Understanding and predicting ciprofloxacin minimum inhibitory concentration in *Escherichia coli* with machine learning

1.1 Author names

Bálint Ármin Pataki[a,b], Sébastien Matamoros[c], Boas C.L. van der Putten[c,f], Daniel Remondini[d], Enrico Giampieri[e], Derya Aytan-Aktug[g], COMPARE ML- AMR group*, Rene S. Hendriksen[g], Ole Lund[h], István Csabai[a,b], Constance Schultsz[c,f]

1.2 Affiliation

[a] Department of Physics of Complex Systems, ELTE Eötvös Loránd University, Budapest, Hungary

[b] Department of Computational Sciences, Wigner Research Centre for Physics of the HAS, Budapest, Hungary

[c] Amsterdam UMC, University of Amsterdam, Department of Medical Microbiology, Amsterdam, The Netherlands

[d] Department of Physics and Astronomy (DIFA), University of Bologna, Bologna, Italy

[e] Department of Experimental, Diagnostic and Specialty Medicine (DIMES), University of Bologna, Bologna, Italy

[f] Amsterdam UMC, University of Amsterdam, Department of Global Health, Amsterdam Institute for Global Health and Development, Amsterdam, The Netherlands

[g] National Food Institute, Technical University of Denmark, Lyngby, Denmark

[h] Department of Bioinformatics, Technical University of Denmark, Lyngby, Denmark

*: see the full list of the COMPARE ML-AMR group members in acknowledgements.

1.3 Corresponding author

corresponding author: Bálint Ármin Pataki, email: patbaa@gmail.com

1.4 Keywords

AMR, MIC, machine learning, antibiotics, personalized medicine, ciprofloxacin

1.5 Repositories:

Our work does not contain new sequence data.

2. Abstract

A possible way to tackle the crisis of antimicrobial resistance development is a strict policy when prescribing antibiotics. Thus, it is important that prescriptions are based on antimicrobial susceptibility data to ensure effective treatment outcomes. The increasing availability of next-generation sequencing (NGS), bacterial whole genome sequencing (WGS) can facilitate a more reliable and faster alternative to traditional phenotyping for the detection and surveillance of AMR.

This work proposes a machine learning approach that can predict the minimum inhibitory concentration (MIC) for a given antibiotic, here ciprofloxacin, on the basis of both genome-wide mutation profiles and profiles of acquired antimicrobial resistance genes (ARG). We analysed 704 *Escherichia coli* genomes combined with their respective MIC measurements for ciprofloxacin originating from different countries. The four most important predictors found by the model, mutations in *gyrA* residues Ser83 and Asp87, a mutation in *parC* residue Ser80 and presence of any *qnrS* gene, have been experimentally validated before. Using only these four predictors in a linear regression model, 65% and 92% of the test samples' MIC were correctly predicted within a two- and a four-fold dilution range, respectively. The presented work goes further than the typical predictions that use machine learning as a black box model concept. The recent progress in WGS technology in combination with machine learning analysis approaches indicates that in the near future WGS of bacteria might become cheaper and faster than a MIC measurement.

3. Impact statement

Whole genome sequencing has become the standard approach to study molecular epidemiology of bacteria. However, the application of WGS in the clinical microbiology laboratory as part of individual patient diagnostics still requires significant steps forward, in particular with respect to prediction of antibiotic susceptibility based on DNA sequence. Whilst the majority of studies of prediction of susceptibility have used a binary outcome (susceptible/resistant), a quantitative prediction of susceptibility, such as the MIC, will allow for earlier detection of trends in increasing resistance as well as the flexibility to follow potential adjustments in definitions of susceptible (wild type) and resistant (non-wild type) categories (breakpoints/ epidemiological cut-off values).

4. Data summary

In this study, 704 *E. coli* genomes combined with MIC measurement for ciprofloxacin were analysed (24). Paired-end sequencing was performed on all isolates and the results were

stored in FASTQ format. The isolates originated from five countries, Denmark, Italy, USA, UK, and Vietnam. The MIC distribution for these isolates is depicted in Table 1. Out of 704, 266 *E. coli* genomes had no country metadata available and were used as an independent test set. All data were deposited in the AMR Data Hub (24) which consists of raw sequencing data, ciprofloxacin minimum inhibitory concentrations, and additional metadata such as the origin of the samples.

Publicly available sequencing data was used from projects PRJEB21131, PRJNA266657, PRJNA292901, PRJNA292904, PRJNA292902, PRJDB7087, PRJEB21880, PRJEB21997, PRJEB14086 and PRJEB16326.

Download and analysis scripts are available at https://github.com/patbaa/AMR_ciprofloxacin . iTOL phylogenetic tree is available at https://itol.embl.de/tree/1451172261149139156948596 9 .

The authors confirm all supporting data, code and protocols have been provided within the article or through supplementary data files.

5. Introduction

Antibiotics are an essential resource in the control of infectious diseases; they have been a major contributor to the decline of infection-associated mortality and morbidity in the 20th century. However, the recent rise of antimicrobial resistance (AMR) threatens this situation (1). Bacterial AMR is associated with a higher likelihood of therapeutic failure in case of infections. Accurate and fast prediction of AMR in bacteria is needed to select the optimal therapy.

With the increasing availability of next-generation sequencing (NGS), bacterial whole genome sequencing (WGS) is becoming a feasible alternative to traditional phenotyping for the detection and surveillance of AMR (2), (3), (4). However, data analysis remains the weak point in this approach; fast and scalable methods are required to transform the ever-growing amount of genomic data into actionable clinical or epidemiological information (5). Machine learning is a promising approach for this kind of data analysis.

AMR can be predicted in numerous ways. In addition to classic and highly standardized phenotypic testing of resistance, several methods of resistance prediction have been developed. Most novel methods use a genetic or genomic approach, although transcriptomic approaches have been investigated as well (6), (7), (8). An important factor in the choice of the resistance prediction method is the microorganism under study. For example, the CRyPTIC consortium managed to predict resistance to four first-line drugs in *Mycobacterium tuberculosis*, using only known mutations

93 extracted from WGS (9). However, *M. tuberculosis* displays little-to-no horizontal gene transfer and
94 low genomic evolution rate (10), which makes it feasible to predict resistance only on the basis of
95 known mutations (11). For other bacteria, more advanced analysis methods such as machine learning
96 need to be applied to allow for accurate prediction.

97 Machine learning has been applied to predict resistance from WGS data in several settings. To
98 date, these methods have been restricted mostly to assign bacteria to binary categories, i.e.
99 susceptible or non-susceptible (12), (13), (14), (8), (15), (16), (17). However, clinical breakpoints used
100 to define susceptible and non-susceptible categories can change and such binary categories do not
101 allow following more trends in subtle changes in susceptibility in time. MIC measures offer an
102 adequate resolution to follow if susceptibility is changing in a population, which is useful for
103 epidemiological purposes. Therefore, a resistance prediction method would preferably output a
104 continuous estimate of resistance similar to MIC, instead of binary classification (S/R), as a number
105 of studies already proposed (18), (19), (20), (21).

106 Additional issues should be considered when developing a reliable and useful prediction model. First,
107 genotypes are often geographically clustered (22). This implies that if a prediction model is trained
108 on data from one country, this model might not be generalizable to data from another country. Data
109 from multiple countries are thus needed. Secondly, complex combinations of chromosomal point
110 mutations and acquired resistance genes influence antimicrobial resistance. Therefore, different data
111 types need to be combined to obtain a biologically relevant set of input data. Lastly, while machine
112 learning is able to analyse highly complex patterns of features, the model would preferably output
113 generally understandable data. K-mer profiles have been used to predict resistance, but these can
114 be difficult to interpret (19) (20).

115 In this study, we focus on predicting a quantitative measure of ciprofloxacin resistance (MIC) for a
116 geographically diverse population of *E. coli* using machine learning.
117 We chose to study ciprofloxacin resistance in *E. coli* because of three reasons:

118 1. Ciprofloxacin resistance in *E. coli* has been studied intensively
119 2. Ciprofloxacin resistance in *E. coli* can be caused by a range of
120 different chromosomal and plasmid-mediated mechanisms (23)
121 3. Ciprofloxacin is commonly used in the treatment of *E. coli* infections
122 across the globe.

123 In our selection of machine learning models, an important criterion was that high-scoring features
124 could be extracted from the model. This would allow us to explore the reasoning behind each
125 prediction and thus to interpret and understand the model.

126 6. Methods

127 6.1 Data preprocessing

128 Raw reads were mapped on the ATCC 25922 reference genome
129 (https://www.ncbi.nlm.nih.gov/assembly/GCF_000743255.1) using BWA-MEM v0.7.17 (25) with
130 default settings. Pileup files were generated with bcftools v1.9 (26) with "–min-MQ 50" settings.
131 Single-nucleotide polymorphisms (SNPs) and insertions-deletions (INDELs) were called using bcftools
132 v1.9 with "–ploidity 1 -m" flags. Further filtering was applied via bcftools v1.9 "%QUAL>=50 & DP>=20"
133 flags. Bcftools output data was expressed as either a SNP (value: 1), an INDEL (value: 5) or no mutation
134 (value: 0) per position in the reference genome. Exact numbers are irrelevant, as tree-based methods
135 are not sensitive to the scale. The intention was to differentiate between reference alleles, SNPs and
136 INDELs at a given position. Acquired resistance genes were identified using ResFinder v3.1.0 (27) with
137 a coverage threshold of 60% and an identity threshold of 90% using a database downloaded on 5th
138 Dec. 2018. ResFinder was used with KMA v1.1.4 (28). The ResFinder output data was expressed as
139 presence (value: 1) or absence (value: 0) of resistance genes. The SNP/INDEL data and ResFinder data
140 were subsequently merged which provided a matrix with more than 830,000 columns representing
141 reference genome positions with at least one mutation and 959 columns representing detected
142 resistance genes. Two more binary columns were added manually, which describe if any *qnr* or *qnrS*
143 gene is present in the given genome or not. Grouping together *qnr* and *qnrS* genes such a way was
144 motivated by their proven role in ciprofloxacin resistance (23). In the ResFinder database these genes
145 have many different variants and for the machine learning models it helps to provide aggregated
146 information too. The different specific gene variants may be present only in a few isolates each.

147 6.2 Phylogenetic tree generation

148 The merged variant call files were converted to a FASTA alignment using vcf2phylip v2.0, retaining
149 positions that were called in at least 50% of isolates (29). The invariant positions were removed from
150 the alignment using snp-sites v2.4.0 (30). The phylogeny was inferred using RAxML v8.2.9 in rapid
151 bootstrap mode (-f a) with 100 bootstraps using a General Time Reversible model with Gamma rate
152 heterogeneity including Lewis ascertainment bias correction (-m ASC_GTRGAMMA) (31). The resulting
153 phylogeny was visualized in iTOL (32).

6.3 Metrics

We used the following metrics for the evaluation of the model:

AUC - area under the receiver operating characteristics curve - we used the clinical breakpoint for ciprofloxacin, 1 mg/L, based on the Clinical & Laboratory Standards Institute guideline (33) to binarize the samples whether they are resistant or not.

R^2 score - coefficient of determination

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

where

- y_i is the true value for sample i,
- $\hat{y}_i$ is the predicted value for sample i,
- $\bar{y}$ is the mean of the true values.

R - Pearson correlation coefficient

$$R_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

where cov is the covariance and σ is the standard deviation.

ME - major error - when the sample is non-resistant by measurement, but it is predicted to be resistant. Non-resistant and resistant labels are derived from MIC via thresholding.

VME - very major error - when the sample is resistant by measurement, but it is predicted to be non-resistant. Non-resistant and resistant labels are derived from MIC via thresholding.

ACC-2 - accuracy within two-fold dilution - the fraction of the samples with MIC properly predicted within a two-fold dilution. If the measured MIC is x, then the prediction is counted as properly predicted within a two-fold dilution if it falls to the [x/2;2x] interval.

ACC-4 - accuracy within four-fold dilution - the fraction of the samples with MIC properly predicted within a four-fold dilution. If the measured MIC is x, then the prediction is counted as properly predicted within a four-fold dilution if it falls to the [x/4;4x] interval.

MAFE - mean absolute fold error - The mean absolute difference between the log2 values of the prediction and the measurements.

6.4 Importance of the validation scheme

Proper validation is a key element in machine learning as most of the models have a large number of parameters. In image recognition, popular convolutional neural networks can have more than 100M parameters (34). This number of parameters is orders of magnitudes larger than the number of pixels

of a single image or even the number of the images in the full usual training data set, such as ImageNet (35). Having that many parameters it is possible to memorize the training data without generalizing any knowledge to the test data or for future use.

However, with having a proper validation scheme it is easy to test the generalization power of a model. In many cases simply randomly splitting the samples into two groups to a test and a validation set is enough. If the data set is small, cross-validation is needed, usually, K-fold cross-validation, where the data set is split into K set, each having the same size. Then, the model is trained on using data from K − 1 set and the predictions are made for the one set that was not used in the training process. Repeating the process, K times, predictions can be generated for the whole data set in a way that the model did not see in training time any of the samples for which it is generating predictions. The weights of the model are reset between any two training.

K-fold cross-validation can produce too optimistic results if the samples are clustered. For example, when the data collection is biased, bacterial isolates from one country are predominantly resistant whilst isolates from other countries are predominantly susceptible to an antibiotic. In addition, genetic signatures are often clustered by country (22). Due to such clustering, the model may predict based on the country of origin of the bacterial isolate, which may be correlated with the MIC, on both the training and the validation data sets, but it is not guaranteed that the same will happen in real-life usage later.

6.4.1 Leave-one-country-out validation

Here we propose a more strict and reliable validation method. Instead of randomly splitting the data into K different folds, we split the folds by country. Using this approach, the model is not rewarded if it only learns country-specific attributes. Leave-one-country-out validation was performed during the selection of the most important features in the data set, see Table S1. The random forest model was fitted K = 5 times leaving out one country each time from the training data set. Then the feature importances were summed over each fold resulting in the final feature importance rankings.

6.5 Random forest model

For tabular data most often tree-like models perform the best. The random forest model is an ensemble of numerous (usually hundreds of) decision trees. In the training process, each tree is trained separately and each of them uses only a random fraction of the data, which ensures that the decision trees will not be identical. For a new sample, the prediction is the average of the prediction of the trees, or for classification the category that was predicted the most often by the individual trees.

216 This ensemble technique ends up an accurate, robust, scalable model. The prediction error is usually
217 large for each individual tree, but as long as the errors of the trees are uncorrelated, averaging their
218 prediction lowers the final error. Random forest regressor was trained on the whole training dataset
219 with mean squared error criterion, n_estimators = 200 for the feature selection. For the final
220 evaluation mean squared error criterion, min_samples_split = 5, and n_estimators = 100 parameters
221 were used. The random seed was fixed for both model to be 42. Other parameters remained default.
222 Scikit-learn v0.21.2 (36) was used for fitting the model in Python 3.6.5.

223 **6.5.1 Random forest feature importance**

224 For decision trees, the input variables can be ranked by their contribution to the predictions. This score
225 is called feature importance. The importance can be defined in various ways; the used scikit-learn
226 v0.21.2 (36) implementation calculates the mean decrease impurity averaged over all the trees in the
227 forest (37) (38). In this approach, the identification of the most important predictors becomes feasible
228 even for cases when there are hundreds of thousands of features.

229 ## 6.6 Model fitting

230 All models were fitted on the log2 values of the MIC, which is the natural scale for the MIC
231 measurement. Later the predicted values were converted back to the MIC units.

232 ## 6.7 Study pipeline

233 The pipeline of this study is shown in Figure 2. First, the raw reads were converted to a numerical table
234 indicating mutations and acquired resistance encoding genes. In the second step, a random forest
235 model is fitted on the train data via leave-one-country cross-validation. Features importances were
236 averaged over each fold. Then the highest-ranking features were kept which significantly reduced the
237 dimensionality of the data. Using this low dimensional training data a random forest model and a
238 linear regression were fitted. Regression models were preferred because they can have a precise
239 susceptibility level estimation instead of classification, which has only a binary susceptibility outcome.
240 For fitting the models always the log2 MIC values were used as a natural scale for the MIC
241 measurements. At the last step, the performance of the models was evaluated on the unseen test
242 data using the same restricted feature set.

243 ## 7. Results

244 Our dataset consists of a phylogenetically diverse collection of *E. coli* strains (Fig. 1). Strains in the test
245 and train group are present throughout the whole phylogeny, although the groups are present
246 predominantly in different parts of the phylogeny.

247 We trained a machine learning model using genome-wide mutation profiles alongside the ResFinder-
248 based profiles of acquired resistance genes. We ranked the predictors proposed by the model itself,
249 see Table S1. The model performed with high accuracy on the training set leave-one-country-out
250 cross-validation using four predictors, see Fig S1. The addition of more features did not improve the
251 cross-validation results, and therefore we kept only the first four, allowing for a simple and
252 understandable model.

253 Using these four predictors, 265 out of the 266 test data samples were correctly classified by our
254 models at susceptible/non-susceptible level, and more than 92% of the corresponding MIC values
255 were correctly predicted within a four-fold dilution, see Table 2.

256 These 4 predictors are the following:
257 1. *gyrA* mutation at amino acid #87
258 2. *gyrA* mutation at amino acid #83
259 3. *parC* mutation at amino acid #80
260 4. presence of any *qnrS* gene

261 All of the predictors above are binary (presence/absence) therefore there are $2^4 = 16$ different possible
262 prediction for any sample based on these features, see Table S2. A linear regression model fitted on
263 the log2 values of the MIC measurements could achieve similar performance as a more complex
264 random forest model, see Figure 3. Linear regression is preferred due to its simplistic nature. Having
265 a random forest regressor with hundreds of decision trees and thousands of genomic features as
266 predictors it is difficult to understand why the model made that particular prediction, leaving doubts
267 of its clinical usefulness.

268 ## 8. Discussion

269 Here, we present a novel method for predicting ciprofloxacin resistance for *E. coli*. With minimal prior
270 knowledge of relevant resistance determinants (that is mainly the use of ResFinder which includes
271 genes known to encode low-level resistance to ciprofloxacin in *E. coli*) and a data-driven approach, we
272 managed to create a machine learning model that was not only accurately predicting the
273 susceptible/non-susceptible labels but also accurately predicting at MIC level. Additionally, the

274 highlighted features of our approach could be narrowed down to four biologically understandable
275 features, making the method simpler and therefore applicable to clinical microbiology practice.

276 It was previously shown that accurate ciprofloxacin resistant/susceptible binary prediction is possible
277 for *E. coli* (17) (12) (3). For some other bacteria-antibiotic combinations even MIC level predictions
278 were performed (19), (20), (18), (21). This study goes beyond by not only predicting MIC level
279 ciprofloxacin resistance for *E. coli*, but also highlighting the underlying reasoning behind the
280 predictions. Furthermore, this study is one of very few that includes the presence or absence of genes
281 located on mobile genetic elements (MGEs), in combination with chromosomal point mutations, in
282 the machine learning algorithm. This is a crucial step since particularly in Gram-negative
283 microorganisms such as *E. coli*, AMR is often encoded by genomic determinants located on MGE, or a
284 combination of chromosomal and MGE encoded determinants, as is demonstrated in our study for
285 ciprofloxacin. In addition, this study used data from different countries and regions thus ensuring
286 potential variation in determinants that may contribute to ciprofloxacin resistance are represented in
287 the data set.

288 Notably, a linear regression model based on only the four most important features of the random
289 forest model performed nearly as well as the full model. These features comprise two *gyrA* mutations,
290 one *parC* mutation and the presence of any *qnrS* gene. All features have been associated with
291 ciprofloxacin resistance before (23). In addition, the presence of a single determinant versus
292 combinations of multiple of the four determinants predicted MIC ranges that were comparable to
293 those observed experimentally and in clinical isolates (23). For example, the single presence of a *qnr*
294 gene predicted a relatively low MIC but the combination of a *qnr* gene with a single mutation in *gyrA*
295 increased the predicted MIC substantially (Table S2). Our results indicate that for prediction of
296 ciprofloxacin susceptibility on the basis of whole-genome sequencing in *E. coli*, the analysis could be
297 limited to only these four determinants.

298 It is worthy to note that the model was trained on all possible mutations present in the data set, not
299 only known mutations or acquired resistance genes from a curated database. Therefore the model
300 could potentially discover novel currently unknown mutation-based resistance encoding mechanisms
301 which may be located in genes that are or are not yet known to contribute to resistance. For *E. coli*
302 the ciprofloxacin resistance determinants that were predicted in our machine learning approach have
303 been experimentally verified, but for other antibiotics, our approach could detect novel genomic
304 variants associated with resistance.

Our study also has some limitations, which mostly pertain to the dataset. For strains with measured MICs in the range of 8-64 mg/L, our model performs worse than for strains with lower MICs. This is most likely due to the fact that the majority of resistant strains in our training data have an MIC of 32 mg/L, with only very few other resistant MICs. This hampers accurate prediction of MIC for more resistant *E. coli*. Additionally, our dataset is not yet diverse and complete enough to be applied on a wide scale. This is a common problem for many studies aiming to predict AMR from WGS data. Solving this would require continuous updating of databases and an adequate database structure, the latter we have addressed previously (24).

In conclusion, we report a novel machine learning approach for a quantitative prediction of antibiotic resistance, which we applied for prediction of ciprofloxacin resistance in *E. coli*. In combination with continuous data base improvements, our approach could allow machine learning methods to enter routine clinical diagnostic and epidemiological practices to continuously improve predictions.

9. Author statements

9.1 Authors and contributors

All authors contributed to the conceptualization of the study. B.A.P., B.C.L.P. and C.S. wrote the manuscript, S.M., B.C.L.P., R.S.H., O.L., C.S. collected the data from various public data sources. B.A.P., E.G., D.A.A. performed machine learning modeling. The COMPARE ML-AMR group contributed data and technical help. All authors contributed to the study with insightful discussion. All authors reviewed the manuscript.

9.2 Conflicts of interest

The authors declare that there are no conflicts of interest regarding the publication of this article.

9.3 Funding information

This study was supported by the COMPARE Consortium, which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 643476.
I.C. acknowledges support from National Research, Development and Innovation Fund of Hungary, Project no. FIEK_16-1-2016-0005

9.4 Acknowledgements

The COMPARE ML-AMR group: S. Matamoros[1]; V. Janes[1]; R. S. Hendriksen[2]; O. Lund[2]; P. Clausen[2]; Frank M. Aarestrup[2]; Marion Koopmans[6]; B. Pataki[3,4]; D. Visontai[3,4]; J. Stéger[3,4]; JM. Szalai-Gindl[3,4]; I. Csabai[3,4]; N. Pakseresht[5]; M. Rossello[5]; N. Silvester[5]; C. Amid[5]; G. Cochrane[5];; C. Schultsz[1,7], F. Pradel[8]; E. Westeel[8]; S. Fuchs[9]; S. Malhotra Kumar[10]; B. Britto Xavier[10]; M. Nguyen Ngoc[10]; D. Remondini[11]; E.

Giampieri[11]; F. Pasquali[12]; L. Petrovska[13]; D. Ajayi[13]; E. M. Nielsen[14]; N. V. Trung[15]; N. T. Hoa[15]; Yoshikazu Ishii[16]; Kotaro Aoki[16]; P. McDermott[17].

[1]: Amsterdam UMC, University of Amsterdam, Department of Medical Microbiology, Amsterdam, The Netherlands.

[2]: National Food Institute, Technical University of Denmark, Lyngby, Denmark.

[3]: Department of Physics of Complex Systems, ELTE Eötvös Loránd University, Budapest, Hungary.

[4]: Department of Computational Sciences, Wigner Research Centre for Physics of the HAS, Budapest, Hungary.

[5]: European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK.

[6]: Department of Viroscience, Erasmus University Medical Center, Rotterdam, The Netherlands.

[7]: Amsterdam UMC, University of Amsterdam, Department of Global Health, Amsterdam Institute for Global Health and Development, Amsterdam, The Netherlands.

[8]: Fondation Mérieux, Lyon, France.

[9]: Department of Infectious Diseases, Robert Koch Institut, Berlin, Germany.

[10]: Department of Medical Microbiology, Vaccine & Infectious Disease Institute, Antwerp University, University Hospital Antwerp, Antwerp, Belgium.

[11]: Department of Physics and Astronomy (DIFA), University of Bologna, Bologna, Italy.

[12]: Department of Agricultural and Food Sciences (DISTAL), University of Bologna, Bologna, Italy.

[13]: Animal and Plant Health Agency, Addlestone, Surrey,United Kingdom.

[14]: Statens Serum Institut, Denmark.

[15]: Oxford University Clinical Research Unit, Centre for Tropical Medicine, Ho Chi Minh City, Vietnam

[16]: Department of Microbiology and Infectious Diseases, Faculty of Medicine, Toho University School of Medicine, 5-21-16 Omorinishi,Ota-ku, Tokyo 143-8540, Japan

[17]: Food and Drug Administration, Center for Veterinary Medicine, Office of Research, Laurel, MD, USA.

10. References

1. Lederberg J. Infectious history. Science 2000; 288(5464):287–293.
2. Otto M. Next-generation sequencing to monitor the spread of antimicrobial resistance. Genome medicine 2017; 9(1):68.
3. Stoesser N, Batty E, Eyre D, Morgan M, Wyllie D, Del Ojo Elias C, Johnson J, Walker A, Peto T, Crook D. Predicting antimicrobial susceptibilities for Escherichia coli and Klebsiella pneumoniae isolates using whole genomic sequence data. J. Antimicrob. Chemother. 2013; 68(10):2234–2244.
4. Su M, Satola SW, Read TD. Genome-based prediction of bacterial antibiotic resistance. J. clinical microbiology 2019; 57(3):e01405–18.
5. Köser CU, Ellington MJ, Peacock SJ. Whole-genome sequencing to control antimicrobial resistance. Trends Genet. 2014;

30(9):401–407.
6. Barczak AK, Gomez JE, Kaufmann BB, Hinson ER, Cosimi L, Borowsky ML, Onderdonk AB, Stanley SA, Kaur D, Bryant KF, et al. RNA signatures allow rapid identification of pathogens and antibiotic susceptibilities. Proc. Natl. Acad. Sci. 2012; 109(16):6217–6222.
7. Khaledi A, Schniederjans M, Pohl S, Rainer R, Bodenhofer U, Xia B, Klawonn F, Bruchmann S, Preusse M, Eckweiler D, et al. Transcriptome profiling of antimicrobial resistance in Pseudomonas aeruginosa. Antimicrob. agents chemotherapy 2016; 60(8):4722–4733.
8. Khaledi A, Weimann A, Schniederjans M, Asgari E, Kuo TH, Oliver A, Cabot G, Kola A, Gastmeier P, Hogardt M, et al. Fighting antimicrobial resistance in Pseudomonas aeruginosa with machine learning-enabled molecular diagnostics. bioRxiv 2019; p. 643676.
9. Consortium C, the 100 GP. Prediction of susceptibility to first-line tuberculosis drugs by DNA sequencing. New Engl. J. Medicine 2018; 379(15):1403–1415.
10. Duchêne S, Holt KE, Weill FX, Le Hello S, Hawkey J, Edwards DJ, Fourment M, Holmes EC. Genome-scale rates of evolutionary change in bacteria. Microb. Genomics 2016; 2(11).
11. Veyrier F, Pletzer D, Turenne C, Behr MA. Phylogenetic detection of horizontal gene transfer during the step-wise genesis of Mycobacterium tuberculosis. BMC evolutionary biology 2009; 9(1):196.
12. Moradigaravand D, Palm M, Farewell A, Mustonen V, Warringer J, Parts L. Prediction of antibiotic resistance in Escherichia coli from large-scale pan-genome data. PLoS computational biology 2018; 14(12):e1006258.
13. Pesesky MW, Hussain T, Wallace M, Patel S, Andleeb S, Burnham CAD, Dantas G. Evaluation of machine learning and rules-based approaches for predicting antimicrobial resistance profiles in gram-negative bacilli from whole genome sequence data. Front. microbiology 2016; 7:1887.
14. Davis JJ, Boisvert S, Brettin T, Kenyon RW, Mao C, Olson R, Overbeek R, Santerre J, Shukla M, Wattam AR, et al. Antimicrobial resistance prediction in PATRIC and RAST. Sci. reports 2016; 6:27930.
15. Yang Y, Niehaus KE, Walker TM, Iqbal Z, Walker AS, Wilson DJ, Peto TE, Crook DW, Smith EG, Zhu T, et al. Machine learning for classifying tuberculosis drug-resistance from DNA sequencing data. Bioinformatics 2017; 34(10):1666–1671.
16. Kouchaki S, Yang Y, Walker TM, Sarah Walker A, Wilson DJ, Peto TE, Crook DW, Consortium C, Clifton DA. Application of machine learning techniques to tuberculosis drug resistance analysis. Bioinformatics 2018; 35(13):2276–2282.
17. Her HL, Wu YW. A pan-genome-based machine learning approach for predicting antimicrobial resistance activities of the Escherichia coli strains. Bioinformatics 2018; 34(13):i89–i95.
18. Eyre DW, De Silva D, Cole K, Peters J, Cole MJ, Grad YH, Demczuk W, Martin I, Mulvey MR, Crook DW, et al. WGS to predict antibiotic MICs for Neisseria gonorrhoeae. J. Antimicrob. Chemother. 2017; 72(7):1937–1947.
19. Nguyen M, Brettin T, Long SW, Musser JM, Olsen RJ, Olson R, Shukla M, Stevens RL, Xia F, Yoo H, et al. Developing an in silico minimum inhibitory concentration panel test for Klebsiella pneumoniae. Sci. reports 2018; 8(1):421.
20. Nguyen M, Long SW, McDermott PF, Olsen RJ, Olson R, Stevens RL, Tyson GH, Zhao S, Davis JJ. Using machine learning to predict antimicrobial MICs and associated genomic features for nontyphoidal Salmonella. J. Clin. Microbiol. 2019; 57(2):e01260–18.
21. Li Y, Metcalf BJ, Chochua S, Li Z, Gertz RE, Walker H, Hawkins PA, Tran T, Whitney CG, McGee L, et al. Penicillin-binding protein transpeptidase signatures for tracking and predicting β-lactam resistance levels in Streptococcus pneumoniae. MBio 2016; 7(3):e00756–16.
22. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, Indap A, King KS, Bergmann S, Nelson MR, et al. Genes mirror geography within Europe. Nature 2008; 456(7218):98.
23. van der Putten BC, Remondini D, Pasquini G, Janes VA, Matamoros S, Schultsz C. Quantifying the contribution of four resistance mechanisms to ciprofloxacin MIC in Escherichia coli: a systematic review. J. Antimicrob. Chemother. 2018; 74(2):298–310.
24. Matamoros S, Hendriksen R, Pataki B, Pakseresht N, Rossello M, Silvester N, Amid C, Cochrane G, Csabai I, Lund O, et al. Accelerating surveillance and research of antimicrobial resistance-an online repository for sharing of antimicrobial susceptibility data associated with whole genome sequences. bioRxiv 2019; p. 532267.
25. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:1303.3997 2013; .
26. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 2011; 27(21):2987–2993.
27. Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, Aarestrup FM, Larsen MV. Identification of acquired antimicrobial resistance genes. J. antimicrobial chemotherapy 2012; 67(11):2640–2644.
28. Clausen PT, Aarestrup FM, Lund O. Rapid and precise alignment of raw reads against redundant databases with KMA. BMC bioinformatics 2018; 19(1):307.

29. Ortiz EM. vcf2phylip v2.0: convert a VCF matrix into several matrix formats for phylogenetic analysis. (Version v2.0). Zenodo 2019 http://doiorg/105281/zenodo2540861; .
30. Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, Keane JA, Harris SR. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. Microb. Genomics 2016; 2(4).
31. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 2014; 30(9):1312–1313.
32. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic acids research 2019;
33. CLSI. Fluoroquinolone Breakpoints for Enterobacteriaceae and Pseudomonas aeruginosa 1st edition. Wayne, PA: Clin. Lab. Standards Inst. 2019; .
34. Simonyan K, Zisserman A. Very deep convolutional networks for largescale image recognition. arXiv preprint arXiv:1409.1556 2014; .
35. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition Ieee; 2009. p. 248–255.
36. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 2011; 12:2825–2830.
37. Louppe G, Wehenkel L, Sutera A, Geurts P. Understanding variable importances in forests of randomized trees. In: Advances in neural information processing systems; 2013. p. 431–439.
38. Breiman L. Classification and regression trees. Routledge; 2017.

11. Supplemental material file list

- **TABLE S1** shows the feature importances over the different leave-one-country-out folds
- **TABLE S2** shows the parameters of the fitted linear regression and the $2^4 = 16$ possible predicted values based on the 4 features.
- **FIGURE S1** shows the leave-one-country-out cross-validation R^2 results based on the number of features.
- **FIGURE S2** shows the results of the models for an additional 100 *E. coli* genomes from Bangladesh. These genomes had only disk diffusion test measurements, which is why they were not discussed in the paper.
- **FIGURE S3** shows data quality control checks for the dataset.

12. Data bibliography

All the used data is publicly available from the SRA EBI database. All used files are listed at https://github.com/patbaa/AMR_ciprofloxacin/blob/master/meta.tsv with URLs provided. Isolates with accession numbers and MIC measurements are also available at https://github.com/patbaa/AMR_ciprofloxacin/blob/master/supplementary_meta_table.csv .

13. Figures and tables

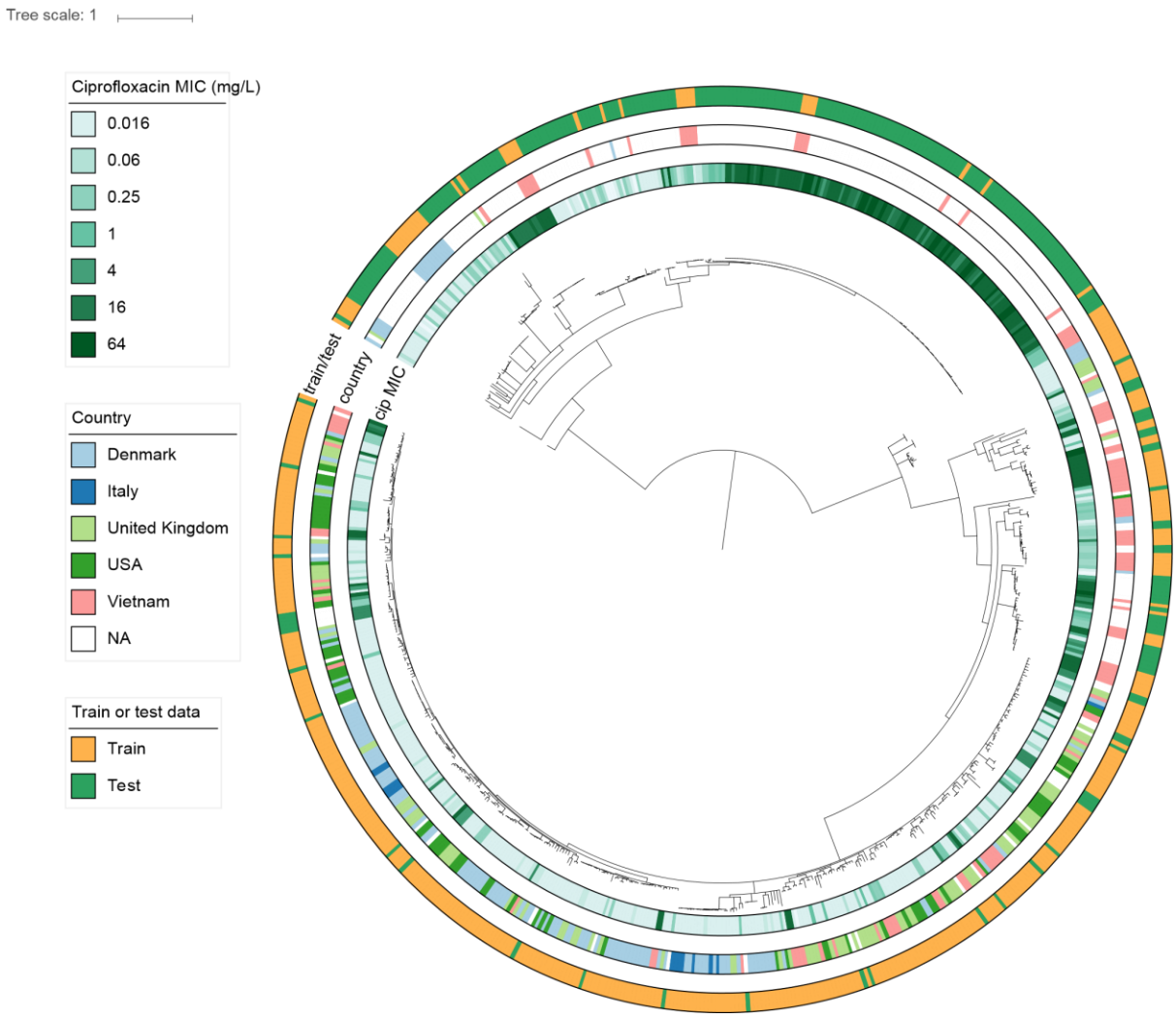

FIG 1 Midpoint-rooted phylogenetic tree of the 704 *E. coli* samples that had ciprofloxacin MIC measurement. It is clearly visible that the test data is clustered separately from the training data suggesting the generalization power of our model. Nodes with lower than 80% bootstrap support are collapsed.

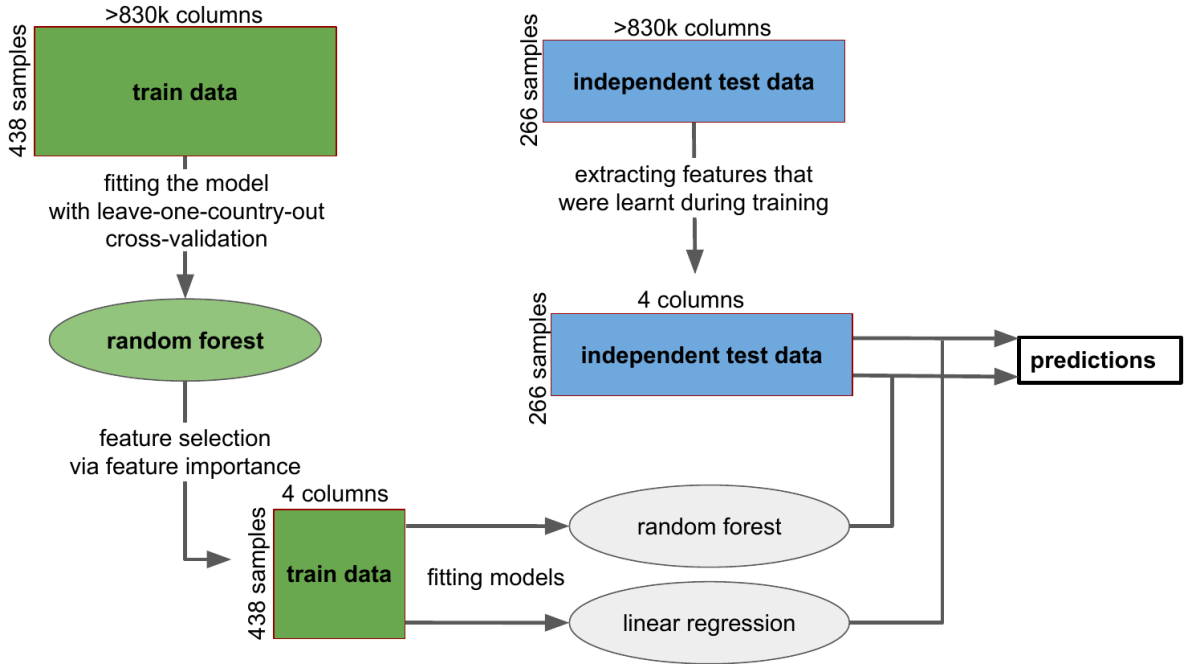

FIG 2 Workflow of the study. First, a random forest model was fitted to the training data with leave-one-country-out validation. Feature importances of the fitted models are averaged over all the folds and the four best features are kept. Then the random forest model and a linear regression model were fitted on all the training samples using only the four best features. And model performances are tested using the independent test dataset.

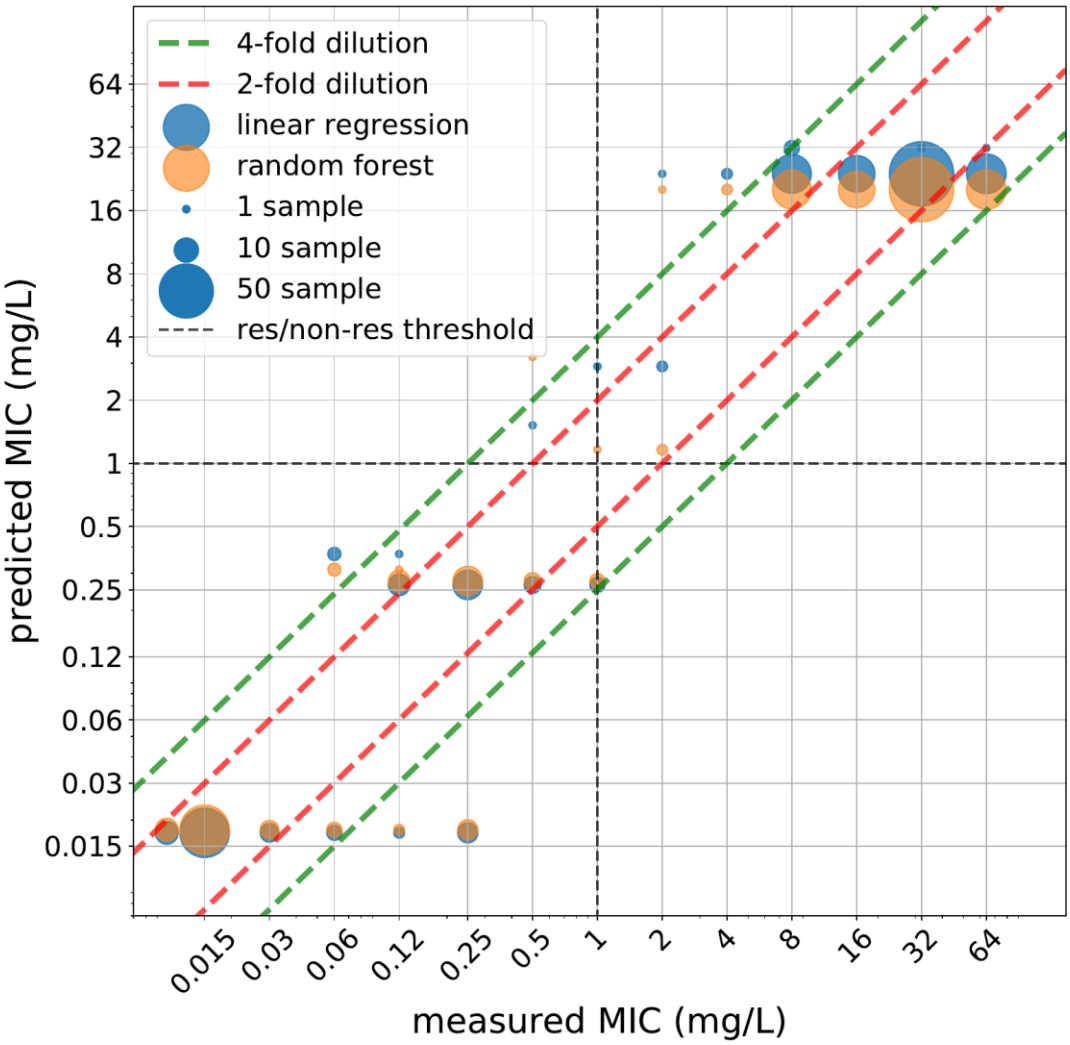

FIG 3 Prediction on the unseen test set was generated via random forest and linear regression model using the best four predictors. It can be clearly seen that the two models do not differ much in terms of predicted values.

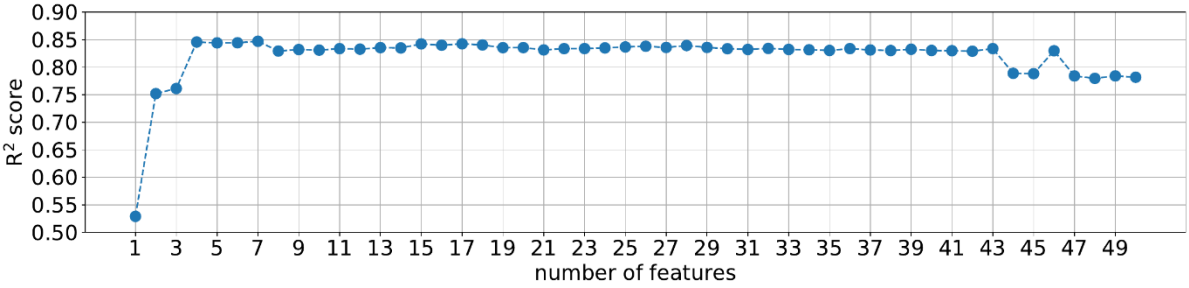

FIG S1 R squared score calculated on the training set using random forest model. The features were ranked based on Table S1 and iteratively a random forest model was fitted on the training set with leave-one-country-out validation.

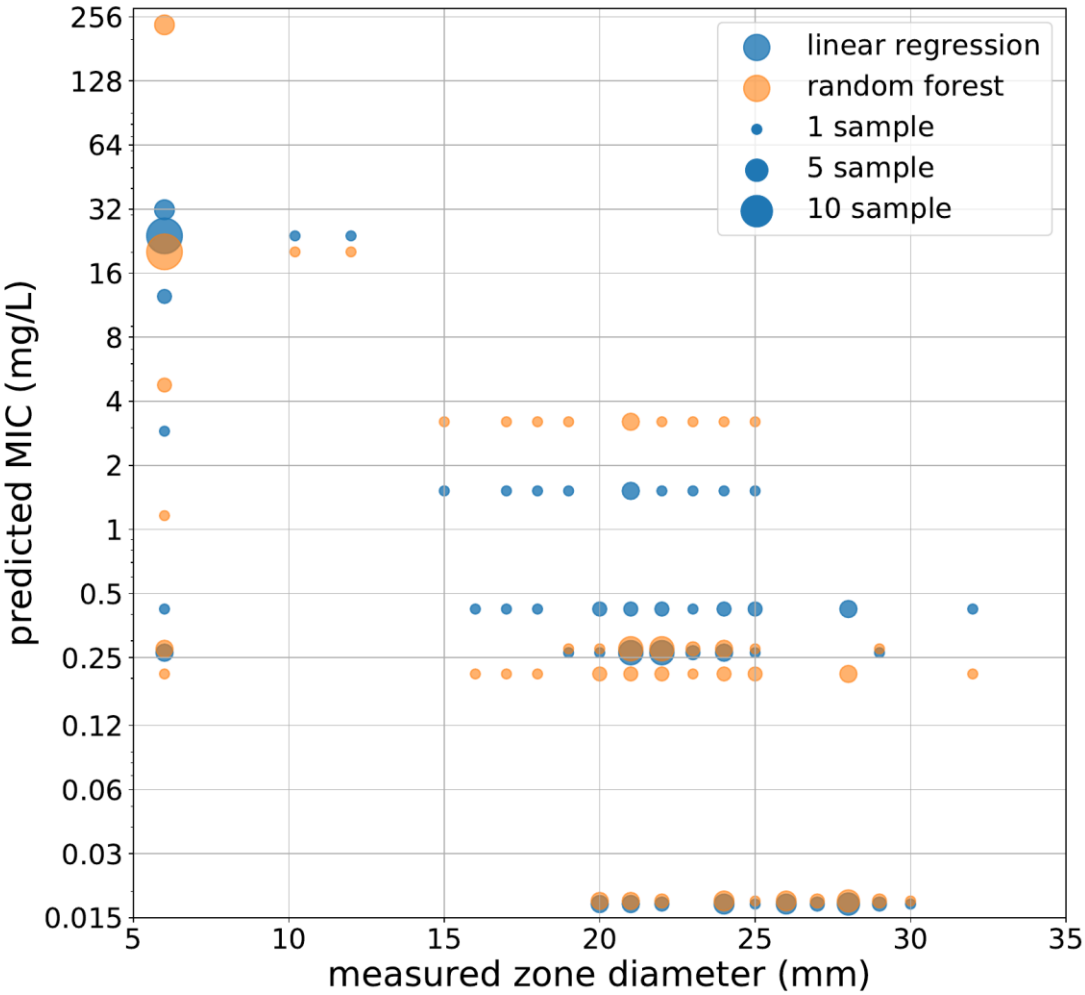

FIG S2 Prediction for samples that had only disk diffusion test measurement. As the larger zone diameter corresponds to smaller MIC values, a negative correlation is desirable on this plot. The same models were used with 4 predictors as it was used for the test set.

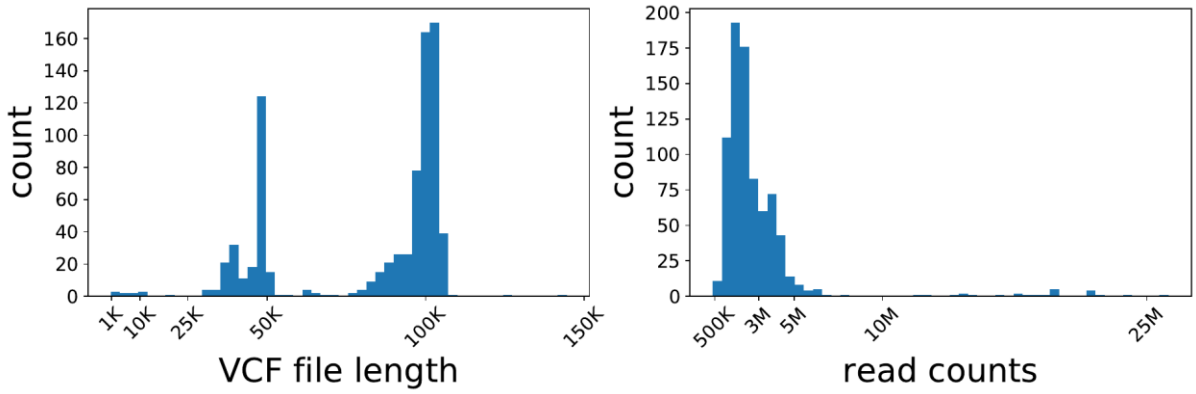

FIG S3 VCF file length distribution and the number of raw reads in the collected dataset.

MIC (mg/L)	Denmark	Italy	NA*	USA	UK	Vietnam	**Total**
0.010	0	0	9	0	0	2	**11**
0.012	0	0	0	0	0	1	**1**
0.015	119	13	42	49	92	0	**315**
0.016	0	0	0	0	0	2	**2**
0.023	0	0	0	0	0	1	**1**
0.030	12	0	6	3	4	0	**25**
0.060	1	0	7	1	0	0	**9**
0.120	0	0	11	2	0	0	**13**
0.125	0	0	0	0	0	6	**6**
0.190	0	0	0	0	0	10	**10**
0.250	6	0	22	11	3	16	**58**
0.380	0	0	0	0	0	5	**5**
0.500	0	0	6	2	0	11	**19**
0.750	0	0	0	0	0	1	**1**
1.000	0	0	5	2	0	5	**12**
2.000	0	0	3	0	0	1	**4**
4.000	0	0	2	6	0	1	**9**
8.000	0	0	30	0	1	2	**33**
12.00	0	0	0	0	0	1	**1**
16.00	0	0	23	0	0	0	**23**
24.00	0	0	0	0	0	1	**1**
32.00	0	0	72	0	0	45	**117**
64.00	0	0	28	0	0	0	**28**
Total	**138**	**13**	**266**	**76**	**100**	**111**	**704**

*country metadata is Not Available

487 **TABLE 1** The collected and used data in the analysis grouped by country and MIC values.

488

489

490

491

model	N_feat	+R^2	+R	#*ME	#*VME	AUC	ACC-2	ACC-4	#MAFE
random forest	4	0.932	0.966	1	0	1.000	0.654	0.944	0.891
random forest	15	0.890	0.944	4	0	0.998	0.684	0.891	1.007
linear regression	4	0.914	0.957	1	0	1.000	0.654	0.921	0.998

*number of samples

+calculated on the log2 values

#the lower the better

492 **TABLE 2** Number of features, R^2 score, Pearson correlation (R), Major Error (ME), Very Major Error (VME), area under the
493 receiver operating curve (AUC), Accuracy within a two/four-fold dilution (ACC-2, ACC-4) and Mean Absolute Fold Error
494 (MAFE) on the unseen test data. For the AUC, ME, VME the data was binarized using 1 mg/L threshold. The number of
495 features were selected according to the performance using leave-one-country-out validation on the training data, see
496 Fig.S1

497

feature	Denmark	Italy	USA	UK	Vietnam	sum
gyrA#87	0.555	0.567	0.570	0.558	0.010	2.260
gyrA#83	0.114	0.148	0.107	0.141	0.690	1.200
parC#80	0.181	0.151	0.199	0.172	0.001	0.705
has_qnrS	0.042	0.039	0.019	0.037	0.004	0.140
qnrS1_1_AB187515	0.012	0.007	0.017	0.004	0.000	0.041
blaCTX-M-55_1_DQ810789	0.002	0.009	0.009	0.013	0.000	0.033
blaVIM-48_1_KY362199	0.004	0.001	0.016	0.002	0.000	0.022
CP009072.1_3517597	0.000	0.000	0.000	0.000	0.013	0.014
CP009072.1_1734215	0.000	0.000	0.000	0.000	0.008	0.009
blaCTX-M-14_1_AF252622	0.000	0.002	0.002	0.003	0.000	0.008
CP009072.1_3517591	0.000	0.000	0.000	0.000	0.007	0.007
CP009072.1_113480	0.000	0.000	0.000	0.000	0.005	0.006
CP009072.1_1205372	0.003	0.002	0.001	0.000	0.000	0.006
has_qnr	0.005	0.000	0.000	0.000	0.000	0.005
CP009072.1_459777	0.001	0.002	0.000	0.002	0.000	0.005

498 **TABLE S1** Feature importances of the fitted random forest models. Random forest model was fitted on the training data
499 using leave-one-country-out validation. Each entry shows the feature importance for the given feature for the validation
500 step when the isolates from the given country were not used to train the model. Sorted by the sum of the feature
501 importances. The features are following the gene # amino acid position naming where possible. For the mutations where
502 there were no genes associated, the naming is chromosome name _ position. For the features coming from ResFinder, the

503 ResFinder naming was kept. has_qnr and has_qnrS are binary features describing if the isolate had any qnr/qnrS entry in
504 the ResFinder results.

prediction (mg/L)	gyrA#87	gyrA#83	parC#80	has_qnrS
0.018	No	No	No	No
0.076	No	No	Yes	No
0.210	No	No	No	Yes
0.275	No	Yes	No	No
0.312	Yes	No	No	No
0.885	No	No	Yes	Yes
1.162	No	Yes	Yes	No
1.317	Yes	No	Yes	No
3.207	No	Yes	No	Yes
3.634	Yes	No	No	Yes
4.771	Yes	Yes	No	No
13.537	No	Yes	Yes	Yes
15.341	Yes	No	Yes	Yes
20.140	Yes	Yes	Yes	No
55.585	Yes	Yes	No	Yes
234.649	Yes	Yes	Yes	Yes

505 **TABLE S2** Parameters of the fitted linear regression model. The interception is −5.796, and the parameters associated with
506 gyrA#87, gyrA#83, parC#80, has_qnrS are 4.116, 3.935, 2.078 and 3.542. Prediction is calculated as 2 to the power of the
507 sum of interception and the present mutation/genes.