Predicting Phenotypic Polymyxin Resistance in *Klebsiella pneumoniae* through Machine Learning Analysis of Genomic Data

Nenad Macesic,a,b Oliver J. Bear Don’t Walk IV,c Itsik Pe’er,d Nicholas P. Tatonetti,c Anton Y. Peleg,b,e Anne-Catrin Uhlemann,f

aDivision of Infectious Diseases, Columbia University Irving Medical Center, New York, New York, USA
bDepartment of Infectious Diseases, The Alfred Hospital and Central Clinical School, Monash University, Melbourne, Australia
cDepartment of Biomedical Informatics, Columbia University, New York, New York, USA
dDepartment of Computer Science, Columbia University, New York, New York, USA
eInfection and Immunity Program, Monash Biomedicine Discovery Institute, Department of Microbiology, Monash University, Clayton, Victoria, Australia
fMicrobiome & Pathogen Genomics Core, Columbia University Irving Medical Center, New York, New York, USA

**ABSTRACT** Polymyxins are used as treatments of last resort for Gram-negative bacterial infections. Their increased use has led to concerns about emerging polymyxin resistance (PR). Phenotypic polymyxin susceptibility testing is resource intensive and difficult to perform accurately. The complex polygenic nature of PR and our incomplete understanding of its genetic basis make it difficult to predict PR using detection of resistance determinants. We therefore applied machine learning (ML) to whole-genome sequencing data from >600 *Klebsiella pneumoniae* clonal group 258 (CG258) genomes to predict phenotypic PR. Using a reference-based representation of genomic data with ML outperformed a rule-based approach that detected variants in known PR genes (area under receiver-operator curve [AUROC], 0.894 versus 0.791, \(P < 0.006\)). We noted modest increases in performance by using a bacterial genome-wide association study to filter relevant genomic features and by integrating clinical data in the form of prior polymyxin exposure. Conversely, reference-free representation of genomic data as \(k\)-mers was associated with decreased performance (AUROC, 0.692 versus 0.894, \(P < 0.015\)). When ML models were interpreted to extract genomic features, six of seven known PR genes were correctly identified by models without prior programming and several genes involved in stress responses and maintenance of the cell membrane were identified as potential novel determinants of PR. These findings are a proof of concept that whole-genome sequencing data can accurately predict PR in *K. pneumoniae* CG258 and may be applicable to other forms of complex antimicrobial resistance.

**IMPORTANCE** Polymyxins are last-resort antibiotics used to treat highly resistant Gram-negative bacteria. There are increasing reports of polymyxin resistance emerging, raising concerns of a postantibiotic era. Polymyxin resistance is therefore a significant public health threat, but current phenotypic methods for detection are difficult and time-consuming to perform. There have been increasing efforts to use whole-genome sequencing for detection of antibiotic resistance, but this has been difficult to apply to polymyxin resistance because of its complex polygenic nature. The significance of our research is that we successfully applied machine learning methods to predict polymyxin resistance in *Klebsiella pneumoniae* clonal group 258, a common health care-associated and multidrug-resistant pathogen. Our findings highlight that machine learning can be successfully applied even in complex forms of antibiotic resistance and represent a significant contribution to the literature that could be used to predict resistance in other bacteria and to other antibiotics.

Citation Macesic N, Bear Don’t Walk OJ, IV, Pe’er I, Tatonetti NP, Peleg AY, Uhlemann A-C. 2020. Predicting phenotypic polymyxin resistance in *Klebsiella pneumoniae* through machine learning analysis of genomic data. mSystems 5:e00656-19. https://doi.org/10.1128/mSystems.00656-19.

Editor Yu-Liang Yang, Agricultural Biotechnology Research Center

Copyright © 2020 Macesic et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International license.

Address correspondence to Nenad Macesic, nenad.macesic@monash.edu, or Anne-Catrin Uhlemann, au2110@columbia.edu.

Received 9 October 2019
Accepted 1 May 2020
Published 26 May 2020
Carbapenem resistance in Enterobacterales (CRE) is a global health challenge that threatens many medical advances. Due to the lack of antimicrobial options for treating CRE infections, polymyxins (including colistin and polymyxin B) have been revived despite their toxicity and are widely used as treatments of last resort (1). Polymyxin resistance (PR) has now become a growing concern and may critically impair our ability to combat CRE infections. Epidemiological studies have indicated PR rates ranging from 5 to 40% in CRE, with Klebsiella pneumoniae accounting for the majority (1).

While improving the diagnosis and treatment of PR infections is an urgent priority, there are several challenges. Phenotypic polymyxin susceptibility testing is resource intensive and difficult to perform accurately (2). Broth microdilution (BMD) testing is recommended by the CLSI and EUCAST but is often available only in reference laboratories. Many clinical laboratories have traditionally relied on gradient diffusion methods such as the Etest, but significant concerns about the accuracy of these methods have been raised (3, 4). Our understanding of the genetic basis of PR also remains limited. Several important genetic loci of PR have been identified in multiple studies (crrAB, mgrB, phoPQ, pmrAB) (5–11) and will henceforth be referred to as PR canonical genes. However, PR is noted in isolates not carrying mutations or disruptions in these canonical genes and may result from mutations in multiple genes (12). Conversely, some susceptible isolates have mutations in canonical genes but do not exhibit PR, raising the possibility of variants that do not cause resistance or compensatory mutations. This makes PR a challenging polygenic trait to diagnose and hampers efforts at rapid molecular detection of PR.

With increasing availability of bacterial whole-genome sequencing (WGS) data, there has been active investigation into using these data for genotype-phenotype prediction of antimicrobial susceptibility testing (AST). This was initially in the form of rule-based approaches that would predict susceptibility through detection of known resistance determinants (e.g., beta-lactamase genes) or known resistance mutations in housekeeping genes (e.g., rpoB conferring rifampin resistance in Staphylococcus aureus) (13). The performance of these approaches has varied depending on the organism and antimicrobial tested, but multiple limitations remain. First, the approach assumes that all determinants of resistance are known and is therefore unable to detect previously uncharacterized determinants. Second, these rule-based models struggle to account for complex interactions between variants in multiple loci. In order to move beyond these limitations, machine learning (ML) methods have been used to predict antimicrobial susceptibility (14–17). Given the incomplete identification of contributing PR mutations and the possible polygenic nature of PR, we hypothesize that ML approaches may be well suited to AST genotype-phenotype prediction in this setting (18) and may ultimately be used to help identify isolates for confirmatory phenotypic testing.

We therefore aimed to use ML for genotype-phenotype prediction of PR in K. pneumoniae clonal group 258 (CG258), as this remains the main CRE clone in North America and a key CRE clone globally (19). We aimed to do this using both a reference-based approach that relied on variant calling and insertion sequence (IS) detection and a reference-free approach using detection of k-mers. We compared these ML approaches to a simple rule-based model that relies on detection of variants in canonical PR genes. Finally, we hypothesized that the ML models could be interpreted to confirm biological plausibility and elucidate novel genetic determinants of PR.

RESULTS
We analyzed 619 previously sequenced K. pneumoniae CG258 genomes (10–12, 20–25). To examine the impact of groups of input genomes on model performance, we conducted analyses on three different data sets: first, on all available genomes (193/619...
Machine Learning Prediction of Polymyxin Resistance

Machine learning approaches outperform rule-based approach for prediction of polymyxin resistance. For our reference-based analyses, we created a binary matrix, with genomes as rows and coding regions as columns (referred to as “instances” and “features,” respectively, in ML literature), with the presence of single nucleotide variants (SNVs) or IS elements in coding regions relative to a CG258 reference genome being recorded, as described previously (12) (Data Set S1, sheet 2). This matrix was then used as an input for a simple rule-based approach which classified isolates as PR if there was a variant present in any of the canonical PR genes (crrAB, mgrB, pmrAB, phoPQ) (5–11) (Data Set S1, sheet 3). This approach had modest performance, resulting in an AUROC of 0.717 to 0.832, depending on input genomes (Table 1).

We then used the same matrix as an input for ML analyses with four different ML algorithms: logistic regression, random forest, support vector machine classifier (SVC), and gradient-boosted trees classifier (GBTC), as implemented in scikit-learn (26) (code is available on GitHub at https://github.com/crowegian/AMR_ML). The best-performing
Publicly available genomes and genomes from Columbia University Irving Medical Center were processed using either a reference-based or a reference-free approach in order to generate a binary matrix. The binary matrix represented either a gene variant matrix of coding regions in the genome (reference-based approach) or individual k-mers (reference-free approach). For machine learning analyses, two approaches were used. In the non-genome-wide association study (non-GWAS) approach, a scikit-learn pipeline was implemented, which included a further feature selection step using a support vector classifier and then comparison between four algorithms with 10-fold cross-validation (CV) for hyperparameter tuning. For the GWAS filtering approach, further feature selection was performed by integrating the results of a bacterial GWAS to prioritize genes. Model training was performed with hyperparameter tuning using the 75% training set, with the same four algorithms evaluated. The best-performing model (Continued on next page)
algorithms achieved mean AUROC of 0.885 to 0.933 (according to input genomes used), and results for these algorithms are shown in Table 1. A significant difference in mean performance when different input genome data sets were compared across metrics was noted in accuracy (analysis of variance [ANOVA] P = 0.0345) but not in any other metrics. The reference-based ML analyses achieved significantly higher AUROC than the rule-based approach (P = 0.014* for CUIMC genomes, P = 0.006* for non-CUIMC genomes, and P = 0.006* for all genomes) For other metrics, ML approaches generally performed as well as or better than the rule-based approach, with the exception of recall (Table 1). However, for CUIMC genomes, worse performance was also noted for balanced accuracy and F1 score (Table 1).

Choice of machine learning algorithm did not impact performance across different input genomes and performance metrics. The ML algorithms used (logistic regression, random forest, SVC, and GBTC) differ substantially in methodology. To assess whether the choice of ML algorithm impacted performance and whether there was an optimal algorithm, we compared ML algorithms according to input genomes and different performance metrics (Fig. 3B). No significant differences in performance were noted between algorithms, regardless of the input genomes and performance metric used.

Impact of feature engineering with GWAS filtering and polymyxin exposure data on performance of ML-based prediction. A key challenge in AST genotype-phenotype prediction is the sparsity of the input genomic data sets due to relatively few genomes in data sets compared to the number of genomic features (14). Appropriate feature selection prior to training ML algorithms is a potential solution to this challenge.

TABLE 1 Comparison of rule-based and reference-based approaches for prediction of PR

| Genomes used | Metric | Rule-based value | Ref-based value using best-performing algorithm indicated | P value (ref-based vs rule-based) | P value (GWAS vs rule-based) |
|--------------|--------|------------------|----------------------------------------------------------|----------------------------------|-----------------------------|
| CUIMC        | AUROC  | 0.832            | 0.885 (0.849, 0.92)                                     | 0.014*                           | 0.004*                      |
|              | bACC   | 0.832            | 0.879 (0.751, 0.827)                                     | 0.049*                           | 0.262                       |
|              | Accuracy | 0.821     | 0.796 (0.762, 0.83)                                     | 0.185                            | 0.052                       |
|              | F1     | 0.819            | 0.755 (0.701, 0.809)                                     | 0.027*                           | 0.919                       |
|              | Precision | 0.738  | 0.799 (0.739, 0.859)                                     | 0.037*                           | 0.006*                      |
|              | Recall | 0.92             | 0.733 (0.635, 0.831)                                     | 0.008*                           | 0.006*                      |
| Non-CUIMC    | AUROC  | 0.717            | 0.933 (0.884, 0.982)                                     | 0.006*                           | 0.002*                      |
|              | bACC   | 0.717            | 0.853 (0.654, 0.853)                                     | 0.141                            | 0.006*                      |
|              | Accuracy | 0.699  | 0.873 (0.822, 0.925)                                     | 0.006*                           | 0.006*                      |
|              | F1     | 0.471            | 0.59 (0.395, 0.785)                                     | 0.185                            | 0.002*                      |
|              | Precision | 0.345  | 0.711 (0.473, 0.949)                                     | 0.184                            | 0.262                       |
|              | Recall | 0.745            | 0.57 (0.362, 0.778)                                     | 0.184                            | 0.262                       |
| All          | AUROC  | 0.791            | 0.894 (0.838, 0.95)                                     | 0.006*                           | 0.002*                      |
|              | bACC   | 0.791            | 0.878 (0.73, 0.838)                                     | 0.61                             | 0.375                       |
|              | Accuracy | 0.761  | 0.827 (0.78, 0.874)                                     | 0.19*                            | 0.006*                      |
|              | F1     | 0.694            | 0.702 (0.623, 0.781)                                     | 0.76                             | 0.02*                       |
|              | Precision | 0.577  | 0.8 (0.675, 0.926)                                     | 0.11*                            | 0.006*                      |
|              | Recall | 0.87             | 0.668 (0.549, 0.788)                                     | 0.006*                           | 0.002*                      |

*, statistical significance with a P value of <0.05 in favor of machine-learning approaches; +, statistical significance with a P value of <0.05 in favor of rule-based approaches. Abbreviations: PR, polymyxin resistance; AUROC, area under receiver-operator curve; bACC, balanced accuracy; CI, confidence interval; CUIMC, Columbia University Irving Medical Center; GBTC, gradient boosted trees classifier; GWAS, genome-wide association study; ref, reference; SVC, support vector machine classifier.
We hypothesized that conducting a bacterial genome-wide association study (GWAS) as a filtering procedure for selecting the most important features by $P$ value would improve the performance of ML models. We therefore conducted a bacterial GWAS using the R package treeWAS in the simultaneous mode, which enables correction for population structure. The results of the GWAS are shown in Data Set S1, sheets 5 to 7, according to the input genomes used. Using these results to prioritize genes resulted in a mean increase of 5.3% in performance when all performance metrics were considered (range, 3% to 13.9%), but this moderate rise was not statistically significant for most metrics when standard results were compared with those using GWAS filtering individually (Table 2 and Fig. 3A). There was less variability in model performance when GWAS filtering was applied (Tables 1 and 2).

The widespread availability of electronic medical records has made it possible to rapidly extract clinical data for use in ML approaches (28). We wanted to assess if

---

**FIG 3** Impact of feature engineering approach and machine learning algorithm on performance of machine learning models for polymyxin resistance prediction. Mean performance with 95% confidence intervals is shown across different performance metrics. The algorithms used were those that achieved the highest area under receiver-operator curve and can be found in Table 3. Histograms show how performance is impacted by the feature engineering approach (A) and the choice of machine learning algorithm (B). Abbreviations: AUROC, area under receiver-operator curve; bACC, balanced accuracy; CUIMC, Columbia University Irving Medical Center; GBTC, gradient boosted trees classifier; GWAS, genome-wide association study; SVC, support vector machine classifier.
integrating clinical data would improve performance compared to that of ML models built only with genomic data. Polymyxin exposure has been recognized as a factor contributing to emergence of PR (1, 12, 29). Data regarding polymyxin exposure in patients prior to culture of K. pneumoniae isolates were available for CUIMC genomes (12). This was used to create an additional binary feature of polymyxin exposure, added to the genome binary matrix, and new ML models were trained as per the reference-based approach. Modest increases in all metrics except recall were noted. Integration of polymyxin exposure data resulted in the best-performing model compared with reference-based and GWAS filtering approaches (AUROC, 0.923 versus 0.885 and 0.893, respectively) (Table 2).

Reference-based approaches outperform reference-free approaches. Reference-based approaches for creating a genomic feature matrix for ML have inherent limitations. Bacteria such as K. pneumoniae with a large accessory genome pose a particular problem, as a reference-based approach cannot evaluate novel genomic content present in the test genomes but lacking in the reference (15). Use of reference-free approaches, such as the use of k-mers (strings of nucleotides of k length) as inputs, may therefore be an attractive solution to overcome this limitation. Furthermore, being able to use reference-free approaches may potentially expand the available data set by including a more diverse collection of genomes and thus improve the performance of ML algorithms.

We tested this hypothesis by generating a reference-free binary matrix based on k-mer profiles from the DSK k-mer counting software (30) using k-mers of 31 nucleotides. This approach resulted in much higher numbers of features being incorporated than in reference-based data sets (415,373 versus 3,054 features for all genomes, 124,954 versus 2,278 features for CUIMC genomes, and 348,157 versus 2,350 features for non-CUIMC genomes, respectively). This matrix was then used as an input into the same ML pipeline as described previously. Despite only focusing on chromosomal changes relative to the reference, reference-based ML models had a higher mean

| Genomes used | Metric | Ref-based value (95% CI)b | Ref-based value with GWAS (95% CI)b | Ref-free value with polymyxin exposure data (95% CI)b | P value (GWAS vs ref based) | P value (ref free vs ref based) | P value (poly exposure vs ref based) |
|--------------|--------|---------------------------|------------------------------------|-----------------------------------------------------|-----------------------------|---------------------------------|----------------------------------|
| CUIMC        | AUROC  | 0.885 (0.849, 0.92)       | 0.893 (0.864, 0.922)               | 0.696 (0.564, 0.828)                                 | 0.923 (0.88, 0.965)        | 0.571                           | 0.241                            |
|              | bACC   | 0.789 (0.751, 0.827)      | 0.841 (0.82, 0.862)                | 0.64 (0.536, 0.743)                                 | 0.796 (0.714, 0.879)       | 0.026*                          | 0.226                            |
|              | Accuracy | 0.796 (0.762, 0.83)    | 0.854 (0.83, 0.878)                | 0.649 (0.541, 0.758)                                 | 0.804 (0.716, 0.892)       | 0.009*                          | 0.91                             |
|              | F1     | 0.755 (0.701, 0.809)      | 0.816 (0.793, 0.84)                | 0.579 (0.454, 0.704)                                 | 0.768 (0.685, 0.85)        | 0.045*                          | 0.734                            |
|              | Precision | 0.799 (0.739, 0.859)   | 0.881 (0.848, 0.914)               | 0.67 (0.506, 0.833)                                 | 0.866 (0.737, 0.996)       | 0.006*                          | 0.023*                           |
|              | Recall | 0.733 (0.635, 0.831)     | 0.763 (0.725, 0.801)               | 0.56 (0.407, 0.714)                                 | 0.732 (0.607, 0.857)       | 1                               | 0.011*                           |
| Non-CUIMC    | AUROC  | 0.933 (0.884, 0.982)      | 0.933 (0.888, 0.979)               | 0.803 (0.692, 0.913)                                 | 0.85                          | 0.677                           |
|              | bACC   | 0.753 (0.654, 0.853)      | 0.82 (0.76, 0.881)                 | 0.5 (0.5, 0.5)                                      | 0.82                          | 0.018*                          | 0.005*                           |
|              | Accuracy | 0.873 (0.822, 0.925)   | 0.917 (0.894, 0.94)                | 0.82 (0.811, 0.83)                                 | 0.82                          | 0.345                           | 0.623                            |
|              | F1     | 0.59 (0.395, 0.785)       | 0.729 (0.648, 0.81)                | 0 (0, 0)                                            | 0.729                         | 0.345                           | 0.623                            |
|              | Precision | 0.711 (0.473, 0.949) | 0.832 (0.759, 0.905)               | 0 (0, 0)                                            | 0.832                         | 0.703                           | 0.569                            |
|              | Recall | 0.57 (0.362, 0.778)      | 0.669 (0.547, 0.791)               | 0 (0, 0)                                            | 0.669                         | 0.88                            | 0*                               |
| All          | AUROC  | 0.894 (0.838, 0.95)       | 0.931 (0.915, 0.947)               | 0.692 (0.546, 0.838)                                 | 0.19                          | 0.015*                          | 0.006*                           |
|              | bACC   | 0.784 (0.73, 0.838)       | 0.801 (0.776, 0.827)               | 0.5 (0.5, 0.5)                                      | 0.801                         | 0.473                           | 0.006*                           |
|              | Accuracy | 0.827 (0.78, 0.874)   | 0.864 (0.84, 0.888)                | 0.688 (0.685, 0.691)                                 | 0.864                         | 0.162                           | 0.045*                           |
|              | F1     | 0.702 (0.623, 0.781)      | 0.741 (0.702, 0.779)               | 0 (0, 0)                                            | 0.741                         | 0.384                           | 0.003*                           |
|              | Precision | 0.8 (0.675, 0.926)  | 0.889 (0.846, 0.932)               | 0 (0, 0)                                            | 0.889                         | 0.363                           | 0.472                            |
|              | Recall | 0.668 (0.549, 0.788)     | 0.638 (0.591, 0.686)               | 0 (0, 0)                                            | 0.638                         | 0.344                           | 0.064                            |

*Statistical significance with a P value of < 0.05. Abbreviations: ML, machine learning; PR, polymyxin resistance; AUROC, area under receiver-operator curve; bACC, balanced accuracy; CI, confidence interval; CUIMC, Columbia University Irving Medical Center; GWAS, genome-wide association study; ref, reference.

Using the best-performing algorithm indicated.

Table 2 Comparison of different feature engineering approaches on performance of ML prediction of PR

Machine Learning Prediction of Polymyxin Resistance

May/June 2020 Volume 5 Issue 3 e00656-19
performance than reference-free models across all metrics in all data sets (Table 2), reaching statistical significance when all genomes were used. The reference-free models for the data sets containing all genomes and non-CUIMC genomes predicted the same phenotype for all samples (Table 2).

In order to assess whether these findings may have been due to the ML algorithms in our pipeline not being specifically designed for k-mer data inputs, we used the same binary matrices as inputs in the KOVER package (31). This was developed for the sparse data sets seen when k-mers are used in AST genotype-phenotype prediction and uses either the set covering machine (SCM) algorithm or the classification and regression trees (CART) algorithm. However, performance in KOVER was similar to that of our pipeline using both SCM and CART (AUROC, 0.525 to 0.770) (Data Set S1, sheet 8).

We also attempted to use GWAS filtering on reference-free data sets. We were able to complete analyses for CUIMC and non-CUIMC genomes but exceeded computing capacity for analyses on all genomes. Similar to the effect of GWAS filtering on reference-based data sets, we saw modest improvements in performance that did not reach statistical significance (AUROC, 0.761 versus 0.696 [P = 0.300] and 0.829 versus 0.803 [P = 0.627] for CUIMC and non-CUIMC genomes, respectively) (Data Set S1, sheet 9). Despite this modest increase, performance using GWAS filtering for reference-free data sets was lower than for reference-based approaches without GWAS filtering (AUROC, 0.885 versus 0.761 [P < 0.001] and 0.933 versus 0.829 [P = 0.003] for CUIMC and non-CUIMC genomes, respectively) (Data Set S1, sheet 10).

Understanding determinants of PR through machine learning. Being able to interpret how ML algorithms arrive at their conclusions remains a central challenge to applying ML approaches in clinical decision-making, including AST genotype-phenotype prediction (18). We therefore wanted to interpret ML models to assess whether they are biologically plausible and incorporate known PR determinants, as well as to use them to identify potential novel determinants of PR. We were able to extract a ranking of genes according to their relative importance to the model (model-specific feature importance) for logistic regression, random forests, and GBTC but not for SVC due to the nature of the model.

We focused on the GBTC model trained on all genomes by using a reference-based approach due to the diversity of the genomes and its high performance. The genes ranked highest in feature importance, and their quantitative feature importance metrics are listed in Table 3. Feature importance results for data sets using CUIMC and non-CUIMC genomes only are included in Data Set S1, sheet 11. With no prior programming, the model identified all canonical PR genes except crrA and ranked mgrB, phoQ, pmrA, and pmrB as the four genes with the highest feature importance.

In addition to identifying these known determinants, we assessed other highly ranked genes as candidate novel determinants of PR by conducting a literature search (Table 3). Several genes have been noted to have potential roles in PR previously, including H239_3063, which encodes a newly identified putative RND-type efflux pump (32), and arnA, which is part of the arn operon that attaches arabinose to lipid A to confer PR (33). Many of the other genes encode outer membrane proteins or have functions that may affect the cell membrane and thus are possibly involved in interactions with polymyxins, including lpdA, ahpF, envZ, pstB, pepN, and pgpB.

DISCUSSION

Our study provides a proof of principle demonstrating the utility of using ML for prediction of phenotypic PR from genomic data and raises important methodological issues that have implications for the use of ML for AST genotype-phenotype prediction more generally. We focused efforts on PR due to the clinical need posed by resistance to this class of last-line antimicrobials, the difficulties associated with performing phenotypic AST for polymyxins, and the complexity of underlying resistance mechanisms that may be uniquely suited to ML approaches. We noted high model performance across a range of metrics by leveraging a large collection of PR isolates from our institution and identifying publicly available genomes with BMD phenotypic susceptibility...
This performance was achieved through the use of reference-based input data sets, the use of GWAS as a filtering procedure, and the addition of polymyxin exposure as a clinical variable. We also were able to interpret ML models to confirm biological plausibility and identify additional potential genetic determinants of PR as candidates for functional testing.

Representation and selection of genomic features, a process termed “feature engineering” in the ML literature, was central to the success of resistance prediction and could affect performance both positively and negatively. First, our reference-based approach had high performance (AUROC, 0.885 to 0.933) and outperformed the rule-based approach that we used as a benchmark. Despite PR being more biologically complex than other forms of antimicrobial resistance in *K. pneumoniae*, it was encouraging that our results were similar to those obtained for other antimicrobials with less complex mechanisms of resistance (e.g., carbapenem resistance) tested with ML approaches (14, 15, 34–36). The performance of this approach falls below the FDA cutoffs for AST tests (37). However, given the problematic nature of phenotypic polymyxin susceptibility testing, it may help to identify a subset of genomes with a high likelihood of PR for confirmatory phenotypic testing.

In the setting of this high performance with our baseline approach, it was difficult to demonstrate a statistically significant improvement using additional feature engineering. With this in mind, we noted a moderate rise across nearly all metrics using GWAS filtering, when both reference-based and reference-free approaches were used. The role of GWAS filtering for improving performance in other AST genotype-
phenotype prediction settings remains to be determined. AST genotype-phenotype
data sets typically have a small number of genomes with a large number of genomic
features; therefore, GWAS filtering provides a biologically consistent way of selecting
genomic features. This approach has not been extensively used in AST genotype-
phenotype prediction, although it has been used to predict resistance in *Streptococcus
c pneumoniae* (38). In addition, it has been used to predict other bacterial phenotypes
(39, 40) and in other nonmicrobiological studies (41–43).

In contrast to filtering unnecessary genomic features, we noted that adding a single
pertinent clinical feature in the form of polymyxin exposure data may increase perfor-
mance by an amount similar to that of the more complex GWAS filtering approach,
although the same caveats regarding a lack of statistical significance apply. MacFadden
et al. used prior antimicrobial exposure data with genotypic data for ML prediction of
phenotype of *Escherichia coli* susceptibility to three different antimicrobial classes and
similarly noted an increase in performance over that of using genotypic data alone (16).

We focused on prior polymyxin exposure, as it is a known epidemiological risk factor
and also as a proof of principle, because drug administration is represented by highly
structured variables in electronic medical records. In a recent retrospective cohort study
of PR isolates at a single institution from 2011 to 2016, neurologic disease, residence in
a skilled nursing facility prior to admission, receipt of carbapenems in the last 90 days,
prior infection with a carbapenem-resistant organism, and receipt of ventilatory sup-
port were risk factors for PR in a multivariate model (44). Potentially, any of these
clinical variables might be added to the model when available. We envision that the
widespread adoption of electronic medical records will allow more extensive and
automated integration of such clinical and genotypic data, thus enabling more accurate
prediction both of AST phenotypes and other outcomes.

Certain forms of representation of genetic data may also have a negative impact on
performance. While the use of reference-free approaches for feature representation
incorporating k-mers has been the focus of much recent work (14, 15, 31, 34, 35, 45–48),
we noted that reference-based approaches using variant calling and IS detection had
significantly better performance overall. This may reflect the biological basis for PR,
with IS elements playing an important role, particularly in association with *mgrB* (6, 8,
12, 49). k-mer-based approaches may have difficulty accurately detecting this, due to
the diversity of IS elements in *K. pneumoniae* genomes, as well as the possibility that
they insert in different sites in individual genes or in regulatory regions upstream (50).
Additionally, reference-free approaches take the accessory genome into account. This
offers putative benefits in terms of detecting various mobile genetic elements and
making it easier to use this approach across diverse genomes. However, these benefits
may be offset by the additional “noise” in the case of organisms with large accessory
genomes, such as *K. pneumoniae*. Hicks et al. also found that ML approaches had worse
performance when attempting to predict ciprofloxacin resistance in organisms with
larger pan-genomes (*K. pneumoniae* and *Acinetobacter*) than in *Neisseria gonorrhoeae*
(15). While k-mer-based approaches may work well when the mechanism largely
depends on a single or a few resistance determinants (e.g., beta-lactamases), our
findings suggest that a more curated approach may be needed for more complex forms
of resistance such as PR.

Beyond these biological issues, use of reference-free input data also has important
computational consequences. First, it increased the number of features by orders of
magnitude in comparison with using reference-based approaches that use genes as
inputs. In order to attenuate that effect, we incorporated feature selection hyperpa-
rameters in our pipeline, but given the limited number of genomes for analysis, the
increased dimensionality nevertheless likely negatively affected algorithm perform-
ance. The large number of features also requires more computationally intensive
analyses, leading to much longer run times and raising questions about how plausible
it would be to run these analyses as part of a clinical workflow. In order to address some
of these computational issues, Drouin et al. adapted the set covering machine algo-
rihthm and implemented it in the KOVER package (31). When we used our data with
KOVER, the run time was much lower than that of our pipeline, but model performance was similarly low.

In contrast to our findings regarding the importance of feature engineering, surprisingly neither the data used to train ML algorithms (in the form of input genomes) nor the choice of specific ML algorithm significantly impacted model performance. With regard to the input genomes used for training, this was in contrast to the sampling bias noted by Hicks et al. (15) and was somewhat surprising, as the CUIMC isolates had a high degree of clonal relatedness and rates of PR varied depending on the origin data set selected (Fig. 1). Similarly, Moradigaravand et al. hypothesized that ensemble algorithms such as GBTC or random forests may be better suited to AST genotype-phenotype prediction (17), but we did not observe these differences. Our different findings may be explained by the fact that these studies focused on different organisms (N. gonorrhoeae and E. coli, respectively) and different antimicrobials. Our approach of representing genomes as binary variants in a reference’s coding regions, rather than focusing on individual variants, may also have led to more robust results. This further underlines prior concerns about attempting to use a “one-size-fits-all” approach, given the diverse biological underpinnings of antimicrobial resistance (15).

The choice of performance metric to be used in AST genotype-phenotype prediction also remains an open question, with no consensus between prior studies (13). Indeed, the choice may be dictated by the expected prevalence of the phenotype of interest and the intended use of the ML model. Our intention was to create a screening method where isolates could be identified for confirmatory phenotypic testing, and hence we chose AUROC, as it is an aggregative metric that looks at the balance between the true-positive and false-positive rates. It has been used in other AST genotype-phenotype studies (16, 34, 45–47, 51–53) and is a metric commonly used to assess clinical test performance. However, it may not be the optimal metric if there is significant imbalance in the data set (54), leading Hicks et al. to advocate for use of multiple metrics (15). We therefore reported multiple metrics, including precision (equivalent to positive predictive value), which we felt was important for our use case, as it helps to assess how many isolates identified as resistant by the ML algorithm are phenotypically resistant. Using GWAS filtering, we achieved a precision of ~90%.

Our best-performing ML models were also interpretable, thus allowing us to confirm that predictions were biologically plausible. The models correctly identified canonical PR genes, with the exception of crrA, which had a low prevalence of variants across all data sets. This is highly encouraging, given that this comprises six genes, with variants in multiple canonical genes often occurring in PR isolates (10, 12). In addition, they also confirmed genes that had been associated with PR in more limited settings, including H239_3063 and areA. Finally, several potential novel determinants were identified through the use of ML, with multiple genes involved in stress responses or maintenance of the cell membrane. Although beyond the scope of this study, these determinants now require functional testing to prove causality, and it is possible that exposure to other antimicrobials may have contributed to the observed genetic changes. This is demonstrated by the fact that the ML model incorporating all genomes identified gyrB, which has been associated with fluoroquinolone resistance.

Our study had several limitations, particularly related to the use of a reference-based approach. First, the representation of genomes as binary variants in a reference’s coding regions is an intentional simplification. An obvious problem is that all variants are treated as equal, whereas our previous work has shown that some variants may not have a functional impact, particularly in phoQ (12). However, this appears to be a reasonable trade-off, given the limited number of genomes available for analysis. Alternative approaches may include using individual alleles or a weighted score that takes into account likelihood of functional impact based on amino acid changes, rather than a simple binary representation. A second limitation is our use of a single reference chromosome. This limits generalizability to CG258 genomes, which in future work may be overcome through attempting to define a K. pneumoniae core genome for analysis. Holt et al. attempted to do this and noted 1,888 “common” genes that were present in
≥95% of 328 K. pneumoniae genomes (55). However, each genome carries thousands of additional accessory genes, likely making it difficult to apply this approach to K. pneumoniae as an entire species. The use of a reference-based approach also entails that nonchromosomal regions are not incorporated into models. While plasmid-mediated PR is rare in K. pneumoniae (and was checked for specifically in our collection), mcr genes or another plasmid-based determinant would not be detected by our reference-based approach. An additional consideration is the impact of population structure, particularly in single-institution data sets which may contain closely related genomes. While this has been well described as a potential confounder in bacterial GWAS (56), it has not been addressed in AST genotype-phenotype prediction but may play a similarly confounding role (15).

In summary, our study demonstrated that ML methods can achieve high performance for prediction of phenotypic PR in K. pneumoniae CG258, even in the face of PR being a remarkably polygenic trait. In contrast to other recent work on AST genotype-phenotype prediction, we noted best performance through the use of a reference-based and curated input data set, which may reflect the underlying biological complexity of PR and be applicable to other complex forms of antimicrobial resistance. We noted that incorporating GWAS as a filtering procedure and addition of clinical data on antimicrobial exposure may improve performance, but these findings need confirmation in other settings. The increasing availability of genomic data makes AST genotype-phenotype prediction an important priority, and use of ML will need to be tailored to specific organisms and antimicrobial agents.

MATERIALS AND METHODS

Genome selection. The study was reviewed and approved by the CUIMC Institutional Review Board. We included K. pneumoniae CG258 isolates, comprising K. pneumoniae ST258 and multilocus sequence type single-locus variants (e.g., ST11 and ST512). CUIMC isolates were selected as previously described and comprised K. pneumoniae CG258 isolates spanning 2011 to 2018 (12). All CUIMC isolates had MIC determination with BMD according to CLSI guidelines (57). We performed WGS on all included CUIMC isolates as described previously (12, 58–60). In total, we included 313 K. pneumoniae CG258 genomes. We then identified publicly available K. pneumoniae genomes by searching PubMed with the terms “colistin resistance” and “polymyxin resistance.” We also searched the CDC/FDA Antibiotic Resistance Isolate Bank and PATRIC databases (20, 25). When possible, we obtained genome raw sequence data from NCBI in preference to draft assemblies. Genomes with non-BMD phenotypic susceptibility testing data were excluded due to concerns about accuracy (2). See Fig. 1 and Data Set S1, sheet 1, for all accession numbers, source data sets, phenotypic susceptibility, and specific multilocus sequence types. We included 306 publicly available K. pneumoniae CG258 genomes. For each analysis, we tested all 619 genomes and then subsets of 313 CUIMC genomes and 306 non-CUIMC genomes. We constructed and visualized a phylogeny of all included isolates, as described previously (12).

MIC ranges varied according to susceptibility testing platform (e.g., SensiTitre versus manual methods). We therefore used a qualitative definition of polymyxin susceptibility, with isolates considered PR if the colistin or polymyxin B MIC was >2 mg/liter (57, 61). We constructed draft de novo assemblies of all genomes with raw sequence data using the Shovill wrapper for SPAdes (62) and then used Kleborate for multilocus sequence type (MLST) and resistance determinant (including mcr genes) detection (63, 64). No mcr genes were detected.

Data set preparation. A reference-based approach and a reference-free approach were used to create input matrices for ML algorithms. The outcome of interest (“label” in ML literature) for both was polymyxin susceptibility as a binary outcome (susceptible/resistant). For the reference-based approach, we had selected a representative polymyxin-susceptible CG258 isolate and created a de novo hybrid assembly (12, 65) that was then used to create a profile of each genome through a combination of variant calling and IS detection, as described previously (12). In brief, variant calling was performed using Snippy 3.2 and running Snippy in the “–contig” mode if only an assembly was available (66). ISseeker was used to identify sites with IS elements present (50). In order to increase the sensitivity of detection for key genes, we also used a BLAST database of PR canonical genes to identify if IS disruption or large-scale deletions of these genes may have occurred.

These analyses allowed us to identify CG258 isolates containing SNVs/IS in coding regions relative to the reference genome. Intergenic regions, synonymous SNVs in coding regions, and SNVs in known mobile genetic elements and repeat regions were not considered. In the data set including all genomes, we identified 14,783 variants in 3,653 genes; however, 7,149 of 14,783 (48.3%) variants occurred in single isolates. Therefore, we incorporated individual genes rather than variants, scoring each of these alleles in each isolate as 1 if it differed from the reference and 0 if it did not. These allele scores were used to create a binary matrix, with isolates as rows and coding regions in the reference as columns. We also integrated clinical data about polymyxin exposure from our prior study for CUIMC isolates and created
a binary variable that indicated polymyxin exposure at any time in that patient prior to culture of the isolate (12).

We constructed a reference-free input matrix based on k-mer profiles generated with the DSK k-mer counting software with k equal to 31, a length commonly used in bioinformatics analyses (14, 15, 30, 31, 34, 67, 68) that may maximize association power in bacterial genomics studies (15, 69, 70). A k-mer presence/absence binary matrix was then created and used as an input for ML analyses.

**Rule-based prediction of polymyxin resistance.** We used rule-based prediction as a baseline for comparison with ML approaches. The SNV/IS detection results described for reference-based data sets described above were used to find variants in any of the canonical PR genes (crrAB, mgrB, pmrAB, phoPQ) that have been established as key contributors to PR (5–11). Isolates were then classified as PR if there was a variant present in any of these genes.

**Use of genome-wide association study for feature filtering.** As an additional feature engineering for the reference-based approach, we filtered features using $P$ values from a GWAS. The reference-based binary matrix was used as input in treeWAS, an R package for bacterial GWAS (27). treeWAS was run in the “simultaneous” mode to account for population structure. The resulting $P$ values were used to rank and select genomic features under a specific $P$ value threshold (see below), with the rest discarded.

**Machine learning analyses.** We created a bespoke pipeline using scikit-learn (26), which performed additional feature selection and hyperparameter tuning (code available on GitHub at https://github.com/crowegian/AMR_ML). Hyperparameters were tuned with 10-fold cross-validation (CV) using AUROC as the performance metric. Reference-based data sets without GWAS filtering and reference-free data sets used a linear support vector classifier to extract important features for feature selection, also with 10-fold CV. As GWAS $P$ values were generated using labeled data, in order to prevent data leakage for reference-based data sets with GWAS filtering, we did not use cross validation. Instead, we used a 75%-25% training-validation split to tune hyperparameters where the GWAS values and models were trained using the 75% split training data. All data sets made use of a linear support vector classifier to filter features. Data sets with GWAS $P$ values also used $P$ values to filter features by making them a hyperparameter tuned from a minimum allowable $P$ value of 0 to the 75th percentile of all $P$ values in steps of 0.01. For data sets with GWAS filtering, features chosen by either step were all selected. Hyperparameters were tuned for each data set across the feature selection step and for each model configuration.

We tested four different ML models: logistic regression, random forests, SVC, and GBTC. The performance metrics used were AUROC, balanced accuracy, accuracy, precision, recall, and F1 score. Best-performing models were chosen based on AUROC, as this is an aggregative metric and has been used in prior studies of AST genotype-phenotype prediction (16, 34, 45–47, 51–53).

We then also compared the results of our ML pipeline to the ML implementation in the KOVER package that utilizes the set covering machine and classification and regression trees algorithms (14, 31). For the KOVER implementation, the best conjunctive and/or disjunctive model was selected using 10-fold cross-validation, testing the suggested broad range of values for the trade-off hyperparameter of 0.1, 0.178, 0.316, 0.562, 1.0, 1.778, 3.162, 5.623, 10.0, 144, and 999,999.0 to determine the optimal rule scoring function with default parameters. The pROC R package was used to calculate AUROC for the predictions made by the rule-based algorithm and KOVER package (71).

**Statistical testing.** Cross-validation results were used to generate means and 95% confidence intervals for all performance metrics. As models trained using the 75%-25% split with GWAS $P$ values do not have sensical confidence intervals, bootstrapping was performed. Ten bootstrapped training and validation sets were created by pooling the split data and repeatedly bootstrapping (sampling $n$ observations with replacement) a training data set from the original data and then using the rest of the data as a validation set. The best model and its hyperparameters for each model class (logistic regression, SVC, etc.) were chosen based on the original 75%-25% split performance and then retrained and evaluated on each of the bootstrapped training-validation pairs, and metrics were collected. Model performance was evaluated by comparing mean performance metrics using two-tailed t tests or Mann-Whitney U tests, as appropriate. Categorical variables were compared using $\chi^2$ or Fisher’s exact tests, as appropriate. ANOVA was used for comparison of means between multiple groups. Statistical analyses were performed in R (v3.4.0).

**Data availability.** All CUIMC genomes have been deposited in the NCBI Sequence Read Archive under NCBI BioProject accession numbers PRJNA557275 and PRJNA445400. For individual accession numbers, please see Data Set S1, sheet 1, in the supplemental material.

**SUPPLEMENTAL MATERIAL**

Supplemental material is available online only.

**FIG S1**, JPG file, 0.1 MB.

**DATA SET S1**, XLSX file, 0.1 MB.

**ACKNOWLEDGMENTS**

This work was supported by the Australian National Health and Medical Research Council (Postgraduate Scholarship APP1169514 to N.M. and Practitioner Fellowship APP1117940 to A.Y.P.) and by the National Institute of Allergy and Infectious Diseases and National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health (grant numbers AI116939 and 5U54DK104309 to A.-C.U.).
The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

N.M. received research support from GlaxoSmithKline, unrelated to the current study. A.Y.P. received research support from Merck Sharp & Dohme (Australia), unrelated to the current study. A.-C.U. received research funding from Merck, GlaxoSmithKline, and Allergan, unrelated to the current study. All other authors declare no conflicts of interest.

REFERENCES

1. Poirel L, Jayol A, Nordmann P. 2017. Polymyxins: antibacterial activity, susceptibility testing, and resistance mechanisms encoded by plasmids or chromosomes. Clin Microbiol Rev 30:557–596. https://doi.org/10.1128/CMR.00064-16

2. Hindler JA, Humphries RM. 2013. Colistin MIC variability by method for contemporary clinical isolates of multidrug-resistant Gram-negative bacilli. J Clin Microbiol 51:1678–1684. https://doi.org/10.1128/JCM.0385-12

3. Humphries RM. 2015. Susceptibility testing of the polymyxins: where are we now? Pharmacotherapy 35:22–27. https://doi.org/10.1002/pha.1505

4. Humphries RM, Green DA, Schuetz AN, Bergman Y, Lewis S, Yee R, Stump S, Lopez M, Macesic N, Uhlemann AC, Kohnler P, Cole N, Simmer PJ. 2019. Multi-center evaluation of colistin broth disk elution and colistin agar test: a report from the Clinical and Laboratory Standards Institute. J Clin Microbiol 57:10128. https://doi.org/10.1128/JCM.01269-19

5. Aires CA, Pereira PS, Asensi MD, Carvalho-Assis AP. 2016. mgb8 mutations mediating polymyxin B resistance in Klebsiella pneumoniae isolates from rectal surveillance swabs in Brazil. Antimicrob Agents Chemother 60:6969–6972. https://doi.org/10.1128/AAC.01456-16

6. Berglund B, Hoang NTB, Tarnberg M, Le NK, Svartstrom O, Khu DTK, Nilsson M, Le VT, Welander J, Olson L, Larsson M, Nilsson LE, Hanberger H. 2018. Inversion sequence transpositions and point mutations in mgb8 causing colistin resistance in a clinical strain of carbapenem-resistant Klebsiella pneumoniae from Vietnam. Int J Antimicrob Agents 51:789–793. https://doi.org/10.1016/j.ijantimicag.2017.11.012

7. Cannatelli A, Di Pilato V, Giani T, Arena F, Ambretti S, Gaibani P, D’Andrea MM, Rosollini GM. 2014. In vivo evolution to colistin resistance by PmrB sensor kinase mutation in KPC-producing Klebsiella pneumoniae is associated with low-dosage colistin treatment. Antimicrob Agents Chemother 58:4399–4403. https://doi.org/10.1128/AAC.02555-14

8. Cannatelli A, COLGRIT Study Group, Giani T, Arena F, Ambretti S, Gaibani P, D’Andrea MM, Rosollini GM. 2014. In vivo evolution to colistin resistance by PmrB sensor kinase mutation in KPC-producing Klebsiella pneumoniae is associated with low-dosage colistin treatment. Antimicrob Agents Chemother 58:4399–4403. https://doi.org/10.1128/AAC.02555-14

9. Cheng YH, Lin TL, Pan YJ, Wang YP, Lin YT, Wang JT. 2015. Colistin resistance mechanisms in Klebsiella pneumoniae strains from Taiwan. Antimicrob Agents Chemother 59:2909–2913. https://doi.org/10.1128/AAC.00763-14

10. Pitt ME, Elliott AG, Cao MD, Ganesamoorthy D, Karaiskis I, Giamarellou D, Mao C, Nordberg EK, Olsen GJ, Murphy-Olson DE, Olson R, Overbeek D, O’Donnell P, Thorson S, Amatya P, Jodh S. 2017. Whole genome sequencing for the molecular characterization of carbapenem-resistant Klebsiella pneumoniae strains isolated at the Italian ASST Fatebenefratelli Sacco Hospital, 2012–2014. BMC Infect Dis 17:666. https://doi.org/10.1186/s12879-017-2760-7

11. Stoezer N, Giess A, Batty EM, Sheppard AE, Walker AS, Wilson DJ, Didelot X, Bashir A, Sebra R, Kasarski A, Shapotit B, Shlaya K, Kelly D, Pollard AJ, Petto TE, Crook DW, Donnelly P, Thorison S, Amatya P, Jodh S. 2017. Whole genome sequencing for the molecular characterization of carbapenem-resistant Klebsiella pneumoniae strains isolated at the Italian ASST Fatebenefratelli Sacco Hospital, 2012–2014. BMC Infect Dis 17:666. https://doi.org/10.1186/s12879-017-2760-7

12. Didelot X, Bashir A, Sebra R, Kasarski A, Shapotit B, Shlaya K, Kelly D, Pollard AJ, Petto TE, Crook DW, Donnelly P, Thorison S, Amatya P, Jodh S. 2017. Whole genome sequencing for the molecular characterization of carbapenem-resistant Klebsiella pneumoniae strains isolated at the Italian ASST Fatebenefratelli Sacco Hospital, 2012–2014. BMC Infect Dis 17:666. https://doi.org/10.1186/s12879-017-2760-7
Machine Learning Prediction of Polymyxin Resistance

27. Collins C, Didelez X. 2018. A phylogenetic method to perform genome-wide association studies in microbes that accounts for population structure and recombination. PLoS Comput Biol 14:e1006958. https://doi.org/10.1371/journal.pcbi.1006958.

28. Rajkomar A, Dean J, Kohane I. 2019. Machine learning in medicine. N Engl J Med 380:1347–1358. https://doi.org/10.1056/NEJMra1814259.

29. Giacobbe DR, ISGIR-SITA (Italian Study Group on Resistant Infections of the Societa Italiana Terapia Antinfettiva), Del Bono V, Trecarichi EM, De Rosa FG, Giannella M, Bassetti M, Bartoloni A, Losito AR, Corcione S, Bartolotti M, Mantegoni E, Saffotti C, Panagi N, Tedeschi S, Spanu T, Rossolini GM, Marchese A, Ambretti S, Cauda R, Viale P, Viscoli C, Tumbarello M, Isgrì S. 2015. Risk factors for bloodstream infections due to colistin-resistant KPC-producing Klebsiella pneumoniae: results from a multicenter case-control-control study. Clin Microbiol Infect 21:1106.E1–1106.E8. https://doi.org/10.1016/j.cmi.2015.08.001.

30. Rzik G, Lavreni D, Chiki R. 2013. DSK: k-mer counting with very low memory usage. Bioinformatics 29:652–653. https://doi.org/10.1093/bioinformatics/btt200.

31. Drouin A, Giguère S, Deraspe M, Marchand M, Tyers M, Loo VG, Bourmaut AM, Laviolette F, Corbeil J. 2016. Predictive computational phenotyping and biomarker discovery using reference-free genome comparisons. BMC Genomics 17:754. https://doi.org/10.1186/s12864-016-3869-6.

32. Chetneet YH, Li N, Wang JT. 2018. Novel putative RND-type efflux pump, H239_3064, contributes to colistin resistance through CrrB in Klebsiella pneumoniae. J Antimicrob Chemother 73:1509–1516. https://doi.org/10.1093/jac/dky054.

33. Gatzeva-Topalova PZ, May AP, Sousa MC. 2005. Structure and mechanisms in AnmA: conformational change implies ordered dehydrogenase mechanism in key enzyme for polymyxin resistance. Structure 13:295–302. https://doi.org/10.1016/j.str.2005.01.018.

34. Long SW, Olsen RJ, Edgar TN, Beres SB, Zhao P, Davis JJ, Brettin T, Xia F, Musser JM. 2017. Population genomic analysis of 1,777 extended-spectrum beta-lactamase-producing Klebsiella pneumoniae isolates, Houston, Texas: unexpected abundance of clonal group 307. mBio 8:e00489-17. https://doi.org/10.1128/mBio.00489-17.

35. Nguyen M, Brettin T, Long SW, Musser JM, Olsen RJ, Olson R, Shukla M, Musser JR, Davis JJ. 2018. Using an in silico minimum inhibitory concentration panel test for Klebsiella pneumoniae. Sci Rep 8:4211. https://doi.org/10.1038/s41598-017-18972-w.

36. Peseyky M, Hussain T, Wallace M, Patel S, Andleeb S, Burnham C, Dantas G. 2016. Evaluation of machine learning and rules-based approaches for predicting antimicrobial resistance profiles in Gram-negative bacilli from whole genome sequence data. Front Microbiol 7:912. https://doi.org/10.3389/fmicb.2016.00912.

37. Humphries RM, Ambler J, Mitchell SL, Castanheira M, Dingle T, Hindler JF, Thompson L, Török ME, Ouadi K, Stevens E, Yokoyama M, Steventon J, Thompson L, Biggs J, Connor TR, Hsu LY, Severin J, Brisse S, Cao H, Wilksch J, Gorrie C, Schultz MB, Edwards DJ, Nguyen KV, Nguyen TV, Dao TT, Mensink M, Mihura TF, Rohde H, Massey R, Mack D, Corander J, Sheppard SK. 2018. Disease- and Klebsiella pneumoniae isolates. J Clin Microbiol 56:e01934-17. https://doi.org/10.1128/JCM.00149-18.

38. Adams MD, Bishop B, Wright MS. 2016. Quantitative assessment of insertion sequence impact on bacterial genome architecture. Microbiom 2:e000062. https://doi.org/10.1093/microbiom/mew0062.

39. Arena F, Henriç De Angelis L, Cannatelli A, Di Pilato V, Amorese M, D’Andrea MM, Giani T, Rossolini GM. 2016. Colistin resistance caused by inactivation of the MgrB regulator is not associated with decreased virulence of sequence type 258 KPC carbapenemase-producing Klebsiella pneumoniae. Antimicrob Agents Chemother 60:2509–2512. https://doi.org/10.1128/AAC.02981-15.

40. Jeni LA, Cohn JF, De La Torre F. 2013. Facing imbalanced data recomposition of TBProfiler and Mykrobe for antibiotic resistance prediction in Mycobacterium tuberculosis. PeerJ 7:e6857-21. https://doi.org/10.7717/peerj.6857.

41. Richter SE, Miller L, Uslan DZ, Bell D, Watson K, Humphries R, McKinnell JA. 2018. Risk factors for colistin resistance among Gram-negative rods and Klebsiella pneumoniae isolates. J Clin Microbiol 56:e00149-18. https://doi.org/10.1128/JCM.00149-18.

42. Davis JJ, Boisvert S, Brettin T, Kenyon RW, Mao C, Olson R, Overbeek R, Santerre J, Shukla M, Wattam AR, Will R, Xia F, Stevens R. 2016. Antimicrobial resistance prediction in PATRIC and RAST. Sci Rep 6:27930. https://doi.org/10.1038/srep27930.

43. Wei Z, Wang W, Bradfield J, Li J, Cardinale C, Frackelton E, Kim C, Mchent F, Van Steen K, Vischer PM, Baldassano RN, Hakonarson H, International IBID Genetics Consortium. 2013. Large sample size, wide variant spectrum, and advanced machine-learning technique boost risk prediction for inflammatory bowel disease. Am J Hum Genet 92:1008–1012. https://doi.org/10.1016/j.ajhg.2013.05.002.

44. Dantas G. 2016. Evaluation of machine learning techniques to tuberculosis drug resistance analysis. Bioinformatics 32:1666–1671. https://doi.org/10.1093/bioinformatics/btw160.

45. Mecas AUX. 2015. Disease-interactions in disease risk prediction—machine learning and network perspectives. BioData Min 6:5. https://doi.org/10.1186/s13338-015-0065-5.

46. Wei Z, Wang K, Qu H-Q, Zhang H, Bradford J, Kim C, Frackelton E, Hou C, Fricker B, Grossner JT, Cao H, Sansores D, Massey R, Hakonarson H. 2009. From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes. PLoS Genet 5:e1000678-11. https://doi.org/10.1371/journal.pgen.1000678.

47. Giddins MJ, Macesic N, Annavajhala MK, Wang Z, McAsce N, Hu Y, Giddins
64. Inouye M, Dashnow H, Raven LA, Schultz MB, Pope BJ, Tomita T, Zobel V, Macesic N, Gomez-Simmonds A, Sullivan SB, Giddins MJ, Ferguson SA, Marti-Ferri X, Iqbal Z, Mottram JL, Brown K, Morgan AW, Stenmark M, Zmolek M, Wilkin D, Deiseroth E, Loo LY, Delwart E, Miller JR, Marra MA, Fraser CM. 2014. SRST2: rapid genomic surveillance for public health and clinical microbiology labs. Genome Med 6:90. https://doi.org/10.1186/s13073-014-0090-6

65. Wick RR, Judd LM, Gorrie CL, Holt KE. 2017. Unicycler: resolving bacterial structure improves power in bacterial association studies. Nat Microbiol 2:1093. https://doi.org/10.1038/s41564-017-0093-1

66. Seemann T. 2018. Snippy: rapid haploid variant calling and core SNPMine the genetic basis of bacterial phenotypes. Nat Commun 7:12797. https://doi.org/10.1038/ncomms12797

67. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, Muller M. 2011. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 12:77. https://doi.org/10.1186/1471-2105-12-77

68. Lees JA, Vehkala M, Qarbar CM, Lent RC, Bou Ghanem EN, Shainheit MG. 2019. PepN is a non-essential, cell wall-localized protein that contributes to neutrophil elastase-mediated killing of Streptococcus pneumoniae. PLoS One 14:e0211632. https://doi.org/10.1371/journal.pone.0211632

69. Jaillard M, Lima L, Tournoud M, Mahe P, van Belkum A, Lacroix V, Jacob SD, Corander J. 2016. Octapeptin C4 and polymyxin resistance occur via distinct pathways in an epidemic XDR Klebsiella pneumoniae ST258 isolate. J Antimicrob Chemother 74:582–593. https://doi.org/10.1093/jac/dky458

70. Hooper DC, Jacoby GA. 2015. Mechanisms of drug resistance: quinolone resistance. Ann N Y Acad Sci 1354:12–31. https://doi.org/10.1111/nyas.12830

71. Tong S, Lin Y, Lu S, Wang M, Bogdanov M, Zheng L. 2016. Structural insight into substrate selection and catalysis of lipid phosphate phosphatase PgpB in the cell membrane. J Biol Chem 291:18342–18352. https://doi.org/10.1074/jbc.M116.737874.

72. Rahavan et al. 2018. Faster SPAdes (or better SKESA/Megahit/Velvet) assembly of Illumina reads. PLoS Comput Biol 13:e1005955. https://doi.org/10.1371/journal.pcbi.1005955

73. Lamarche MG, Wanner BL, Crepin S, Harel J. 2008. The phosphate regulon and bacterial virulence: a regulatory network connecting phosphate homeostasis and pathogenesis. FEMS Microbiol Rev 32:461–473. https://doi.org/10.1111/j.1574-6976.2008.00101.x

74. Miethke M, Marahiel MA. 2007. Siderophore-based iron acquisition and pathogen control. Microbiol Mol Biol Rev 71:413–451. https://doi.org/10.1128/MMBR.00122-07

75. Bhagirath AY, Li Y, Patidar R, Yerex K, Ma X, Kumar A, Duan K. 2019. Two component regulatory systems and antibiotic resistance in Gram-negative pathogens. Int J Mol Sci 20:E1781. https://doi.org/10.3390/ijms20071781

76. Dorman MJ, Feltwell T, Goulding DA, Parkhill J, Short FL. 2018. The capsule phosphate PgpB in the cell membrane. J Biol Chem 291:18342–18352. https://doi.org/10.1074/jbc.M116.737874.

77. Lamarche MG, Wanner BL, Crepin S, Harel J. 2008. The phosphate regulon and bacterial virulence: a regulatory network connecting phosphate homeostasis and pathogenesis. FEMS Microbiol Rev 32:461–473. https://doi.org/10.1111/j.1574-6976.2008.00101.x

78. Tong S, Lin Y, Lu S, Wang M, Bogdanov M, Zheng L. 2016. Structural insight into substrate selection and catalysis of lipid phosphate phosphatase PgpB in the cell membrane. J Biol Chem 291:18342–18352. https://doi.org/10.1074/jbc.M116.737874.

79. Pitt ME, Cao MD, Butler MS, Ramu S, Ganesamoorthy D, Blaskovich MAT, Coin LJM, Cooper MA. 2019. Octapeptin C4 and polymyxin resistance occur via distinct pathways in an epidemic XDR Klebsiella pneumoniae ST258 isolate. J Antimicrob Chemother 74:582–593. https://doi.org/10.1093/jac/dky458

80. Hooper DC, Jacoby GA. 2015. Mechanisms of drug resistance: quinolone resistance. Ann N Y Acad Sci 1354:12–31. https://doi.org/10.1111/nyas.12830

81. Klein G, Raina S. 2019. Regulated assembly of LPS, its structural alterations and cellular response to LPS defects. Int J Mol Sci 20:E1781. https://doi.org/10.3390/ijms20071781

82. Tong S, Lin Y, Lu S, Wang M, Bogdanov M, Zheng L. 2016. Structural insight into substrate selection and catalysis of lipid phosphate phosphatase PgpB in the cell membrane. J Biol Chem 291:18342–18352. https://doi.org/10.1074/jbc.M116.737874.