a new genome, the k-mers with the
404 highest importance values are the most informative for the MIC prediction. Thus, by analyzing
405 the feature importance values of each k-mer, we can use the models as a tool for
406 understanding the genomic regions that differentiate MICs. For each antibiotic-specific model,
407 we parsed the XGBoost decision trees from each fold of the ten-fold cross validation to extract
408 the importance values for each k-mer. To understand the relationship between known AMR
409 genes and the important k-mers that were chosen by each model, we then searched for k-mers
410 with high importance values within AMR genes that occur in close proximity to an AMR gene (in
411 this case, we consider a window of 3kb, approximately 3 genes, to be a close proximity). Table
412 6 lists the highest-ranking k-mers from each model that occur within or in close proximity to an
413 AMR gene. In most cases, the k-mers correspond to known AMR genes including class A and C
414 beta-lactamases for the beta lactam antibiotics, aminoglycoside nucleotidyl- and
415 acetyltransferases for the aminoglycosides, DNA gyrase and QnrB for the fluoroquinolones,
416 TetA and TetR for tetracycline, and dihydrofolate reductase and dihydropteroate synthase for
417 co-trimoxazole and sulfisoxazole. In the case of azithromycin, the collection contains mostly
418 susceptible genomes (Table 1), so the first macrolide resistance gene observed corresponds
419 with the eighth ranking k-mer. The top ten k-mers with the highest feature importance values
420 from each of the ten folds used in model training are listed in Tables S7-S21. In addition to the
421 top AMR k-mers displayed in Table 6, these tables show other highly ranking k-mers from the
422 same AMR genes as well as k-mers from related genes that are known to confer resistance to
423 the given antibiotic. In some cases, k-mers matching regions or genes from unrelated AMR
424 mechanisms have high importance values, suggesting a pattern of co-occurrence on
425 horizontally transferred genetic elements.

426
427 Since each model is predicting the entire range of MICs, some of the highly ranking k-mers will
428 be used to predict susceptible MICs. To assess this, we computed the fraction of susceptible
429 and resistant genomes with each k-mer from Tables S7-21. The set of k-mers that are most
430 enriched in the susceptible genomes is shown in Table 7. Overall, seven of the top ten k-mers
431 represent significantly different SNPs (P-value < 0.001) in both the complete set of 5,278
432 genomes and in the set of 1,000 diverse genomes used to build the models (Figure S3). The top
433 k-mer associated with susceptibility is from the nalidixic acid model and occurs in the DNA
434 gyrase *gyrA* gene. This is also the top k-mer that was found in an AMR gene for nalidixic acid
435 from Table 6. In this case, the model is relying more heavily on the “wild type” version of the k-
436 mer rather than any of the resistant versions (the remaining k-mers from Table 6 occur almost
437 exclusively in resistant genomes). The same *gyrA* k-mer is also found as a highly ranking k-mer
438 in the case of ciprofloxacin (Table S12). Two significant *gyrA* SNPs are captured by this k-mer
439 (Figure S3). These are missense mutations in the resistant genomes occurring at Ser-83 and
440 Asp-87, and changes at these positions have been shown to confer quinolone resistance in *E.*
441 *coli* [66, 67]. The remaining significant mutations from Figure S3 that occur in the protein-
442 encoding genes are same sense (not amino acid changing) mutations. In the cases of *eptA* (Ser,
443 TCG to TCA), *oadA* (Ala, GCC to GCA), the AraJ precursor gene (Leu, CTG to CTA), and the second
444 *gcd* mutation (Thr, ACG to ACA), the codon changes from a commonly used codon in the
445 susceptible genomes to the least preferred codon in the resistant genomes. In the cases of the

446 *nrfE/nrfF* mutation (Asn, AAT to AAC) and the first *gcd* mutation (Asp, GAC to GAT), the
447 resistant genomes have the preferred codon of the pair. Whether these SNPs have a
448 modulating effect on protein translation or contribute to the fitness of the resistant organisms
449 requires further analysis.

450

451

452 Discussion

453 In this study, we have built machine learning-based MIC prediction models for nontyphoidal
454 *Salmonella* genomes using XGBoost[41] that achieve overall accuracies of 95-96% within ± 1
455 two-fold dilution factor. To our knowledge, this is one of the largest and most accurate MIC
456 prediction models to be published to date. Importantly, it provides a model strategy for
457 performing MIC prediction directly from genome sequence data that could be applied to other
458 human or veterinary pathogens.

459

460 The success of our MIC prediction model was dependent on the large, publicly available,
461 population-based collection of genomes with associated metadata. Since researchers often
462 focus on collecting highly resistant or otherwise unusual strains, the opportunities to generate
463 balanced models are rare. We demonstrate the many benefits from comprehensive sampling
464 for the entire range of possible MICs. First, diverse and balanced data sets improve model
465 accuracies because there is better sampling across all MIC dilutions. Second, having balanced
466 data enabled us to achieve acceptable ME and VME rates for 7 of the 15 antibiotics used in the
467 study. Third, compared with our recent model for *Klebsiella pneumoniae*, the larger and more
468 balanced data set in nontyphoidal *Salmonella* enabled us to build models for individual
469 antibiotics that had similar accuracies to the combined model. This enabled us to begin to
470 disambiguate the important genomic regions relating to resistant and susceptible MICs. Finally,
471 we show that MICs in the susceptible range can be accurately predicted with the algorithm
472 using all genomic data rather than scoping it to known AMR genes or gene polymorphisms.
473 This contrasts with prior work correlating MICs to known resistance mechanisms in
474 *Salmonella*[68]. In future studies, our strategy could be used as a starting point for identifying
475 the subtle genomic changes that result in different MICs.

476

477 For each single-antibiotic model, we analyzed the k-mers that had high feature importance
478 values and were important to the models for predicting MICs. The highly ranking k-mers that
479 were enriched in the resistant genomes mainly occurred within or in close proximity to well-
480 known AMR genes. With the exception of the *gyrA* k-mer, the highly ranking k-mers that were
481 enriched in the susceptible genomes were significant in several cases, but more difficult to
482 interpret. Some of these susceptibility k-mers hint at a possible relationship between AMR and
483 oxidative stress or electron transport, such as the k-mers matching components of the nitrate
484 and nitrite reductases and pqq-dependent glucose dehydrogenase, which is consistent with the
485 known link between antibiotics to oxidative stress[69, 70]. Determining the molecular
486 mechanisms underlying the susceptibility k-mers and AMR phenotypes should be further
487 investigated.

488

489 The genomes in this study were collected over a 15-year period from 41 U.S. states. By building
490 models encompassing ranges of earlier dates, we demonstrated stable and accurate MIC
491 prediction for genomes collected at later dates. Presently, we are not aware of any large
492 publicly available collections of *Salmonella* genomes with MIC data from other countries. Since
493 AMR gene content may vary across pathogen populations, validation of the *Salmonella* models
494 using strains from other countries is important to its application in global health. Nevertheless,
495 the present analysis clearly demonstrates that current model provides accurate MIC predictions
496 for United States isolates. Similarly, an analysis of this model on *Salmonella typhi* strains would
497 provide information about the utility of the model over broader phylogenetic distances.
498
499

500 **Acknowledgements**

501 This work was supported by the National Institute of Allergy and Infectious Diseases, National
502 Institutes of Health, Department of Health and Human Service [Contract No.
503 HHSN272201400027C]. We thank Emily Dietrich for her helpful edits.
504

505 **Author contribution statement**

506 MN: study design, experiments, data generation manuscript preparation, SWL: study design,
507 PFM: study design, data generation, RJO: study design, RO: software engineering, RLS: study
508 design, GHT: study design, data generation, SZ: study design, data generation, JJD: study design,
509 data generation, manuscript preparation
510

511 **Additional Information**

512 **Accession Codes:** Data are available under bioprojects PRJNA292661 and
513 PRJNA292666. SRA run accession for each genome are displayed in (Table S1).
514

515 **Competing financial interests:** The authors claim no competing financial interests.
516

517 **Disclaimer**

518 The views expressed in this article are those of the authors and do not necessarily reflect the
519 official policy of the Department of Health and Human Services, the U.S. Food and Drug
520 Administration, and Centers for Disease Control and Prevention or the U.S. Government.
521 Mention of trade names or commercial products in this publication is solely for the purpose of
522 providing specific information and does not imply recommendation or endorsement by the U.S.
523 Department of Agriculture or Food and Drug Administration.
524
525

526 References

- 527 1. Centers for Disease Control and Prevention (CDC). Surveillance for Foodborne Disease
528 Outbreaks, United States, 2015, Annual Report. Atlanta, Georgia: US Department of Health and
529 Human Services, CDC. 2017. Available from:
530 https://www.cdc.gov/foodsafety/pdfs/2015FoodBorneOutbreaks_508.pdf.
- 531 2. Crim SM, Griffin PM, Tauxe R, Marder EP, Gilliss D, Cronquist AB, et al. Preliminary
532 incidence and trends of infection with pathogens transmitted commonly through food-
533 Foodborne Diseases Active Surveillance Network, 10 US sites, 2006-2014. MMWR Morbidity
534 and mortality weekly report. 2015;64(18):495-9.
- 535 3. Scallan E, Hoekstra RM, Widdowson M, Hall A, Griffin P. Foodborne illness acquired in
536 the United States. Emerging Infectious Diseases. 2011;17(7):1339-40.
- 537 4. World Health Organization. WHO estimates of the global burden of foodborne diseases:
538 foodborne disease burden epidemiology reference group 2007-2015. 2015.
- 539 5. Andino A, Hanning I. *Salmonella enterica*: survival, colonization, and virulence
540 differences among serovars. The Scientific World Journal. 2015;2015.
- 541 6. Aserkoff B, Bennett JV. Effect of antibiotic therapy in acute salmonellosis on the fecal
542 excretion of salmonellae. N Engl J Med. 1969;281(12):636-40. Epub 1969/09/18. doi:
543 10.1056/NEJM196909182811202.
- 544 7. Crump JA, Sjölund-Karlsson M, Gordon MA, Parry CM. Epidemiology, clinical
545 presentation, laboratory diagnosis, antimicrobial resistance, and antimicrobial management of
546 invasive *Salmonella* infections. Clinical microbiology reviews. 2015;28(4):901-37.
- 547 8. Varma JK, Mølbak K, Barrett TJ, Beebe JL, Jones TF, Rabatsky-Ehr T, et al. Antimicrobial-
548 resistant nontyphoidal *Salmonella* is associated with excess bloodstream infections and
549 hospitalizations. The Journal of infectious diseases. 2005;191(4):554-61.
- 550 9. Varma JK, Greene KD, Ovitt J, Barrett TJ, Medalla F, Angulo FJ. Hospitalization and
551 antimicrobial resistance in *Salmonella* outbreaks, 1984–2002. Emerging Infectious Diseases.
552 2005;11(6):943.
- 553 10. Krueger AL, Greene SA, Barzilay EJ, Henao O, Vugia D, Hanna S, et al. Clinical outcomes
554 of nalidixic acid, ceftriaxone, and multidrug-resistant nontyphoidal *Salmonella* infections
555 compared with pansusceptible infections in FoodNet sites, 2006–2008. Foodborne pathogens
556 and disease. 2014;11(5):335-41.
- 557 11. Angulo FJ, Mølbak K. Human health consequences of antimicrobial drug-resistant
558 *Salmonella* and other foodborne pathogens. Clinical infectious diseases. 2005;41(11):1613-20.
- 559 12. Karp BE, Tate H, Plumblee JR, Dessai U, Whichard JM, Thacker EL, et al. National
560 Antimicrobial Resistance Monitoring System: Two Decades of Advancing Public Health Through
561 Integrated Surveillance of Antimicrobial Resistance. Foodborne pathogens and disease.
562 2017;14(10):545-57.
- 563 13. Food and Drug Administration (FDA). NARMS Now. Rockville, MD: 2018. Available from:
564 <https://www.fda.gov/AnimalVeterinary/SafetyHealth/AntimicrobialResistance/NationalAntimicrobialResistanceMonitoringSystem/ucm416741.htm>.
- 565
566 14. Abrams AJ, Trees DL. Genomic sequencing of *Neisseria gonorrhoeae* to respond to the
567 urgent threat of antimicrobial-resistant gonorrhea. Pathogens and disease. 2017;75(4).

- 568 15. Goldberg B, Sichtig H, Geyer C, Ledebner N, Weinstock GM. Making the leap from
569 research laboratory to clinic: challenges and opportunities for next-generation sequencing in
570 infectious disease diagnostics. *MBio*. 2015;6(6):e01888-15.
- 571 16. Didelot X, Bowden R, Wilson DJ, Peto TE, Crook DW. Transforming clinical microbiology
572 with bacterial genome sequencing. *Nature Reviews Genetics*. 2012;13(9):601.
- 573 17. Brown EW, Gonzalez-Escalona N, Stones R, Timme R, Allard MW. The Rise of Genomics
574 and the Promise of Whole Genome Sequencing for Understanding Microbial Foodborne
575 Pathogens. *Foodborne Pathogens: Springer*; 2017. p. 333-51.
- 576 18. McArthur AG, Tsang KK. Antimicrobial resistance surveillance in the genomic age. *Annals
577 of the New York Academy of Sciences*. 2017;1388(1):78-91.
- 578 19. Opota O, Croxatto A, Prod'homme G, Greub G. Blood culture-based diagnosis of
579 bacteraemia: state of the art. *Clinical Microbiology and Infection*. 2015;21(4):313-22.
- 580 20. Reller LB, Weinstein M, Jorgensen JH, Ferraro MJ. Antimicrobial susceptibility testing: a
581 review of general principles and contemporary practices. *Clinical infectious diseases*.
582 2009;49(11):1749-55.
- 583 21. Saha SK, Darmstadt GL, Baqui AH, Hanif M, Ruhulamin M, Santosham M, et al. Rapid
584 identification and antibiotic susceptibility testing of *Salmonella enterica* serovar Typhi isolated
585 from blood: implications for therapy. *Journal of clinical microbiology*. 2001;39(10):3583-5.
- 586 22. Llor C, Bjerrum L. Antimicrobial resistance: risk associated with antibiotic overuse and
587 initiatives to reduce the problem. *Therapeutic advances in drug safety*. 2014;5(6):229-41.
- 588 23. Kumar A, Roberts D, Wood KE, Light B, Parrillo JE, Sharma S, et al. Duration of
589 hypotension before initiation of effective antimicrobial therapy is the critical determinant of
590 survival in human septic shock. *Critical care medicine*. 2006;34(6):1589-96.
- 591 24. Palmer H, Palavecino E, Johnson J, Ohl C, Williamson J. Clinical and microbiological
592 implications of time-to-positivity of blood cultures in patients with Gram-negative bacilli
593 bacteremia. *European journal of clinical microbiology & infectious diseases*. 2013;32(7):955-9.
- 594 25. McDermott PF, Tyson GH, Kabera C, Chen Y, Li C, Folster JP, et al. Whole-genome
595 sequencing for detecting antimicrobial resistance in nontyphoidal *Salmonella*. *Antimicrobial
596 agents and chemotherapy*. 2016;60(9):5515-20.
- 597 26. Hunt M, Mather AE, Sánchez-Busó L, Page AJ, Parkhill J, Keane JA, et al. ARIBA: rapid
598 antimicrobial resistance genotyping directly from sequencing reads. *Microbial genomics*.
599 2017;3(10).
- 600 27. Niehaus KE, Walker TM, Crook DW, Peto TE, Clifton DA, editors. Machine learning for
601 the prediction of antibacterial susceptibility in *Mycobacterium tuberculosis*. 2014 IEEE-EMBS
602 International Conference on Biomedical and Health Informatics (BHI); 2014: IEEE.
- 603 28. Stoesser N, Batty E, Eyre D, Morgan M, Wyllie D, Del Ojo Elias C, et al. Predicting
604 antimicrobial susceptibilities for *Escherichia coli* and *Klebsiella pneumoniae* isolates using whole
605 genomic sequence data. *Journal of Antimicrobial Chemotherapy*. 2013;68(10):2234-44.
- 606 29. Pesesky MW, Hussain T, Wallace M, Patel S, Andleeb S, Burnham C-AD, et al. Evaluation
607 of machine learning and rules-based approaches for predicting antimicrobial resistance profiles
608 in Gram-negative Bacilli from whole genome sequence data. *Frontiers in microbiology*.
609 2016;7:1887.

- 610 30. Lipworth SIW, Hough N, Leach L, Morgan M, Jeffrey K, Andersson M, et al. Whole
611 genome sequencing for predicting *Mycobacterium abscessus* drug susceptibility. *bioRxiv*.
612 2018:251918.
- 613 31. Bradley P, Gordon NC, Walker TM, Dunn L, Heys S, Huang B, et al. Rapid antibiotic-
614 resistance predictions from genome sequence data for *Staphylococcus aureus* and
615 *Mycobacterium tuberculosis*. *Nature communications*. 2015;6:10063.
- 616 32. Harrison OB, Clemence M, Dillard JP, Tang CM, Trees D, Grad YH, et al. Genomic
617 analyses of *Neisseria gonorrhoeae* reveal an association of the gonococcal genetic island with
618 antimicrobial resistance. *Journal of Infection*. 2016;73(6):578-87.
- 619 33. Grad YH, Harris SR, Kirkcaldy RD, Green AG, Marks DS, Bentley SD, et al. Genomic
620 epidemiology of gonococcal resistance to extended-spectrum cephalosporins, macrolides, and
621 fluoroquinolones in the United States, 2000–2013. *The Journal of infectious diseases*.
622 2016;214(10):1579-87.
- 623 34. Coelho JR, Carriço JA, Knight D, Martínez J-L, Morrissey I, Oggioni MR, et al. The use of
624 machine learning methodologies to analyse antibiotic and biocide susceptibility in
625 *Staphylococcus aureus*. *PLoS One*. 2013;8(2):e55582.
- 626 35. Eyre DW, De Silva D, Cole K, Peters J, Cole MJ, Grad YH, et al. WGS to predict antibiotic
627 MICs for *Neisseria gonorrhoeae*. *Journal of Antimicrobial Chemotherapy*. 2017;72(7):1937-47.
- 628 36. Drouin A, Giguère S, Déraspe M, Marchand M, Tyers M, Loo VG, et al. Predictive
629 computational phenotyping and biomarker discovery using reference-free genome
630 comparisons. *BMC genomics*. 2016;17(1):754.
- 631 37. Davis JJ, Boisvert S, Brettin T, Kenyon RW, Mao C, Olson R, et al. Antimicrobial resistance
632 prediction in PATRIC and RAST. *Scientific reports*. 2016;6:27930.
- 633 38. Nguyen M, Brettin T, Long SW, Musser JM, Olsen RJ, Olson R, et al. Developing an in
634 silico minimum inhibitory concentration panel test for *Klebsiella pneumoniae*. *Scientific reports*.
635 2018;8(1):421.
- 636 39. Metcalf BJ, Chochua S, Gertz R, Li Z, Walker H, Tran T, et al. Using whole genome
637 sequencing to identify resistance determinants and predict antimicrobial resistance phenotypes
638 for year 2015 invasive pneumococcal disease isolates recovered in the United States. *Clinical
639 Microbiology and Infection*. 2016;22(12):1002. e1- e8.
- 640 40. Li Y, Metcalf BJ, Chochua S, Li Z, Gertz RE, Walker H, et al. Penicillin-binding protein
641 transpeptidase signatures for tracking and predicting β -lactam resistance levels in
642 *Streptococcus pneumoniae*. *MBio*. 2016;7(3):e00756-16.
- 643 41. Chen T, Guestrin C, editors. XGBoost: A scalable tree boosting system. *Proceedings of
644 the 22nd acm sigkdd international conference on knowledge discovery and data mining; 2016:
645 ACM*.
- 646 42. US Food and Drug Administration (FDA). National Antimicrobial Resistance Monitoring
647 System—Enteric Bacteria (NARMS): 2011 executive report. US Department of Health and Human
648 Services. Food and Drug Administration, Rockville, MD. 2013.
- 649 43. Wattam AR, Davis JJ, Assaf R, Boisvert S, Brettin T, Bun C, et al. Improvements to
650 PATRIC, the all-bacterial bioinformatics database and analysis resource center. *Nucleic acids
651 research*. 2016;45(D1):D535-D42.
- 652 44. Nikolenko SI, Korobeynikov AI, Alekseyev MA, editors. BayesHammer: Bayesian
653 clustering for error correction in single-cell sequencing. *BMC genomics; 2013: BioMed Central*.

- 654 45. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a
655 new genome assembly algorithm and its applications to single-cell sequencing. *Journal of*
656 *computational biology*. 2012;19(5):455-77.
- 657 46. Brettin T, Davis JJ, Disz T, Edwards RA, Gerdes S, Olsen GJ, et al. RASTtk: a modular and
658 extensible implementation of the RAST algorithm for building custom annotation pipelines and
659 annotating batches of genomes. *Scientific reports*. 2015;5:8365.
- 660 47. Antonopoulos DA, Assaf R, Aziz RK, Brettin T, Bun C, Conrad N, et al. PATRIC as a unique
661 resource for studying antimicrobial resistance. *Briefings in bioinformatics*. 2017.
- 662 48. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7:
663 improvements in performance and usability. *Molecular biology and evolution*. 2013;30(4):772-
664 80.
- 665 49. Price MN, Dehal PS, Arkin AP. FastTree 2—approximately maximum-likelihood trees for
666 large alignments. *PloS one*. 2010;5(3):e9490.
- 667 50. Letunic I, Bork P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree
668 display and annotation. *Bioinformatics*. 2006;23(1):127-8.
- 669 51. Deorowicz S, Kokot M, Grabowski S, Debudaj-Grabysz A. KMC 2: fast and resource-frugal
670 k-mer counting. *Bioinformatics*. 2015;31(10):1569-76.
- 671 52. Jorgensen JH. Selection criteria for an antimicrobial susceptibility testing system. *Journal*
672 *of clinical microbiology*. 1993;31(11):2841.
- 673 53. US Food and Drug Administration (FDA). Class II Special Controls Guidance Document:
674 Antimicrobial Susceptibility Test (AST) Systems. Rockville, MD: US FDA. 2009.
- 675 54. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn:
676 Machine Learning in Python. *Journal of Machine Learning Research*. 2011;12:2825--30.
- 677 55. Bellman R. Dynamic programming. Princeton: Princeton University Press; 2013.
- 678 56. Shalev-Shwartz S, Ben-David S. Understanding machine learning: From theory to
679 algorithms: Cambridge University Press; 2014.
- 680 57. Aggarwal CC, Hinneburg A, Keim DA, editors. On the surprising behavior of distance
681 metrics in high dimensional space. *International conference on database theory*; 2001:
682 Springer.
- 683 58. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+:
684 architecture and applications. *BMC bioinformatics*. 2009;10(1):421.
- 685 59. Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ. Jalview Version 2—a
686 multiple sequence alignment editor and analysis workbench. *Bioinformatics*. 2009;25(9):1189-
687 91.
- 688 60. Crooks GE, Hon G, Chandonia J-M, Brenner SE. WebLogo: A Sequence Logo Generator.
689 *Genome Research*. 2004;14(6):1188-90. doi: 10.1101/gr.849004. PubMed PMID: PMC419797.
- 690 61. Davis JJ, Olsen GJ. Modal codon usage: assessing the typical codon usage of a genome.
691 *Molecular biology and evolution*. 2009;27(4):800-10.
- 692 62. Davis JJ, Olsen GJ. Characterizing the native codon usages of a genome: an axis
693 projection approach. *Molecular biology and evolution*. 2010;28(1):211-21.
- 694 63. Ranieri ML, Shi C, Switt AIM, Den Bakker HC, Wiedmann M. Comparison of typing
695 methods with a new procedure based on sequence characterization for *Salmonella* serovar
696 prediction. *Journal of clinical microbiology*. 2013;51(6):1786-97.

- 697 64. Zhu L, Olsen RJ, Nasser W, Beres SB, Vuopio J, Kristinsson KG, et al. A molecular trigger
698 for intercontinental epidemics of group A *Streptococcus*. The Journal of clinical investigation.
699 2015;125(9):3545-59.
- 700 65. Nasser W, Beres SB, Olsen RJ, Dean MA, Rice KA, Long SW, et al. Evolutionary pathway
701 to increased virulence and epidemic group A *Streptococcus* disease derived from 3,615 genome
702 sequences. Proceedings of the National Academy of Sciences. 2014;111(17):E1768-E76.
- 703 66. Yoshida H, Kojima T, Yamagishi J-i, Nakamura S. Quinolone-resistant mutations of the
704 *gyrA* gene of *Escherichia coli*. Molecular and General Genetics MGG. 1988;211(1):1-7.
- 705 67. Yoshida H, Bogaki M, Nakamura M, Nakamura S. Quinolone resistance-determining
706 region in the DNA gyrase *gyrA* gene of *Escherichia coli*. Antimicrobial agents and chemotherapy.
707 1990;34(6):1271-2.
- 708 68. Tyson GH, Zhao S, Li C, Ayers S, Sabo JL, Lam C, et al. Establishing genotypic cutoff values
709 to measure antimicrobial resistance in *Salmonella*. Antimicrobial agents and chemotherapy.
710 2017;61(3):e02140-16.
- 711 69. Kohanski MA, Dwyer DJ, Hayete B, Lawrence CA, Collins JJ. A common mechanism of
712 cellular death induced by bactericidal antibiotics. Cell. 2007;130(5):797-810.
- 713 70. Foti JJ, Devadoss B, Winkler JA, Collins JJ, Walker GC. Oxidation of the guanine
714 nucleotide pool underlies cell death by bactericidal antibiotics. Science. 2012;336(6079):315-9.
715
716

717 **Tables**

718

719 **Table 1.** The number of susceptible, intermediate and resistant genomes across the 15
720 antibiotics for the 5278 *Salmonella* genomes used in this study.

Antibiotic	Susceptible genomes	Intermediate genomes	Resistant genomes
AMP	3682	2	1593
AUG	4145	355	778
AXO	4508	1	769
AZI	2409	0	7
CHL	5026	87	164
CIP	5217	53	7
COT	5219	0	58
FIS	3356	0	1573
FOX	4501	98	679
GEN	4577	68	633
KAN	837	3	84
NAL	5233	0	45
STR	872	0	1919
TET	2364	28	2885
TIO	4517	8	753

721

722

723
724
725
726
727

Table 2. Very major error (VME) rate, defined as resistant genomes predicted as being susceptible, and major error (ME) rate, defined as susceptible genomes predicted as being resistant, for the 4500-genome model.

Antibiotic	VME Avg ¹	VME 95% CI ²	ME Avg ¹	ME 95% CI ²	Resistant Samples	Susceptible Samples
All	0.027	[0.024-0.030]	0.001	[0.001-0.002]	10979	47366
AMP	0.028	[0.022-0.033]	0.000	[0.000-0.001]	1442	3054
AUG	0.012	[0.000-0.025]	0.000	[0.000-0.000]	746	3449
AXO	0.022	[0.011-0.032]	0.000	[0.000-0.001]	740	3758
AZI	0.857	[0.508-1.207]	0.000	[0.000-0.000]	7	2040
CHL	0.000	[0.000-0.000]	0.000	[0.000-0.001]	149	4271
CIP	0.417	[-0.099-0.933]	0.000	[0.000-0.000]	7	4445
COT	0.670	[0.515-0.825]	0.000	[0.000-0.001]	55	4443
FIS	0.039	[0.026-0.053]	0.000	[0.000-0.000]	1479	2757
FOX	0.009	[-0.001-0.020]	0.000	[0.000-0.000]	651	3754
GEN	0.090	[0.066-0.113]	0.000	[0.000-0.000]	579	3862
KAN	0.074	[0.012-0.136]	0.000	[0.000-0.000]	82	662
NAL	0.917	[0.819-1.014]	0.000	[0.000-0.001]	39	4460
STR	0.014	[0.008-0.020]	0.027	[0.013-0.040]	1703	744
TET	0.000	[0.000-0.000]	0.018	[0.012-0.025]	2575	1901
TIO	0.004	[-0.001-0.009]	0.000	[0.000-0.000]	725	3766

728 ¹ Reported within ± 1 two-fold dilution step

729 ² 95% confidence interval

730
731
732

733
734
735

Table 3. Model accuracy for the genomes from each sample collection year.

Collection Date	Accuracy	Genomes	Bins [*]
2002	0.97	55	624
2003	0.95	159	1809
2004	0.96	235	2850
2005	0.95	274	3384
2006	0.95	313	3880
2007	0.94	258	3192
2008	0.95	388	4821
2009	0.95	436	5367
2010	0.94	230	2820
2011	0.95	214	2968
2012	0.96	257	3694
2013	0.97	265	3793
2014	0.95	506	7100
2015	0.95	689	9646
2016	0.96	83	1161

736 ^{*}The total number of MICs available for the genomes isolated in that year
737
738
739

Table 4. Model accuracy for the genomes isolated from various sources.

Source	Accuracy	Genomes	Bins [*]
Chicken	0.96	1981	25869
Cow/Beef	0.94	419	5688
Pig/Pork	0.95	448	6144
Turkey	0.94	1651	21260

740 ^{*}The total number of MICs available for the genomes of each category
741

Table 5. The ability of models trained on genomes from prior years to predict MICs for genomes collected in later years.

Training set years	Test set years	Accuracy	95% CI	Training Bins*	Testing Bins*	Training Genomes	Testing Genomes
2002-2008	2009-2016	0.88	[0.88-0.89]	36563	22412	1819	2681
2002-2009	2010-2016	0.88	[0.88-0.89]	31196	27779	2255	2245
2002-2010	2011-2016	0.88	[0.88-0.88]	28376	30599	2485	2015
2002-2011	2012-2016	0.88	[0.88-0.89]	25408	33567	2699	1801
2002-2012	2013-2016	0.88	[0.87-0.88]	21714	37261	2956	1544
2002-2013	2014-2016	0.86	[0.86-0.87]	17921	41054	3221	1279
2002-2014	2015-2016	0.92	[0.92-0.92]	10807	48168	3728	772

*The total number of genome/antibiotic combinations

Table 6. The highest-ranking AMR-related protein function (or genomic region) with a matching k-mer from the XGBoost models.

Antibiotic	K-mer Rank	Type of match ¹	k-mer	PATRIC Annotation(s)
AMP	1	direct	CTTAATCAGTGAGGC	Class A beta-lactamase (EC 3.5.2.6) => TEM family
AUG	1	direct	AAACGTCTTACTAAC	Class C beta-lactamase (EC 3.5.2.6) => CMY/CMY-2/CFE/LAT family
AXO ²	1	proximity	AAAGAGAAAGAAAGG	Class C beta-lactamase (EC 3.5.2.6) => CMY/CMY-2/CFE/LAT family
AZI	8	direct	CCCATTTCGCCGCC	Macrolide 2'-phosphotransferase => Mph(A) family
CHL ²	1	proximity	AGACAAGTAAGCCGC	Chloramphenicol/florfenicol resistance, MFS efflux pump => FloR family
CIP	1	proximity	ACAGTCCATCCAGGA	Pentapeptide repeat protein QnrB family => Quinolone resistance protein QnrB10
COT ²	1	proximity	AAAAACGATAGCTGC	Dihydrofolate reductase (EC 1.5.1.3)
FIS ²	1	proximity	CGCAACGGCTCAAGC	Dihydropteroate synthase type-2 (EC 2.5.1.15) @ Sulfonamide resistance protein
FOX	1	direct	AAAAAAACCTTGCA	Class C beta-lactamase (EC 3.5.2.6) => CMY/CMY-2/CFE/LAT family
GEN	1	proximity	AGTTAAGCCGCGCCG	Aminoglycoside 3''-nucleotidyltransferase (EC 2.7.7.-) => APH(3'')-Ia (AadA family); Aminoglycoside N(3)-acetyltransferase (EC 2.3.1.81) => AAC(3)-II,III,IV,VI,VIII,IX,X
KAN	1	direct	AAAAAGCCGTTTCTG	Aminoglycoside 3'-phosphotransferase (EC 2.7.1.95) => APH(3')-I
NAL	1	direct	ATTCCGCAGTGATG	DNA gyrase subunit A (EC 5.99.1.3)
STR	1	direct	ATTTGTACGGCTCCG	Aminoglycoside 3''-nucleotidyltransferase (EC 2.7.7.-) => APH(3'')-Ia (AadA family)
TET	1	proximity	CGTTCTGCCTTGCGC	Tetracycline resistance regulatory protein TetR; Tetracycline resistance, MFS efflux pump => Tet(A)
TIO ²	1	proximity	AAAGAGAAAGAAAGG	Class C beta-lactamase (EC 3.5.2.6) => CMY/CMY-2/CFE/LAT family

¹A "direct" match means that the k-mer is an exact match to the protein-encoding gene, a "proximity" match means that an AMR gene occurs within 3kb of the k-mer. The k-mer sequences and the top k-mers for each antibiotic are shown in Tables S7-21.

²Has other AMR genes within 3kb listed in Tables S7-21.

Table 7. Important k-mers used by the individual antibiotic models for predicting susceptible MICs.

Antibiotic	k-mer	Sus ¹	Res ¹	Frac Sus ²	Frac Res ²	Genomic region ³	PATRIC annotation or genomic region
NAL	ATCCGCAGTGTATG	5233	45	1.00	0.38	PEG	DNA gyrase subunit A (EC 5.99.1.3)
AXO	TGGTATTCGCATCAA	4508	769	0.78	0.48	PEG	Phosphoethanolamine transferase EptA
KAN	CTGGCTTTTTTTTTT	837	84	0.30	0.00	RNA	RyhB RNA
STR	CCCTTATCCAACACG	872	1919	0.85	0.55	PEG	Respiratory nitrate reductase delta chain (EC 1.7.99.4)
AXO	CAGAACCAGAATTTG	4508	769	0.74	0.46	PEGs	Formate-dependent nitrite reductase complex subunit Nrff, and Cytochrome c-type heme lyase subunit nrfE, nitrite reductase complex assembly
TIO	AGAGAAGCCTGCCGC	4517	753	0.68	0.40	PEG	Oxaloacetate decarboxylase alpha chain (EC 4.1.1.3)
AXO	ATCCCCGCCATTACA	4508	769	0.73	0.46	PEG	Tagatose-1,6-bisphosphate aldolase GatY (EC 4.1.2.40)
AXO	TGCTGCAAAACGCCA	4508	769	0.69	0.45	PEG	Protein AraJ precursor
AXO	GAAAACAGGGTGTAG	4508	769	0.47	0.23	INT	Upstream of llvGMEDA operon leader peptide
FOX	GGATACCACGCCGGG	4501	679	0.58	0.35	PEGs	Glucose dehydrogenase, PQQ-dependent (EC 1.1.5.2), and IncF plasmid conjugative transfer protein TraP

¹ Total number of susceptible and resistant genomes in the entire collection

² Fraction of susceptible and resistant genomes with the k-mer in the entire collection

³ PEG is protein-encoding gene, RNA is RNA gene, INT is intergenic region

Figure Legends

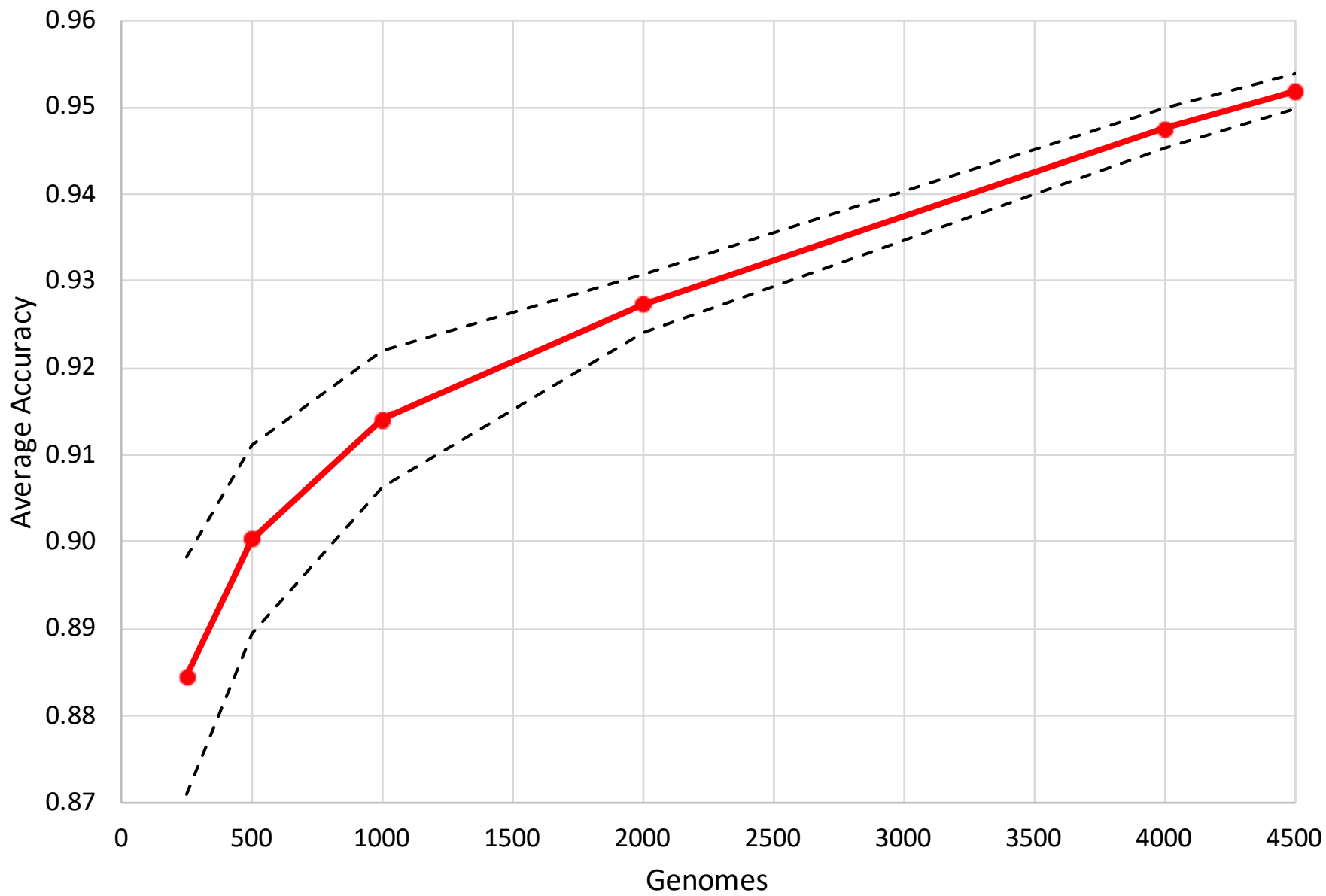
Figure 1. MIC prediction model accuracy for subsamples of genomes. Diverse subsamples of genomes were chosen and the model accuracy within ± 1 two-fold dilution step based on a 10-fold cross validation is shown with the red plot line. The dashed line represents the high and low values for the 95% confidence interval for the average accuracy at each given plot point.

Figure 2. Box plot of the overall accuracies within ± 1 two-fold dilution step for each antibiotic in the 4500-genome model. The Y-axis depicts each antibiotic (abbreviations are defined in Materials and Methods). The X-axis depicts the accuracy. Each vertical red line represents the median accuracy over the holdout sets for each fold in the ten-fold cross validation. The blue box encompasses the data of the first and third quartiles. The dashed blue horizontal lines bounded by black vertical lines (or “whiskers”) depict the entire distribution of accuracies for each fold and antibiotic. The accuracy of the entire 4500 genome model over all antibiotics and folds is depicted in the row marked “ALL”.

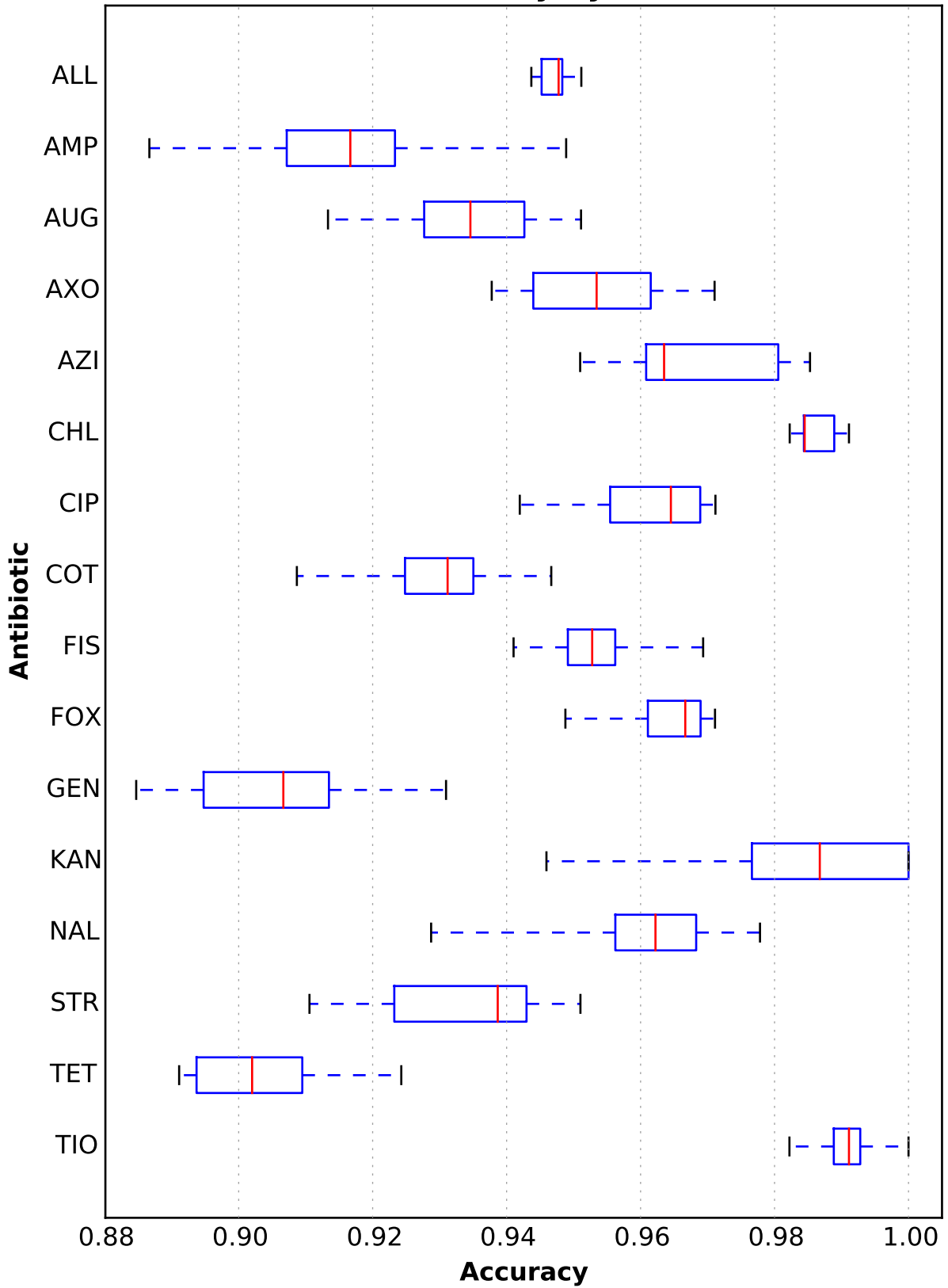
Figure 3. The accuracy of the MIC prediction model based on 4,500 diverse genomes. The heat map depicts the accuracy within ± 1 two-fold dilution step of the laboratory-derived MIC. The X-axis shows the MIC ($\mu\text{g/ml}$) and each antibiotic is shown on the Y-axis. The accuracy for each antibiotic-MIC 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.

Figure 4. The average accuracy of the model based on 4,500 diverse genomes for predicting MICs for the *Salmonella* genomes from each state. Light blue is most accurate and dark blue/black is least accurate. Note that the scale starts at an accuracy of 0.90. Each state is labeled with the number of genomes collected from that state. States without a label contain no samples and are colored in grey; no genomes exist in the collection from Alaska and Hawaii.

Model Accuracy vs. Genomes Used



Overall Accuracy by Antibiotic



4500 Genome Model

