A Genome-based Model to Predict the Virulence of *Pseudomonas aeruginosa* Isolates

Nathan B. Pincus¹, Egon A. Ozer², Jonathan P. Allen^{1#}, Marcus Nguyen^{3,4}, James J. Davis^{3,4,5},
Deborah R. Winter⁶, Chih-Hsien Chuang⁷, Cheng-Hsun Chiu⁸, Laura Zamorano⁹, Antonio
Oliver⁹, Alan R. Hauser^{1,2*}

¹Department of Microbiology-Immunology, Northwestern University Feinberg School of
Medicine, Chicago, Illinois, USA

²Department of Medicine, Division of Infectious Diseases, Northwestern University Feinberg
School of Medicine, Chicago, Illinois USA

³University of Chicago Consortium for Advanced Science and Engineering, University of
Chicago, Chicago, Illinois USA

⁴Division of Data Science and Learning, Argonne National Laboratory, Argonne, Illinois, USA

⁵Northwestern-Argonne Institute of Science and Engineering, Evanston, Illinois, USA

⁶Department of Medicine, Division of Rheumatology, Northwestern University Feinberg School
of Medicine, Chicago, USA

⁷School of Medicine, College of Medicine, Fu-Jen Catholic University, New Taipei, Taiwan

⁸Molecular Infectious Disease Research Center, Chang Gung Memorial Hospital, Chang Gung
University, Taoyuan, Taiwan

⁹Servicio de Microbiología y Unidad de Investigación, Hospital Universitari Son Espases, Institut
d'Investigació Sanitaria Illes Balears, Palma de Mallorca, Spain

*Corresponding Author: Alan R. Hauser, ahauser@northwestern.edu

#Present Address: Department of Microbiology and Immunology, Loyola University Chicago,
Maywood, Illinois USA

ABSTRACT

Variation in the genome of *Pseudomonas aeruginosa*, an important pathogen, can have dramatic impacts on the bacterium's ability to cause disease. We therefore asked whether it was possible to predict the virulence of *P. aeruginosa* isolates based upon their genomic content. We applied a machine learning approach to a genetically and phenotypically diverse collection of 115 clinical *P. aeruginosa* isolates using genomic information and corresponding virulence phenotypes in a mouse model of bacteremia. We defined the accessory genome of these isolates through the presence or absence of accessory genomic elements (AGEs), sequences present in some strains but not others. Machine learning models trained using AGEs were predictive of virulence, with a mean nested cross-validation accuracy of 75% using the random forest algorithm. However, individual AGEs did not have a large influence on the algorithm's performance, suggesting instead that the virulence prediction derives from a diffuse genomic signature. These results were validated with an independent test set of 25 *P. aeruginosa* isolates whose virulence was predicted with 72% accuracy. Machine learning models trained using core genome single nucleotide variants and whole genome k-mers also predicted virulence. Our findings are a proof of concept for the use of bacterial genomes to predict pathogenicity in *P. aeruginosa* and highlight the potential of this approach for predicting patient outcomes.

IMPORTANCE

Pseudomonas aeruginosa is a clinically important gram-negative opportunistic pathogen. As a species, *P. aeruginosa* has a large degree of heterogeneity both through variation in sequences found throughout the species (core genome) and the presence or absence of sequences in different isolates (accessory genome). *P. aeruginosa* isolates also differ markedly in their ability to cause disease. In this study, we used machine learning to predict the virulence level of *P. aeruginosa* isolates in a mouse bacteremia model based on genomic content. We show that

51 both the accessory and core genome are predictive of virulence. This study provides a machine
52 learning framework to investigate relationships between bacterial genomes and complex
53 phenotypes such as virulence.

INTRODUCTION

Pseudomonas aeruginosa is a ubiquitous gram-negative opportunistic pathogen that can infect a variety of hosts. Its ability to cause severe acute infections in susceptible patients and chronic infections in individuals with cystic fibrosis, coupled with increasing rates of antimicrobial resistance, make it an organism of particular concern to the medical community¹⁻³. The *P. aeruginosa* species, however, is not monolithic but rather shows a large amount of genomic diversity both through polymorphisms and differences in gene content⁴⁻⁶. As routine whole genome sequencing becomes increasingly feasible, understanding how these genomic differences impact the pathogenicity of *P. aeruginosa* may allow clinicians to rapidly identify high-risk infections and researchers to select the most high-yield strains for further study.

As with other bacteria, the genome of *P. aeruginosa* can be divided into a core genome, made up of sequences common to the species, and an accessory genome, made up of sequences present in some strains but not others^{6,7}. While only 10-15% of a typical strain's genome is accessory, these sequences when combined from all strains comprise the vast majority of the *P. aeruginosa* pangenome^{4,7,8}. Variations in both the core and accessory genomes play an important role in the virulence of any given *P. aeruginosa* strain. Core genome mutations that accumulate in *P. aeruginosa* strains during chronic infection of cystic fibrosis patients lead to decreased markers of *in vitro* virulence⁹, and these strains have attenuated virulence in animal models of acute infection¹⁰. Genomic islands, major components of the accessory genome, are enriched for predicted virulence factors¹¹. Several genomic islands in *P. aeruginosa*, including those containing the type III secretion system (T3SS) effector gene *exoU*, have been shown to enhance pathogenicity in multiple infection models¹²⁻¹⁴. We recently mined the accessory genome and identified multiple novel factors important in virulence in a mouse model of bacteremia¹⁵. Conversely, a study using a *Caenorhabditis elegans* model identified several *P. aeruginosa* accessory genes whose presence reduced virulence¹⁶. Further, the presence of active CRISPR systems was associated with increased virulence¹⁶, supporting the

hypothesis that many horizontally transferred elements are genetic parasites with respect to the host bacterium¹⁷. Because of its role in both increasing and decreasing the pathogenicity of individual *P. aeruginosa* strains, the accessory genome may serve as a useful predictor of an isolate's virulence. This is not necessarily as simple as detecting individual virulence or anti-virulence factors. For example, *exoU* is a recognized virulence factor whose disruption dramatically attenuates a strain's ability to cause disease^{18,19}, but some strains naturally lacking *exoU* are more virulent than those possessing the gene¹⁵. As virulence is a complex and combinatorial phenotype, the strategy taken to study it must be appropriately robust to that complexity.

Machine learning is an increasingly important tool in bacterial genomics and has been extensively applied to the prediction of antimicrobial resistance and identification of potential resistance determinants. This has proven successful in a variety of species and using a variety of genomic features²⁰⁻²⁶. These studies have benefited from readily available whole genome sequencing and resistance data, as well as an often easily explainable phenotype. Researchers have also begun to apply machine learning techniques to predict bacterial pathogenicity. Examples include using discriminatory single nucleotide variants (SNVs) to predict *Staphylococcus aureus* *in vitro* cytotoxicity²⁷, using variation in core genome loci to predict patient mortality in specific *S. aureus* clones²⁸, and using predicted perturbations in protein coding sequences to classify *Salmonella* strains as causing either gastrointestinal or extraintestinal infections²⁹. A support vector machine approach has been used to distinguish the transcriptomes of *P. aeruginosa* in human infection compared to *in vitro* growth³⁰. However, to our knowledge there has been no study directly modeling *P. aeruginosa* pathogenicity from genomic content.

In this study, we utilized a machine learning approach to predict *P. aeruginosa* virulence in a mouse model of bloodstream infection based on genomic content. We found that there is signal within the accessory genome predictive of virulence, a finding validated using an

independent test set of isolates. This appears to be through the detection of a diffuse genetic fingerprint rather than individual virulence or anti-virulence genes. The core genome also showed predictive signal for virulence.

RESULTS

Genomic and Virulence Characterization of *P. aeruginosa* Strains

To assess the ability of the *P. aeruginosa* genome to predict a given isolate's virulence, we needed a large number of *P. aeruginosa* isolates with known whole genome sequences and *in vivo* virulence data. For this purpose, we used two previously reported collections: 98 archived isolates from adults with bacteremia at Northwestern Memorial Hospital (NMH) in Chicago, USA³¹ and 17 isolates from children with Shanghai fever, a *P. aeruginosa* infection presenting with sepsis and gastrointestinal symptoms, at Chang Gung Children's Hospital in Taiwan³² (Supplementary Table 1). Together these formed a training set of 115 isolates. We performed whole genome sequencing for each of the isolates that had not been previously sequenced. Likewise, previously reported virulence data^{15,32} were supplemented with new experiments (Supplementary Table 2) to approximate the colony forming units (CFU) of each bacterial isolate necessary to cause pre-lethal illness in 50% of mice using a bacteremia model. From these data, a modified 50% lethal dose (mLD₅₀) was estimated for each of the 115 *P. aeruginosa* isolates (Supplementary Table 3). The isolates showed a median mLD₅₀ of 6.9 log₁₀ CFU but with a wide range of pathogenicity in mice, varying by over 100-fold in the dose required to cause severe disease, as was previously reported for the NMH isolates¹⁵. For the purpose of this study, we classified those isolates with an estimated mLD₅₀ below the median value for the group as "high virulence" and the remainder as "low virulence" (Figure 1A). These results provided a large collection of *P. aeruginosa* isolates with known whole genome sequences and virulence in a mouse bacteremia model.

We performed a phylogenomic analysis to assess the diversity of the core genomes of all 115 isolates in the training set (Figure 1B). The core genome phylogenetic tree showed that the isolates are largely nonclonal and were found in both major clades of the species, which are mainly differentiable by the near-mutually exclusive presence of the T3SS effector genes *exoS* or *exoU*^{4,5}. One distinct outlier isolate from the PA7-like clade was also present in the collection⁴. The *exoU*+ clade contained a larger proportion of highly virulent isolates than the *exoS*+ clade. Although some clusters of closely related isolates shared the same virulence class, both major clades contained high- and low-virulence isolates.

We next defined the accessory genome of each of the 115 isolates in the training set. The accessory genome can be divided into accessory genomic elements (AGEs), discrete sequences found in the genomes of some isolates but not others⁷. For the purpose of this study, perfectly correlated sets of accessory sequences were grouped and considered as a single AGE. This means that noncontiguous sequences are treated as a single element if they are present and absent from the same isolates in the training set. Sets of accessory sequences totaling less than 200 bp were excluded from analysis. Using this approach, a total of 3,013 AGEs, with mean length 4,059 bp, median length 672 bp, and forming a pan-accessory genome of 12.2 Mb, were identified in these isolates (Supplementary Table 4). A Bray-Curtis dissimilarity heatmap of AGE presence/absence, weighted by the length of each AGE, shows that there is considerable accessory genomic variability in our collection (Fig. 1C). Consistent with previous findings⁴, the clade containing *exoS* and the clade containing *exoU* largely separate based on accessory genomic content, as evidenced by both Bray-Curtis dissimilarity and multiple correspondence analysis. Similar to the core genome phylogenetic analysis, some clusters of isolates with similar accessory genomes share a virulence rank, but both high and low virulence isolates show diverse AGE content (Figure 1C and Supplementary Figure 1A and B).

Evaluating Machine Learning Models Predicting *P. aeruginosa* Virulence Based on Accessory Genome Content

We hypothesized that as the *P. aeruginosa* accessory genome is variable between strains^{6,7,33} and includes multiple known virulence determinants^{12,13,15}, it would contain information predictive of strain virulence in mice. To test this hypothesis, we took a 10-fold nested cross-validation approach (Supplementary Figure 2). Using this approach, we tested the performance of four commonly used machine learning algorithms: random forest, l2-regularized logistic regression, elastic net logistic regression, and support vector classifier. Accessory genome content, in the form of AGE presence/absence, was used as features and virulence level (high or low) was used as labels during modeling. Nested cross-validation does not return a final machine learning model, instead examining how multiple models (with hyperparameters tuned through grid-search cross-validation) perform against held-out data. This process provides an estimate of how well a model trained through a given strategy will generalize to new data. All four algorithms performed similarly, with mean nested cross-validation accuracies of 0.75 (95% CI 0.69-0.80) for random forest, 0.75 (95% CI 0.65-0.85) for l2-regularized logistic regression, 0.72 (95% CI 0.65-0.79) for elastic net logistic regression, and 0.74 (95% CI 0.67-0.81) for support vector classifier. Other performance metrics showed similar ranges of values (Figure 2). Notably, the accuracy of all four algorithms was substantially higher than the null accuracy of simply predicting all isolates to be the majority class, which in this case was the prevalence of low virulence isolates (0.51). This indicates that there is signal in the accessory genome predictive of virulence in *P. aeruginosa*. Since all four machine learning algorithms performed similarly in nested cross-validation, we chose the random forest approach for further investigation.

We next evaluated whether sample size limited the performance of the random forest approach. We tested how accuracy of a model changed with increasing training set size, both against training and cross-validation examples (Figure 3A). While the training and cross-validation performance for the random forest model did not completely converge as more training examples were added, the learning curve showed that we are unlikely to see substantial

improvement in cross-validation accuracy with additional training isolates. A caveat to this result is that the learning curve assumes our training set is adequately representative of *P. aeruginosa* genetic diversity.

To further probe the characteristics of the random forest approach, we built a final random forest model using all 115 isolates in the training set. The out-of-bag accuracy (performance on the out-of-bag samples not included in each of the 10,000 decision trees making up the random forest) of this model was 0.75 (Supplementary Table 6), which is consistent with our nested cross-validation results. When assessed against the training isolates, the model showed an accuracy of 0.79, consistent with the trend in training accuracies observed in the learning curve (Supplementary Table 6 and Figure 3A). The training accuracy can be thought of as an idealized maximal performance and supports the conclusion that additional training examples are unlikely to substantially improve the model.

We next investigated which AGEs were most critical in making a prediction of high or low virulence in this model. We calculated the permutation importance (the mean decrease in model accuracy if a given feature is randomly permuted) for each AGE. To do this, we randomly permuted each AGE 100 times and determined the impact on out-of-bag accuracy. Overall, individual features showed low importance in the predictions made by the model, with permutation of the most important AGE causing only a mean 1% drop in model accuracy (Figure 3B). The vast majority of features (2,979/3,013) had no impact on out-of-bag accuracy when randomly permuted (Supplementary Table 7), indicating that the machine learning model bases decisions on a genomic signature predictive of virulence level rather than on identifying individual virulence or anti-virulence factors. If a given AGE is randomly permuted, it appears that other correlated features compensate for it. Each individual AGE was included as a feature in a minority of the 10,000 decision trees, with the most prevalent AGE appearing in only 148 trees in the final model (Figure 3C). As such, it is not possible for a single AGE to have a large impact on the prediction of virulence. To further assess the apparent redundancy in our feature

set, we randomly divided the 3,013 AGEs in the training set into 2, 4, and 10 subsets and evaluated the performance of random forest models built using only these subsets through nested cross-validation. We found that even when training on only a smaller subset of the accessory genomic features, model accuracy remained mostly unchanged (Supplementary Figure 3). This provides further evidence that a broad genetic fingerprint, rather than individual virulence or anti-virulence factors, is being used to classify strains as high or low virulence.

With the low permutation importance of any individual AGE, one must be cautious in drawing conclusions about their role in virulence. However, looking at the AGEs most predictive of virulence class and how they relate to one another may provide insights into genomic characteristics that are associated with, though not necessarily causative of, differences in pathogenicity. All of the ten most predictive AGEs in the random forest model were more prevalent in low virulent isolates (Table 1, Supplementary Data File 1). This is consistent with the finding that horizontally acquired genetic elements, major components of the accessory genome^{6,17}, can incur a fitness cost on the host bacterium¹⁷. While some genomic islands encode virulence factors¹¹, many horizontally acquired elements can have a parasitic relationship with the bacterium¹⁷. The AGE with the highest permutation importance aligns to a gene encoding for the conjugative protein TraD, perhaps suggesting a general association of conjugative elements with reduced virulence. Four of the top ten AGEs are comprised of sequences from the same “bin” in clustAGE analysis. This indicates that in at least some strains they are located near each other on the genome (i.e. part of a single, larger element). One of these four AGEs encodes an integrative and conjugative element (ICE) protein. These findings suggest that these AGEs are markers for a larger variable element common in low virulence strains. Two other AGEs are part of the same gene encoding a hypothetical protein. Finally, genes encoding for arsenic resistance are highly prevalent in low virulence isolates, perhaps suggesting either that this resistance comes at a cost or that strains adapted to survive heavy metal exposure are less able to cause disease in animals.

Assessing Model Performance with an Independent Test Set

The nested cross-validation performance of our random forest model provided an estimate of how well it would generalize to new *P. aeruginosa* isolates. To follow up on this, we applied the final random forest model built using all 115 training isolates to an independent test set of *P. aeruginosa* isolates to examine how well it predicted their virulence. As our test set, we selected 25 genetically diverse *P. aeruginosa* isolates previously cultured from patients with bacteremia in Spain between 2008-2009³⁴ and which we have whole genome sequenced (Supplementary Table 1 and Figure 4A). The virulence of each isolate was assessed in the mouse model of bacteremia, and isolates were classified as high or low virulence using the same threshold (estimated mLD50 of 6.9 log₁₀ CFU) defined for the training set (Figure 4B and Supplementary Tables 2-3). The test set was somewhat more pathogenic than the training set, with 15/25 (60%) of isolates classified as high virulence. We identified which of the 3,013 AGEs used as training features were present in each of the test isolates (Supplementary Table 8). Adding these isolates to a Bray-Curtis dissimilarity heatmap of AGE presence/absence showed that the test set is also relatively diverse in accessory genomic content (Figure 4C), a finding supported by multiple correspondence analysis (Supplementary Figure 1C-E).

We used the random forest model built with the training set accessory genomic and virulence information to predict the virulence of each isolate in the test set based on AGE presence or absence. Model performance on the test set (Table 2 and Figure 4D) was comparable to the estimates made through nested cross-validation. For example, the test set accuracy of 0.72 was comparable to the mean nested cross-validation accuracy of 0.75 (95% CI 0.69-0.80). This suggests that our predictive model of virulence is broadly applicable, even when tested against geographically distinct isolates. Several of the misclassified isolates in the test set appear to be exceptions in virulence when compared to their closest neighbor(s) in the core genome phylogenetic tree and the accessory genome heatmap (Figure 4A and C). Difficulty classifying these exceptional isolates is consistent with the notion that the model

predictions are based on genomic signatures which perhaps approximate phylogenetic relationships. Closely related isolates that differ in virulence from the majority of their genomic neighbors would therefore be expected to be misclassified.

While it was reassuring that the random forest model performed similarly against the test set as in nested cross-validation, we wanted to ensure that the accuracy observed did not simply occur by chance. We used permutation testing to model the null distribution of test set accuracies we would expect if no link between accessory genome content and virulence existed in the test set. Using this approach, we determined the probability of observing a test set accuracy of at least 0.72. Randomly permutating the predicted virulence of the 25 test isolates one million times resulted in an accuracy of at least 0.72 in 53,476 cases (one-sided $p = 0.053$) (Figure 4E). The test set performance observed is therefore unlikely if the accessory genome does not predict virulence. Limiting factors include the small sample size of the independent test set, as is evident from the discrete possible accuracies when the predictions were permuted, and that we would not expect the model to perform better against new data than it did during nested cross-validation.

Addressing Model Limitations by Removing Isolates with Intermediate Levels of Virulence

While the models generated thus far show that the accessory genome is predictive of *P. aeruginosa* virulence in mice, limitations inherent to our binary classification of virulence may constrain their performance. The first lies in the resolution of the mLD₅₀ estimates used as the basis for these classes. Because of the practical limitations of testing over 100 isolates in mice, many isolates were tested with only two or three doses. This leads to uncertainty in the dose required to cause severe disease (Supplementary Tables 2-3). Second, isolates with mLD₅₀ estimates close to the cutoff may actually be quite similar, both in their virulence and in their genomic makeup, but still be assigned to different virulence classes. To assess the extent to which this ambiguity influenced the results, we repeated the machine learning pipeline using the

random forest algorithm after removing intermediate virulence isolates (the middle third of estimated mLD₅₀ values). This enforced a greater separation of isolates classified as high and low virulence (Figure 5A). Even with a third fewer training isolates, nested cross-validation performance was similar to when all training isolates were included, with a mean accuracy of 0.76 (95% CI 0.67-0.85) (Figure 5B). The learning curve, however, showed a greater distance between the training and cross-validation scores (Figure 5C). This suggests a higher potential performance when intermediate virulence isolates are removed. The benefit of having a clearer boundary between high and low virulence would likely become apparent with a larger training set, though the number needed and degree of improvement is unclear.

Incorporating Test Set Isolates into the Accessory Genome Model

After using the 25 additional isolates as an independent test set, we next examined their impact on nested cross-validation performance if they were included in the training set. As this changes the median estimated mLD₅₀, we performed the modeling using both the median of the 115 training set isolates and the median of all 140 isolates as the cutoff for high/low virulence. These models performed similarly, both to each other and to the results seen with only the original training set. The mean nested cross-validation accuracy was 0.72 (95% CI 0.65-0.79) when using the median mLD₅₀ cutoff of the 115 training isolates and 0.69 (95% CI 0.60-0.78) when using the median mLD₅₀ cutoff of all 140 isolates (Supplementary Figure 4C and E). The learning curves, however, showed greater overfitting of the model when the all-isolates median cutoff was used, with a larger separation between the training and cross-validation accuracies (Supplementary Figure 4D and F). This suggests the choice of cutoff between high and low virulence isolates may become more important with increasing training set sizes. Removing intermediate virulence isolates resulted in similar nested cross-validation performance and learning curves as seen when performing this analysis on the original training isolates (Supplementary Figure 4B, G, and H).

Modeling *P. aeruginosa* Virulence with Features Incorporating Core Genome Information

Thus far we have shown that the accessory genome of *P. aeruginosa* is predictive of strain virulence. The accessory genome and core genome are correlated with each other, as can be seen from previous reports⁴ and by comparing core and accessory genome measures of strain relatedness (Figure 1B and C). As such, the accessory genome contains implicit information about the core genome. Still, it is possible that our focus on the accessory genome misses important core features predictive of virulence. To address this possibility, we defined our feature set in two additional ways and examined the performance of random forest models trained using these features. First, we considered core genome SNVs, using one-hot encoding in our machine learning pipeline to convert the SNVs into binary variables interpretable by the algorithm. Second, we used whole genome k-mer counts, which encode information about variability in both the accessory and core genome, considering a k-mer length of both 8 and 10 bp. Unlike the AGE feature set used previously, which considered the presence and absence of accessory elements, the k-mer feature sets additionally capture polymorphisms within these elements. We estimated the performance of approaches using these feature sets through nested cross-validation and then assessed how well final models built with each were able to predict the virulence of the 25 independent test set isolates.

A random forest approach using core genome SNVs as features performed worse on average in nested cross-validation than when using accessory genomic features, with a mean accuracy of 0.65. However, its 95% confidence interval (0.55-0.75) still overlapped with those seen for the accessory genomic models (Figure 6A). Therefore, some information important for determining virulence level may be missed by not considering the accessory genome. Another explanation is that more strains may be needed to model this substantially more complex feature set, as there were 440,116 core genome SNV loci detected in our training set. As the confidence intervals overlap, we must be careful drawing conclusions about the relative predictive power of the core and accessory genomes. The final model trained with core genome SNV features showed an accuracy of 0.72 on the independent test set (Supplementary Table

9). This was identical to the test set accuracy seen for the accessory genomic model but was associated with a lower sensitivity and higher specificity (Table 2 and Supplementary Table 9). Despite its lower nested cross-validation accuracy, we therefore cannot say whether the accessory genome or core genome are superior in predicting virulence.

The random forest approach using k-mer counts as features performed similarly to the accessory genome models in nested cross-validation, with a nested cross-validation accuracy of 0.71 (95% CI 0.58-0.83) when 8-mer counts were used and 0.69 (95% CI 0.63-0.76) when 10-mer counts were used (Figure 6B and D). This suggests that no additional predictive information was gained from incorporating core genome features, and that AGE presence/absence encodes the same information in a smaller feature set. The learning curve for the model trained on 8-mer counts showed overfitting, with a large discrepancy between the training and cross-validation accuracies (Figure 6C). This suggests that performance would improve with a larger training set, and perhaps that the increased complexity of the 8-mer feature set makes it more difficult to learn from than the presence or absence of AGEs. The final model trained with 8-mer features showed an accuracy of 0.60 on the test set, while the final model trained on the 10-mer feature set showed an accuracy of 0.68 (Supplementary Table 9). The performance of the 8-mer feature set was more variable in nested cross-validation, with a wider range in its 95% confidence interval, and it is possible that lower model stability contributed to its poorer performance against the test set.

DISCUSSION

In this study, we have shown that a signal exists in the *P. aeruginosa* accessory genome that is predictive of an isolate's virulence in a mouse model of infection. This finding was consistent across a variety of machine learning algorithms. Results for the random forest approach were validated using an independent test set of clinical isolates collected from a geographically distinct source, showing the broad applicability of the *P. aeruginosa* accessory

genome in predicting virulence. We additionally showed that the core genome, alone or in combination with the accessory genome, is also predictive of virulence, but the ability of models trained on this information to generalize to the independent test set was less conclusive. These types of genetic features were substantially more complex, and models trained from them may benefit from increasing sample size. The machine learning analyses conducted here serve as a framework to further investigate the relationship between the genome of a bacterium and its phenotype.

The random forest model trained on accessory genomic information classified isolates as high or low virulence based on a diffuse genomic signature rather than by detecting a small number of virulence or anti-virulence factors. These results suggest that lineage may predict phenotype, echoing the recent finding that genomic neighbors are highly predictive of antimicrobial resistance in *Streptococcus pneumoniae* and *Neisseria gonorrhoeae*²⁶. As such the genomic signature detected may approximate lineage. These findings also highlight the complexity of *P. aeruginosa* virulence compared to antimicrobial resistance, which is often dictated by one or a few genes or mutations. Still, it is relevant that all of the ten most important AGEs included in our model were associated with low virulence. This supports the earlier finding that *P. aeruginosa* accessory genes can reduce virulence in *C. elegans* and that active CRISPR systems, which would limit acquisition of foreign DNA and new AGEs, are associated with higher virulence in that model¹⁶. While certain AGEs enhance virulence¹⁵, the majority of AGEs (e.g. parasitic phages, plasmids, or ICEs) may decrease virulence. We focused on virulence in a mouse model of acute infection, and as such certain bacterial genetic factors important in the hospital setting may not apply. Antimicrobial resistance, for example, can be an important prognostic factor for patient outcomes^{35,36} but would not be relevant in this model. Future studies should examine the types of AGEs that are associated with, and ultimately causal of, both increased and decreased virulence and how this varies between infection models.

Our random forest model built on accessory genomic features showed similar performance in nested cross-validation as when applied to an independent test set of 25 isolates. By looking at the test set isolates that were classified incorrectly, we can learn why the model sometimes failed. Some incorrect predictions may be because of mLD₅₀s near the threshold between high and low virulence, leading to ambiguity in their true virulence level. An example of this scenario is the isolate PASP518, whose estimated mLD₅₀ of 7.0 log₁₀ CFU is near the cutoff of less than 6.9 log₁₀ CFU for high virulence. This highlights an inherent limitation of this study, in that virulence exists on a continuum not neatly divided into binary classes. To address this limitation, we examined how the model performs when excluding intermediate virulence isolates. In this condition, a random forest approach performed similarly in nested cross-validation with a third less samples and learning curve analysis showed a potential for higher accuracy with increasing sample size (Figure 5). On the other hand, some of the incorrect predictions in the test set were exceptions in virulence compared to closely related isolates. For example, PASP251 has an estimated mLD₅₀ of 6.3 log₁₀ CFU, while its nearest four phylogenetic neighbors all have estimated mLD₅₀s of greater than 7 log₁₀ CFU. This could be because PASP251 possesses additional virulence determinants in the form extra accessory genes or distinct core genome polymorphisms. In either case, PASP251 is a particularly interesting isolate for future study. An alternative explanation is that, while geography does not seem to play a role in model performance as a whole, the closely related isolates from Chicago have acquired common mutations or genes modifying their virulence. An increased sample size may ameliorate the problem of isolates being misclassified by allowing for finer resolution of subgroups that are associated with high or low virulence, especially if the model were able to learn new and more discriminatory patterns of features. Learning curve analysis for the random forest approach (Figure 3A) suggests the impact of adding more isolates would be limited, but this cannot account for new or more predictive features that could arise from increasing the amount of genetic data available.

As whole genome sequencing becomes an increasingly routine component of clinical microbiology practice, it will create the opportunity to risk-stratify patients based on the genome of an infecting bacteria and influence treatment decisions in real-time. The ability of the genome to predict antibiotic resistance has been established^{20,21,23,25,26}, opening the door for sequencing to supplement or replace traditional antimicrobial susceptibility testing. This study serves as a proof of concept that the *P. aeruginosa* genome can be used to predict its pathogenicity. Future studies are needed to expand beyond virulence in mice and provide a more complete understanding of the role genetic variation plays the ability of *P. aeruginosa* to cause disease. An area of particular interest is in predicting patient outcomes from the genome of an infecting isolate. Large retrospective studies using archived isolates with corresponding clinical data would allow for exploration of the relative importance that bacterial and patient factors play in predicting patient outcomes, as has been shown for specific *S. aureus* clones²⁸. This could improve the sophistication of current diagnostics and allow clinicians to rapidly identify patients at highest risk for poor outcomes.

METHODS

Bacterial Isolates

A training set of *P. aeruginosa* isolates for use in the machine learning analyses was established as follows. A total of 98 isolates previously collected at NMH in Chicago, USA from 1999-2003 from adults with *P. aeruginosa* bacteremia³¹ were selected after exclusion of 2 isolates that had been collected from patients with a history of cystic fibrosis. An additional 17 isolates from pediatric patients with Shanghai Fever collected at Chang Gung Children's Hospital in Taiwan from 2003-2008³² were included. This yielded a training set size of 115 isolates. A genetically diverse independent test set of 25 isolates was selected from a larger cohort of isolates collected from patients with bacteremia in Spain between 2008-2009³⁴ (Supplementary Table 1).

Mouse Model of Bacteremia

Female 6- to 9-week-old BALB/c mice were infected via tail-vein-injection in a model of bacteremia as previously described³². Isolates were plated from freezer stocks onto lysogeny broth (LB) agar, and single colonies were inoculated into MINS broth³⁷ and grown overnight at 37 °C. Overnight cultures were then subcultured in fresh MINS broth for approximately 3 hours at 37 °C. Cultures were resuspended in PBS before dilution to the target dose, and 50 µL was injected into each mouse via the tail vein. Inocula, in CFUs, were then determined by serial dilution, plating, and colony counts. Mice were monitored for the development of severe disease over 5 days, with mice exhibiting endpoint disease euthanized and scored as dead. Each isolate was tested at a minimum of 2 doses, with 3-5 mice per dose (minimum 9 total mice per isolate) (Supplementary Table 2). Many of the mouse experiments included in this study were previously reported as part of other studies. In particular, the majority of experiments with the NMH strains were performed as part of Allen *et al.*, 2020¹⁵. Some experiments with the Taiwan isolates PAC1 and PAC6 were performed as part of Chuang *et al.*, 2014³².

A modified 50% lethal dose (mLD₅₀) for each isolate was estimated from the above experiments using the drc package (v3.0-1)³⁸ in R (v3.6.1)³⁹. One outlier experiment for strain S2, which caused 20% mortality at a dose of ~7.2 log₁₀ CFU, was excluded as doses of ~6.3 and ~6.8 log₁₀ CFU caused 80% and 100% mortality, respectively, in other experiments. Percent mortality as a function of dose (in units of log₁₀ CFU) was modeled using a two-parameter log-logistic function and binomial data type. These models were used to estimate the mLD₅₀ for each isolate, which was then rounded to the nearest tenth (Supplementary Table 3). Isolates with rounded mLD₅₀ estimates below the median were classified as high virulence, with the remainder classified as low virulence.

All experiments were approved by the Northwestern University Institutional Animal Care and Use Committee in compliance with all relevant ethical regulations for animal testing and research.

Whole genome Sequencing and Assembly

Short-read whole genome sequencing was performed for all isolates using either Illumina HiSeq or Miseq platforms to generate paired-end reads. Reads were trimmed using Trimmomatic (v0.36)⁴⁰, with Nextera adapter removal, a sliding window size of 4 bp with average quality threshold of 15, and a minimum trimmed read length of 36 bp. Draft genomes were assembled from trimmed paired-end reads using SPAdes (v3.9.1)⁴¹ with the careful and automatic read coverage cutoff options. Draft genomes were further filtered to remove contigs shorter than 200 bp, with less than 5-fold mean read coverage, or with alignment to phiX. Even using only trimmed reads, the mean coverage of each filtered assembly was at least 24-fold. Many of the whole genome sequences used in this study were previously reported as parts of other studies^{15,42,43}. Draft genomes originally assembled through different methodologies were re-assembled as described above.

For several genomes (PABL012, PABL017, PABL048, PAC1, and PAC6), long-read sequencing and hybrid assembly was performed. Briefly, genomes were sequenced on the

PacBio RS II platform. Raw data were assembled using the HGAP assembler (SMRT Analysis v2.3.0), Canu assembler (v1.2)⁴⁴, and Celera assembler (v8.2)⁴⁵, all using default settings. Contigs were combined and circularized using Circlator (v1.5.1)⁴⁶. Assemblies were polished using Quiver (SMRT Analysis v2.3.0). Indel errors were corrected using Pilon (v1.21)⁴⁷ using paired-end reads generated on Illumina HiSeq or MiSeq platforms. The complete genome for PABL048 was generated as part of a previous study⁴².

Phylogenetic Analysis

kSNP (v3.0.21) was used to generate 95% core genome parsimony phylogenetic trees for both 115 isolates in the training set and all 140 isolates in the training and test sets, using fasta files as input. The Kchooser program was used to select the optimum k-mer size of 21, and SNP loci present in at least 95% of input genomes were used to make the trees⁴⁸. The phylogenetic trees were annotated and plots generated using iTOL (v4)⁴⁹.

Accessory Genome Determination

Accessory genomes for the 115 *P. aeruginosa* isolates in the training set were determined using the programs Spine (v0.3.2), AGEnt (v0.3.1), and ClustAGE (v0.8)^{7,50}. Spine was used with Prokka⁵¹-annotated genbank files for each isolate as input to generate a core genome of sequences present in at least 95% of isolates. AGEnt was then used to determine the accessory genome of each isolate based on comparison to the core genome. The accessory genomes of all 115 isolates were then compared using ClustAGE to identify shared sequences using an 85% identity cutoff. ClustAGE identifies the longest continuous accessory sequences as “bins” and the portions of these bins that differ from isolate to isolate as “subelements”^{15,50}. As part of this process, the read correction feature of ClustAGE was used to identify sequences present in the original sequencing reads that were missed during genome assembly. All perfectly correlated subelements identified through clustAGE were collapsed into a single feature, termed a “unique group (of subelements)”. For the purpose of this study, accessory genomic elements (AGEs) were defined as all unique groups totaling ≥ 200 bp. A

dataframe of all AGEs in the training isolates served as the accessory genome feature set in subsequent machine learning analyses (Supplementary Tables 4-5). To generate AGE features present in all genomes (both the original training and test sets), this process was repeated using all 140 *P. aeruginosa* isolates as input (Supplementary Table 10).

To determine which AGEs from the training set were present in the test set, clustAGE was run using the training set read-corrected subelement sequences (for all subelements ≥ 50 bp) from the 115 training isolates as a reference AGE set with the "--AGE" option and comparing to the draft genomes of all isolates in the test set, with read correction to identify any sequences present that were not included in draft genome assembly. This identified which portions of each subelement were found in the test set with an 85% identity cutoff. An AGE (defined as a unique group of subelements) was called as present if at least 85% of the screened length was detected (Supplementary Table 8).

To examine the relationships between accessory genomes in the training isolates, their AGE content was compared using the subelement_to_tree.pl utility from ClustAGE. This calculated the Bray-Curtis dissimilarity between each isolate based on AGE presence or absence, with the impact of each AGE weighted by its length. A neighbor joining tree was constructed from 1000 bootstrap replicates using the matrix of Bray-Curtis dissimilarities. For consistency with the definition of AGE used in this study, unique groups of subelements were used as input. The neighbor joining tree and associated heatmap of Bray-Curtis dissimilarities were annotated and visualized with iTOL (v4)⁴⁹. To examine the accessory genomic relatedness of the 25 test set isolates based on training-set derived AGEs, the training set AGE calls defined above were added and Bray-Curtis dissimilarity calculations, and neighbor joining tree construction were repeated. To further evaluate the relationships between accessory genomes, multiple correspondence analysis (MCA) was performed based on the presence or absence of AGEs in the 115 training isolates. Additionally, MCA was performed considering which of the

training isolate AGEs were identified all 140 isolates. MCA was performed in R (v3.6.1)³⁹ using the FactoMineR (v2.3)⁵² package and visualized using the factoextra (v1.0.6) package.

Sequence Alignment and Core SNV Calling

Sequence alignment of paired-end Illumina reads for each genome to the reference genome PAO1 (RefSeq accession NC_002516) was performed as previously described⁴². Briefly, reads were trimmed with Trimmomatic (v0.36)⁴⁰ and aligned to PAO1 with BWA (v0.7.15)⁵³. Loci passing inclusion criteria were called as having the PAO1 base or a SNV base for each genomic position, with the remainder of positions converted to gaps. PAO1 alignments for all 115 training isolates were concatenated, SNV positions present in fewer than 95% of genomes were filtered, and invariant sites were then removed. This core variant SNV alignment was used as the SNV feature set in subsequent machine learning analyses, with a one-hot-encoding step added to the pipeline to convert SNV loci into multiple binary variables. This feature set was defined in the test set by considering the genomic positions identified as variant in the training set. By extracting the sequence present at these variant positions in the PAO1 alignments for each of the 25 test set isolates, we created a SNV feature set corresponding to that used in the training set.

K-mer Counts

K-mer counts (using either 8 or 10 bp k-mers) were determined for each genome using KMC3 (v3.0.0)⁵⁴. All k-mers occurring at least once in each genome's fasta file were identified using the kmc application, and a count file was generated using the kmc_dump application. All unique k-mers identified in the training set of 115 *P. aeruginosa* genomes were used to construct a dataframe of k-mer counts for each genome. This served as k-mer feature set in subsequent machine learning analyses. This feature set was defined in the 25 test set isolates by considering the counts of all k-mers previously identified in the training set.

Predicting Virulence Based on Genomic Features

Machine learning analyses were performed using the sci-kit learn library (v0.21.2)⁵⁵ in Python (v3.6.9). The general workflow for the machine learning pipeline is described in Supplementary Figure 2. A training dataset of features (AGEs, k-mers, or core SNVs) and labels (high/low virulence) was defined. A machine learning algorithm (random forest, l2-regularized logistic regression, elastic net logistic regression, or support vector classifier) was chosen, and a grid of relevant hyperparameters to test were defined. A machine learning model was then trained using the selected algorithm, with hyperparameter tuning performed through grid-search cross-validation. A 10-fold stratified cross-validation strategy was used. This generated a final model which can be used to predict the virulence class of new isolates. Concurrently the generalization performance of this model was estimated through nested cross-validation. In this process, grid-search cross-validation was performed within an outer 10-fold stratified cross-validation loop. The performance of a grid-search cross-validation tuned model against each cross-validation fold was determined (including accuracy, sensitivity, specificity, positive predictive value, area under the receiver operating characteristic curve, and F1 score). The median, mean, and 95% confidence interval of the nested cross-validation results were determined and plotted with the values for each fold using R (v3.6.1)³⁹ with the tidyverse library suite (v1.2.1)⁵⁶.

For the random forest algorithm, the number of trees was set to 10,000 and “max_features”, “min_samples_split”, “min_samples_leaf”, “criterion”, and “max_depth” were varied as hyperparameters during grid-search cross-validation. The logistic regression algorithm was considered using l2 regularization (penalty = “l2”) and elastic net regularization (penalty = “elasticnet”) separately. For l2-regularized logistic regression, the “lbfgs” solver was used, “max_iter” was set to 10,000, and “C” was varied as a hyperparameter during grid-search cross-validation. For elastic net logistic regression, the “saga” solver was used, “max_iter” was set to 10,000, and “C” and “l1_ratio” were varied as hyperparameters. For the support vector classifier

algorithm, the radial basis function kernel was used, and “C” and “gamma” were varied as hyperparameters during grid-search cross-validation.

In some cases, learning curves were created to examine how training and nested cross-validation accuracy varied with increasing training test size. For this, the dataset was split into training and cross-validation folds through 10-fold stratified cross-validation. Subsets of examples were then drawn from each training fold ranging from 25% to 100% of the training fold size. On each subset, a model was trained through the grid-search cross-validation approach described above. The mean and 95% confidence interval for training and cross-validation accuracies at each number of examples were then determined and plotted.

Random Forest Permutation Importance

Out-of-bag permutation importance for the random forest model of virulence based on accessory genomic content trained on the complete training set of 115 *P. aeruginosa* isolates was determined using the rfimp (v1.3.4) Python package (<https://github.com/parrt/random-forest-importances>). This measures the decrease in accuracy in predicting out-of-bag samples (samples not used to train a given decision tree in the random forest) if a feature is randomly permuted. As the impact of permuting a given feature on model accuracy may depend on how it is permuted, this process was repeated a total of 100 times to determine a mean permutation importance (Supplementary Table 7). The putative annotation of the top 10 AGEs identified by permutation importance was determined by blast search of subelement sequences against the Pseudomonas Genome Database⁵⁷ and including the annotation of any ORF for which at least 50 bp were contained in the AGE.

Evaluating Random Forest Model Performance with an Independent Test Set

The random forest model trained on AGE presence/absence in the 115 training isolates were tested against the independent test set of 25 isolates. The training-set AGEs identified in these 25 isolates were used as features, and the predicted virulence classes were compared to the actual virulence for these isolates. This was used to estimate testing accuracy, sensitivity,

specificity, positive predictive value, area under the receiver operating characteristic curve, and F1 score and to plot the receiver operating characteristic curve. This approach was also used to assess the performance of random forest models trained on core genome SNVs, 8-mers, and 10-mers against the independent test set of 25 isolates.

For the accessory genome model, the probability of seeing the observed test set accuracy by chance if there were no true association between the predicted virulence (and therefore accessory genome) of an isolate and its true virulence was estimated through permutation testing. The predicted virulence classes for the 25 test isolates were randomly permuted 1 million times and used to create a null distribution of possible model accuracies. The observed test set accuracy was compared to this null distribution to estimate a one-sided p value.

DATA AVAILABILITY

BioSample accessions for all isolates used in this study are listed in Supplementary Table 1. For all isolates, the version of the genome assemblies used in this study are available on GitHub. Input data for machine learning analyses are also available on GitHub (https://github.com/nathanpincus/PA_Virulence_Prediction).

CODE AVAILABILITY

Code used for machine learning analyses in this study, including details on hyperparameters used during grid-search cross-validation, and for plotting the results are available on GitHub (https://github.com/nathanpincus/PA_Virulence_Prediction).

REFERENCES

- 1 Talbot, G. H. *et al.* Bad bugs need drugs: an update on the development pipeline from the Antimicrobial Availability Task Force of the Infectious Diseases Society of America. *Clinical Infectious Diseases* **42**, 657-668, doi:10.1086/499819 (2006).
- 2 Gellatly, S. L. & Hancock, R. E. W. *Pseudomonas aeruginosa* : new insights into pathogenesis and host defenses. *Pathogens and Disease* **67**, 159-173, doi:10.1111/2049-632X.12033 (2013).
- 3 Silby, M. W., Winstanley, C., Godfrey, S. A. C., Levy, S. B. & Jackson, R. W. *Pseudomonas* genomes: diverse and adaptable. *FEMS Microbiology Reviews* **35**, 652-680, doi:10.1111/j.1574-6976.2011.00269.x (2011).
- 4 Ozer, E. A., Nnah, E., Didelot, X., Whitaker, R. J. & Hauser, A. R. The population structure of *Pseudomonas aeruginosa* is characterized by genetic isolation of *exoU*⁺ and *exoS*⁺ lineages. *Genome Biol Evol* **11**, 1780-1796, doi:10.1093/gbe/evz119 (2019).
- 5 Freschi, L. *et al.* The *Pseudomonas aeruginosa* pan-genome provides new insights on its population structure, horizontal gene transfer, and pathogenicity. *Genome Biol Evol* **11**, 109-120, doi:10.1093/gbe/evy259 (2019).
- 6 Kung, V. L., Ozer, E. A. & Hauser, A. R. The accessory genome of *Pseudomonas aeruginosa*. *Microbiology and Molecular Biology Reviews : MMBR* **74**, 621-641, doi:10.1128/MMBR.00027-10 (2010).
- 7 Ozer, E. A., Allen, J. P. & Hauser, A. R. Characterization of the core and accessory genomes of *Pseudomonas aeruginosa* using bioinformatic tools Spine and AGEnt. *BMC Genomics* **15**, 737, doi:10.1186/1471-2164-15-737 (2014).
- 8 Mosquera-Rendón, J. *et al.* Pangenome-wide and molecular evolution analyses of the *Pseudomonas aeruginosa* species. *BMC Genomics* **17**, 45, doi:10.1186/s12864-016-2364-4 (2016).
- 9 LaFayette, S. L. *et al.* Cystic fibrosis–adapted *Pseudomonas aeruginosa* quorum sensing *lasR* mutants cause hyperinflammatory responses. *Science Advances* **1**, e1500199, doi:10.1126/sciadv.1500199 (2015).
- 10 Lorè, N. I. *et al.* Cystic fibrosis-niche adaptation of *Pseudomonas aeruginosa* reduces virulence in multiple infection hosts. *PLOS ONE* **7**, e35648, doi:10.1371/journal.pone.0035648 (2012).
- 11 Ho Sui, S. J., Fedynak, A., Hsiao, W. W. L., Langille, M. G. I. & Brinkman, F. S. L. The association of virulence factors with genomic islands. *PLoS ONE* **4**, e8094, doi:10.1371/journal.pone.0008094 (2009).
- 12 He, J. *et al.* The broad host range pathogen *Pseudomonas aeruginosa* strain PA14 carries two pathogenicity islands harboring plant and animal virulence genes. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 2530-2535, doi:10.1073/pnas.0304622101 (2004).
- 13 Battle, S. E., Meyer, F., Rello, J., Kung, V. L. & Hauser, A. R. Hybrid pathogenicity island PAGI-5 contributes to the highly virulent phenotype of a *Pseudomonas aeruginosa* isolate in mammals. *Journal of Bacteriology* **190**, 7130-7140, doi:10.1128/jb.00785-08 (2008).
- 14 Harrison, E. M. *et al.* Pathogenicity islands PAPI-1 and PAPI-2 contribute individually and synergistically to the virulence of *Pseudomonas aeruginosa* strain PA14. *Infection and Immunity* **78**, 1437-1446, doi:10.1128/IAI.00621-09 (2010).
- 15 Allen, J. P. *et al.* A comparative genomics approach identifies contact-dependent growth inhibition as a virulence determinant. *Proceedings of the National Academy of Sciences*, 201919198, doi:10.1073/pnas.1919198117 (2020).
- 16 Vasquez-Rifo, A., Veksler-Lubinsky, I., Cheng, Z., Ausubel, F. M. & Ambros, V. The *Pseudomonas aeruginosa* accessory genome elements influence virulence towards

- 681 *Caenorhabditis elegans*. *Genome Biology* **20**, 270, doi:10.1186/s13059-019-1890-1
682 (2019).
- 683 17 Brockhurst, M. A. *et al.* The ecology and evolution of pangenomes. *Current Biology* **29**,
684 R1094-R1103, doi:<https://doi.org/10.1016/j.cub.2019.08.012> (2019).
- 685 18 Hauser, A. R., Kang, P. J. & Engel, J. N. PepA, a secreted protein of *Pseudomonas*
686 *aeruginosa*, is necessary for cytotoxicity and virulence. *Molecular Microbiology* **27**, 807-
687 818, doi:10.1046/j.1365-2958.1998.00727.x (1998).
- 688 19 Shaver, C. M. & Hauser, A. R. Relative contributions of *Pseudomonas aeruginosa* ExoU,
689 ExoS, and ExoT to virulence in the lung. *Infection and Immunity* **72**, 6969-6977,
690 doi:10.1128/iai.72.12.6969-6977.2004 (2004).
- 691 20 Davis, J. J. *et al.* Antimicrobial resistance prediction in PATRIC and RAST. *Scientific*
692 *Reports* **6**, 27930, doi:10.1038/srep27930 (2016).
- 693 21 Nguyen, M. *et al.* Using machine learning to predict antimicrobial MICs and associated
694 genomic features for nontyphoidal *Salmonella*. *Journal of Clinical Microbiology* **57**,
695 e01260-01218, doi:10.1128/JCM.01260-18 (2019).
- 696 22 Kavvas, E. S. *et al.* Machine learning and structural analysis of *Mycobacterium*
697 *tuberculosis* pan-genome identifies genetic signatures of antibiotic resistance. *Nature*
698 *Communications* **9**, 4306, doi:10.1038/s41467-018-06634-y (2018).
- 699 23 Chen, M. L. *et al.* Beyond multidrug resistance: Leveraging rare variants with machine
700 and statistical learning models in *Mycobacterium tuberculosis* resistance prediction.
701 *EBioMedicine* **43**, 356-369, doi:10.1016/j.ebiom.2019.04.016 (2019).
- 702 24 Hyun, J. C., Kavvas, E. S., Monk, J. M. & Palsson, B. O. Machine learning with random
703 subspace ensembles identifies antimicrobial resistance determinants from pan-genomes
704 of three pathogens. *PLOS Computational Biology* **16**, e1007608,
705 doi:10.1371/journal.pcbi.1007608 (2020).
- 706 25 Khaledi, A. *et al.* Predicting antimicrobial resistance in *Pseudomonas aeruginosa* with
707 machine learning-enabled molecular diagnostics. *EMBO Molecular Medicine* **12**,
708 e10264, doi:10.15252/emmm.201910264 (2020).
- 709 26 Brinda, K. *et al.* Rapid inference of antibiotic resistance and susceptibility by genomic
710 neighbour typing. *Nature Microbiology*, doi:10.1038/s41564-019-0656-6 (2020).
- 711 27 Laabei, M. *et al.* Predicting the virulence of MRSA from its genome sequence. *Genome*
712 *Research* **24**, 839-849, doi:10.1101/gr.165415.113 (2014).
- 713 28 Recker, M. *et al.* Clonal differences in *Staphylococcus aureus* bacteraemia-associated
714 mortality. *Nature Microbiology* **2**, 1381-1388, doi:10.1038/s41564-017-0001-x (2017).
- 715 29 Wheeler, N. E., Gardner, P. P. & Barquist, L. Machine learning identifies signatures of
716 host adaptation in the bacterial pathogen *Salmonella enterica*. *PLOS Genetics* **14**,
717 e1007333, doi:10.1371/journal.pgen.1007333 (2018).
- 718 30 Cornforth, D. M. *et al.* *Pseudomonas aeruginosa* transcriptome during human infection.
719 *Proceedings of the National Academy of Sciences* **115**, E5125,
720 doi:10.1073/pnas.1717525115 (2018).
- 721 31 Scheetz, M. H. *et al.* Morbidity associated with *Pseudomonas aeruginosa* bloodstream
722 infections. *Diagnostic Microbiology and Infectious Disease* **64**, 311-319,
723 doi:<http://dx.doi.org/10.1016/j.diagmicrobio.2009.02.006> (2009).
- 724 32 Chuang, C.-H. *et al.* Shanghai fever: a distinct *Pseudomonas aeruginosa* enteric
725 disease. *Gut* **63**, 736-743, doi:10.1136/gutjnl-2013-304786 (2014).
- 726 33 Mathee, K. *et al.* Dynamics of *Pseudomonas aeruginosa* genome evolution. *Proceedings*
727 *of the National Academy of Sciences* **105**, 3100-3105, doi:10.1073/pnas.0711982105
728 (2008).
- 729 34 Peña, C. *et al.* Prospective multicenter study of the impact of carbapenem resistance on
730 mortality in *Pseudomonas aeruginosa* bloodstream infections. *Antimicrobial Agents and*
731 *Chemotherapy* **56**, 1265-1272, doi:10.1128/aac.05991-11 (2012).

- 35 Peña, C. *et al.* Influence of virulence genotype and resistance profile in the mortality of *Pseudomonas aeruginosa* bloodstream infections. *Clinical Infectious Diseases* **60**, 539-548, doi:10.1093/cid/ciu866 (2015).
- 36 Morata, L. *et al.* Influence of multidrug resistance and appropriate empirical therapy on the 30-day mortality rate of *Pseudomonas aeruginosa* bacteremia. *Antimicrobial Agents and Chemotherapy* **56**, 4833-4837, doi:10.1128/aac.00750-12 (2012).
- 37 Nicas, T. I. & Iglewski, B. H. Isolation and characterization of transposon-induced mutants of *Pseudomonas aeruginosa* deficient in production of exoenzyme S. *Infection and Immunity* **45**, 470 (1984).
- 38 Ritz, C., Baty, F., Streibig, J. C. & Gerhard, D. Dose-response analysis using R. *PLOS ONE* **10**, e0146021, doi:10.1371/journal.pone.0146021 (2016).
- 39 R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing, Vienna, Austria, 2019).
- 40 Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120, doi:10.1093/bioinformatics/btu170 (2014).
- 41 Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* **19**, 455-477, doi:10.1089/cmb.2012.0021 (2012).
- 42 Pincus, N. B. *et al.* Long-term persistence of an extensively drug resistant subclade of globally distributed *Pseudomonas aeruginosa* clonal complex 446 in an academic medical center. *Clin Infect Dis*, doi:10.1093/cid/ciz973 (2019).
- 43 Bacht, K. E. R., Allen, J. P., Cheung, B. H., Chiu, C.-H. & Hauser, A. R. Systemic infection facilitates transmission of *Pseudomonas aeruginosa* in mice. *Nature Communications* **11**, 543, doi:10.1038/s41467-020-14363-4 (2020).
- 44 Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research* **27**, 722-736, doi:10.1101/gr.215087.116 (2017).
- 45 Myers, E. W. *et al.* A whole-genome assembly of *Drosophila*. *Science* **287**, 2196, doi:10.1126/science.287.5461.2196 (2000).
- 46 Hunt, M. *et al.* Circlator: automated circularization of genome assemblies using long sequencing reads. *Genome Biology* **16**, 294, doi:10.1186/s13059-015-0849-0 (2015).
- 47 Walker, B. J. *et al.* Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLOS ONE* **9**, e112963, doi:10.1371/journal.pone.0112963 (2014).
- 48 Gardner, S. N., Slezak, T. & Hall, B. G. kSNP3.0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome. *Bioinformatics* **31**, 2877-2878, doi:10.1093/bioinformatics/btv271 (2015).
- 49 Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Research* **47**, W256-W259, doi:10.1093/nar/gkz239 (2019).
- 50 Ozer, E. A. ClustAGE: a tool for clustering and distribution analysis of bacterial accessory genomic elements. *BMC Bioinformatics* **19**, 150, doi:10.1186/s12859-018-2154-x (2018).
- 51 Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068-2069, doi:10.1093/bioinformatics/btu153 (2014).
- 52 Lê, S., Josse, J. & Husson, F. FactoMineR: An R package for multivariate analysis. *2008* **25**, 18, doi:10.18637/jss.v025.i01 (2008).
- 53 Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997* (2013).
- 54 Kokot, M., Długosz, M. & Deorowicz, S. KMC 3: counting and manipulating k-mer statistics. *Bioinformatics* **33**, 2759-2761, doi:10.1093/bioinformatics/btx304 (2017).

782 55 Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *Journal of Machine*
783 *Learning Research* **12**, 2825-2830 (2011).
784 56 Wickham, H. *et al.* Welcome to the Tidyverse. *Journal of Open Source Software* **4**, 1686
785 (2019).
786 57 Winsor, G. L. *et al.* Enhanced annotations and features for comparing thousands of
787 *Pseudomonas* genomes in the *Pseudomonas* genome database. *Nucleic acids research*
788 **44**, D646-D653, doi:10.1093/nar/gkv1227 (2016).
789

ACKNOWLEDGEMENTS

This work was supported by the National Institute of General Medical Sciences [T32 GM008061 and T32 GM008152 (N.B.P)], the American Cancer Society [MRSG-13-220-01-MPC (E.A.O)], and the National Institute of Allergy and Infectious Diseases [R01 AI118257, R21 129167, K24 AI104831, and U19 AI135964 (A.R.H.)]. J.J.D. and M.N. are supported by the United States Defense Advanced Research Projects Agency Friend or Foe program iSENTRY award [Contract No. HR0011937807 (J.J.D.)], and by the United States National Institute of Allergy and Infectious Diseases Bacterial and Viral Bioinformatics Resource Center award [Contract No. 75N93019C00076 (PI Rick Stevens)]. A.O. is supported by Instituto de Salud Carlos III, Subdirección General de Redes y Centros de Investigación Cooperativa, Ministerio de Economía y Competitividad, Spanish Network for Research in Infectious Diseases (REIPI RD16/0016/0004), co-financed by European Development Regional Fund "A way to achieve Europe" and operative program Intelligent Growth 2014-2020. This research was supported in part through the computational resources and staff contributions provided by the Genomics Compute Cluster, which is jointly supported by the Feinberg School of Medicine, the Center for Genetic Medicine, and Feinberg's Department of Biochemistry and Molecular Genetics, the Office of the Provost, the Office for Research, and Northwestern Information Technology. The Genomics Compute Cluster is part of Quest, Northwestern University's high-performance computing facility, with the purpose to advance research in genomics. We acknowledge the University of Maryland School of Medicine Institute for Genome Sciences for performance of PacBio whole-genome sequencing.

AUTHOR CONTRIBUTIONS

N.B.P. and A.R.H. conceived the study. N.B.P., A.R.H., J.J.D., M.N., and D.R.W. designed the experiments. N.B.P. and E.A.O. performed bioinformatic analyses. N.B.P. performed computational experiments. N.B.P., E.A.O., and J.P.A. performed *in vivo* experiments. C.-H.

815 Chuang, C.-H. Chiu, L.Z., and A.O. provided instrumental resources. All authors provided
816 intellectual contributions and reviewed the paper.

817 **COMPETING INTERESTS**

818 A.R.H. serves on the Scientific Advisory Board and as a consultant for Microbiotix, Inc.
819 (Worcester, Massachusetts). The remaining authors declare no competing interests.

820 **MATERIALS AND CORRESPONDENCE**

821 Correspondence and requests for materials should be addressed to A.R.H.

TABLES

Table 1. AGEs Most Predictive of Virulence in the Accessory Genome Random Forest Model

AGE	Mean OOB Permutation Importance	Subelements	Total Length (bp)	Total Prevalence	Prevalence High Virulence	Prevalence Low Virulence	Putative Annotation ^a
unique_grp_5582	0.0100	bin364_se00006	433	0.417	0.161	0.661	TraD
unique_grp_6841	0.0069	bin610_se00004	902	0.304	0.107	0.492	Hypothetical protein
unique_grp_1425	0.0063	bin20_se00056	1717	0.330	0.125	0.525	TetR/AcrR family Transcriptional regulator, Short chain dehydrogenase
unique_grp_6842	0.0063	bin610_se00005	369	0.296	0.089	0.492	Hypothetical protein
unique_grp_6989	0.0063	bin654_se00007	436	0.313	0.107	0.508	Intergenic region
unique_grp_1437	0.0062	bin20_se00073 bin20_se00075	2009	0.339	0.125	0.542	SoxR, MerR family DNA-binding transcriptional regulator, ICE relaxase PFGI-1 class, Hypothetical protein
unique_grp_8120	0.0058	bin987_se00001 bin1807_se00001	2821	0.339	0.125	0.542	AsrR family transcriptional regulators, Arsinic transporter, Arsenate reductase, ArsH, Hypothetical protein
unique_grp_1423	0.0057	bin20_se00054 bin20_se00057	1278	0.348	0.125	0.559	Type II glyceraldehyde-3- phosphate dehydrogenase
unique_grp_1435	0.0057	bin20_se00069	509	0.365	0.143	0.576	Hypothetical protein
unique_grp_5112	0.0057	bin258_se00005	419	0.357	0.143	0.559	ArsH

^aBased on annotation of any ORF with at least 50 bp overlap with the AGE sequence when blasted against the Pseudomonas Genome Database⁵⁷

Table 2. Performance of the Accessory Genome Random Forest Model Against the 25 Test Set Isolates

Prevalence of High Virulence	Accuracy	Sensitivity	Specificity	PPV	AUC	F1
0.6	0.72	0.8	0.6	0.75	0.77	0.77

PPV: positive predictive value, AUC: area under the receiver operating characteristic curve

830 FIGURES AND FIGURE LEGENDS

831

Figure 1. Virulence and genomic characteristics of the training set of 115 *P. aeruginosa* isolates. (A) Cumulative density function of estimated mLD₅₀s for the 115 isolates in a mouse model of bacteremia. Isolates with estimated mLD₅₀s less than the median value (red dashed line) were designated as high virulence, with the remainder designated as low virulence. (B) Mid-point rooted core genome phylogenetic tree of the 115 training isolates constructed from SNV loci present in at least 95% of genomes, annotated with T3SS genotype and virulence level. (C) Bray-Curtis dissimilarity heatmap comparing AGE presence in the 115 training isolates, weighted by AGE length, and accompanying neighbor joining tree. Isolates are annotated (from left to right) by T3SS genotype, virulence level, and the dissimilarity heatmap. A higher value indicates that two isolates have more similar accessory genomes.

Figure 2. Nested 10-fold cross-validation performance of machine learning algorithms in predicting *P. aeruginosa* virulence in mice based on accessory genomic content. (A) Random forest, (B) l2-regularized logistic regression, (C) elastic net logistic regression, and (D) support vector classifier algorithms were tested. Accuracy, sensitivity, specificity, positive predictive value (PPV), area under the receiver operating characteristic curve (AUC), and F1 score were determined for each cross-validation fold (black dots). The mean and 95% confidence interval of each statistic are indicated in red.

Figure 3. Evaluation of the random forest algorithm in predicting *P. aeruginosa* virulence based on accessory genomic content. (A) Learning curve showing change in mean training accuracy (red line) and cross validation accuracy (green line) in predicting *P. aeruginosa* virulence as increasing numbers of isolates are used to train the random forest model. Shading indicates the 95% confidence interval. Assessments at each number of training examples were through 10-fold nested cross-validation. (B) Out-of-bag permutation importance for the 10 most important AGEs in the random forest model, showing decrease in accuracy when these AGEs were randomly permuted. Permutation importance testing was performed 100 times, with the results of each test represented by the width of the black lines and the mean and 95% confidence interval indicated in red for each AGE. (C) Histogram indicating how many trees within the random forest model contained each AGE (feature), out of a total of 10,000 trees.

865

Figure 4. Characteristics of a random forest model trained on the accessory genomic content of the 115 *P. aeruginosa* training isolates to predict the virulence of an independent test set of 25 isolates. (A) Mid-point rooted core genome phylogenetic tree of the 115 training isolates and 25 test isolates constructed from SNV loci present in at least 95% of genomes, annotated (from inner to outer rings) with dataset, T3SS genotype, virulence level, and accuracy of prediction by the random forest model in test set. Arrowheads indicate examples of incorrectly classified test set strains whose closest core and accessory genomic neighbor(s) show a discordant virulence phenotype. (B) Cumulative density function of estimated mLD₅₀ values for the 25 *P. aeruginosa* isolates making up the independent test set in a mouse model of bacteremia. Isolates with estimated mLD₅₀ values less than the median estimated mLD₅₀ of the training set (red dashed line) were designated as high virulence, with the remainder designated as low virulence. (C) Bray-Curtis dissimilarity heatmap comparing presence of the 3,013 AGEs identified in the training set in all 140 isolates, weighted by AGE length, and accompanying neighbor joining tree. Isolates are annotated (from left to right) by collection, T3SS genotype, virulence level, accuracy of prediction by the random forest model in test set isolates (arrowheads highlighting specific incorrectly classified test set strains as in (A)), and the dissimilarity heatmap. A higher value indicates that two isolates have more similar accessory genomes. (D) Receiver operating characteristic curve for predictions of the 25 test set isolates using the random forest model (AUC = 0.77). (E) Permutation analysis showing the likelihood of predicting test virulence with an accuracy of at least 0.72 if no true link between virulence and accessory genomic content existed. The predicted virulence of the 25 test isolates were randomly permuted 1 million times, and the resulting null distribution of possible model accuracies is shown. The vertical red line indicates the true accuracy of the random forest model in predicting test set virulence (one-sided p = 0.053).

Figure 5. Performance of the random forest algorithm in predicting *P. aeruginosa* virulence from accessory genomic content when intermediate virulence isolates (middle 3rd of estimated mLD₅₀s) were removed. (A) Cumulative density function of estimated mLD₅₀ values after removing intermediate virulence isolates. Isolates with estimated mLD₅₀ values less than the median value in the complete training set (red dashed line) were designated as high virulence, with the remainder designated as low virulence. (B) Nested 10-fold cross-validation performance of the random forest model, including accuracy, sensitivity, specificity, positive predictive value (PPV), area under the receiver operating characteristic curve (AUC), and F1 score. The results for each cross-validation fold are shown in black with the mean and 95% confidence interval of each statistic indicated in red. (C) Learning curve showing change in mean training accuracy (red line) and cross validation accuracy (green line) with increasing training set sizes. Shading indicates the 95% confidence interval. Assessments at each number of training examples were through 10-fold nested cross-validation.

Figure 6. Performance of the random forest algorithm in predicting *P. aeruginosa* virulence when 8-mer counts, 10-mer counts, or core genome SNVs were used as model features. Nested cross-validation performance when using (A) core genome SNVs, (B) 8-mer counts, and (D) 10-mer counts, including accuracy, sensitivity, specificity, positive predictive value (PPV), area under the receiver operating characteristic curve (AUC), and F1 score. The results for each cross-validation fold are shown in black with the mean and 95% confidence interval of each statistic indicated in red. (C) Learning curve showing change in mean training accuracy (red line) and cross validation accuracy (green line) when using 8-mer counts as features as increasing numbers of isolates are used to train the random forest model. Shading indicates the 95% confidence interval. Assessments at each number of training examples were through 10-

917 fold nested cross-validation. Learning curves were not constructed when using core genome
918 SNV or 10-mer counts as features for reasons of computational feasibility.