PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Predicting antimicrobial resistance using conserved genes

Marcus Nguyen[1,2], **Robert Olson**[1,2], **Maulik Shukla**[1,2], **Margo VanOeffelen**[3], **James J. Davis**[1,2,3,4]*

1 Division of Data Science and Learning, Argonne National Laboratory, Argonne Illinois, United States of America, **2** Consortium for Advanced Science and Engineering, University of Chicago, Chicago, Illinois, United States of America, **3** Fellowship for Interpretation of Genomes, Burr Ridge, Illinois, Illinois, United States of America, **4** Northwestern Argonne Institute for Science and Engineering, Evanston, Illinois, United States of America

* jjdavis@anl.gov

OPEN ACCESS

Citation: Nguyen M, Olson R, Shukla M, VanOeffelen M, Davis JJ (2020) Predicting antimicrobial resistance using conserved genes. PLoS Comput Biol 16(10): e1008319. https://doi.org/10.1371/journal.pcbi.1008319

Editor: Jason A. Papin, University of Virginia, UNITED STATES

Received: May 11, 2020

Accepted: September 7, 2020

Published: October 19, 2020

Data Availability Statement: All genome sequence data are available from NCBI and PATRIC, and genome IDs, gene IDs and PubMed IDs are provided to these data sources where appropriate in the supplemental tables and main text. The underlying data, genes, and models corresponding to the alignment based models described in this study are available at our github site for this project: https://github.com/jimdavis1/Core-Gene-AMR-Models. K-mer based models are too large to host in this way, but we show that results from the alignment and k-mer based models are equivalent in the main text.

Abstract

A growing number of studies are using machine learning models to accurately predict antimicrobial resistance (AMR) phenotypes from bacterial sequence data. Although these studies are showing promise, the models are typically trained using features derived from comprehensive sets of AMR genes or whole genome sequences and may not be suitable for use when genomes are incomplete. In this study, we explore the possibility of predicting AMR phenotypes using incomplete genome sequence data. Models were built from small sets of randomly-selected core genes after removing the AMR genes. For *Klebsiella pneumoniae*, *Mycobacterium tuberculosis*, *Salmonella enterica*, and *Staphylococcus aureus*, we report that it is possible to classify susceptible and resistant phenotypes with average F1 scores ranging from 0.80–0.89 with as few as 100 conserved non-AMR genes, with very major error rates ranging from 0.11–0.23 and major error rates ranging from 0.10–0.20. Models built from core genes have predictive power in cases where the primary AMR mechanisms result from SNPs or horizontal gene transfer. By randomly sampling non-overlapping sets of core genes, we show that F1 scores and error rates are stable and have little variance between replicates. Although these small core gene models have lower accuracies and higher error rates than models built from the corresponding assembled genomes, the results suggest that sufficient variation exists in the core non-AMR genes of a species for predicting AMR phenotypes.

Author summary

Machine learning models for predicting AMR phenotypes from sequence data are often built using features derived from well-studied sets of AMR genes, or from whole genome sequences. In this study, we build models using core genes that are held in common among the members of a species and that are not known to confer antimicrobial resistance based on their annotations. We find that there is sufficient variation in these core conserved genes to produce models with accuracies greater than or equal to 80% in four

Funding: This work is funded by the United States Defense Advanced Research Projects Agency iSENTRY Friend or Foe program award [Contract No. HR0011937807], (https://www.darpa.mil) to JJD, and by the United States National Institute of Allergy and Infectious Diseases Bacterial and Viral Bioinformatics Resource Center award [Contract No. 75N93019C00076], (https://www.niaid.nih.gov) to PI Rick Stevens. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

species, using as few as 100 genes. However, we note that these models are less accurate than models built from whole genomes or lists of AMR genes. The results of this study suggest that variations relating to, or co-occurring with AMR are extensive, and that it is possible to use conserved non-AMR genes to predict AMR phenotypes.

Introduction

The discovery and use of antimicrobial agents for the treatment of bacterial infections revolutionized medicine in the twentieth century. In 1900, the top three causes of death in the United States were pneumonia, tuberculosis, and diarrhea/enteritis [1]. Antimicrobial therapy, coupled with medical advancements and sanitation improvements, has resulted in a marked shift in these statistics, with the top three causes of death in the United States now being heart disease, cancer, and unintentional injuries [2, 3]. The rise of antimicrobial resistance (AMR), along with the sluggish development of new antimicrobial drugs, threatens to jeopardize this achievement.

Antimicrobial susceptibility testing is the gold standard for determining which antibiotics will be effective against bacterial pathogens. This requires culturing the organisms in the presence of a panel of antibiotics [4, 5]. Since culturing can be slow, clinicians often rely on empirical judgement when administering antibiotics. When the administration of antibiotics is incorrect or inappropriate, it can increase mortality rates and exacerbate the spread of AMR [6, 7]. Developing diagnostics that can determine the AMR phenotypes of bacterial pathogens in real time is crucial for reducing morbidity and mortality in patients, and for lowering endemic levels of AMR through more precise antibiotic prescription and stewardship practices [8, 9].

With the continued reductions in cost and the development of devices that are more suitable for point of care use, genome sequencing has received attention for its potential value as a diagnostic tool [10–12]. Many bioinformatic techniques have been developed for making sequence-based comparisons against large comprehensive databases of known AMR genes, proteins, and variants making the prediction of resistant phenotypes possible [13–15]. These predictions are typically made using either rules-based or machine learning models [16–20]. Several studies have also built machine learning models for predicting AMR phenotypes by using assembled genomes or pan genomes as training sets [21–27]. In these cases, the machine learning algorithm detects the most discriminating features (typically short nucleotide k-mers) from a training set with laboratory-derived AMR phenotypes. Both the AMR gene- and whole genome-based approaches have the limitation that they require either a complete genome or the complete set of AMR genes from a genome to provide accurate AMR phenotype predictions.

Although whole genome sequencing of bacterial isolates provides extensive information about AMR, pathogenicity, and epidemiology, it also requires a culturing step, which means that it is not much faster than conventional susceptibility testing. Thus, culture-free diagnostic techniques, such as shotgun metagenomics and PCR-based amplification of AMR markers, represent appealing alternatives, but these also come with challenges. For instance, in metagenomics, where the pathogen DNA is sequenced directly from an infection source, there can be difficulty in eliminating contaminating host DNA, accurately binning reads or contigs into individual genomes, determining if binned genomes are complete, and assessing the risk posed by incomplete genomes found in the sample. Furthermore, while whole genome sequencing of pure cultures enables accurate source attribution for mobile genetic elements

carrying AMR genes, this becomes more difficult in a metagenomic sample [28]. PCR-based approaches, including the direct amplification of AMR genes, or amplification paired with sequencing, have similar challenges including the difficulties in amplifying a comprehensive set of AMR genes or regions, and the subsequent ability to attribute these detected AMR genes to specific pathogens. Given the obvious appeal and drawbacks of culture-free diagnostic approaches, developing strategies for predicting AMR phenotypes from incomplete genome sequence data presents an interesting technical challenge.

In a previous study, we used the complete assembled genomes of 1667 *Klebsiella pneumoniae* clinical isolates to build machine learning models for predicting AMR phenotypes for 20 antibiotics [21]. During that study, we built a similar model from the same set of genomes, except that we excluded the known AMR genes based on their annotations. To our surprise, the resulting model had nearly identical accuracies and error rates across all antibiotics compared with the model built from the full genomes (approximately 92%). Our results suggested that it may be possible to build accurate models with partial genome sequence data. In this study, we explore this finding in greater detail.

Results

AMR models based on core genes have predictive power

In previous work, we observed that is possible to build accurate AMR phenotype prediction models from whole genomes without using the AMR genes [21]. In this study, in order to explore the possibility of building models from limited genome sequence data, we chose to build models from core genes that are held in common among the members of a species, and which are not annotated as having a direct role in AMR [29, 30]. By being nearly universally conserved, core genes are less likely to be horizontally transferred, and are also useful for assessing genome completeness and phylogeny. We built machine learning models using the core gene sets for *K. pneumoniae*, *M. tuberculosis*, *S. enterica*, and *S. aureus*, which have a large number of publicly available genomes with laboratory-derived AMR metadata (Tables 1 and 2 and Tables A-F in S2 File). For all species, classifiers were built for predicting susceptible and resistant (SR) phenotypes. We used the XGBoost (XGB) [31] machine learning algorithm as described previously and 15-mer oligonucleotide k-mers from the core gene sets along with the SR phenotypes as features to train each model [21, 22].

For each species, we started by randomly selecting subsets of core genes ranging in size from 25–500 genes. We then built SR classifiers for each set, tuning the XGB parameter for tree depth, which has been shown previously to have the most influence on models of this type [21] (Figure A in S1 File). A tree depth of 16 was chosen for the models because the F1 scores tend to plateau beyond this point. In most cases, we see little improvement beyond depths of 16 regardless of gene set size, so it is likely that we are nearing the maximum accuracy that

Table 1. Data sets used in this study.

Species	Genomes	Antibiotics	Species-specific Core Genes[a]
K. pneumoniae	1667	18	3856
M. tuberculosis	5353	11	1670
S. enterica	1999[b]	10	2991
S. aureus	1274	6	1501

[a]Does not include genes with AMR-related annotations

[b]Down selected for diversity from a larger set of 5278 genomes published previously [22].

https://doi.org/10.1371/journal.pcbi.1008319.t001

Table 2. Counts of susceptible and resistant genomes used in this study, data are displayed as (Susceptible|Resistant).

Antibiotic	Abv.	*K. pneumoniae*	*M. tuberculosis*	*S. enterica*	*S. aureus*
Amikacin	AMK	1320|103	868|230		
Amoxicillin/Clavulanate	AMC			1489|400	
Ampicillin	AMP			1291|706	
Ampicillin/Sulbactam	SAM	90|1455			
Aztreonam	ATM	216|1407			
Capreomycin	CAP		846|214		
Cefazolin	CFZ	97|1570			
Cefepime	FEP	418|963			
Cefoxitin	FOX	667|828		1599|348	
Ceftazidime	CAZ	136|1488			
Ceftiofur	TIO			1602|393	
Ceftriaxone	CRO	80|1528		1601|397	
Cefuroxime sodium	CXM	91|1469			
Chloramphenicol	CHL			1864|88	
Ciprofloxacin	CIP	201|1424			752|243
Clindamycin	CLI				444|144
Erythromycin	ERY				1016|257
Ethambutol	EMB		3889|484		
Ethionamide	ETO		167|123		
Fusidic acid	FA				1156|83
Gentamicin	GEN	926|683		1639|329	
Imipenem	IPM	1160|478			
Isoniazid	INH		4098|1204		
Kanamycin	KAN		614|167		
Levofloxacin	LVX	349|1287			
Meropenem	MEM	1134|481			
Methicillin	MET				771|215
Moxifloxacin	MXF		593|85		
Ofloxacin	OFX		182|176		
Penicillin	PEN				175|1063
Piperacillin/Tazobactam	TZP	432|1048			
Pyrazinamide	PZA		3605|428		
Rifampin	RIF		4438|828		
Streptomycin	STR		2140|684	381|772	
Sulfisoxazole	FIS			1108|772	
Tetracycline	TET	739|778		867|1124	
Tobramycin	TOB	589|723			
Trimethoprim/Sulfamethoxazole	SXT	416|1251			

https://doi.org/10.1371/journal.pcbi.1008319.t002

these gene sets can provide given the genes in the sets, set sizes, and experimental design. For all four species, models built from 25 genes and optimized to a tree depth of 16, range in their average F1 scores from 0.75 [0.73–0.77, 95%CI] for *S. enterica* to 0.80 [0.78–0.81, 95%CI] for *K. pneumoniae* (Fig 1). The F1 scores increase as the set size increases, with the models built from 500 genes having F1 scores ranging from 0.84 [0.81–0.86, 95%CI] for *M. tuberculosis* to 0.89 [0.86–0.90, 95%CI] for *S. aureus*. The average very major error rate (VME), which is defined as resistant genomes that are erroneously predicted to be susceptible, and the average major error rate (ME), which is defined as susceptible genomes that are erroneously predicted

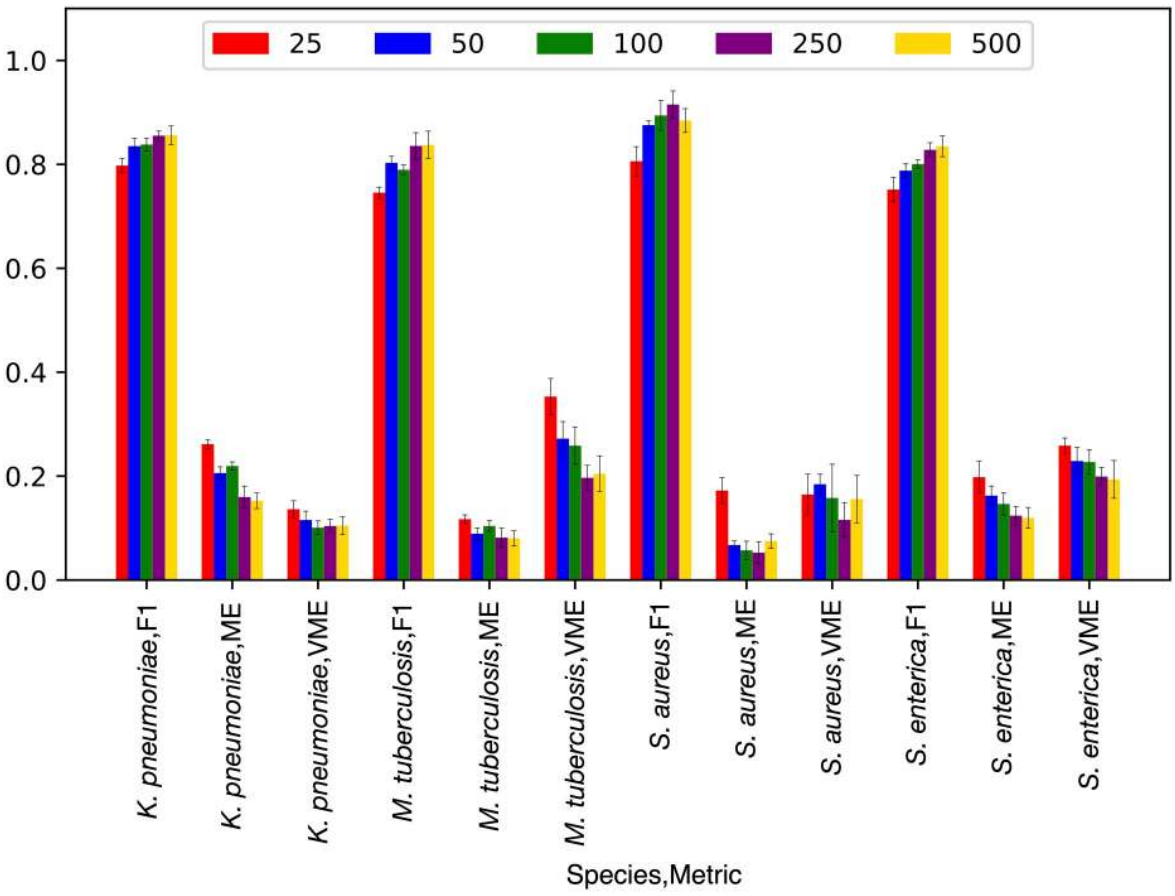

Fig 1. F1 scores, major error (ME), and very major error (VME) rates for AMR phenotype classifiers built from core gene sets. Randomly selected core gene sets ranging in size from 25–500 genes are shown. Error bars depict 95% confidence intervals.

https://doi.org/10.1371/journal.pcbi.1008319.g001

to be resistant, tend to go down as gene set size increases. Although the core gene set models described in Fig 1 have lower F1 scores and higher error rates than full-genome models that have been published previously [21–24, 27, 32], their accuracies are striking given the small sizes of the input data sets and the removal of well-annotated AMR genes.

Models built from randomly sampled core gene sets are consistent

To assess the variability that could be expected from building models from core gene sets, we built models by randomly selecting non-overlapping sets of genes. Ten models, each containing 100 non-overlapping core genes, were computed for each species and the accuracies, F1 scores, and error rates were averaged over all 10 models (Fig 2, Table G in S2 File). The average F1 scores for these 100-gene models range from 0.80 in *M. tuberculosis* and *S. enterica* to 0.89 for *S. aureus*, and are 5–17% lower than the accuracies for models built from the same set of genomes using the whole assembled genomes as input. Within a species, we observe little variation in the accuracies or F1 scores for each gene set. In each of the four species, the average 95% confidence intervals for each model differ by only 1–2% for all ten replicates. This indicates that most randomly-selected subsets of 100 core genes will have accuracies within this range, regardless of the functions encoded in the underlying gene sequences.

Although the F1 scores are higher for models built from whole genomes, the models built from core genes show a similar pattern across antibiotics. That is, when a F1 score for an

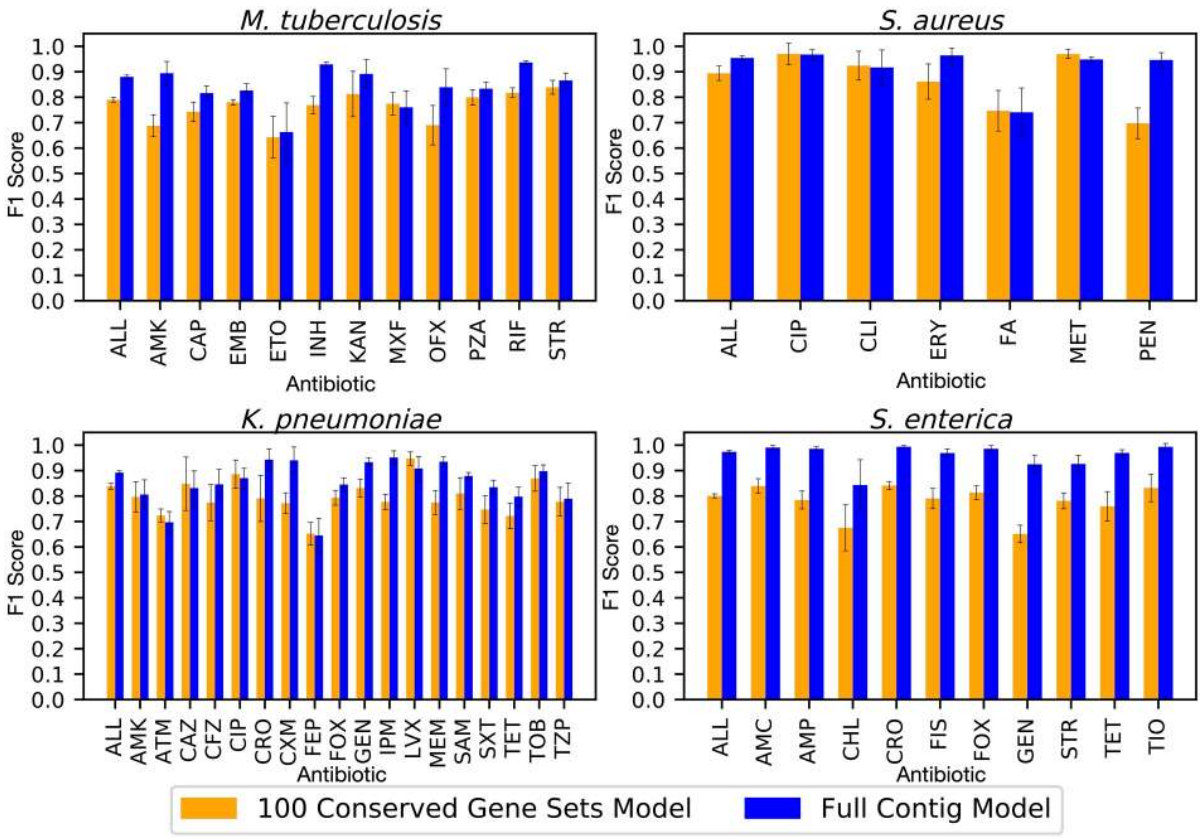

Fig 2. Average F1 scores by antibiotic for ten models built from non-overlapping sets of 100 core genes (orange bars) compared with the F1 scores for models built from whole assembled contigs for the same genomes (blue bars). Error bars represent 95% confidence intervals for the whole genome model, and the minimum and maximum confidence interval observed in all ten replicates for the core gene models. Antibiotic abbreviations are defined in Table 2.

https://doi.org/10.1371/journal.pcbi.1008319.g002

antibiotic is high in a whole genome model, it also tends to be high in the 100-gene set models and vice versa (Fig 2). There are also many examples where the accuracy is high regardless of whether the AMR mechanism is known to result from chromosomally-encoded SNPs in core genes, or horizontal gene transfer. For instance, the 100 core gene models have average F1 scores of 0.90 [0.88–0.91, 95%CI] and 0.96 [0.96–0.97, 95%CI] for predicting ciprofloxacin resistance in *K. pneumoniae* and *S. aureus*, respectively. Ciprofloxacin resistance is often caused by SNPs in the *gyrA* gene [33], but this gene is not used in any of the ten subsamples for either species. This means that other core genes carry sufficient information for making the prediction. This is also true in cases where the AMR mechanism is the result of horizontal gene transfer. One notable example is methicillin resistance in *S. aureus*, where resistance genes are carried by SCC*mec* elements [34], which had an F1 score of 0.96 [0.95–0.97, 95% CI]. Another example is cephalosporin resistance in *S. enterica*, which has F1 scores ranging from 0.82–0.84, and is known to be plasmid-mediated [35] (Table G in S2 File). The annotated functions for each of the ten core gene subsets are listed in Table H of S2 File for each species.

In nearly all cases, the error rates are higher for the core gene models than they are for the whole genome models. The average VME rates for the whole genome models range from 0.04 [0.03–0.05, 95%CI] in *S. enterica* to 0.12 [0.11–0.12, 95%CI] for *M. tuberculosis*, and in the core gene models, they range from 0.11 [0.10–0.12, 95%CI] in *K. pneumoniae* to 0.23 [0.21–0.24, 95%CI]for *M. tuberculosis* (Fig 3, Table G in S2 File). Likewise, the ME rates are mostly

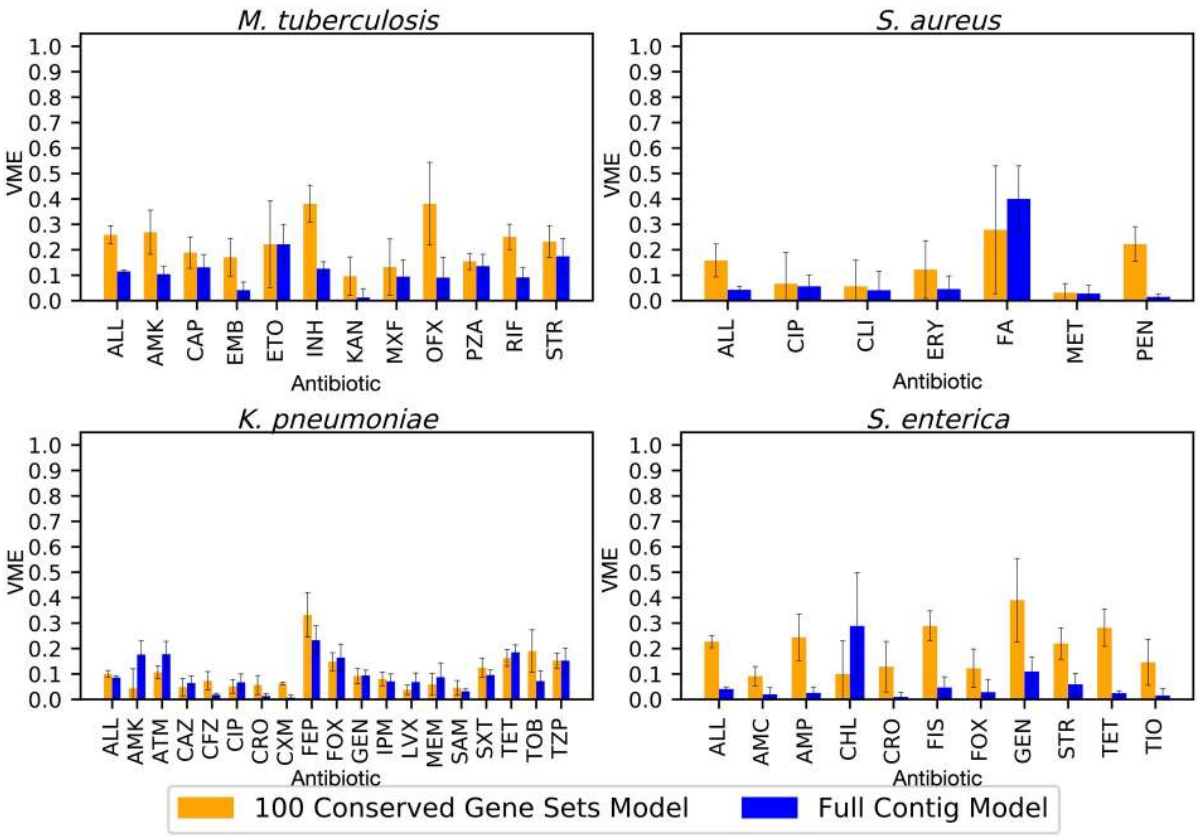

Fig 3. Average very major error rate (VME) by antibiotic for ten models built from non-overlapping sets of 100 randomly-selected core genes (orange bars) compared with the VME for models built from whole assembled contigs for the same genomes (blue bars). VMEs are defined as resistant genomes being incorrectly classified as susceptible. Error bars represent 95% confidence intervals for the whole genome model, and the minimum and maximum confidence interval observed in all ten replicates for the core gene models. Antibiotic abbreviations are defined in Table 2.

https://doi.org/10.1371/journal.pcbi.1008319.g003

higher in the core gene models than the whole genome models (Fig 4). The average ME rates for the whole genome models range from 0.01 [0.01–0.02, 95%CI] for *S. enterica* to 0.12 [0.11–0.13, 95%CI] for *K. pneumoniae*. In the 100 gene models, the average ME rates range from 0.06 [0.05–0.07, 95%CI] in *S. aureus* to 0.20 [0.18–0.22, 95%CI] in *K. pneumoniae* (Table G in S2 File, Fig 4). Overall, the antibiotics with poor F1 scores also tend to have higher error rates. There are also larger confidence intervals, and thus more variability predicting S or R phenotypes for antibiotics with an underrepresented class. Based on Fig 1, we would expect the VME and ME rates to go down as the core gene set size is increased beyond 100 genes.

Experimental approach has little effect on core gene models

Although we monitored all models using a validation set to prevent overfitting, it is possible that some aspect of the algorithm or approach, rather than the underlying nucleotide sequences, could be causing the high accuracies observed in the core gene set models. One possibility is that the models are capable of memorizing the data set, rather than learning the nucleotide variation associated with AMR phenotypes. If this were true, we would observe high accuracies regardless of how the genomes are labeled. To test this, we built ten models for ten randomly selected non-overlapping sets of 100 core genes as described above. We then shuffled the labels (i.e., the phenotypes) prior to training the models, and measured the

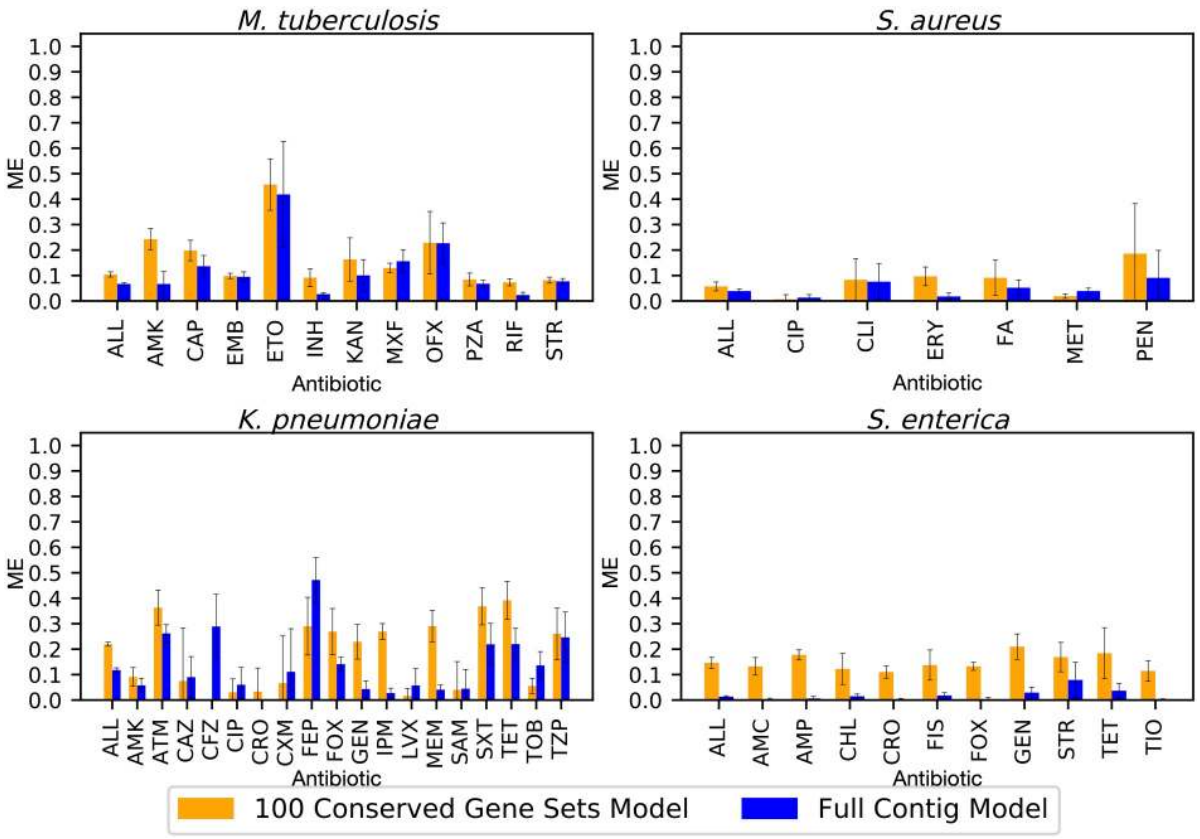

Fig 4. Average major error rate (ME) by antibiotic for ten models built from non-overlapping sets of 100 randomly-selected core genes (orange bars) compared with the ME for models built from whole assembled contigs (blue bars). MEs are defined as susceptible genomes being incorrectly classified as resistant. Error bars represent 95% confidence intervals for the whole genome model, and the minimum and maximum confidence interval observed in all ten replicates for the core gene models. Antibiotic abbreviations are defined in Table 2.

https://doi.org/10.1371/journal.pcbi.1008319.g004

resulting F1 scores (Fig 5). The average F1 scores for the models built from shuffled labels fall to approximately 50%, which is what would be expected from a random guess. This indicates that sequences associated with each phenotype are important for generating an accurate prediction and that the model is not memorizing the data set.

It is possible that some unforeseen aspect of the experimental design, such as using k-mers as features or XGB as the machine learning algorithm, could be resulting in the high accuracies observed for the core gene models. To control for this, we built a concatenated alignment of 100 randomly-selected core genes for each species. To convert the categorical nucleotide alignment data into numeric values for machine learning, the alignment was one-hot encoded (i.e., converted to a binary vector) and used to build a matrix with the one-hot encoded antibiotics for each species. Models were trained using both the XGB and Random Forest (RF) algorithms. The resulting F1 scores and accuracies for the alignment-based models built using XGB and RF are nearly identical, with overlapping 95% confidence intervals (Fig 6). The F1 scores from the k-mer-based and alignment-based XGB models built from the same set of genes are also nearly identical (Fig 6). The strong agreement between the XGB models built using k-mers or alignment columns as features indicates that feature selection is not biasing the outcome, and that both approaches are using the underlying nucleotide sequence to make the prediction. Furthermore, the close agreement between the XGB and RF models indicates

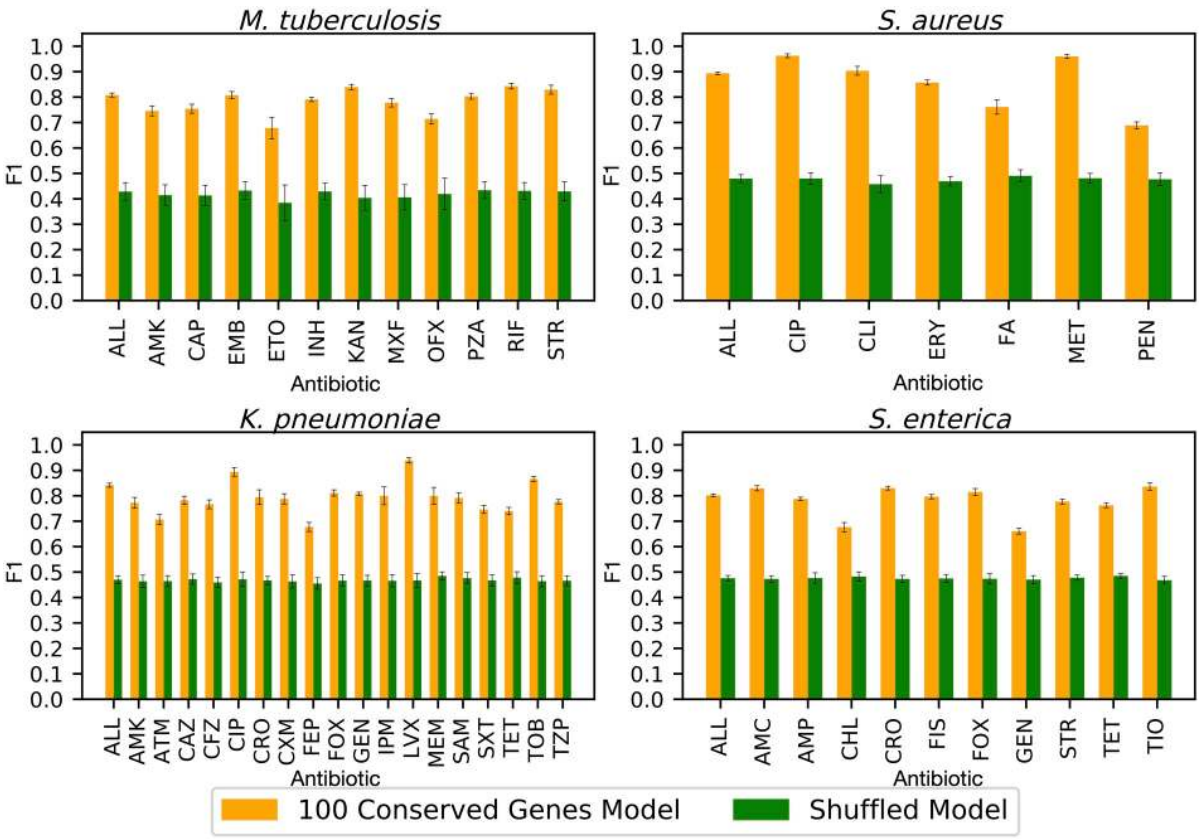

Fig 5. Average F1 scores by antibiotic for ten models built from non-overlapping sets of 100 core genes (orange bars) versus the same model where the AMR phenotypes have been randomized (green bars). Error bars represent the minimum and maximum 95% confidence observed in all ten replicates. Antibiotic abbreviations are defined in Table 2.

https://doi.org/10.1371/journal.pcbi.1008319.g005

that the high accuracies are not an anomaly relating the algorithm choice, with the slight variation in F1 scores, ME, and VME rates being more consistent with what would be expected from comparing results from different machine learning algorithms.

One aspect of the experimental design that could be increasing the reported accuracy of the core gene models is our decision to build a single model for predicting all AMR phenotypes, rather than building an individual model for each antibiotic. In previous work, we observed that combining antibiotics may result in slightly more accurate models, presumably because related antibiotics lend support to each other in the combined model [22]. To evaluate this, we built individual models for each antibiotic using the same set of 100 randomly selected genes that was used in Fig 1 (Figures B-D in S1 File). The F1 scores averaged over all individual antibiotic models is 2–8% lower than the average F1 scores for the merged models. However, when the 95% confidence intervals are compared for each antibiotic, they mostly overlap between the merged and individual models. Overall, combining all antibiotics into a single model slightly improves the F1 scores in some cases, but this is not sufficient to explain why the small non-AMR core gene models have accuracies exceeding 80%.

In previous work, we established methods for predicting minimum inhibitory concentrations (MICs) from whole genome sequence data [21, 22]. Although we lack MIC data for *M. tuberculosis* and S. aureus, we also built MIC models using a randomly selected subset of 100 core genes for *K. pneumoniae* and *S. enterica*. In the case of *K. pneumoniae*, the average

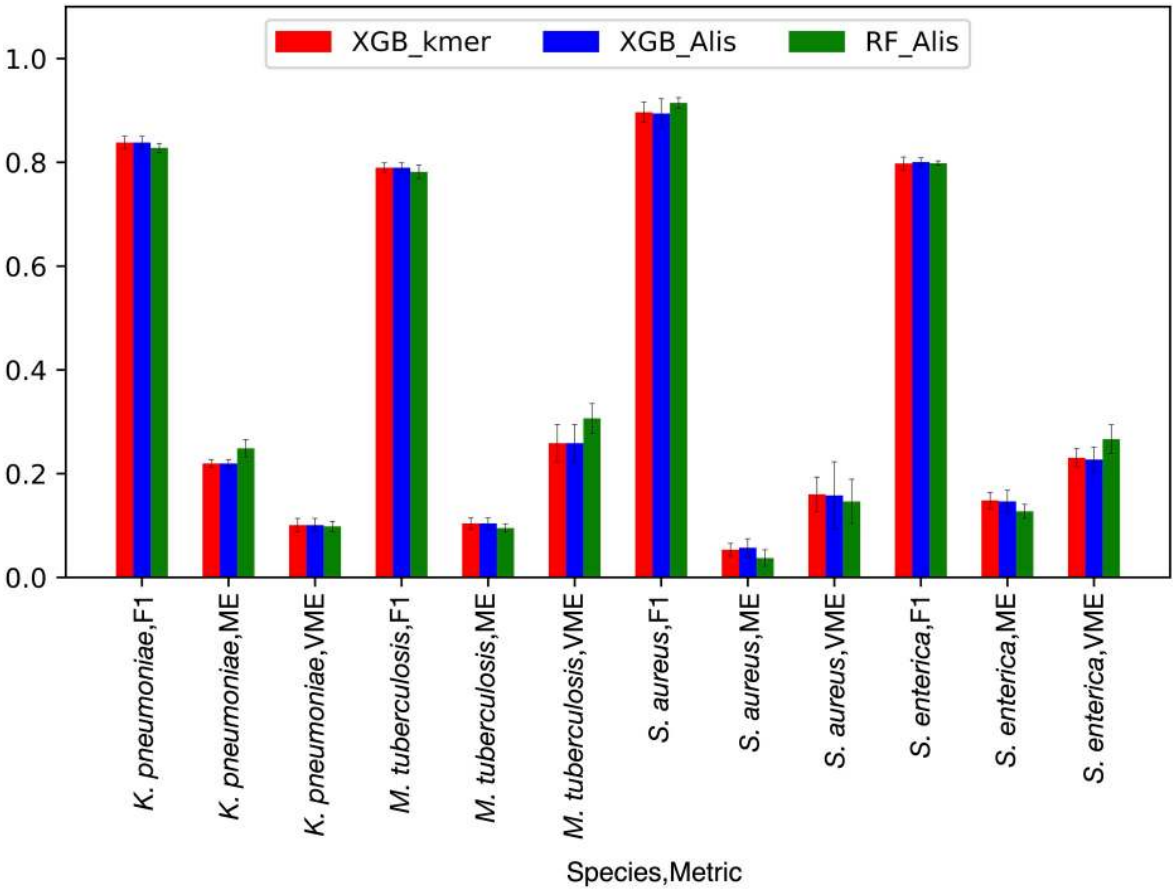

Fig 6. A comparison of algorithms and features. The same set of 100 randomly-selected core genes was used to build each model. Each set of bars represent the F1, ME, and VME scores for *K. pnuemoniae*, *M. tuberculosis*, *S. enterica*, and *S. aureus*, respectively. The red, blue, and green bars represent the accuracy metrics (F1, ME, or VME) for the XGBoost model built from k-mers for the gene set, XGBoost model built using a concatenated alignment where the columns were one-hot encoded, and random forest model built using a concatenated alignment where columns were one-hot encoded, respectively. Error bars depict the 95% confidence interval.

https://doi.org/10.1371/journal.pcbi.1008319.g006

accuracy of the model within ± 1 two-fold dilution step is 0.87 [0.86–0.87, 95%CI], and is 5% lower than the corresponding whole genome model 0.92 [0.91–0.92, 95%CI] (Table I in S2 File). In the case of *S. enterica*, the core gene MIC prediction model has an accuracy of 0.74 [0.73–0.74, 95%CI], and is 17% lower than the accuracy the corresponding whole genome model, which is 0.91 [0.90–0.93, 95%CI]. For both species, the corresponding error rates are also higher for the MIC models built from 100 core genes (Table I in S2 File). The 5% drop in accuracy between the whole genome and core gene MIC models for *K. pneumoniae*, and the 17% drop in accuracy for *S. enterica*, are both consistent with the difference in accuracies observed for the SR models described in Table G in S2 File, which are 3% and 15% lower for *Klebsiella* and *Salmonella*, respectively (F1 scores are shown in Fig 2). Thus, the data requirements for predicting MICs and susceptible and resistant phenotypes appear to be similar in these cases. The *Salmonella* data set differs from the other data sets because the susceptible genomes are slightly over represented and because the genomes were down selected prior to analysis. A combination of these factors, or some other aspect of *Salmonella* biology, may be contributing to the larger drop in accuracy going from the whole genome models to the core gene models.

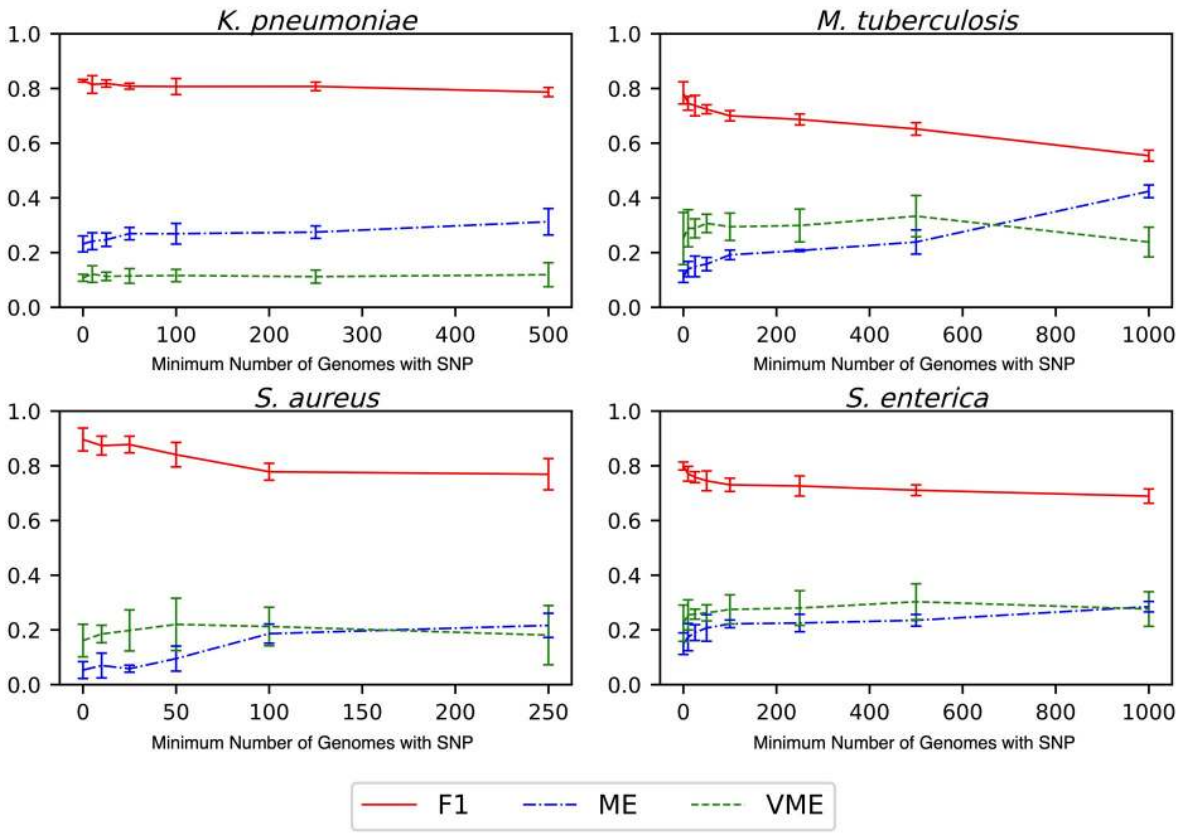

Fig 7. The effect of removing strain-specific SNPs prior to building models. SNP-containing alignment columns with less than 10, 25, 50 etc. nucleotides of the same type were successively removed and a model was generated. The X-axis depicts the SNP conservation and the Y-axis depicts the F1 score. The F1 score (red line), ME rate (blue line), and VME rate (green line) are shown. Error bars depict the 95% confidence interval. ME, or major error, is defined as a susceptible genome that is misclassified as being resistant. VME, or very major error, is defined as a resistant genome that is misclassified as being susceptible.

https://doi.org/10.1371/journal.pcbi.1008319.g007

The models are making predictions from conserved SNPs

The machine learning models described in this study work by using decision trees to make the AMR phenotype predictions. Each node in a decision tree represents a feature that contributes to the overall prediction for a given genome. It is difficult to know how well a model copes with using highly conserved SNPs that are abundant versus strain specific SNPs that occur infrequently in the population. A model that makes its predictions from infrequently occurring SNPs may be biased toward outlier data and not generalizable. On the other hand, a model that is able to make predictions based on more conserved SNPs may be more generalizable.

To evaluate this, we used the alignments of 100 core genes described in the previous section. Passing over each column in the alignment, we asked if the column contains a SNP, and if so, how frequently each nucleotide occurred in that position. We then eliminated the SNP containing columns, starting with those that contained nucleotides occurring most infrequently, reasoning that the infrequently occurring nucleotides represented strain-specific SNPs. This was then repeated eliminating alignment columns containing nucleotides that occurred fewer than 10, 25, 50, etc. times; the F1 scores, ME, and VME rates were recorded for each run (Fig 7). For each species, the average F1 scores slightly decrease, and the error rates slightly

increase as more SNP-containing columns are removed. However, since this trend is so gradual, we would conclude that SNPs that are conserved still contain a large amount of predictive power, and strain-specific SNPs are not required for making predictions with high accuracies and low error rates.

Clade size and the distribution of phenotypes within clades has little influence on the models

Strain diversity may play a role in influencing the accuracy of the models. In the context of a phylogenetic tree, this could happen if many strains are closely related, resulting some clades being much larger than others, or if strains with susceptible or resistant phenotypes are over-represented in certain clades. Indeed, it has been shown previously that population structure can provide predictive information to AMR models [36, 37]. Because of the potential bias that strain diversity could impart on the models, previous studies have used MLST-based [38] and k-mer similarity-based weighting schemes [39] to both normalize models and assess generalizability. In this case, rather than defining a population at a fixed distance, by traversing the phylogenetic tree and defining clades at various depths, we are normalizing the effect that strain relationships have on the models ranging from superficial strain similarity to deep evolutionary patterns within the species.

To account for biases due to clade size and phenotype distribution within clades, phylogenetic trees were computed for each species using 100 core genes that were not used in any of the AMR models (Figures E-H in S1 File). We defined clades of varying sizes based on tree distance, and generated k-mer based models, weighting the influence of each genome in the model based on 1) how many genomes are in a given clade, and 2) how many susceptible and resistant genomes occur in a given clade. Models were computed using each weighting scheme and a combination of both, using a random sample of 100 core genes.

Overall, for each of the four species, there are no observable differences between the unweighted models and the models that were built from genomes that were weighted according to clade size at various tree distances (Table J in S2 File). Similarly, in most cases the 95% confidence intervals overlap for unweighted models and models that were weighted according to the distribution of susceptible and resistant genomes within each clade at each tree distance (Table J in S2 File). Combining both weighting schemes also has a nearly identical outcome to weighting according to the distribution of S and R genomes within clades.

Since training and testing at the same tree distance could potentially mask biases, each model that was trained at a given tree distance was also used to evaluate the test set of genomes defined at every other tree distance for a given species, and the F1 scores were averaged by clade (Table K in S2 File). For clade-size weighted models, the minimum and maximum 95% confidence intervals for the F1 scores observed for all comparisons in this analysis were 0.79–0.89 for *K. pneumoniae*, 0.78–0.91 for *M. tuberculosis*, 0.75–0.86 for *S. enterica*, and 0.84–0.96 for *S. aureus*. SR weighting had a similar outcome, with the minimum and maximum 95% confidence intervals for any comparison being 0.74–0.87 for *K. pneumoniae*, 0.72–0.90 for *M. tuberculosis*, 0.75–0.86 for *S. enterica*, and 0.83–0.94 for *S. aureus*. The range in F1 scores indicates that there is variation between clades, with the phenotypes of the genomes in some clades being easier to predict that than others. However, the results do not indicate extreme drops in accuracy or imbalances that would indicate that the input data are biasing the model based on clade size or S and R phenotype distributions. In other words, the ability to predict S and R phenotypes is relatively consistent over the tree.

Genes with high feature importance values are distributed over the genome

Because of the relatively high accuracy of the AMR models built from small sets of core genes, and their consistency across random samples, we wanted to see if the models were relying on genes occurring in similar chromosomal regions, or on genes with related functions. To do this, we used the ten models that were built from 100 randomly sampled non-overlapping core gene sets described in Figs 2–4. Each k-mer used in a model has an associated feature importance value that describes its contribution to the model. The feature importance for the k-mers of each gene were summed to approximate the relative contribution of each gene to each model.

We plotted each gene based on its chromosomal coordinates in a reference strain, coloring each gene by its total feature importance (Fig 8). The core genes tend to distribute evenly over the entire chromosome, but there are several regions that lack important genes. These regions will contain non-coding regions, such as the ribosomal RNA operons, and unconserved regions, such as prophage, which would not have been sampled. Overall, for any individual model, we do not observe clustering of important genes on the chromosome, and we also observe no clustering of important genes across models.

Many genes contribute to the accuracy of the models

For each set of 100 genes, the relative contribution of each k-mer, and thus each gene, to the model is not uniform. A small number of genes tend to contribute most of the predictive k-mers to each model. This is followed by a relative decline in feature importance over the set (Figure I in S1 File), and in some sets, there are genes that do not have a k-mer that is used in the model (Table H in S2 File). The ensemble machine learning methods can be "greedy" in selecting important features. This means that feature lists may not include a comprehensive set of k-mers that could have provided signal. Indeed, when we generate a new model for each species by combining the 10 genes with the lowest feature importance values from each of the original ten 100 core gene models, we observe average F1 scores that are 3–7% lower and average error rates that are 2–7% higher (Table L in S2 File). This indicates that there is some signal in the k-mers of the genes that were not originally used heavily by the machine learning algorithm. In both cases, we observe that the ribosomal protein encoding genes tend to have very low total feature importance scores, presumably due to their slower mutation rates.

Overall, the k-mers with high feature importance values appear to indicate the presence of genes with important biological properties that correlate with the presence or absence of AMR in the sampled population. The three genes with the highest total feature importance values from the ten 100 core gene models are shown for each species in Table 3. Most of these genes do not have an obvious role in AMR (Table 3, Table H in S2 File), although this is not surprising because the well annotated AMR genes were discarded in this study. Ascertaining if these genes have a role in AMR is challenging, but some clues can be gleaned from web resources and the literature. For example, we used the MycoBrowser [40] web resource to examine the top gene from each of the ten core gene models for *M*. *tuberculosis*. Nine of the top ten genes were listed as non-essential in laboratory transposon mutagenesis studies [41–43], which is surprising given their strong conservation in over 5,000 strains. One *M*. *tuberculosis* gene, OpcA (Rv1446c), was listed as essential [42, 44] and was found to be upregulated in isoniazid resistant clinical isolates in a proteomic study [45], which may imply a role in AMR. Six of the of the top ten *M*. *tuberculosis* genes also encode proteins that are associated with the cell membrane or wall [46], which could imply roles in biofilm formation, immune system avoidance, transport, or virulence. This is also true for several of the other genes listed in Table 3. For example, the *wzi* gene (annotated as 55.8kDa ORF3) in *K*. *pneumoniae* is involved in the K-

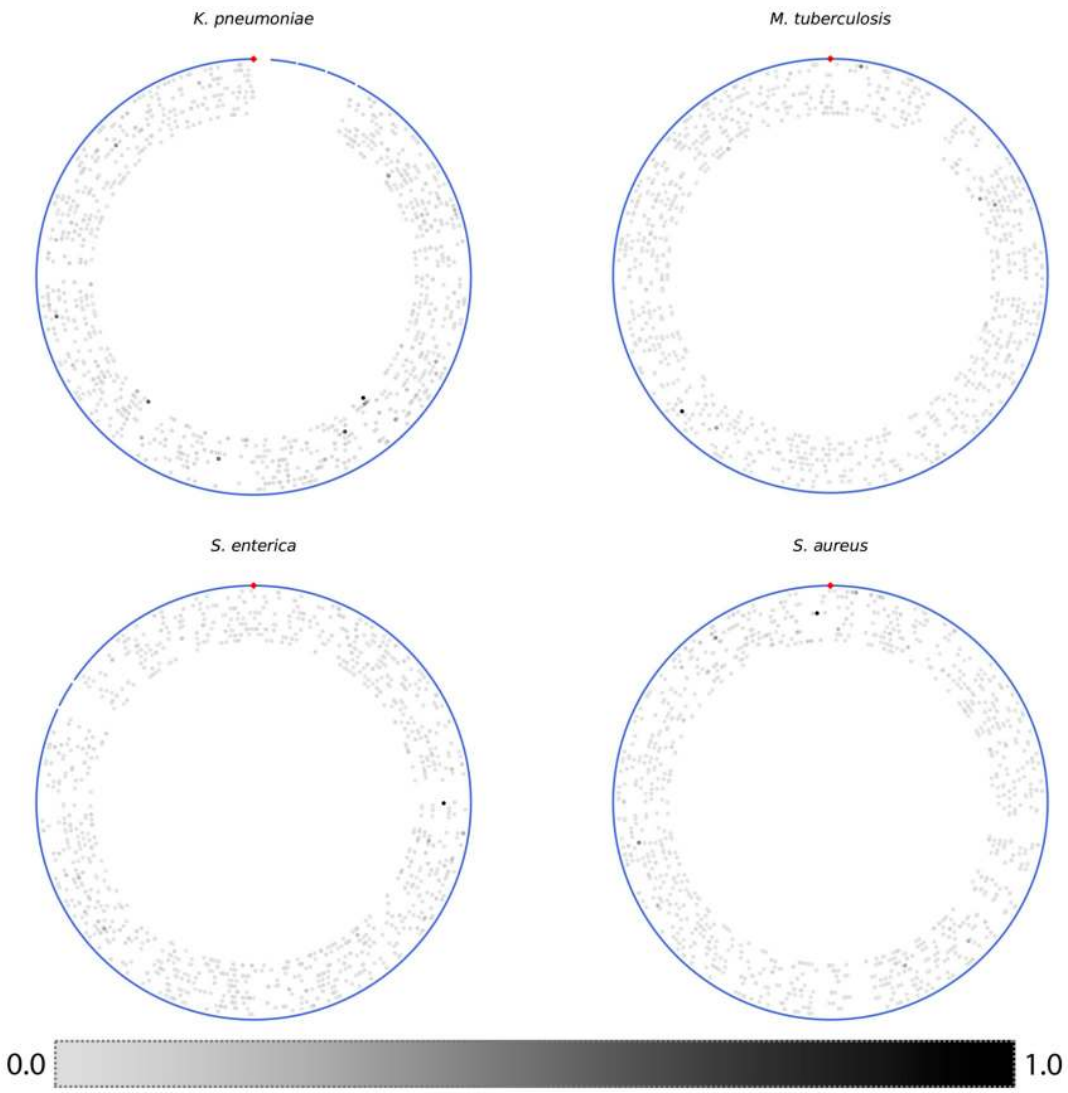

Fig 8. Plots showing the position and total feature importance of each core gene based on its coordinates in a reference genome. The contigs from the reference genome are depicted as blue lines on the outside of each plot, and the origin of replication is shown at the top with a red diamond. Each gene is depicted as a gray dot, and the genes with higher total feature importance in each model are colored in darker shades of gray. The genes of each model are shown as concentric circles from outside to inside. The reference strains used were *K. pneumoniae* HS11286, *M. tuberculosis* H37Rv, *S. enterica* serovar Typhimurium LT2, and *S. aureus* NCTC 8325.

https://doi.org/10.1371/journal.pcbi.1008319.g008

antigen capsular polysaccharide formation, and is an important virulence factor [47–50]; the staphylococcal secretory antigen *ssaA*, is involved in biofilm formation [51, 52] and is also a likely virulence factor in Staphylococcus strains [53]; and the large *Salmonella bapA* gene (annotated as T1SS secreted agglutinin RTX) is also a virulence factor that is necessary for biofilm formation [54]. Although perhaps not an AMR gene *per se*, the *S. aureus rlmH* gene (annotated as 23S rRNA (pseudouridine(1915)-N(3))-methyltransferase (EC 2.1.1.177)), contains the integration site for SCC*mec* elements, which confer methicillin resistance [34, 55, 56]. Indeed, the highest-ranking k-mer for this gene, "AAGCATATCATAAAT", corresponds with the integration site and has a feature importance score of 389, which is 80% of the total feature importance for this gene. Upon integration, the attachment site, *attB*, is split into two similar

Table 3. The top three genes with the highest total feature importance from all k-mers used in each model. Results are based on ten separate models built from 100 randomly sampled non-overlapping core gene sets.

Protein family	Average gene length	Cumulative feature importance	PATRIC annotation
K. pneumoniae			
PLF_570_00002496	1422.3	818.8	Uncharacterized 55.8 kDa protein in cps region (ORF3)
PLF_570_00001044	1419.6	669.2	Alpha,alpha-trehalose-phosphate synthase [UDP-forming] (EC 2.4.1.15)
PLF_570_00003328	1247.9	533.3	N-carbamoyl-L-amino acid hydrolase (EC 3.5.1.87)
M. tuberculosis			
PLF_1763_00001229	909.9	1157.0	OpcA, an allosteric effector of glucose-6-phosphate dehydrogenase, actinobacterial
PLF_1763_00001419	697.0	649.9	putative phosphoglycerate mutase
PLF_1763_00001681	1580.9	518.2	DNA methylase
S. enterica			
PLF_590_00006292	11345.6	1828.8	T1SS secreted agglutinin RTX
PLF_590_00001168	1593.0	616.6	Bis-ABC ATPase YbiT
PLF_590_00001455	906.0	571.3	GTP-binding protein Era
S. aureus			
PLF_1279_00000411	477.3	484.8	23S rRNA (pseudouridine(1915)-N(3))-methyltransferase (EC 2.1.1.177)
PLF_1279_00002023	687.0	250.0	Phage-encoded chromosome degrading nuclease YokF
PLF_1279_00001560	899.8	232.0	Secretory antigen precursor SsaA

https://doi.org/10.1371/journal.pcbi.1008319.t003

sequences, *attL* and *attR* [55], and this k-mer is picking up this key difference between susceptible and resistant strains.

Discussion

In this work, we have shown that AMR phenotypes can be predicted from sets of core genes that are not annotated to be involved in AMR. These gene sets can be quite small, less than 100 genes, and still have predictive power. Building models from randomly sampled non-overlapping sets of 100 core genes results in model accuracies that are quite stable with little variation between samples. This result is not influenced by the choice of features (nucleotide k-mers or alignments) or machine learning algorithms (XGB or RF), and the high accuracies do not appear to be the result of overfitting, memorization, strain-specific SNPs, or imbalances in sampling or phylogeny. This effect is seen in the Gram-negative *K. pneumoniae* and *S. enterica* genomes, and the Gram-positive *M. tuberculosis* and *S. aureus* genomes, indicating that this is likely widespread. These small core gene models also have predictive power in the cases where the primary AMR mechanism results from SNPs or horizontal gene transfer. These results are consistent with two previous studies that have used data from whole genome sequences to show that strain similarity can be used to predict AMR [36, 37]. This study extends this finding to show that accurate predictions can be made without the accessory genes that vary between strains, and that signal can be found in small arbitrarily chosen subsets of core genes lacking a direct annotated role in AMR.

Although the small core gene models described in this study do not meet the accuracy or error rate requirements of clinical diagnostic tools [57], the results have implications for the development of bioinformatic strategies aimed at tracking and understanding AMR. The models demonstrate that the prediction of AMR phenotypes from incomplete genome sequence data is possible, even when the AMR genes are missing. This means that it may be possible to develop this into a strategy for making AMR phenotype predictions for contigs that are binned from a metagenomic assembly, but do not represent a complete chromosome, or lack the corresponding plasmid contigs that carry important AMR genes. This assertion comes with the

obvious caveat that using more sequence data in the training set results in better models with higher accuracies and lower error rates. In order to demonstrate the predictive power of these small core gene models, we specifically chose to avoid using well-known AMR genes, and we focused on smaller gene sets. However, it is possible to envision a variety of approaches that could improve the accuracy of the models or result in more rapid AMR prediction. For instance, accuracies might be improved by training on larger core gene sets, training on select chromosomal regions, or by finding bespoke conserved gene sets that optimize the predictive power. Predictions could be made more rapid by employing read mapping strategies against specific gene sets to avoid *de novo* assembly. It is also possible to envision a strategy of nesting predictions from core genes within well-known bioinformatic typing schemes that are already in use. For instance, 74% of the genes used for modeling in this study and shown in Table H in S2 File correspond with genes used in the core genome cgMLST typing schemes (www.cgmlst.org) [58].

One consideration when interpreting the results of this study or any other study that has published a ML-based AMR prediction model is that the model accuracy statistics are scoped to the diversity of the training set. In other words, the genomes used in the study define the decision space for the ML algorithm. In this study, the data sets were large (at the time of writing), but the data for each of the four species came from a small number of studies that sequenced a large number of isolates. These studies would therefore be expected to carry sampling biases based on the context and location of sampling, and the selection criteria from the original study. If a model reported in this study were applied to a new genome, and the genome was encompassed by the diversity of the training set, then the resulting accuracy should fall within the reported bounds. If a prediction is desired for a more diverse genome occurring outside of the reported diversity of the training set, then it would be preferable to retrain a new model by including members from this new lineage. For most organisms and antibiotics, we expect sampling to improve over time, so iteratively retraining the models as the data collection improves should lead to subsequent improvements in the generalizability, and perhaps accuracy, of the core gene AMR models reported in this study.

Although this study was not designed to be an exhaustive search for uncharacterized AMR genes, it does highlight the value of the large publicly-available genomic data sets that have corresponding antimicrobial susceptibility test metadata, and the power of using artificial intelligence methods for identifying important information within them. In particular, the analysis of important features in the core gene models revealed at least two genes, *rlmH* and *optA*, with a known or possible role in AMR. Several of the other genes that were identified as contributing important k-mers were characterized virulence factors including *wzi*, *bapA*, and *ssaA*. It is difficult to know if the genes encoding these virulence factors have a direct biological role in AMR, or if certain alleles of the gene simply correlate with resistant or susceptible genomes. In other words, the resistant genomes might also be expected to be the more virulent genomes and vice versa [59].

Another consideration is the potential of the core gene models for identifying compensatory or epistatic changes throughout the genome. These mutations occur in non-AMR genes in order to accommodate the potentially reduced fitness cost of maintaining the primary AMR conferring genes or SNPs [60–63]. The results of this study, and previous artificial intelligence studies focusing on AMR [25, 27], suggest that these changes could be widespread throughout the genome. Indeed, although each model had a few genes with very high total feature importance values, most genes were contributing k-mers to the decision trees for each model. In general, AMR-related epistatic changes remain poorly understood, and are perhaps combinatoric, but an incisive use of artificial intelligence methods may help to establish a more mechanistic understanding of this phenomenon.

Materials and methods

Data acquisition

Whole genome sequence data and paired laboratory-derived antimicrobial susceptibility test data were downloaded from the PATRIC FTP server (ftp://ftp.patricbrc.org/RELEASE_NOTES/PATRIC_genomes_AMR.txt) on or after December 1, 2018 [32, 64]. We chose to analyze *Klebsiella pneumoniae*, *Mycobacterium tuberculosis*, *Salmonella enterica*, and *Staphylococcus aureus*, which had the largest collections of genomes paired with antimicrobial susceptibility test data in this collection at the time (Table 1, Tables A-F in S2 File). The corresponding genomic data were also downloaded from the PATRIC FTP server (ftp://ftp.patricbrc.org/genomes) [65]. The PATRIC command line interface was used for all other annotation and metadata acquisition tasks [66]. In the case of *M*. *tuberculosis*, we also used data from a study that had not yet been integrated into the PATRIC collection [67]. The reads for these genomes were downloaded from the European Nucleotide Archive [68], assembled using the full spades pipeline at PATRIC [66, 69, 70], and annotated using the PATRIC annotation service, which is a variant of RAST [29].

AMR phenotype data in the PATRIC collection usually exists in two forms, the first being a laboratory derived value such as a minimum inhibitory concentration (MIC), and the second being the susceptible, intermediate, or resistant determinations that were made by the authors at the time of the study. When only the susceptible or resistant (SR) determination was available for a genome, we did not mix data based on breakpoint values from the Clinical and Laboratory Standards Institute (CLSI) and the European Committee on Antimicrobial Susceptibility Testing (EUCAST) because their recommended breakpoint values can differ [57, 71]. When AMR phenotypes were published as MICs, they were converted to SR determinations using either CLSI or EUCAST guidelines to create the largest possible set of genomes with SR phenotypes [57, 71]. The use of CLSI or EUCAST breakpoints for a given dataset was determined on the total size of the dataset, with the larger dataset being chosen. Classification was not performed on intermediate phenotypes because they are underrepresented. All SR data used in this study are shown in Tables A-D in S2 File, and MICs are shown in Tables D and E in S2 File.

Core conserved gene sets

In order to work with clean subsets of genomes, we chose to base analyses on the protein-encoding genes that are shared among members of the same species. We used the "PATtyFam" collection, which is a set of protein families that cover the ~230,000 publicly available genomes in the PATRIC database [30]. Protein similarity for building these families is based on the RAST signature k-mer collection [29], and all proteins must have the same annotation in order to be members of the same family. Specifically, we used the genus-specific PATRIC local families (PLFs). All genomes were reannotated so that they had the same set of protein family calls based on the May 31, 2018 protein family build, using the PATRIC annotation server [29]. All families with a PATRIC annotation associated with AMR were removed from consideration in this study [64].

We used two criteria when defining core gene sets. First, for each family, the average nucleotide length was computed for the corresponding genes. Any family member that had a total nucleotide length that was less than half of the average length, or that was 50% longer than the average length, was excluded. This helped to eliminate duplicate genes, partial genes, and mixtures of genes encoding single and multi-subunit proteins. Next, we excluded any family whose members represented less than 99% of the genomes of the set. This criterion was relaxed

to 75% to build sufficient core family sets for *S. aureus*. After the core gene sets were computed, sets of 25, 50, 100, 250, and 500 core genes were randomly selected for each model.

Model generation

K-mer-based models were generated using XGBoost (XGB) version 0.81 [31] to classify susceptible and resistant phenotypes, or to predict MICs following protocols described previously [21, 22]. Briefly, the sequences of core genes or whole assembled genomes were converted into nucleotide k-mers, with a step size of one nucleotide, and lexicographically sorted and counted using the k-mer counting program KMC [72]. Partial k-mers at the ends of sequences, and sequences containing ambiguous nucleotides were not considered. 15-mers were used to compute models for core genes, and 10-mers were used to build models from assembled genomes to ensure that all models would fit in memory [22]. For the *S. enterica* models, a diverse set of 1999 genomes was chosen for analysis out of an original set of 5278 genomes as described previously [22], so that models could be computed efficiently.

Matrix files were constructed by merging the k-mer counts and the one-hot encoded antibiotic and phenotype. That is, each row of the matrix contains all of the k-mer counts for a given genome's core genes, the encoded antibiotic, and the phenotype for a single genome-antibiotic pair. Because 10-mers are more likely to be redundant in a genome than 15-mers, we use k-mer counts when building 10-mer based models and presence versus absence of a given k-mer when building 15-mer-based models.

We have shown previously that in this context, the XGB model parameter that has the greatest effect on model accuracy is tree depth [21]. This parameter was varied to tune the models. Unless otherwise stated, core gene models reported used a depth of sixteen, and whole genome models were tuned to a depth of four as previously described [21, 22]. The learning rate was set to 0.0625, and column and row subsampling was set to 1.0 as described previously [21, 22]. The number of rounds of boosting was limited to 1000. Since susceptible and resistant classes are not balanced, we weighted the counts for each class using the formula, $W = 1 - \frac{N}{T}$, where N is count of a given class for an antibiotic, and T is the total counts for all classes for the given antibiotic. This weighting was not used for the MIC models or tree-based analyses described below.

Models were evaluated by dividing the set of genomes for each species into non-overlapping training (80%), testing (10%), and validation (10%) sets. Unless otherwise stated, statistics including the accuracy and macro-averaged F1 scores are shown for the first five folds of a 10-fold cross validation with 95% confidence intervals (CI). This was done to reduce the large compute time over the many sample replicates and parameter optimizations depicted in this study. To prevent overfitting, the accuracy of the model on the held-out validation set was monitored and training was stopped if the validation set accuracy did not improve after 25 rounds of boosting. Models were assessed by computing the accuracies, $Acc = \frac{C}{T}$, where C is the correctly predicted samples and T is the total samples tested, or the macro-averaged F1 score, $F1_{macro} = \frac{\sum_C F1_C}{N}$, where C is a class and N is the number of classes. The very major error rate (VME), which is the fraction of resistant test set genomes that is predicted to be susceptible, and the major error rate (ME), which is the fraction of susceptible test set genomes that are predicted to be resistant, were also used to evaluate models [73]. For MIC models, accuracy is depicted within ± 1 two-fold dilution step, which is the limit of resolution for most automated laboratory AST devices [73].

To control for unforeseen effects relating to the methodology or algorithm choice, we also built models from nucleotide alignments following a similar protocols to those that have been described previously [25, 26]. One random subset of 100 core genes per species was chosen for

the analysis. Each of the 100 core genes sets was aligned using MAFFT version 7.13 [74]. Alignment columns where ≥95% of the characters were dashes (i.e., gaps) were masked. Start positions were trimmed to the first column that had no dashes in order to eliminate variability from start site calling, rather than true nucleotide differences. Any genome missing one or more of the 100 core genes was removed. Columns with 100% conservation were eliminated because they lack discriminating information. Alignments were then concatenated and one-hot encoded so that each nucleotide, the N character, and the dash character were encoded in a 6-digit string. The one-hot encoded alignments and AST phenotypes were then merged to form a matrix. XGB and Random Forest (RF) [75] were then used to build the classifiers. For XGB, the same model parameters were used as described above for the k-mer-based core gene models. RF parameters were set to bagging of 1000 decision trees each, at a depth of 16, subsampling 75% of rows, and subsampling 75% of features. Accuracies and error rates were generated using cross validations as described above. To assess the impact that strain-specific SNPs had on these models, alignment columns with varying fractions of conservation were incrementally removed, and the models were recomputed with accuracy statistics. The scripts and links to data files from this study are available on our GitHub page: https://github.com/jimdavis1/Core-Gene-AMR-Models.

Tree-based analyses

Using too many closely related strains, or having closely related strains with skewed S or R phenotype distributions, could result in biased models. To assess this, we generated phylogenetic trees for each of the four species and computed models that were weighted based on the total number of tips in a subtree, and the distributions of S and R classes within each subtree. By defining subtrees at different tree depths, potential biases can be assessed in more closely or distantly related sets of organisms.

Trees were generated for each species by randomly selecting a set of 100 leftover core genes that were not used for computing the core gene models. Concatenated nucleotide alignments were generated as described above, and computed with FastTree version 2.1.7 using the nucleotide and generalized time reversible model options [76]. The trees were then divided into all possible subtrees in order to define clades of related genomes at various sizes. Each subtree containing more than one tip was midpoint rooted, and the tree distance was used to define the size of the subtree. Trees were manipulated using tools from the PATRIC command line interface application version 1.025 [66]. At a small distance, subtrees will be comprised of nearly identical tips. As the distance increases, subtree membership becomes more inclusive, with more diverse sequences being represented within a subtree, until ultimately the distance becomes so large that all the tips in the tree are represented by a single subtree. In this way, by defining the most inclusive subtrees at increasing tree distances, we can measure the effect of imbalances in diversity and SR distribution on the models at varying levels of phylogenetic resolution. A submodule describing the code and analysis can be found in the GitHub repository for this project: https://github.com/jimdavis1/Core-Gene-AMR-Models. Trees were rendered using the iTOL web server [77].

K-mer based XGB models from core genes were created as described above. To weight genomes based on subtree size, the number of genomes in the subtree was divided by the total number of genomes in the model, and the corresponding fraction was assigned to each genome as a weight, using the equation, $W_1 = 1 - \frac{S}{T}$, where S is the size of the subtree a genome belonged to and T is the total number of genomes analyzed. Likewise, the fraction of susceptible and resistant genomes was computed for each subtree, and a value of 1 minus the appropriate fraction was assigned to each genome as a weight, using the equation, $W_2 = 1 - \frac{C}{T}$,

where C is the number of genomes of the same class within the subtree as the genome being weighted and T is the total number of genomes within a given subtree. However, if $C = T$, then the weight assigned was 1 (no weighting) to avoid a 0-weight sample. In this way, frequently occurring phenotypes within a clade had low weights, and rarely occurring phenotypes had high weights. Finally, to assess both features together, the subtree size weight was multiplied by the SR weight.

Analysis of top features

Ten k-mer-based XGB models were built from 100 randomly-selected, non-overlapping core genes as described above. The k-mers that were used by each model, and their associated feature importance values were obtained. For each gene in the model, the corresponding feature importance of each corresponding k-mer was summed to generate a total feature importance for the gene. The location of each gene was plotted using its coordinates in a high-quality reference genome. The choice of reference genomes was based on a curated list of high-quality genomes maintained by NCBI: ftp://ftp.ncbi.nlm.nih.gov/genomes/GENOME_REPORTS/prok_reference_genomes.txt.

Supporting information

S1 File. Supplemental Figures A-I.
(PDF)

S2 File. Supplemental Tables A-L.
(XLSX)

Acknowledgments

We thank the members of the BV-BRC PATRIC, iSENTRY, NARMS and Houston Methodist teams, as well as Dion Antonopoulos, Philippe Noirot, and Sarah Owens for their helpful comments. We thank NARMS and Houston Methodist for the *Salmonella* and *Klebsiella* data sets, respectively. We also thank Emily Dietrich for her careful editing.

Author Contributions

Conceptualization: Marcus Nguyen, James J. Davis.

Data curation: Robert Olson, Maulik Shukla, Margo VanOeffelen, James J. Davis.

Formal analysis: Marcus Nguyen, James J. Davis.

Funding acquisition: James J. Davis.

Investigation: Marcus Nguyen, James J. Davis.

Methodology: Marcus Nguyen.

Project administration: James J. Davis.

Resources: Robert Olson, James J. Davis.

Software: Marcus Nguyen.

Supervision: Robert Olson, James J. Davis.

Validation: Marcus Nguyen.

Visualization: Marcus Nguyen.

Writing – original draft: James J. Davis.

Writing – review & editing: Marcus Nguyen, Robert Olson, Maulik Shukla, Margo VanOeffelen, James J. Davis.

References

1. Centers for Disease Control and Prevention. Achievements in Public Health, 1900–1999. Morbidity and Mortality Weekly Report. 1999; 48(29):621–9.
2. Heron MP. Deaths: Leading causes for 2016. National Vital Statistics Reports. 2018; 67(6). PMID: 30248017
3. Adedeji W. The treasure called antibiotics. Annals of Ibadan postgraduate medicine. 2016; 14(2):56. PMID: 28337088
4. Reller LB, Weinstein M, Jorgensen JH, Ferraro MJ. Antimicrobial susceptibility testing: a review of general principles and contemporary practices. Clinical infectious diseases. 2009; 49(11):1749–55. https://doi.org/10.1086/647952 PMID: 19857164
5. Opota O, Croxatto A, Prod'hom G, Greub G. Blood culture-based diagnosis of bacteraemia: state of the art. Clinical Microbiology and Infection. 2015; 21(4):313–22. https://doi.org/10.1016/j.cmi.2015.01.003 PMID: 25753137
6. Llor C, Bjerrum L. Antimicrobial resistance: risk associated with antibiotic overuse and initiatives to reduce the problem. Therapeutic advances in drug safety. 2014; 5(6):229–41. https://doi.org/10.1177/2042098614554919 PMID: 25436105
7. Weinstein RA. Controlling antimicrobial resistance in hospitals: infection control and use of antibiotics. Emerging infectious diseases. 2001; 7(2):188. PMID: 11294703
8. Palmer H, Palavecino E, Johnson J, Ohl C, Williamson J. Clinical and microbiological implications of time-to-positivity of blood cultures in patients with Gram-negative bacilli bacteremia. European journal of clinical microbiology & infectious diseases. 2013; 32(7):955–9.
9. Kumar A, Roberts D, Wood KE, Light B, Parrillo JE, Sharma S, et al. Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. Critical care medicine. 2006; 34(6):1589–96. https://doi.org/10.1097/01.CCM.0000217961.75225.E9 PMID: 16625125
10. Boolchandani M, D'Souza AW, Dantas G. Sequencing-based methods and resources to study antimicrobial resistance. Nature Reviews Genetics. 2019: 1.
11. Jeukens J, Freschi L, Kukavica-Ibrulj I, Emond-Rheault JG, Tucker NP, Levesque RC. Genomics of antibiotic-resistance prediction in Pseudomonas aeruginosa. Annals of the New York Academy of Sciences. 2019; 1435(1):5–17. https://doi.org/10.1111/nyas.13358 PMID: 28574575
12. Lo SW, Kumar N, Wheeler NE. Breaking the code of antibiotic resistance. Nature Publishing Group; 2018.
13. Alcock BP, Raphenya AR, Lau TTY, Tsang KK, Bouchard M, Edalatmand A, et al. CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. Nucleic Acids Research. 2019. https://doi.org/10.1093/nar/gkz935 PMID: 31665441
14. Feldgarden M, Brover V, Haft DH, Prasad AB, Slotta DJ, Tolstoy I, et al. Using the NCBI AMRFinder Tool to Determine Antimicrobial Resistance Genotype-Phenotype Correlations Within a Collection of NARMS Isolates. bioRxiv. 2019:550707. https://doi.org/10.1101/550707
15. Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, et al. Identification of acquired antimicrobial resistance genes. Journal of Antimicrobial Chemotherapy. 2012; 67(11):2640–4. https://doi.org/10.1093/jac/dks261 PMID: 22782487
16. McDermott PF, Tyson GH, Kabera C, Chen Y, Li C, Folster JP, et al. Whole-genome sequencing for detecting antimicrobial resistance in nontyphoidal *Salmonella*. Antimicrobial agents and chemotherapy. 2016; 60(9):5515–20. https://doi.org/10.1128/AAC.01030-16 PMID: 27381390
17. Niehaus KE, Walker TM, Crook DW, Peto TE, Clifton DA, editors. Machine learning for the prediction of antibacterial susceptibility in Mycobacterium tuberculosis. 2014 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI); 2014: IEEE.
18. Stoesser N, Batty E, Eyre D, Morgan M, Wyllie D, Del Ojo Elias C, et al. Predicting antimicrobial susceptibilities for *Escherichia coli* and *Klebsiella pneumoniae* isolates using whole genomic sequence data. Journal of Antimicrobial Chemotherapy. 2013; 68(10):2234–44. https://doi.org/10.1093/jac/dkt180 PMID: 23722448
19. Pesesky MW, Hussain T, Wallace M, Patel S, Andleeb S, Burnham C-AD, et al. Evaluation of machine learning and rules-based approaches for predicting antimicrobial resistance profiles in Gram-negative

Bacilli from whole genome sequence data. Frontiers in microbiology. 2016; 7:1887. https://doi.org/10.3389/fmicb.2016.01887 PMID: 27965630

20. Eyre DW, De Silva D, Cole K, Peters J, Cole MJ, Grad YH, et al. WGS to predict antibiotic MICs for *Neisseria gonorrhoeae*. Journal of Antimicrobial Chemotherapy. 2017; 72(7):1937–47. https://doi.org/10.1093/jac/dkx067 PMID: 28333355

21. Nguyen M, Brettin T, Long SW, Musser JM, Olsen RJ, Olson R, et al. Developing an in silico minimum inhibitory concentration panel test for *Klebsiella pneumoniae*. Scientific reports. 2018; 8(1):421. https://doi.org/10.1038/s41598-017-18972-w PMID: 29323230

22. Nguyen M, Long SW, McDermott PF, Olsen RJ, Olson R, Stevens RL, et al. Using machine learning to predict antimicrobial MICs and associated genomic features for nontyphoidal *Salmonella*. Journal of Clinical Microbiology. 2019; 57(2):e01260–18. https://doi.org/10.1128/JCM.01260-18 PMID: 30333126

23. Drouin A, Giguère S, Déraspe M, Marchand M, Tyers M, Loo VG, et al. Predictive computational phenotyping and biomarker discovery using reference-free genome comparisons. BMC genomics. 2016; 17(1):754. https://doi.org/10.1186/s12864-016-2889-6 PMID: 27671088

24. Hicks AL, Wheeler N, Sánchez-Busó L, Rakeman JL, Harris SR, Grad YH. Evaluation of parameters affecting performance and reliability of machine learning-based antibiotic susceptibility testing from whole genome sequencing data. PLOS Computational Biology. 2019; 15(9):e1007349. https://doi.org/10.1371/journal.pcbi.1007349 PMID: 31479500

25. Hyun JC, Kavvas ES, Monk JM, Palsson BO. Machine learning with random subspace ensembles identifies antimicrobial resistance determinants from pan-genomes of three pathogens. PLoS computational biology. 2020; 16(3):e1007608. https://doi.org/10.1371/journal.pcbi.1007608 PMID: 32119670

26. Aytan-Aktug D, Clausen PTLC, Bortolaia V, Aarestrup FM, Lund O. Prediction of Acquired Antimicrobial Resistance for Multiple Bacterial Species Using Neural Networks. Msystems. 2020; 5(1).

27. Kavvas ES, Catoiu E, Mih N, Yurkovich JT, Seif Y, Dillon N, et al. Machine learning and structural analysis of *Mycobacterium tuberculosis* pan-genome identifies genetic signatures of antibiotic resistance. Nature communications. 2018; 9(1):4306. https://doi.org/10.1038/s41467-018-06634-y PMID: 30333483

28. Forbes JD, Knox NC, Ronholm J, Pagotto F, Reimer A. Metagenomics: The Next Culture-Independent Game Changer. Frontiers in Microbiology. 2017; 8(1069). https://doi.org/10.3389/fmicb.2017.01069 PMID: 28725217

29. Brettin T, Davis JJ, Disz T, Edwards RA, Gerdes S, Olsen GJ, et al. RASTtk: a modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. Scientific reports. 2015; 5:8365. https://doi.org/10.1038/srep08365 PMID: 25666585

30. Davis JJ, Gerdes S, Olsen GJ, Olson R, Pusch GD, Shukla M, et al. PATtyFams: protein families for the microbial genomes in the PATRIC database. Frontiers in microbiology. 2016; 7:118. https://doi.org/10.3389/fmicb.2016.00118 PMID: 26903996

31. Chen T, Guestrin C, editors. XGBoost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining; 2016: ACM.

32. Davis JJ, Boisvert S, Brettin T, Kenyon RW, Mao C, Olson R, et al. Antimicrobial resistance prediction in PATRIC and RAST. Scientific reports. 2016; 6:27930. https://doi.org/10.1038/srep27930 PMID: 27297683

33. Yoshida H, Kojima T, Yamagishi J-i, Nakamura S. Quinolone-resistant mutations of the gyrA gene of Escherichia coli. Molecular and General Genetics MGG. 1988; 211(1):1–7. https://doi.org/10.1007/BF00338386 PMID: 2830458

34. Katayama Y, Ito T, Hiramatsu K. A new class of genetic element, staphylococcus cassette chromosome mec, encodes methicillin resistance in *Staphylococcus aureus*. Antimicrobial agents and chemotherapy. 2000; 44(6):1549–55. https://doi.org/10.1128/aac.44.6.1549-1555.2000 PMID: 10817707

35. Miriagou V, Tassios P, Legakis N, Tzouvelekis L. Expanded-spectrum cephalosporin resistance in non-typhoid *Salmonella*. International journal of antimicrobial agents. 2004; 23(6):547–55. https://doi.org/10.1016/j.ijantimicag.2004.03.006 PMID: 15194124

36. 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. https://doi.org/10.1371/journal.pcbi.1006258 PMID: 30550564

37. Břinda K, Callendrello A, Ma KC, MacFadden DR, Charalampous T, Lee RS, et al. Rapid inference of antibiotic resistance and susceptibility by genomic neighbour typing. Nature microbiology. 2020; 5(3):455–64. https://doi.org/10.1038/s41564-019-0656-6 PMID: 32042129

38. Khaledi A, Weimann A, Schniederjans M, Asgari E, Kuo TH, Oliver A, et al. Predicting antimicrobial resistance in *Pseudomonas aeruginosa* with machine learning-enabled molecular diagnostics. EMBO molecular medicine. 2020; 12(3):e10264. https://doi.org/10.15252/emmm.201910264 PMID: 32048461

39. Lees JA, Mai TT, Galardini M, Wheeler NE, Corander J. Improved inference and prediction of bacterial genotype-phenotype associations using pangenome-spanning regressions. BioRxiv. 2019:852426.

40. Kapopoulou A, Lew JM, Cole ST. The MycoBrowser portal: a comprehensive and manually annotated resource for mycobacterial genomes. Tuberculosis. 2011; 91(1):8–13. https://doi.org/10.1016/j.tube.2010.09.006 PMID: 20980200

41. Sassetti CM, Boyd DH, Rubin EJ. Genes required for mycobacterial growth defined by high density mutagenesis. Molecular microbiology. 2003; 48(1):77–84. https://doi.org/10.1046/j.1365-2958.2003.03425.x PMID: 12657046

42. DeJesus MA, Gerrick ER, Xu W, Park SW, Long JE, Boutte CC, et al. Comprehensive essentiality analysis of the *Mycobacterium tuberculosis* genome via saturating transposon mutagenesis. MBio. 2017; 8 (1):e02133–16. https://doi.org/10.1128/mBio.02133-16 PMID: 28096490

43. Griffin JE, Gawronski JD, DeJesus MA, Ioerger TR, Akerley BJ, Sassetti CM. High-resolution phenotypic profiling defines genes essential for mycobacterial growth and cholesterol catabolism. PLoS pathogens. 2011; 7(9).

44. Rengarajan J, Bloom BR, Rubin EJ. Genome-wide requirements for *Mycobacterium tuberculosis* adaptation and survival in macrophages. Proceedings of the National Academy of Sciences. 2005; 102 (23):8327–32.

45. Jiang XIN, Zhang W, Gao F, Huang Y, Lv C, Wang H. Comparison of the proteome of isoniazid-resistant and-susceptible strains of *Mycobacterium tuberculosis*. Microbial drug resistance. 2006; 12(4):231–8. https://doi.org/10.1089/mdr.2006.12.231 PMID: 17227207

46. Kelkar DS, Kumar D, Kumar P, Balakrishnan L, Muthusamy B, Yadav AK, et al. Proteogenomic analysis of Mycobacterium tuberculosis by high resolution mass spectrometry. Molecular & cellular proteomics. 2011; 10(12).

47. Rahn A, Beis K, Naismith JH, Whitfield C. A novel outer membrane protein, Wzi, is involved in surface assembly of the *Escherichia coli* K30 group 1 capsule. Journal of bacteriology. 2003; 185(19):5882–90. https://doi.org/10.1128/jb.185.19.5882-5890.2003 PMID: 13129961

48. Brisse S, Passet V, Haugaard AB, Babosan A, Kassis-Chikhani N, Struve C, et al. wzi gene sequencing, a rapid method for determination of capsular type for *Klebsiella* strains. Journal of clinical microbiology. 2013; 51(12):4073–8. https://doi.org/10.1128/JCM.01924-13 PMID: 24088853

49. Gomez-Simmonds A, Uhlemann A-C. Clinical implications of genomic adaptation and evolution of carbapenem-resistant Klebsiella pneumoniae. The Journal of infectious diseases. 2017; 215(suppl_1): S18–S27. https://doi.org/10.1093/infdis/jiw378 PMID: 28375514

50. Liu Y, Liu P-p, Wang L-h, Wei D-d, Wan L-G, Zhang W. Capsular polysaccharide types and virulence-related traits of epidemic KPC-producing *Klebsiella pneumoniae* isolates in a Chinese university hospital. Microbial Drug Resistance. 2017; 23(7):901–7. https://doi.org/10.1089/mdr.2016.0222 PMID: 28437231

51. Lang S, Livesley MA, Lambert PA, Littler WA, Elliott TS. Identification of a novel antigen from Staphylococcus epidermidis. FEMS Immunology & Medical Microbiology. 2000; 29(3):213–20.

52. Aguila-Arcos S, Ding S, Aloria K, Arizmendi J, Fearnley I, Walker J, et al. A commensal strain of *Staphylococcus epidermidis* overexpresses membrane proteins associated with pathogenesis when grown in biofilms. The Journal of membrane biology. 2015; 248(3):431–42. https://doi.org/10.1007/s00232-015-9801-1 PMID: 25837994

53. Resch A, Rosenstein R, Nerz C, Götz F. Differential gene expression profiling of Staphylococcus aureus cultivated under biofilm and planktonic conditions. Appl Environ Microbiol. 2005; 71(5):2663–76. https://doi.org/10.1128/AEM.71.5.2663-2676.2005 PMID: 15870358

54. Latasa C, Roux A, Toledo-Arana A, Ghigo JM, Gamazo C, Penadés JR, et al. BapA, a large secreted protein required for biofilm formation and host colonization of *Salmonella enterica* serovar Enteritidis. Molecular microbiology. 2005; 58(5):1322–39. https://doi.org/10.1111/j.1365-2958.2005.04907.x PMID: 16313619

55. Misiura A, Pigli YZ, Boyle-Vavra S, Daum RS, Boocock MR, Rice PA. Roles of two large serine recombinases in mobilizing the methicillin-resistance cassette SCCmec. Molecular microbiology. 2013; 88 (6):1218–29. https://doi.org/10.1111/mmi.12253 PMID: 23651464

56. Noto MJ, Kreiswirth BN, Monk AB, Archer GL. Gene acquisition at the insertion site for SCCmec, the genomic island conferring methicillin resistance in *Staphylococcus aureus*. Journal of bacteriology. 2008; 190(4):1276–83. https://doi.org/10.1128/JB.01128-07 PMID: 18083809

57. European Committee on Antimicrobial Susceptibility Testing Breakpoint Tables for Interpretation of MICs and Zone Diameters. 2019;9. http://www.eucast.org.

58. Jünemann S, Sedlazeck FJ, Prior K, Albersmeier A, John U, Kalinowski J, et al. Updating benchtop sequencing performance comparison. Nature biotechnology. 2013; 31(4):294–6. https://doi.org/10.1038/nbt.2522 PMID: 23563421

59. Schroeder M, Brooks BD, Brooks AE. The complex relationship between virulence and antibiotic resistance. Genes. 2017; 8(1):39.

60. Mira PM, Meza JC, Nandipati A, Barlow M. Adaptive landscapes of resistance genes change as antibiotic concentrations change. Molecular biology and evolution. 2015; 32(10):2707–15. https://doi.org/10.1093/molbev/msv146 PMID: 26113371

61. Trindade S, Sousa A, Xavier KB, Dionisio F, Ferreira MG, Gordo I. Positive epistasis drives the acquisition of multidrug resistance. PLoS genetics. 2009;5(7).

62. Lindgren PK, Marcusson LL, Sandvang D, Frimodt-Møller N, Hughes D. Biological cost of single and multiple norfloxacin resistance mutations in Escherichia coli implicated in urinary tract infections. Antimicrobial agents and chemotherapy. 2005; 49(6):2343–51. https://doi.org/10.1128/AAC.49.6.2343-2351.2005 PMID: 15917531

63. Marcusson LL, Frimodt-Møller N, Hughes D. Interplay in the selection of fluoroquinolone resistance and bacterial fitness. PLoS pathogens. 2009; 5(8).

64. Antonopoulos DA, Assaf R, Aziz RK, Brettin T, Bun C, Conrad N, et al. PATRIC as a unique resource for studying antimicrobial resistance. Briefings in bioinformatics. 2017.

65. Wattam AR, Abraham D, Dalay O, Disz TL, Driscoll T, Gabbard JL, et al. PATRIC, the bacterial bioinformatics database and analysis resource. Nucleic acids research. 2013; 42(D1):D581–D91.

66. Davis JJ, Wattam AR, Aziz RK, Brettin T, Butler R, Butler RM, et al. The PATRIC Bioinformatics Resource Center: expanding data and analysis capabilities. Nucleic Acids Research. 2019. https://doi.org/10.1093/nar/gkz943 PMID: 31667520

67. Coll F, Phelan J, Hill-Cawthorne GA, Nair MB, Mallard K, Ali S, et al. Genome-wide analysis of multi- and extensively drug-resistant *Mycobacterium tuberculosis*. Nature genetics. 2018; 50(2):307. https://doi.org/10.1038/s41588-017-0029-0 PMID: 29358649

68. Amid C, Alako BTF, Balavenkataraman Kadhirvelu V, Burdett T, Burgin J, Fan J, et al. The European Nucleotide Archive in 2019. Nucleic acids research. 2019. https://doi.org/10.1093/nar/gkz1063 PMID: 31722421.

69. Nikolenko SI, Korobeynikov AI, Alekseyev MA, editors. BayesHammer: Bayesian clustering for error correction in single-cell sequencing. BMC genomics; 2013: BioMed Central.

70. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. Journal of computational biology. 2012; 19(5):455–77. https://doi.org/10.1089/cmb.2012.0021 PMID: 22506599

71. Weinstein MP, et al. Performance Standards for Antimicrobial Susceptibility Testing 2019; 29.

72. Deorowicz S, Kokot M, Grabowski S, Debudaj-Grabysz A. KMC 2: fast and resource-frugal k-mer counting. Bioinformatics. 2015; 31(10):1569–76. https://doi.org/10.1093/bioinformatics/btv022 PMID: 25609798

73. US Food and Drug Administration (FDA). Class II Special Controls Guidance Document: Antimicrobial Susceptibility Test (AST) Systems. Rockville, MD: US FDA. 2009.

74. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular biology and evolution. 2013; 30(4):772–80. https://doi.org/10.1093/molbev/mst010 PMID: 23329690

75. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011; 12:2825–30.

76. Price MN, Dehal PS, Arkin AP. FastTree 2–approximately maximum-likelihood trees for large alignments. PloS one. 2010; 5(3):e9490. https://doi.org/10.1371/journal.pone.0009490 PMID: 20224823

77. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic acids research. 2019; 47(W1):W256–W9. https://doi.org/10.1093/nar/gkz239 PMID: 30931475