NORTHWESTERN UNIVERSITY

Genomic Interrogation of *Pseudomonas aeruginosa* Virulence and Antimicrobial Resistance

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Life Sciences

By

Nathan Benjamin Pincus

EVANSTON, ILLINOIS

December 2020

ABSTRACT

Genomic Interrogation of *Pseudomonas aeruginosa* Virulence and Antimicrobial Resistance

Nathan Pincus

Pseudomonas aeruginosa is an important gram-negative opportunistic pathogen whose large genome allows it to thrive in diverse environments. There is a wide range of phenotypic variation within the species, which can be attributed both to variation in sequences present in most isolates (the core genome) or the presence or absence of sequences found in only some isolates (the accessory genome). In this dissertation, I present two bacterial genomics studies examining the relationship between the *P. aeruginosa* genome and phenotypes, one focusing on antimicrobial resistance and the other on virulence.

Antimicrobial resistance is a major barrier to treatment of *P. aeruginosa* infections, with multidrug-resistant infections disproportionally caused by globally distributed sequence types (ST) known as “high-risk clones”. Examining bacterial collections from Northwestern Memorial Hospital, we identified a number of isolates belonging to ST298 which showed substantial drug resistance. ST298, along with the closely related ST446, is part of a larger clonal complex (CC) 446, which has been previously identified as responsible for multidrug-resistant infections around the world. Genomic and phylogenetic analyses identified a subclade of ST298, which we named ST298*, that has caused repeated infections at our institution for at least 16 years and has thus far only been found at our institution. The estimated last common ancestor of this subclade was in 1980, suggesting that it may have been a problem for even longer than appreciated. Many isolates within this subclade harbored a large (~415 kb) plasmid, which contributed to antimicrobial resistance through the presence of a novel class 1 integron. We found that this plasmid was part of a family of large *Pseudomonas* genus plasmids. In this project, we both

uncover a prolonged local epidemic of highly drug-resistant *P. aeruginosa* and propose that CC446 is an emerging high-risk clone in need of further study.

P. aeruginosa isolates show a wide range of virulence in infection models, but it is a complex and combinatorial phenotype with many contributing factors. We took a machine learning approach to predict virulence (high or low) of *P. aeruginosa* isolates based on genomic content. Using a training set of 115 isolates, we found that the accessory genome could be used to predict virulence level, with nested cross-validation accuracy ranging from 72-75% depending on the algorithm used. We confirmed this finding using a test set of 25 isolates where an accessory genome-based random forest model was able to correctly identify virulence level 72% of the time. Individual accessory genomic elements showed low importance in the accessory genome-based random forest model, which appears to be learning a diffuse genomic fingerprint. We also showed that core genome single nucleotide variants and whole-genome k-mers could be used to predict virulence. While genomic content could be used to predict virulence in *P. aeruginosa*, it was not predictive of persistence in a collection of early cystic fibrosis isolates. In sum, we found that there is signal within the *P. aeruginosa* genome that is predictive of an isolate's virulence in mice. This project can serve as a starting point for future machine learning studies examining the relationship between bacterial genomics and diverse phenotypes.

ACKNOWLEDGEMENTS

I would like to thank all of the current and former members of the Hauser laboratory for providing a welcoming community and insightful discussion (both on and off topic). Science is always more gratifying when you can talk to someone about it. Thank you to Katie Murphy for helping me get started in the lab and for teaching me how to how prepare sequencing libraries. Thank you to Dr. Jonathan Allen for training me on how to do the mouse model of bacteremia. Thank you to Dr. Jonathan Allen and Dr. Kelly Bachta for all of their help troubleshooting bacterial cloning (even if none of those mutants made it into my dissertation). In particular, I would like to thank Dr. Egon Ozer for his guidance and mentorship in bacterial genomics. Without your advice and the foundation provided by the tools you have developed (even if they are written in Perl), my PhD research would have been much less computational. Thank you to Dr. Jim Davis and Marcus Nguyen at Argonne National Laboratory for their input and advice on the virulence prediction project. I would also like to acknowledge NUIT Research Computing Services and the Genomics Compute Cluster on Quest, as this resource was absolutely essential for my computational experiments. The Research Computing Services workshops were also invaluable, and I use coding tips I picked up in them to this day.

I would like to thank my advisor, Dr. Alan Hauser, for his support throughout my PhD. Thank you for the open door you always had to discuss my results and next steps. I also appreciate the freedom you gave me to follow my project in directions I am sure neither of us originally imagined. Your mentorship has helped me grow as a scientist, and I hope to take everything I learned in your lab with me to my next steps.

I would like to thank my committee, Drs. Karla Satchell, Patrick Seed, Deborah Winter, and Richard Wunderink for their feedback and thoughtful discussion as I moved my project

forward. In particular I would like to thank Dr. Winter for inviting me to her functional genomics work group and giving me an avenue to both discuss the computational aspects of my work and learn from what others are doing.

I would like to thank the Northwestern University Medical Scientist Training Program and the Driskill Graduate Program for their continued support throughout my time at Northwestern. I would also like to acknowledge the support of the Northwestern University MSTP (T32GM008152) and Cellular and Molecular Basis of Disease Training Program (T32GM008061) training grants. I also appreciate the opportunities that the CMBD gave me to hone my presentation skills and see the diverse biological research happening across both campuses.

LIST OF ABBREVIATIONS

AGE – Accessory genomic element

AMR – Antimicrobial resistance

AUC – Area under the receiver-operator characteristic curve

BWH – Brigham and Women's Hospital

CC – Clonal complex

CFU – Colony forming unit

CLSI - Clinical and Laboratory Standards Institute

GWAS – Genome-wide association study

ICE – Integrative and conjugative element

MCA – Multiple correspondence analysis

MIC – Minimum inhibitory concentration

MDR – Multidrug resistant

mLD_{50} – Modified 50% lethal dose

MLST – Multilocus sequence typing

NMH – Northwestern Memorial Hospital

NCBI – National Center for Biotechnology Information

ORF – Open reading frame

PPV – Positive predictive value

RGP – Region of genomic plasticity

SNV – Single nucleotide variants

ST – Sequence type

T3SS – Type III secretion system

XDR – Extensively drug resistant

STATEMENT OF PUBLICATION

An excerpt from Chapter 1 has been modified for inclusion in a forthcoming review article:

Allen, J. P., Pincus, N. B. & Hauser, A. R. Comparative genomic approaches to identifying bacterial virulence determinants. *Submitted*.

The majority of Chapter 2 (and relevant methods in Chapter 5) has been published in:

Pincus, N. B., Bachta, K. E. R., Ozer, E. A., Allen, J. P., Pura, O. N., Qi, C., Rhodes, N. J., Marty, F. M., Pandit, A., Mekalanos, J. J., Oliver, A. & Hauser, A. R. 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).

The majority of Chapter 3 (and relevant methods in Chapter 5) has been published in:

Pincus, N. B., Ozer, E. A., Allen, J. P., Nguyen, M., Davis, J. J., Winter, D. R., Chuang, C.-H., Chiu, C.-H., Zamorano, L., Oliver, A. & Hauser, A. R. A genome-based model to predict the virulence of *Pseudomonas aeruginosa* isolates. *mBio* **11**, e01527-01520, doi:10.1128/mBio.01527-20 (2020).

DEDICATION

To Kathy, for your love and support. I wouldn't have made it through without you.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

SUPPLEMENTARY TABLES – AVAILIBLE ONLINE

Supplementary Table 3.3 Sequences of Subelements >50 bp Making up the 10 AGEs Most Predictive of Virulence Class in the Random Forest Model Trained on AGE Content of the 115 Training Isolates

Supplementary Table 3.4 Collection of 207 Early Cystic Fibrosis *P. aeruginosa* Isolates Classified as Persistent or Eradicated

Supplementary Table 3.5 Accessory Genomic Elements (perfectly correlated subelements >200bp) in the 207 Early Cystic Fibrosis *P. aeruginosa* Isolates

CHAPTER 1

Introduction

Pseudomonas aeruginosa genome structure

Pseudomonas aeruginosa is a clinically relevant gram-negative bacterium from the class Gammaproteobacteria[1]. As an opportunistic pathogen, *P. aeruginosa* can cause a wide variety of infections in susceptible patients, including skin and soft tissue infections, osteomyelitis, keratitis, urinary tract infections, pneumonia, and bacteremia[2,3]. Patients supported with mechanical ventilation are especially vulnerable to *P. aeruginosa* infection[3,4]. In fact, between 2011-2014 *P. aeruginosa* was second only to *Staphylococcus aureus* as a cause of ventilator-associated pneumonia in the United States[5]. Severe *P. aeruginosa* infections, such as pneumonia and bacteremia, are associated with substantial mortality[6-9]. In a 13-year prospective study examining outcomes from thousands of bloodstream infections, infections caused by *P. aeruginosa* were associated with higher mortality than those caused by *S. aureus* or other gram-negative bacteria in a multivariable analysis[7]. In addition to severe acute infections, *P. aeruginosa* causes chronic lung infections, particularly in patients with cystic fibrosis, bronchiectasis, or chronic obstructive pulmonary disease[2,3,10]. For cystic fibrosis patients, a single clone can colonize a patient for decades[11,12] despite standard treatment. Carriage of *P. aeruginosa* is associated with a decline in lung function and progression towards poor outcomes in cystic fibrosis[13,14].

P. aeruginosa, however, is not simply associated with human infection. It can be found ubiquitously in the environment, particularly in wet settings. In the natural environment, *P. aeruginosa* has been isolated from rivers and other bodies of water[15,16], plants[16,17], and soil[16-18] (even oil-contaminated soils[19,20]). *P. aeruginosa* can also be found in manmade environments,

particularly plumbing and water systems[16,17,21-23]. It can cause infections in a variety of hosts, including mammals[24-26], insects[27-30], nematodes[27,29,31,32], plants[24,30], and amoebae[33]. Unsurprisingly given its ability to survive in such myriad environments, the diverse metabolism of *P. aeruginosa* has made it of interest for industrial purposes such as bioremediation[34-36] and surfactant production[37]. Additionally, the ability of *P. aeruginosa* to cause disease in a given host may be secondary to evolutionary pressures exerted in a different environment. For example, it is hypothesized that amoebae may be the original target of the *P. aeruginosa* type III secretion system (T3SS), an important virulence factor in mammalian infection[4,38,39]. While the role of *P. aeruginosa* as an opportunistic pathogen has been discussed, *Pseudomonas* species do not appear to be a major component of the normal human microbiome[40]. It can, however, at least transiently colonize patients, increasing risk for infection[41,42]. For example, in a prospective study Cohen et al. noted *P. aeruginosa* carriage in a large fraction of ICU patients (38% [13/34 patients] on admission and rising to 52% [11/21 patients] after one week). The majority of infections observed in this study in this study were caused by colonizing strains[42].

A question then arises in how *P. aeruginosa* is so successful as a generalist, particularly when some other human pathogens, such as *Helicobacter pylori*[43], *Neisseria gonorrhoeae*[44], and *Mycobacterium tuberculosis* complex[45], show a far narrower host ranges. The answer to this question likely lies in the large genome of *P. aeruginosa*, which is diverse both in gene content (within and between strains) and encoded functions[46-49].

The P. aeruginosa genome sequence

The first complete genome sequence of *P. aeruginosa* was described by Stover et al. in 2000 for the well-known laboratory strain PAO1[46]. They found that the PAO1 genome was approximately 6.3 Mb and contained 5,570 open reading frames (ORFs), quite large compared to

other bacteria that had been sequenced at the time. Of these, 2,531 (45.4%) were annotated as hypothetical genes with no known function. Their annotations predicted that 468 (8.4%) ORFs had regulatory or sensing functions. This was substantially higher (both in number and percentage of the genome) than other bacteria that had been sequenced at the time. The authors posited that large genome and abundant regulatory machinery of *P. aeruginosa* contribute to its success as a generalist[46].

Advances in whole-genome sequencing technology have led to a rapid rise in the number of available *P. aeruginosa* sequences and, with this, advances in our understanding of the *P. aeruginosa* genome and how it varies between strains. As of July 21, 2020, there are 4,660 *P. aeruginosa* genomes included in the *Pseudomonas* genome database, 206 of which are complete genomes[50]. Recent large studies show that the genome size of *P. aeruginosa* ranges from 5.5-7.6 Mb, with an mean genome size of approximately 6.6 Mb[51,52]. The genome of PAO1 is therefore somewhat smaller than average. Correspondingly, a *P. aeruginosa* genome has 6-6.2 thousand genes on average, with the some variation due to differences in study populations[48,52-54]. The wide range of genome sizes is indicative of the sequence diversity within the species. Still, much of the *P. aeruginosa* genome remains poorly characterized[48], highlighting a need for further study.

Transposon sequencing studies have been conducted to look at what proportion of genome are important in different environments, including rich and minimal laboratory media, urine, sputum (patient-derived and synthetic), fetal bovine serum, and several mouse models[55-59]. One study by Poulsen et al. examined genes essential for growth on five solid media preparations (including media derived from urine and fetal bovine serum) in nine *P. aeruginosa* strains. They found that between 354-737 genes were essential for growth in these different conditions, with

variation both between strains and between conditions in the same strain. They defined 321 genes as "core essential," required in all strains in all conditions. Unsurprisingly, almost all of these are involved in metabolism, macromolecular synthesis, or cell structure and division[58]. From this one can conclude that only a minority of *P. aeruginosa* genes are required in any given environment, and that these requirements can vary between environments. This is consistent with the idea that the large genome of *P. aeruginosa* provides it with a toolbox to succeed as a generalist. There are even different requirements in different infection models. For example, flagella are important during an acute burn model but not a chronic wound model of mouse infection[57]. Further, differences in essential genes between strains show that *P. aeruginosa* strains are not functionally identical. Variation in either gene presence or functionality may make a given gene or pathway redundant in one strain but absolutely essential in another.

The P. aeruginosa pangenome and its components

As stated above, the genome size of *P. aeruginosa* varies from strain to strain with a range of approximately 5.5-7.6 Mb[51,52]. This necessitates that *P. aeruginosa* genomes must also vary in gene and sequence content. This brings up a concept important in bacterial genomics, that of the bacterial pangenome[60-64]. Unlike higher eukaryotes, bacteria can participate in horizontal gene transfer, both within and between species. This creates the potential for extensive variation in gene content within a bacterial species, a finding that challenged the traditional species concept[62,63]. With this variable gene presence in mind, the bacterial genome can be split into two main components. Genes or sequences present in all (or almost all) strains make up a species' "core genome" and can be thought of as the defining features of that species. On the other hand, genes or sequences that are present in only some strains make up a species' "accessory genome" (sometimes also called the "flexible" or "dispensable" genome). Together,

all genes or sequences present in the core and accessory genomes make up a species' pangenome[48,52,61,63,64].

In a given study, one can only examine the portion of the pangenome captured in the collection of genomes being analyzed. It is, however, possible to extrapolate from the genomes considered to predict how the pangenome size would change if additional isolates were sequenced. A bacterial pangenome can be considered to be "open" if it continues grow as new genomes are obtained, and conversely can be considered to be "closed" if it reaches a maximum size as new genomes are added[61,64]. Species that inhabit a highly specialized niche (particularly obligate symbionts or parasites) or have limited contact with other bacteria tend to have closed pangenomes, while generalists and members of complex microbial communities tend to have open pangenomes[61,63,64]. The concept of the bacterial pangenome was first introduced by Tettelin et al. in 2005 as part of a comparative study of eight *Streptococcus agalactiae* genomes[60]. In 2008, Tettelin et al. proposed a mathematical framework to determine whether a species has an open or closed pangenome by fitting the change in pangenome or new genome size as additional strains are sequenced to a power law function. Specifically, the number of new genes identified as each additional genome is sequenced can be fit to the equation $n = \kappa N^{-\alpha}$, where n is the number of new genes added and N is the number of genomes considered (Figure 1.1). If the resulting α value is ≤ 1, it is projected that new genes will continue to be identified indefinitely (albeit at a slowing rate) as additional genomes are added, indicating that the species has an open pangenome. On the other hand, if $\alpha > 1$, it is projected that all genes present in the species will eventually be captured. Applying this method to available genomes from a number of species, they determined that the pangenomes of *Bacillus cereus*, *Streptococcus pneumoniae*, and *Escherichia coli* were open, while the pangenomes of *S.* aureus and *Bacillus anthracis* were

closed. In each of these examples, Tettelin et al. examined what is now a relatively small number of genomes[61]. As more genomes become available, the concept of an open or closed pangenome may need to evolve. For example, should one consider a pangenome closed if it would require thousands of genomes before no new genes are identified? Still, as described below it is clear that even when comparing over 700 genomes the *P. aeruginosa* pangenome has not been fully captured[51].

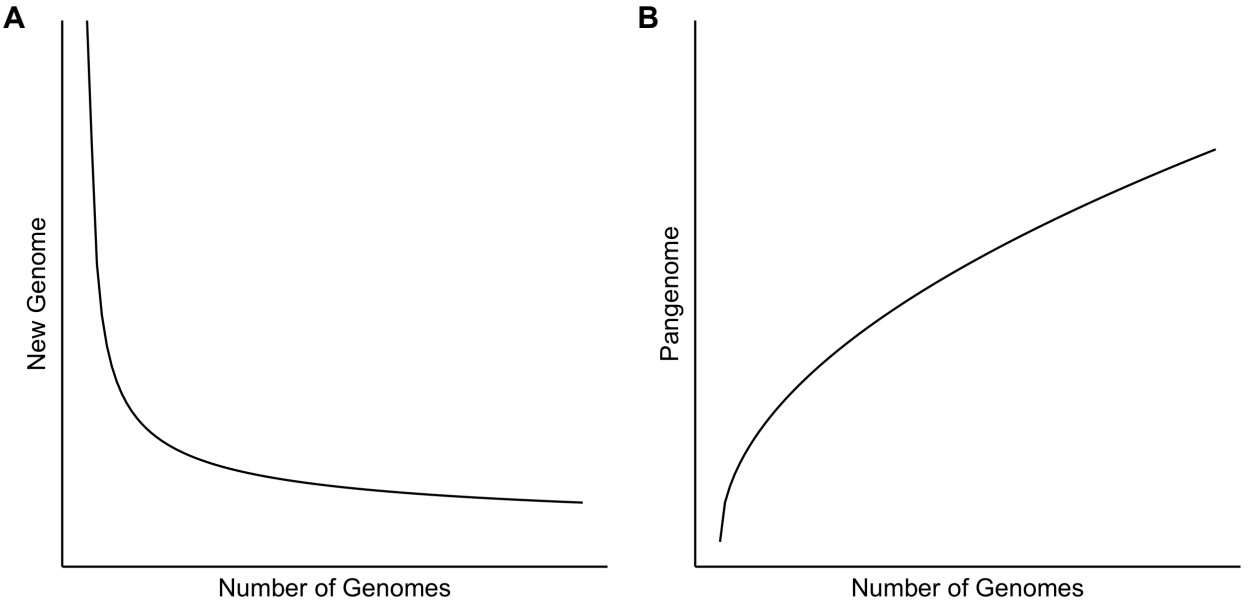


Figure 1.1 Model of new genome and pangenome size as additional genome sequences are considered. (A) The amount of new genome added can be modeled using the power law function $n = \kappa N^{-\alpha}$, where n is the number of new genes or amount of sequence (in bp) added and N is the number of genomes considered. (B) The change in pangenome size can be modeled using the power law function $n = \kappa N^{\gamma}$, where n is the pangenome size (in number of genes or total sequence length) and N is the number of genomes considered. Equations are from Tettelin et al., 2008[61]. Plots assume an open pangenome ($\alpha \leq 1$), consistent with the conclusions of Ozer et al., who examined the change in new genome size (in bp) within increasing numbers of genomes for a set of 739 *P. aeruginosa* genomes[51].

To examine a species' pangenome or its components, the core and accessory genomes, one must first define these components in a collection of bacterial genomes. Multiple approaches have been developed to accomplish this. One method that is commonly used is to identify genes within each genome and then cluster them into gene or protein families (or directly compare gene sequences between each genome to see whether they meet a minimum threshold to be categorized as the same gene)[31,53,65,66]. Programs used for this purpose include CD-HIT[67] and Roary[68]. The set of all gene clusters would then make up the pangenome for the study population, and individual clusters can be assigned to the core or accessory genomes based on their prevalence[31,53,65]. The Hauser laboratory, working with Dr. Egon Ozer, has taken a sequence-based (and therefore gene-agnostic) approach to defining the core and accessory genomes, and with them the pangenome. Here, a contiguous genetic sequence present in some genomes is defined as an accessory genomic element (AGE). The rationale behind this decision is that variation between strains is not limited to individual genes. A contiguous accessory sequence could range in size from dozens of genes (such as a genomic island, defined below) down to only part of a gene. This approach is also not dependent on annotation and can examine variable intergenic regions (such as promoters)[48,69,70]. Dr. Ozer has developed a suite of bioinformatic tools to define bacterial core and accessory genomes. The first of these, Spine, aligns all genomes within a collection to each other and identifies a backbone core genome of sequences present in a specified proportion of isolates (e.g. 100% or 95%). Next, the tool AGEnt is used to identify the accessory genome of each strain by identifying sequences that do not align back to the core genome[48]. Finally, clustAGE is used to compare the accessory genomes of each strain to each other and define all AGEs making up the pan-accessory genome of that population[69]. The pangenome could then be considered as the combination of the core genome

and pan-accessory genome. Spine can also directly calculate and output a pangenome after aligning all input genomes[48]. In all of these tools, a default 85% sequence identity cutoff is used to determine whether sequences from two genomes are part of the same element (be it core or accessory).

Now that the general concept of the pangenome, its components, and how it can be defined have been introduced, the pangenome of *P. aeruginosa* will be discussed. While several studies have examined the *P. aeruginosa* pangenome[30,31,48,51-54,65,66,71], two in particular have looked into the question of whether it possesses an open or closed pangenome. Mosquera-Rendón et al took a gene-based approach to analyze 181 genomes, defining a pangenome for this population of 16,820 genes. This analysis found that *P. aeruginosa* had a closed pangenome, with an estimated α of 2.36[54]. Ozer et al. took a sequence-based approach to analyze 739 *P. aeruginosa* genomes, finding a pangenome size of 32 Mb[51], almost five times the amount of sequence in an average *P. aeruginosa* isolate. They, conversely, report that the species' pangenome is open, with an estimated α of 0.65[51]. As such, there is disagreement regarding whether the *P. aeruginosa* pangenome will continue to grow as additional genomes are considered. This may have to do with the different ways the studies defined the pangenome (gene vs. sequence-based) or the specific populations they examined. With the finding that the pangenome continues to grow at over 700 *P. aeruginosa* isolates, I would argue that, from a practical perspective, it can be considered open. Regardless, these studies agree that the *P. aeruginosa* pangenome is much larger than the sequence contained in a single *P. aeruginosa* genome. Other large studies support this finding, calculating pangenome sizes of 54,272 and 28,793 genes when considering collections of 1,311 and 1488 genomes respectively[31,52]. This large amount of diversity within the species likely explains a substantial proportion of

phenotypic differences between strains. Gene prevalence in the *P. aeruginosa* pangenome has a roughly U-shaped distribution, with a large number of genes that are common (in all or most genomes) and rare (in one or few genomes), but relatively few genes of moderate prevalence. As the number of genomes considered increases, new rare genes continue to be found and outnumber the common ones[31,53,54].

The size of the *P. aeruginosa* core genome is dependent on a combination of the strictness by which it is defined and the number of genomes. An early study examining five genomes found that 5,021 genes were universally conserved[71]. However, as study populations increases in size, the number of genes found in all isolates decreases drastically. In a more recent study looking at a total of 181 genomes, the size of the core genome approximately halved to 2,503 genes[54]. Looking at 1,311 genomes and using a strict core genome definition, Freschi et al. defined the core genome as being made up of only 665 genes. This is not much more than 10% of the average *P. aeruginosa* genome. The authors say that a strength of this core genome definition is that number of genes resembles the number essential for survival in media in transposon sequencing studies but acknowledge that it is very conservative[52]. Still, there are drawbacks to this strict core genome definition, such as the loss of many genes that would be thought of as largely characteristic of the species but are missing in a few outliers. Examples include the exotoxin A gene, *toxA*, which is missing in the outlier strain PA7, or large chromosomal deletions that occur in chronic cystic fibrosis isolates[48,53]. Further, when working with draft genomes, certain sequences or genes may be excluded from the core genome simply because of issues related to the assembly (e.g. they are fragmented over multiple contigs or not captured) even if they are actually present in that strain's genome[65]. With less strict core genome definitions (for example considering genes or sequences present in 90-95% of genomes), various

studies have defined core genomes ranging from 5,081-5,316 genes[31,48,53,65]. It is notable that this value is consistent with early studies comparing smaller numbers of genomes with a strict core definition[71] and appears to be fairly stable to increases in number of genomes. For example, using a 90% core genome definition and examining 1488 *P. aeruginosa* genomes, Vasquez-Rifo et al. calculated a core genome size of 5170 genes[31]. For these reasons (better capturing genes generally characteristic of *P. aeruginosa*, less susceptibility to missing sequence in draft genomes, more stability as new genomes are added), I find the less conservative core genome definition to be generally more useful. This ensures that the core genome size is comparable when analyzing different populations and, as I have primarily worked with draft genomes, is less effected by imperfect sequencing coverage and assembly. However, the most appropriate core genome definition may vary depending on the population being studied and the specific question being asked.

Using the less conservative core genome definition, we can begin to examine its properties and what features are generally characteristic of the species. As stated, *P. aeruginosa* can be thought to have a core genome that is approximately 5,081-5,316 genes[31,48,53,65] or 5.8 Mb[48,51] in size. Considering an average genome size of 6.6 Mb, approximately 88% of the sequence in any given strain would be considered core, though this would vary based on the size of that strain's accessory genome. Ozer et al. found that the core genome's nucleotide content is 67%, guanine and cytosine[48], consistent with the understanding that *P. aeruginosa* is a high G+C content organism[46]. Ozer et al. also compared core gene sequences to the Clusters of Orthologous Groups of proteins (COG) database[72]. Through this, they found that approximately 80% of core genes could be assigned to a COG group. Even so, close to 40% of core genes were poorly characterized (with no match in the COG database or in a COG group with no clear

functional prediction)[48]. This sizable proportion of genes with no known function even in the parts of the genome well conserved throughout the species illustrates how much remains to be learned about *P. aeruginosa* genome. Unsurprisingly, genes providing functions necessary for general bacterial life can be found in the core genome. In their early study of the *P. aeruginosa* pangenome, Mathee et al. found that most housekeeping genes are part of the core genome[71]. This agrees with analysis performed by Valot et al., who found that the core genome was enriched for functions related to metabolism, signal transduction, post-translational modification, transcription, and translation[53].

As with the core genome, the proportion of the genome for any given isolate called as accessory is dependent on the strictness of the core genome definition used. Using a less strict core genome definition, the median accessory genome size was 912 kb in a collection of 739 *P. aeruginosa* isolates. While this is minority of sequence in any given *P. aeruginosa* genome, it is clear that accessory genome contributes the majority of sequence in the pangenome (32 Mb in this collection). The size of the accessory genome is strain-dependent, varying from 277 kb – 2.2 Mb in these genomes[51]. This accessory sequence can include genes involved in various functions, such as antimicrobial resistance (AMR), bacterial pathogenicity, or metabolism[24,47,48,65,70,71,73-76]. Compared to the core genome, even less of the accessory genome has a predicted or known function, with Ozer et al. finding that almost 70% of accessory genes were poorly characterized. It also has a lower G+C content (averaging 61%)[48], consistent with the idea that much of this sequence has origins outside of the species (perhaps originating in species with lower G+C content)[47]. Indeed, the accessory genome is enriched for signatures of horizontal gene transfer, such as genes associated with integrative and conjugative elements (ICEs), integrons, and phages[48,52,71]. This makes sense given the central role of horizontal gene

transfer in bacterial pangenomes in general[62,63]. However, it is important to note that not all accessory genes need be mobile in origin, and accessory regions can also be the result of genetic deletions in a portion of *P. aeruginosa* strains[52,71,77,78]. While accessory sequences can be found scattered throughout the *P. aeruginosa* chromosome (and can also be extra-chromosomal in the form of plasmids), they are often concentrated at locations termed "regions of genomic plasticity" (RGPs). These RGPs are often the result of insertion of accessory sequence at tRNA gene sites. Large, contiguous (10 kb or larger) accessory sequences present at RGPs are referred to as genomic islands[47,48,71,79,80].

Accessory genomic elements in P. aeruginosa

Many types of genetic elements can contribute to the *P. aeruginosa* accessory genome. Often these have a mobile origin and show evidence of horizontal gene transfer[47,48,71]. Some of these AGEs are still be capable of horizontal gene transfer, while others have become fixed within their resident genomes[81-83]. The likely origins of these horizontally acquired AGEs appear to be from species that *P. aeruginosa* would interact with in both natural and clinical environments. In an analysis of transcribed *P. aeruginosa* accessory genes, Pohl et al. identified homologues in a variety of species, including other *Pseudomonas* species, Enterobacteriaceae (such as *K. pneumoniae* and *S. enterica*), *Burkholderia vietnamiensis*, and *Acidovorax* species[84]. Deletions can also lead to sequences being classified as accessory in the remaining strains[52,71,77,78]. In this section, I will briefly review the types of elements that contribute to the *P. aeruginosa* accessory genome. Although the functions of many AGEs are poorly understood, the accessory genome can influence the biology of *P. aeruginosa* in a variety of ways[47,48]. However, the context in which these elements has been studied has often been from the perspective of examining bacterial pathogenicity or antimicrobial resistance. As such, the examples noted here

will focus on AGEs involved in these phenotypes. Additionally, it is not necessarily obvious if a given accessory sequence is part of a larger mobile element. This can require genomic context that may not be available. For example, with only a draft genome, it is difficult to say whether a given sequence is chromosomal or part of a plasmid.

Genomic islands are often formed by ICEs. Complete ICEs form large genomic islands that can mobilize by excision, mediate conjugation into a new host, and then reintegrate into the chromosome[47,85,86]. Integration often (but not always) occurs at tRNA genes[87], leading to the association Mathee et al. observed between these sites and RGPs[71]. However, some ICEs have been degraded and have lost some or all of the machinery required for transfer, resulting in smaller genomic islands fixed within the genome[81]. Several ICE-derived genomic islands have been implicated in *P. aeruginosa* virulence in a variety of infection models[24,75,76,82]. These include genomic islands carrying the important T3SS effector gene *exoU*, which are perhaps derivatives of single original island[77]. *P. aeruginosa* ICEs harboring AMR genes have also been identified[88,89].

Bacteriophages can facilitate horizontal gene transfer through transduction and can themselves encode important accessory genes. As such, prophages (lysogenic phages integrated into the host chromosome) and their remnants are major contributors to the *P. aeruginosa* accessory genome[47,82,83,90-93]. As has been noted, phage genes are enriched in the accessory genome[48,52], with Ozer et al. finding an average of 124 predicted phage genes in the accessory genomes of 12 isolates[48]. As with ICEs, several prophages have been shown to contribute to virulence in *P. aeruginosa*[82,92,93], including the cytotoxin-encoding φCTX[93]. There are also loci that are evolutionarily derived from bacteriophages but have lost components required for

complete phage production. An example here are the antibacterial proteins known as pyocins, which are related to bacteriophage tails[83].

Transposable elements (including insertion sequences, composite transposons, and complex transposons) allow for the movement of sequence to different sites within the genome. Intact transposable elements are able to move within a genome through the action of a transposase enzyme, which mediates excision (at inverted repeats on either end of the element) and insertion into another site. The most basic type of these is an insertion sequence, which consists of a transposase flanked by inverted repeats. While insertion sequences possess only the machinery required for transposition, transposons can also carry additional genes unrelated to mobility[47,94]. Transposable elements are mobile in the sense that they can move within the genome but can be found inserted into other elements potentially capable of horizontal gene transfer, such as an ICE or plasmid[89,95]. In *P. aeruginosa*, a number of transposable elements have been shown to be involved in AMR[89,95-97].

Integrons are accessory elements commonly associated with AMR. At their core, they possess an integrase gene followed by multiple gene cassettes (on the opposite strand from the integrase gene). The integrase gene contains an internal promoter driving the expression of these gene cassettes. Gene cassettes can exist as circular DNA fragments, with insertion into an integron driven by the action of the integrase[98,99]. While integrons themselves are not mobile, they are generally associated with transposons (intact or defective)[98,99] and can be commonly found as parts of larger mobile elements[88,95,100,101]. There are multiple classes of mobile integron, each characterized by the presence of a different integrase gene[98]. Integrons from classes 1 and 2 have been detected in *P. aeruginosa* strains[102], but those of class 3 have also been found in other *Pseudomonas* species[103,104]. Mobile integrons are evolutionarily linked with chromosomal

superintergrons found in a number of bacterial species (such as *Vibrio* species). These superintegrons can be >100 kb long and contain numerous cargo genes[99,105,106]. In *P. aeruginosa*, class 1 integrons are widespread and have been implicated in resistance to multiple classes of antibiotics[88,95,100,101]. An example from this dissertation is illustrated in Figure 2.3.

The role of plasmids in *P. aeruginosa* is less well studied than in Enterobacteriaceae[94], but they are still an important component of the accessory genome. *P. aeruginosa* plasmids have been long implicated in AMR and heavy metal resistance[107-109]. They can range in size from several kb[110] to megaplasmids well over 100 kb[74,111], and often harbor other mobile elements such as AMR integrons and transposons[74,95,112,113]. In some cases *P. aeruginosa* plasmids possess conjugation machinery allowing for horizontal transfer[74], while others lack their own conjugation machinery but can be mobilized by the presence of other plasmids[113]. Draft assemblies derived from short-read sequencing can often obscure the presence of plasmids in a given genome[114]. As long-read sequencing and complete genome assembly become more common, it is likely that many new *P. aeruginosa* plasmids will be described.

There are also genetic loci, known as replacement islands, which encode features common to most *P. aeruginosa* strains but possess variable gene content, placing them as part of the accessory genome[47,115]. Loci that exhibit this type of variation include those involved in synthesis of the lipopolysaccharide O-antigen[116], the siderophore pyoverdine[115], and the type IV pilus pilin[117]. These loci are under selective pressure to diversify (e.g. to avoid phage predation)[115] and can show signatures of horizontal gene transfer[115,117]. Another example that would fall into this category would be the contact-dependent inhibition gene *cdi1A*. Contact-dependent inhibition systems play a role in bacterial competition by delivering toxic effectors into neighboring cells. While the N-terminal portion of Cdi1A is highly conserved, it can possess

a variety of C-terminal toxin domains, which can presumably be swapped through horizontal gene transfer and recombinaiton[118]. One of these toxin domains was recently implicated in virulence in a mouse model of infection[70]. This type of AGE (with a variable domain in a larger conserved gene) would likely be missed in a gene-based examination of the accessory genome, highlighting the strength of taking a sequence-based approach to these analyses.

Genetic deletions can also contribute to the accessory genome. Chromosomal deletions, including large deletions (>100 kb), are known to occur in *P. aeruginosa*, a phenomenon that has been specifically (but not exclusively) noted in strains colonizing patients with cystic fibrosis[71,119,120]. Whether a sequence with variable presence secondary to chromosomal deletions would be considered as part of the accessory genome is dependent on both the prevalence of the deletion in the population being studied and the strictness of the core genome definition being used. As has been described, *P. aeruginosa* can withstand the disruption of all but several hundred genes[58] and using a very strict core genome definition only 665 genes were classified as core[52]. Many of the remaining genes common to *P. aeruginosa* were likely classified as accessory simply because they were missing in a few genomes. Still, there are important examples of gene loss contributing to variation between strains. A likely case where this has occurred is in the accessory gene *exoS*, which encodes a T3SS effector. It is well appreciated that the T3SS effector genes *exoS* and *exoU* are nearly mutually exclusive. Most strains possess one of these effector genes, but few carry both[121,122]. It is hypothesized that this is due to the excision of *exoS* in a recombination event that coincided with the acquisition of the original *exoU* genomic island, perhaps due to partial homology observed between repeats flanking the *exoS* gene and multiple *exoU* islands[77]. PA7-like strains, taxonomic outliers in the global *P. aeruginosa* phylogeny[123,124], also show evidence of gene loss compared to other members of the

species. These strains completely lack a T3SS locus, and, in one group, there is evidence of a scar containing remnants of T3SS loci genes[78]. It is notable that PA7-like strains are rare in *P. aeruginosa* collections[51,52], and if none or few of these are present in a given collection the T3SS locus would likely be classified as part of the core genome. A subset of these outlier strains possess the outer membrane channel-encoding gene *oprA*, but genomic evidence (conservation of surrounding genes and presence of remnant sequence in other strains) suggests that its absence in the rest of the species is through gene loss rather than specific acquisition by these strains[52,125].

P. aeruginosa population structure

Two large-scale phylogenetic studies have recently been conducted to examine the population structure of *P. aeruginosa* at a species-wide scale. In both of these studies, it was clear that the *P. aeruginosa* species can be divided into two major clades, with a small proportion of strains belonging to two to three outlier clades[51,52]. These clades also largely segregate by accessory genomic content, highlighting that the accessory and core genomes are not independent of each other[51]. While the accessory genome includes many elements with a mobile origin, and therefore would not necessarily be expected to follow the core genome phylogeny, it makes sense that this horizontal gene transfer has to occur in evolutionary time. Therefore, more closely related isolates would, to a certain extent, have a shared history of AGE gains and losses. This is an important concept to keep in mind when studying the accessory genome, showing that it cannot necessarily be considered as independent of the background genetic environment. Ozer et al. showed that a small subset of AGEs were highly discriminative between the two major groups. The most notable of these are two T3SS effector genes *exoU and exoS*. In one of these clades 98% of genomes are *exoS*+*exoU*-, while in the other 95% of isolates are *exoU*+*exoS*-. For the two major clades, within-clade core genome recombination also occurred at higher rate than

recombination between clades. This recombination barrier may be secondary to the clades occupying different niches. While Ozer et al. identified both core genome SNVs (including in signal transduction and metabolic genes) and AGEs (such as the T3SS effectors) that may be niche-adaptive, the extent to which these clades are environmentally separated is not clear at this time[51]. A recent study found that while *exoS*+ isolates were found in both natural and man-made environments, *exoU*+ isolates were found predominantly in man-made environments[16], perhaps supporting the conclusion that these respective clades occupy distinct but overlapping niches. The most distinct outlier clade showed an average nucleotide identity of 93-94% with other groups[51,52]. While they are more closely related to the other *P. aeruginosa* clades than to other *Pseudomonas* species, this places them on the border of being classified as a separate species[52]. PA7 is the prototypical isolate for this clade, which is notable for a deleted (or otherwise absent) T3SS[78,123,124]. The two major *P. aeruginosa* clades can clearly be seen in a phylogenetic tree of 115 isolates considered in Chapter 3. This collection also contains one PA7-like strain, which is apparent as an outlier in the tree (Figure 1.2). It is important to note that mid-point rooted phylogenetic trees (as shown here) make an underlying assumption that the genomes considered are evolving at the same constant rate[126]. In reality this may not be true, with evolutionary rate varying depending on the environment in which a given bacterium resides in combination with its genetic background. As such, while these trees are useful for comparing genetic similarity between isolates, caution should be taken in inferring evolutionary relationships.

In clinical practice, *P. aeruginosa* is often typed through a technique called multi-locus sequence typing (MLST), in which a strain is assigned to a given sequence type (ST) based on the alleles present in seven housekeeping genes[127]. Based on MLST, *P. aeruginosa* has been described as having a “nonclonal epidemic population structure”, indicating that while many

infections are caused by isolates from rare STs, a disproportionate number are caused by a few important clones[127-129]. In particular, specific clones are enriched as causes of acute healthcare-associated infections, while others are overrepresented in chronic infections in cystic fibrosis patients. As will be elaborated in the following section, this includes STs that are enriched for high levels of AMR[129-133].

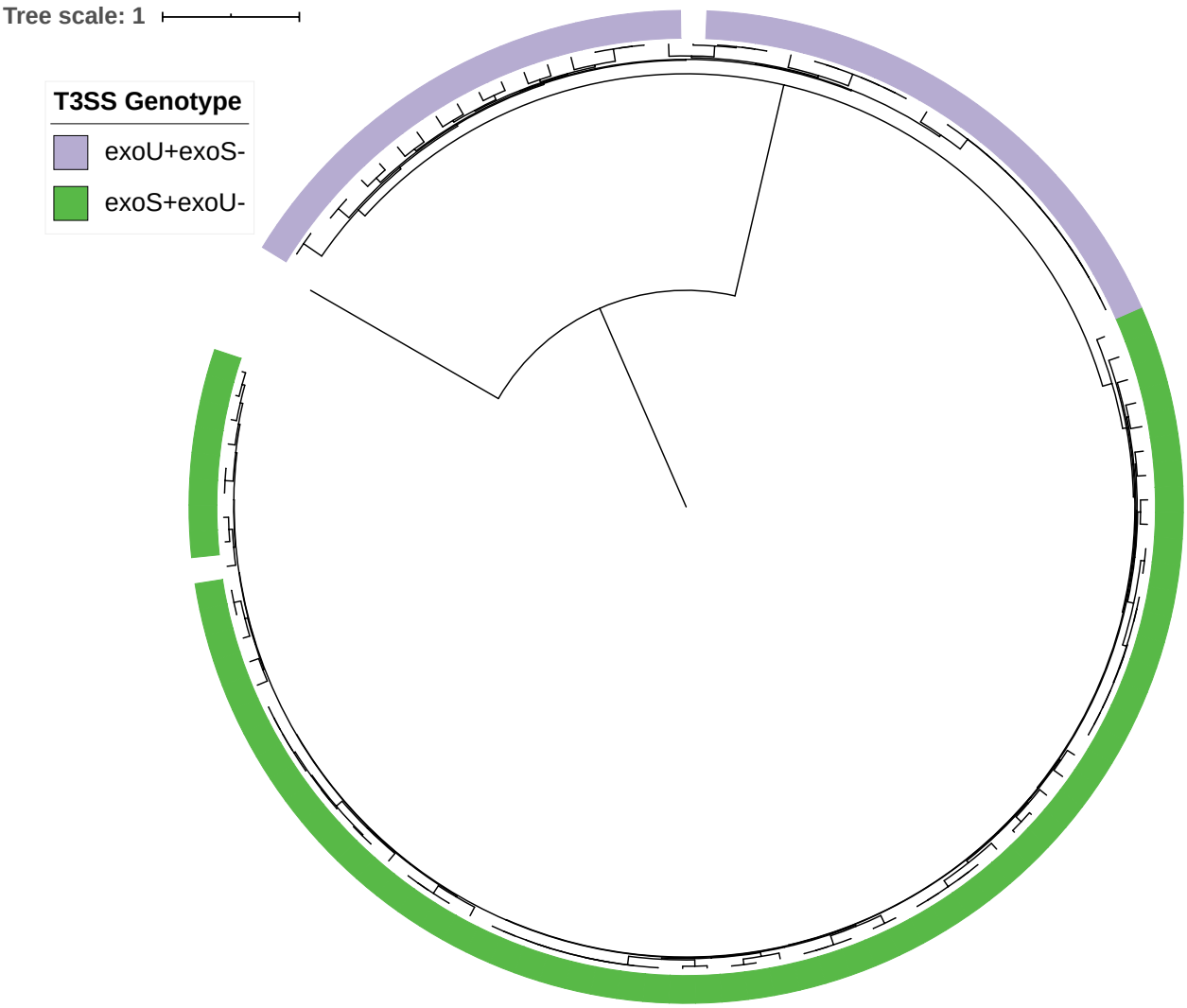


Figure 1.2 Major clades of *P. aeruginosa*. Mid-point rooted core genome phylogenetic tree of *P. aeruginosa* isolates constructed from SNV loci present in at least 95% of genomes, annotated with T3SS genotype. The two main clades of *P. aeruginosa* are apparent by their characteristic T3SS genotypes. One PA7-like outlier isolate is also present. The underlying phylogenetic tree here is also presented in Figure 3.2 and was constructed from the genomes of the 115 *P. aeruginosa* isolates forming the training set for machine learning analysis in Chapter 3.

Genetic determinants of antimicrobial resistance in *Pseudomonas aeruginosa*

Not only is *P. aeruginosa* able to cause severe or chronic infections as previously described, but it also shows a high propensity for AMR. *P. aeruginosa* has shown resistance to each of the classes of antibiotics used to treat the organism. Isolates can be classified as multidrug resistant (MDR) if nonsusceptible to at least one agent in ≥3 antipseudomonal classes, and extensively drug resistant (XDR) if nonsusceptible to at least one agent in all but ≤2 of the tested classes[134]. There have also been reports of panresistant organisms, which are not susceptible to any tested drugs[135-137]. In fact, there has even been evidence of resistance to a novel combination agent (ceftazidime-avibactam, a cephalosporin and β-lactamase inhibitor) in banked isolates collected before this drug was used clinically[138]. AMR poses as a challenge for treatment and increases the healthcare burdens of *P. aeruginosa* infections as MDR phenotypes are associated with worse patient outcomes and increased mortality[8,139-141]. Given all of this, it is unsurprising that in 2006 the Infectious Diseases Society of America named *P. aeruginosa* one of the pathogens in most dire need of new therapeutic options[142]. In 2017, the World Health Organization followed suit, classifying carbapenem-resistant *P. aeruginosa* as a top-priority pathogen for new drug development. In their recent 2019 report, the Centers for Disease Control and Prevention (CDC) stated that MDR *P. aeruginosa* is a serious public health threat, with an estimated 32,600 cases and 2,700 deaths in 2017. It is, however, promising that the disease burden of MDR *P. aeruginosa* is improving (with cases down 29% from 2012), perhaps due to rigorous stewardship and infection control measures[143]. AMR in *P. aeruginosa* can be intrinsic (an innate characteristic of the species), mutational, or develop through the acquisition of AMR genes[94,144-147]. As such, both the core and accessory genomes of *P. aeruginosa* play important roles in drug resistance.

Intrinsic resistance

The *P. aeruginosa* genome encodes for a variety of factors which reduce its susceptibility to antibiotics. One of these is the chromosomally encoded β-lactamase AmpC, whose gene is part of the *P. aeruginosa* core genome. AmpC is able to hydrolyze cephalosporins and its expression is induced in the presence of some β-lactams but at wild-type levels of expression does not confer clinical resistance to antipseudomonal β-lactams[53,144,148,149]. Another chromosomal β-lactamase, OXA-50, provides additional low-level resistance[150]. Additional chromosomal resistance genes found in *P. aeruginosa* include the *catb7* (chloramphenicol resistance)[151], *aph(3')-IIb* (aminoglycoside resistance)[152,153], and *fosA* (fosfomycin resistance)[154].

To exert bactericidal or bacteriostatic actions, antibiotics need to reach their targets in the bacterial periplasm or cytoplasm. One characteristic that makes *P. aeruginosa* less susceptible to antibiotics than other gram-negatives, such as *E. coli*, is relatively poor outer membrane permeability secondary to its slower primary outer membrane porin OprF[146,155]. Yoshimura et al. showed that transport across the outer membrane is the rate limiting step for *P. aeruginosa* (strain PAO1) to hydrolyze the cephalosporins cephacetrile and cephaloridine, and that this occurred at least 100 times slower than in *E. coli* K12[156]. *P. aeruginosa* also possesses a number of efflux systems that limit the accumulation of antibiotics (and other toxic compounds) within the bacterial cell. A notable example is the MexAB-OprM efflux pump, which decreases susceptibility to β-lactams and multiple other antibiotic classes (such as fluoroquinolones and tetracycline)[144,146,157,158].

Mutational resistance

One way in which *P. aeruginosa* strains can develop increased resistance to β-lactam antibiotics (particularly antipseudomonal penicillins and cephalosporins) is through mutations

affecting the structure or expression of the chromosomal β-lactamase AmpC. Mutations in the AmpC regulatory pathway can lead to substantial increases in gene expression, resulting in clinically significant levels of resistance[144,145]. A major cause here is loss-of-function mutations disrupting the negative regulator AmpD, as well as its homologues AmpDh2 and AmpDh3[159-162]. Mutations in the *ampC* gene itself can also confer resistance to these antibiotics[145]. For example, Berrazeg et a. examined 23 AmpC variants identified in 35 clinical isolates. When expressed in an AmpC deficient strain background, 20 of them resulted in elevated MICs against the cephalosporin ceftazidime compared to the wildtype protein[163].

The outer membrane porin OprD serves as the main channel by which carbapenems diffuse into *P. aeruginosa*[144,164,165]. It is therefore unsurprising that mutations which disrupt the function of OprD can increase carbapenem resistance and that these mutations develop spontaneously in the laboratory setting upon exposure to carbapenems[166]. Loss of OprD function is a major source of carbapenem resistance clinically and can occur through truncations, frameshifts, insertions (e.g. of an insertion sequence) and inactivating point mutations[145,167-170]. Upregulation of *mexT* (whose product drives expression of the MexEF-OprN efflux system) can also promote carbapenem resistance as it additionally represses *oprD* expression[144,171]. Along with mutational causes of reduced drug entry, mutations leading to overexpression of efflux pumps can also contribute to resistance to multiple classes of antibiotics, including carbapenems[144-146]. Indeed, MexAB-OprM efflux pump overexpression together with OprD disruption provides higher meropenem MICs than either change alone[166]. Examples of mutations resulting in MexAB-OprM overexpression include those causing loss-of-function of the genes encoding the negative regulators MexR and NalC[144,172,173].

The most common way that *P. aeruginosa* develops resistance to fluoroquinolone antibiotics is through mutations in their targets, specifically the DNA gyrase gene *gyrA* and topoisomerase IV gene *parC*[144,174,175]. Resistance-conferring mutations occur in a specific portion of each gene referred to as the quinolone resistance determining region (QRDR), a phenomenon that is conserved across different bacterial species[176,177]. Mutations in *parC* (e.g. S87L) are less common and often occur in combination with *gyrA* mutations (e.g. T83I)[175,178,179]. Experimentally, *parC* S87L and S87W mutations were not alone sufficient to raise the ciprofloxacin MIC of PA14, but when added onto a *gyrA* T83I background resulted in substantially higher resistance[178]. While overexpression of efflux pumps contributes to fluoroquinolone resistance, this is also more effective when coupled with existing *gyrA* mutations[178].

Acquired resistance

Horizontal gene transfer is a major source of AMR in *P. aeruginosa*. Integrons and transposons often carry genes encoding for a variety of resistance functions[88,89,95-97,100,101]. These acquired resistance genes can also be part of larger mobile elements, such as plasmids or ICEs[74,88,89,95,112,113]. This is notable because these acquired resistance genes often travel together, potentially allowing a previously susceptible strain to become MDR or even XDR in a single event. For example, the large conjugative plasmid pOZ176 harbors two class 1 integrons carrying cassettes encoding for β-lactam, aminoglycoside, and chloramphenicol resistance. Conjugal transfer of this plasmid was shown to increase MICs against carbapenems, cephalosporins, and an antipseudomonal penicillin/β-lactamase inhibitor combination in the recepient[95].

Acquired β-lactamases are an important source of resistance to multiple classes of β-lactams. The metallo-β-lactamases (including the various IMP, VIM, and NDM enzymes) are of particular concern as they confer resistance to carbapenems and are commonly found as part of accessory elements (such as integrons and plasmids) in carbapenem resistant isolates[95,102,112,129,132,180-186]. Acquired class D OXA β-lactamases are also found in *P. aeruginosa* in addition to the chromosomal OXA-50 described previously[150]. These can be narrower spectrum, with OXA-10, for example, providing resistance to piperacillin, but not ceftazidime or cefepime. However, certain variants (including OXA-10 derivates) show broader spectrum activity, decreasing susceptibility to antipseudomonal cephalosporins (like cefepime and ceftazidime)[187-190]. OXA-198, detected on a *P. aeruginosa* plasmid, was shown to have carbapenemase activity[100,191]. While the KPC enzymes are associated with carbapenem resistance in Enterobacteriaceae (particularly *Klebsiella pneumoniae*), they have also been found in *P. aeruginosa* and have been identified on transposons and plasmids[96,97,110,129,184,192]

Acquired aminoglycoside resistance is often conferred by integron or transposon-borne aminoglycoside-modifying enzymes, which inactivate aminoglycoside antibiotics with the addition of functional groups. Multiple classes of these enzymes exist, each modifying aminoglycoside antibiotics in a different way (phosphorylation, acetylation, or adenylation)[153]. Many class 1 and 2 integrons found in *P. aeruginosa* possess at least one acquired aminoglycoside resistance gene[102,129,182]. Integrons and transposons also often contain acquired resistance genes conferring resistance to compounds other than β-lactams and aminoglycosides. In fact, the 3' conserved sequence common to many class 1 integrons contains *sul1* (encoding a sulphonamide resistance protein) and *qacE*Δ1 (encoding a truncated but partially active small multidrug efflux protein conferring quaternary ammonium compound resistance)[98,99,193,194].

Other types of acquired resistance genes found on integrons and transposons include those involved in resistance to chloramphenicol, fosfomycin, and disinfectants (other *qac* genes)[47,101,129,182]. While resistance to fluoroquinolones generally develops from mutations in DNA gyrase and topoisomerase IV genes[144,174,175], this can also be conferred through acquired fluoroquinolone resistance genes. Examples include *qnrVC1* and *qnrVC6*, which have been detected in *P. aeruginosa* on an integron and plasmid respectively[195,196]. The mobile colistin resistance genes *mcr-1* and *mcr-5* have also been detected in *P. aeruginosa*, a concerning finding as colistin is considered to be one of the last lines of treatment for highly resistant infections[197,198].

Antimicrobial resistance and high-risk clones

As previously discussed, in clinical practice *P. aeruginosa* shows a nonclonal epidemic population structure. In particular, a disproportionate number of highly drug-resistant *P. aeruginosa* infections are caused by a small number of globally distributed STs. These are referred to as "high-risk clones". Some of the most successful high-risk clones in acute infection are ST235, ST111, and ST175[129,131,140,186,192,199-201]. Even within a high-risk clone a variety of resistance mechanisms exit (both mutational and acquired)[129,132,174], and not all isolates identified as part of each clone are MDR[65,130,199,200]. This suggests that it is not simply the acquisition of resistance that allows a given clone to become epidemiologically dominant. This conclusion is supported by the fact that isolates from rare STs can also show MDR or XDR phenotypes[65,174] while few gain the prevalence needed to be classified as high-risk clones. As a single acquisition of AMR does not appear to be sufficient for the establishment of a high-risk clone, the success of these strains may instead be due to a propensity to acquire resistance, factors which make them

better able to cause infections or persist in the hospital environment, or other unknown characteristics.

ST235 is perhaps the best studied high-risk clone responsible for MDR acute infections. It a common (if not the most common) cause of MDR or XDR *P. aeruginosa* infections identified in studies from various countries[130,182,192,200-202]. This ST is *exoU*+, and accordingly can be found in the corresponding major *P. aeruginosa* clade[70]. In 2018, Treepong et al. published a genomic analysis of a collection of geographically and temporality diverse ST235 isolates. Through phylogenetics, they estimated that the last common ancestor of this clone occurred in approximately 1984. Additionally, these isolates possessed a wide variety of acquired resistance genes (particularly those conferring aminoglycoside and extended spectrum β-lactam resistance)[132]. This is consistent with the literature as a whole, where ST235 isolates have been observed with incredibly diverse sets of acquired resistance genes[129]. While a large proportion of the ST235 isolates in Treepong et al.'s study possessed conserved mutations in *gyrA* and *parC* contributing to fluoroquinolone resistance, mutational causes of β-lactam resistance were more diverse. The genes *ampD* and *oprD* were commonly affected, but through a number of different mutations[132].

ST175 has caused regional epidemics of XDR *P. aeruginosa* infections in Europe, particularly in Spain and France[140,199-201], though isolates have also been detected in Japan[182] and the United States[65]. In fact, ST175 isolates made up a majority (62/81) of XDR isolates in a large collection of bacteremia isolates from Spain[129]. In contrast to ST235, ST175 genomes contain *exoS*[140], placing them in the other major *P. aeruginosa* clade. Unlike ST235, where both mutations and acquired genes play a major role in AMR, much of the resistance in ST175 appears to be mutational in origin. In an analysis of 22 ST175 isolates from Spain and France,

Cabot et al. observed largely conserved causes of mutational resistance (with many shared between 80-100% of isolates). This included mutations resulting in AmpC overexpression, truncation of OprD, fluoroquinolone resistance (*gyrA* and *parC* mutations), and MexXY efflux pump overexpression. These isolates did not possess acquired β-lactamases, but did all possess at least one acquired aminoglycoside resistance gene[168]. On the other hand, a more recent study of carbapenemase-producing *P. aeruginosa* in Spain identified 33 ST175 isolates possessing at least one acquired β-lactamase. The genomes of these isolates contained an average of 5.9 acquired resistance genes[184]. Together with high levels of preexisting mutational resistance, increasing acquired resistance mechanisms may make ST175 even more clinically problematic.

Genetic determinants of virulence in *Pseudomonas aeruginosa*

A major impetus to study *P. aeruginosa* is its ability to cause severe, and even fatal, disease in humans. As has been stated, bloodstream infections with *P. aeruginosa* are associated with higher mortality than either *S. aureus* or other gram-negative bacteria[7]. In many studies of *P. aeruginosa* bacteremia or pneumonia, mortality rates are near or above 30%[6,8,9,203]. This raises the question of how *P. aeruginosa* is able to cause such severe disease. In other words, what are the factors which determine its virulence? Better understanding the pathogenicity of *P. aeruginosa* may allow for the development of anti-virulence therapies (such as T3SS or quorum sensing inhibitors), which could serve as adjuncts to traditional antibiotics[204-206].

P. aeruginosa possesses a variety of systems that enable it to infect a eukaryotic host[2-4]. Virulence, however, is not a simple phenotype. *P. aeruginosa* strains have been shown vary widely in their pathogenicity in different infection models, including mice, *Drosophila melanogaster*, *Galleria mellonella*, *Caenorhabditis elegans*, and plants[24,30,31,70,207]. Studies show

that virulence in *P. aeruginosa* is complex and combinatorial, with multiple genetic factors (in both the core and accessory genomes) playing a role in increasing or decreasing its pathogenicity[24,28,31,57,70,75,76,207,208]. Further, the ability of a given strain to cause disease can vary between infection models[24,30,209], and genes important in some models are dispensable in others[29,57]. This shows that while bacterial factors are obviously important in pathogenesis, interactions with hosts are necessary components of disease. From a clinical context, a mechanically ventilated patient in the intensive care unit, a neutropenic patient, and a patient with cystic fibrosis may all be predisposed to *P. aeruginosa* infections for different reasons, and different *P. aeruginosa* virulence factors may be important in each of these scenarios. For example, while a functional T3SS is associated with increased mortality in acute infection[203], chronic infection in cystic fibrosis patients is associated with the loss of secretion[210,211]. Clearly, large-scale genomics studies are needed to disentangle how different components of the *P. aeruginosa* pangenome control its pathogenicity.

As indicated, the ability of *P. aeruginosa* to cause disease is not simply due to factors intrinsic to the bacterium itself. Infection involves complex interactions between the bacterium, the host, and the broader environment. This is clear in cystic fibrosis, where impaired mucociliary clearance creates a permissive environment for colonization by *P. aeruginosa*. Mechanical trauma, such as occurs during mechanical ventilation, can provide a route of entry by which *P. aeruginosa* can establish infection and medical devices can serve as foci for biofilms. Historically, neutropenic patients are at high risk for *P. aeruginosa* infection, highlighting the role of the innate immune system in defense against this bacterial pathogen[2-4] While acknowledging the importance of the host in infection, this dissertation focuses on role of factors encoded by the *P. aeruginosa* genome. In this section, I will provide a brief overview of several

of the factors influencing virulence in *P. aeruginosa*, with an emphasis on those contained within the accessory genome.

P. aeruginosa virulence factors

As indicated, there are a number of factors that contribute to the pathogenicity of *P. aeruginosa*[2-4]. Two of these, type IV pili and flagella, play important roles in both adhesion and motility. The average *P. aeruginosa* strain produces multiple type IV pili, which mediate twitching motility across solid surfaces and act as adhesins[2,212-214]. Pili-mediated adhesion and motility are both involved in biofilm formation[2,212,215,216], and pili can act as mechanosensors (modulating downstream gene expression upon surface attachment)[217,218]. Nonpiliated mutants are attenuated in both murine pneumonia and keratitis models[219,220]. Flagella mediate swimming motility and are also involved in initial surface attachment (and therefore biofilm formation)[2,215,221-223]. Flagella are required for virulence in a murine pneumonia model[224] and are likewise important in an acute burn model but are dispensable in a chronic wound model[57]. In the acute burn wound model both flagellar motility and glycosylation (which contributes to the high immunogenicity of these structures) play a role in virulence[225]. It is notable that both of these systems involve replacement islands, with pilin genes and flagellar glycosylation loci varying between strains[47,117,226]. Another virulence factor produced by *P. aeruginosa* is pyocyanin, a secondary metabolite which induces oxidative damage to host cells and which has been shown to increase virulence in a murine pneumonia model[2,227,228].

The *P. aeruginosa* genome encodes for a variety of secretion systems, many of which are implicated in pathogenicity[229]. The Xcp type II secretion system (T2SS) of *P. aeruginosa* secretes several toxins, including exotoxin A, two elastases, and several other proteases[229-232]. A functional Xcp T2SS was sufficient to cause lethal disease in a Toll-like receptor 2 and 4

knockout mouse model, though T2SS-mediated mortality was delayed relative to T3SS-mediated mortality[233]. Exotoxin A is an adenosine diphosphate-ribosyltransferase which compromises host cell processes by deactivating elongation factor 2 and has itself been associated with increased mortality in mice[2,229,234]. Two other T2SS have been identified in *P. aeruginosa*. The first of these is the Hxc system, which is involved in the bacterium's response to low phosphate conditions[229,235,236]. The third T2SS, termed Txc, is found in PA7-like strains and secretes a chitin-binding protein[124,237]. *P. aeruginosa* also produces three type VI secretion systems (T6SS), termed H1, H2, and H3. Type VI secretion is known to be important for interbacterial competition, but the H2- and H3-T6SS have also been shown to play roles in virulence[229,238-242]. For example, the copper-scavenging effector Azu of the H2-T6SS is important for full virulence in a mouse model of pneumonia, potentially by increasing bacterial fitness in a copper-starved environment[241].

The T3SS is a potent and well-studied *P. aeruginosa* virulence factor[2,4,38]. As previously noted, it is present in most *P. aeruginosa* strains, but absent in the PA7-like outlier clade[16,52,78,121,122,124]. Whether it would be considered a core or accessory virulence factor therefore depends on the strictness of the core genome definition being used. The T3SS is a needle-like structure (evolutionarily linked with flagella) that injects effector proteins directly into the cytoplasm of a host cell[38,243]. Secretion is thought to be contact-mediated, but in vitro can be induced by calcium depletion[38,244,245]. The *P. aeruginosa* T3SS is important to pathogenicity in multiple model systems, including mice, *G. mellonella*, and amoebae[25,39,246-249]. Functional secretion appears to be important in acute infection outcomes in humans, with Hattemer et al. finding that secretion (as defined by detection of at least one effector or the secretion-pore forming PopB or PopD in vitro) was associated with higher mortality in a

multivariate analysis[203]. Four major T3SS effectors have been characterized: ExoU, ExoS, ExoT, and ExoY[38]. The *exoT* and *exoY* genes are present in almost all T3SS-possessing strains. The *exoU and exoS* genes, on the other hand, are clearly contained in the *P. aeruginosa* accessory genome (being present in approximately 70-80% and 20-30% of strains respectively) and will be discussed further with other accessory genomic virulence factors[16,38,121,122]. ExoT possesses GTPase activating protein and ADP-ribosyltransferase domains, which ultimately result in cytoskeletal disruption and proliferation inhibition in the target cell[38,250,251] . It plays a smaller role in virulence than either ExoU or ExoS. In a mouse model of pneumonia, its presence was sufficient for dissemination to organs but showed a similar 50% lethal dose as a secretion-negative mutant[25]. It does, however, mediate apoptosis and inhibition of would repair in vitro[252,253]. ExoY is an adenylate cyclase and minor virulence factor[38,254]. In the absence of other known T3SS effectors, it contributes to virulence in a mouse keratitis and rat pneumonia model[255,256].

Virulence and the accessory genome

As has been discussed previously, a number of genomic islands have been implicated in *P. aeruginosa* virulence[24,47,75,76,82,92,93]. Two notable examples are PAPI-1 and PAPI-2, which were first identified in the well-studied strain PA14[75]. Deletion of PAPI-2 decreases virulence in mouse models of both bacteremia and pneumonia. While PAPI-1 was only independently required for virulence in the bacteremia model, a dual PAPI-1/PAPI-2 deletion mutant showed further attenuation beyond the single PAPI-2 deletion in pneumonia[76]. PAPI-1 is a large (108 kb) transmissible ICE while PAPI-2 is a smaller (11 kb) *exoU*-containing island[75,77,86]. While much of the virulence conferred by PAPI-2 is attributed to *exoU*, studies suggest that it is not the only important gene in the island[75,76]. Multiple ORFs within PAPI-1 have been shown to be involved

in pathogenicity against mice or plants, but many of these are poorly characterized[75]. A chaperone-usher pilus and two two-component regulatory systems on this island are involved in both pathogenicity and biofilm regulation (both formation and dispersal) in PA14[75,257,258]. Ferrara et al. recently identified a PAPI-1-encoded small RNA (PesA) that enhances pyocin S3 production and increases cytotoxicity (with increased target cell viability observed in a *pesA* deletion mutant)[259]. Battle et al. identified an additional PAPI-1-like island, termed PAGI-5, that also contributes to pathogenicity in a mouse model of pneumonia[24]. Other genomic islands with known roles in virulence include the cytotoxin-encoding prophage φCTX[93] and three prophages and one ICE-derived element identified in the cystic fibrosis isolate LES that were required for full virulence in a rat chronic pneumonia model (measured through competitive index between wildtype and mutant strains in coinfection)[47,82].

Along with being clade-defining accessory genes[51], *exoU* and *exoS* both encode for T3SS effectors and play an important role in the virulence of strains that contain them[38]. The *exoU* gene is found in a variable genomic island (with variants including PAPI-2 and ExoU islands A-C)[77]. ExoU is a phospholipase which causes membrane disruption in the target cell, leading to high levels of cytotoxicity[38,260-264]. Unsurprisingly, this results in increased injury and mortality in an infected host[25,248,261,264,265]. It also may play a role in outcomes during human infection, with Peña et al. finding that, in a cohort of patients with *P. aeruginosa* bacteremia, *exoU*+ isolates were associated with early morality (a higher proportion of deaths within 5 days)[140]. The *exoS* gene, on the other hand, is not in a genomic island and, as previously discussed, may have been lost in *exoU*+ isolates as part of a recombination event during the acquisition of the *exoU* island[47,77,266]. It is homologous with ExoT and, like ExoT, also functions as both a GTPase activating protein and an ADP ribosyl transferase[38,267-269]. While ExoS is independently

associated with virulence in mice, its impact is smaller than that of ExoU in both a pneumonia and abscess model[25,248]. However, *exoS*+ strains can be just as or more virulent than *exoU*+ strains in a mouse model of bacteremia, suggesting that there must be other factors as play shaping their pathogenicity[70].

While PA7-like strains lack a T3SS, certain members of this clade possess their own characteristic toxin, exolysin A[78,124,270]. Exolysin A is a potent pore-forming toxin that was first recognized in a PA7-like isolate from a patient with hemorrhagic pneumonia[270,271]. This isolate was highly virulent in mice (causing 100% mortality in a pneumonia model at a dose of 5 x 10^6 CFU) and showed in vitro cytotoxicity similar to that of an *exoU*+ strain, challenging the idea that *P. aeruginosa* requires a T3SS for full virulence[270]. Exolysin A is a two-partner secreted protein, and insertion of the pore requires *P. aeruginosa* to be in contact with the target cell (as evidenced by necessity of type IV pili for cytotoxicity and lack of cytotoxicity in a transwell experiment)[270,271]. When the two-partner system (*exlBA*) was cloned into a T3SS-deficient PAO1, it increased both cytotoxicity in vitro and mortality in mice[270]. Orthologs of the *exlA* gene have also been found in other *Pseudomonas* species[272].

With the large size of the *P. aeruginosa* pangenome[31,51,52], it is almost certain that there are virulence-influencing AGEs that have yet to be described. Dr. Jonathan Allen in the Hauser laboratory recently completed a large-scale pan-accessory genome screen to identify novel AGEs contributing to virulence in a mouse model of bacteremia. Of 15 AGEs he identified as potential virulence factors, 11 were shown to be truly involved in virulence (with deletion mutants attenuated in the mouse model). One of these, a C-terminal toxin domain of the contact-dependent inhibition protein Cdi1A, was selected for detailed analysis. This domain was shown to have tRNase activity, which was necessary for its role in pathogenicity. Further, the AGE was

important for virulence in multiple mouse models (bacteremia, pneumonia, and subcutaneous infection). The other virulence AGEs identified in this study were largely poorly characterized, but had features suggestive of association with genomic islands, phages, or transposons. One virulence AGE had homology to rearrangement hot spot-type polymorphic toxins[70]. Vasquez-Rifo et al. recently performed a similar screen for accessory genes influencing virulence in a *C. elegans* infection model, examining genes both positively and negatively associated with virulence. They identified two genes which lowered virulence in this model (*qsrO* and *tegN*), with longer *C. elegans* survival when these genes were overexpressed. They also found that the presence of an active CRISPR system correlated with increased virulence. As CRISPR systems would act to limit the acquisition of horizontally acquired DNA, this may suggest that, while certain AGEs confer increased virulence, they may have a general tendency to reduce virulence during acute infection[31].

Machine learning analyses for the prediction and exploration of bacterial phenotypes

In Chapter 3, I take a machine learning approach to predict the virulence level of *P. aeruginosa* isolates based on the content of their genomes. As has been discussed, virulence is a complex phenotype with many contributing factors and varies between strains within the species[24,30,31,70,207]. With that in mind, we chose machine learning as a tool to study the relationship between the bacterial genome and virulence, as it provides a way to deal with this complexity and has been successfully applied to a number of problems in bacterial genomics[66,273-283]. In this section, I provide an overview of basic machine learning concepts and discuss its applications in the field of bacterial genomics.

Machine learning concepts

Simply, machine learning is a process of building computational models that are "fit" to a training dataset. This training dataset contains a number of samples (e.g. bacterial genomes), each with features (e.g. gene presence/absence) and potentially labels based on the condition one is trying to predict (e.g. AMR). In supervised machine learning, a model is fit using a labeled training dataset which can then be used to predict the labels of new samples based on their features. In unsupervised machine learning, a model is fit using an unlabeled training dataset. Common uses of unsupervised machine learning include clustering and dimensionality reduction. Here I will focus on supervised machine learning as it is generally more relevant to the problem of phenotype prediction (with the phenotype serving as the label for each sample). Supervised machine learning can be broken down further into classification, in which the labels are a categorical variable (e.g. susceptible vs. nonsusceptible in AMR prediction), and regression, in which the labels are a continuous variable (e.g. minimum inhibitory concentration in AMR prediction)[273,284,285].

For a machine learning model to be useful, one must have an idea of how well it performs. For a classification model, it is possible to ask the question of how well the model is able to separate the training samples into their true classes. While this training performance is useful to know, it may overestimate how well the model would predict the class of new samples. As the model was trained to optimally separate training samples into classes based on their features, it may be "overfit" to intricacies of the training set that do not hold true for other populations (fitting on noise, rather than signal). The best way to assess model performance is with an external test set, a labeled dataset independent of the training dataset. Predicting the labels of the test set and comparing to the true labels allows one to determine how well the model

generalizes to new data (at least the data contained in the test set). An alternative (or adjunct) to an external test set is cross-validation. In k-fold cross-validation, the training set is split into k (e.g. 10) non-overlapping subsets (termed cross-validation folds). For each cross-validation fold, all training set samples not set aside as cross-validation are used to train a model. The labels of each sample in the cross-validation fold are then predicted with the model and compared to their true values. By considering performance across all cross-validation folds, one can estimate how well a model built using this training dataset would generalize to new data[284,285]. Two caveats of cross-validation are that it does not build a final model that can be applied to new samples (only providing an estimate on how it would perform) and all cross-validation samples are drawn from the same population as the training samples (i.e. the complete training dataset). If test samples originated from a population that differs from the training dataset, model performance may also differ (likely decreasing).

In Chapter 3, I make use of the common supervised machine learning algorithms random forest, logistic regression (with L2 and elastic net regularization), and support vector machine. In each of these algorithms there are hyperparameters, variables which cannot be learned during model fitting and must instead be defined by the user (such as the regularization component "C" in logistic regression). The optimal (highest performing) combination of hyperparameters may vary depending on the dataset being used[284,286]. One way to identify this optimal combination is by testing all possible combinations of hyperparameters. In order to prevent overfitting to a specific dataset influencing hyperparameter choice, this can be performed within a cross-validation loop in a process termed "grid-search cross-validation". A final model can then be built using all training data and the best performing hyperparameters set from grid-search cross-validation. The performance of this final model can then be tested against an independent test

set. Alternatively, grid-search cross-validation could be performed inside an outer cross-validation loop (a process termed "nested cross-validation") to estimate how well this approach would generalize to new data[284].

When both training and evaluating a model, it is important to have a performance metric in mind. Accuracy can seem a natural choice, as one of course wants to build a highly accurate model. However, there may be cases where another metric is more important (e.g. high sensitivity so as to not miss patients with a rare disease in a medical screening test). Additionally, there may be times when accuracy is a poor metric of model performance, such as when there is substantial class imbalance in the dataset. In an AMR classification problem, if almost all samples in the training dataset were susceptible, a model which simply predicts "susceptible" in all cases would be highly accurate in cross-validation even though it had learned no meaningful information about genetic contributions to AMR. It would likely generalize poorly to a test set with more nonsusceptible isolates. An alternative metric that can be used in cases of class imbalance is the F1 score, the harmonic mean of sensitivity and positive predictive value[284,285]. For the F1 score to be high, a model must both identify the majority of samples in the class of interest (high sensitivity) and the majority of samples predicted to be in the class of interest must truly be in that class (high positive predictive value). The F1 score can range from near 0 (either or both sensitivity and positive predictive value are very low) to 1 (both sensitivity and positive predictive value are perfect). F1 score is technically undefined if both sensitivity and positive predictive value are 0, but in machine learning implementations[287] is often set to 0 in this case to allow analysis to continue. Another commonly used performance metric is the area under the receiver operating characteristic curve (AUC), which examines the relationship between the true positive rate and false positive rate in a model if the threshold for choosing which class to predict

is varied. A completely random model would show an AUC of 0.5 while a perfect model would show an AUC of 1[284].

In this dissertation I make use of the scikit-learn[287] suite of machine learning tools implemented in the Python programming language, which has been employed in multiple other bacterial genomics studies[276,277,288,289]. Computational biologists also often make use of a variety of other machine learning libraries, including the Python-based XGboost (for gradient-boosted trees)[275,290-292] and keras (for neural networks)[277,288] and the R-based caret (multiple algorithms)[281,293,294] and randomForest (for random forest models)[282,295,296]. Further detail on the machine learning approach I take in this dissertation is described in Chapters 3 and 5, with a graphical summary in Figure 3.4.

Machine learning for AMR prediction

A major focus of studies using machine learning to predict bacterial phenotypes from genomic data has been in the field of AMR prediction, with goals of classifying strains as resistant or susceptible to a panel of drugs, identifying novel resistance determinants, or both. These studies have approached the problem of AMR prediction from different angles, utilizing a number of different machine learning algorithms (support vector machines, logistic regression, neural networks, and gradient-boosted trees) and encoding genetic information in a number of different formats (such as gene presence or absence, SNVs, k-mers, and even gene expression levels)[66,274-279]. Here, I highlight a few examples that are notable for their methodology and what they reveal about how machine learning can be applied to bacterial genomics problems.

Khaledi et al. used a support vector machine approach to build predictive models of AMR in a collection of 414 *P. aeruginosa* isolates. They constructed models to predict resistance or susceptibility to ciprofloxacin (a fluoroquinolone), tobramycin (an aminoglycoside),

ceftazidime (a cephalosporin), and meropenem (a carbapenem). To train these models, they used three types of genomic information both individually and in combination: SNVs (capturing core genome information), gene presence or absence (capturing pangenome variation), and gene expression. They took a nested cross-validation approach to estimate the performance of models built with each antibiotic and feature set combination. Models trained on the best-performing combination of genomic features all had F1 scores between 0.82 and 0.92[66]. There are two areas of potential weakness in this study. First, it would have been ideal to validate the performance of a final model for each antibiotic (with the best combination of genomic features) against an independent test set drawn from a different population. This would confirm that their models are truly generalizable. Second, they excluded isolates with intermediate resistance from analysis, simplifying the problem and potentially overestimating the performance of their models in a real-life clinical scenario. Still, Khaledi et al. show that machine learning techniques can be successfully applied to predict phenotypes in *P. aeruginosa*.

The types of genomic feature sets that were most predictive of resistance, and the individual features that were most important in making those predictions (i.e. with the highest feature weights in the support vector machine model), in Khaledi et al.'s study can further cement our understanding of AMR in *P. aeruginosa*. SNVs, for example, were most effective in predicting resistance to ciprofloxacin, with SNVs in *gyrA* and *parC* weighted highest. As appreciated by the authors[66], mutations in these genes are the major source of fluoroquinolone resistance in *P. aeruginosa*[174,175]. For the other drugs, a combination of gene presence and gene expression was most predictive of resistance. For meropenem, disruptions in and gene expression of the outer membrane porin-encoding *oprD* were both important in predicting resistance[66], consistent with the important role OprD plays in carbapenem entry into the bacterial cell[165].

Chen et al. built predictive models of AMR using a collection of 3601 *Mycobacterium tuberculosis* isolates, considering a relatively small set of 222 genomic features. Models built using a complex "wide and deep neural network" strategy showed high performance, with a mean AUC of 0.953 across all 10 drugs tested. However, model performance was almost identical using the far simpler L2-regularized logistic regression algorithm (mean AUC 0.949)[277]. This shows that a more complex algorithm does not necessarily yield a better-performing model, an important point to consider when embarking on a machine learning project.

While the majority of studies have focused on classifying strains as susceptible vs. nonsusceptible (or resistant), it is also possible to treat AMR prediction as a regression problem. Nguyen et al. have applied this approach to predict minimum inhibitory concentrations (MICs) of antibiotics for both *Klebsiella pneumoniae* and *Salmonella* species. For both of these populations, they were able to predict the correct MIC (within 1 two-fold dilution, which they note is consistent with clinical practices) over 90% of the time for must drugs (15/20 for *K. pneumoniae* and 15/15 for *Salmonella)*[275,292].

Additionally, Nguyen et al. recently showed that susceptibility or resistance can be accurately predicted in multiple bacterial species (*K. pneumoniae*, *M. tuberculosis*, *Salmonella enterica*, and *S. aureus*) even when genes known to be involved in AMR are explicitly excluded from the analysis. Further, these results held true even when clonality of the training genomes (i.e. reducing the weight of training isolates in common clades) and the frequency of resistance within each clade were taken into account during the model building process[291]. This is significant, as it shows that the genomic signal predictive of AMR is not limited simply to features directly causing resistance. Instead, genomic features causing resistance are likely

correlated with a number of noncausal features that can still be used for the purpose of prediction.

Břinda et al. took a different approach to predict AMR in *S. pneumoniae* and *N. gonorrhoeae*. They first constructed a database of genomes for each species based on their genomic content and AMR class. Rather than directly predicting a new isolates resistance from its genomic content, they identified closely related strains in the database ("genomic neighbors") and assigned resistance predictions based on the labels of its closest neighbors. At least in these species, simply identifying the closest neighbors of an isolate generally allowed accurate prediction of resistance (with AUCs ranging from 0.80 to 0.98 for different species-drug combinations). A proposed benefit of this approach is that it could be applied rapidly even while isolate is still being sequenced (using long-read nanopore technology). This could allow resistance predictions to be made in a clinically actionable time-frame[279]. Along with Nguyen et al.'s findings, this shows that a model does not have to directly capture the cause of AMR to be an effective predictor[279,291].

Machine learning for other bacterial phenotypes

To our knowledge, no previous machine learning study has examined genomic contributions to virulence in *P. aeruginosa*. Still, machine learning approaches have been applied to bacterial genomics problems outside of AMR prediction, including those directly relating to bacterial pathogenicity. A relatively early example is a study published in 2014 by Laabei et al. looking at cytotoxicity in 90 ST239 *S. aureus* isolates. Using a pre-defined set of 52 SNVs and indels that were identified as highly associated with cytotoxicity in their collection, they found that a random forest approach was successful in classifying isolates by level of cytotoxicity (high, medium, or low). A model trained on 60 isolates was able to correctly identify the class of

27/30 test isolates (only misclassifying 3 medium cytotoxicity isolates as "low")[296]. However, as their features had been pre-selected to separate strains by cytotoxicity in their dataset before it was split into two training and test sets, it is likely that their model was overfit to that dataset and that performance would decrease if the model was applied to truly novel isolates. In my opinion, this highlights the care that needs to be taken in interpreting the results of machine learning analyses reported in the literature.

An area where machine learning techniques have been repeatedly applied is to questions of host tropism and pathogenic potential of bacterial isolates[280,283,297-299]. Lupolva et al., for example, used pangenome content to train support vector machine models to classify the host source of *E. coli* (human or bovine) and *S. enterica* serovar Typhimurium (human, avian, bovine, or swine) isolates. The models developed could predict the proper host of *E. coli* isolates 83% of the time, while for *S. enterica* isolates accuracy ranged from 67% to 90% depending on the host source[283]. Something that has likely contributed to success of these studies is that host associations (or type of infection caused) in the studied organisms often cluster phylogenetically, resulting in a clear genetic signal that could be detected by the machine learning approaches employed. At a more high-level scale, investigators have examined whether it is possible to use genome content to separate pathogens from nonpathogens[298,299]. A particularly ambitious study by Barash et al. examined the genomes of 17,881 pathogenic and 3274 nonpathogenic human-colonizing bacteria. They found that a support-vector machine approach was highly successful in predicting pathogenicity, with an average F1 score of 0.897 in cross-validation. Their definition of pathogenicity, however, was quite broad and defined based on purely metadata associated with each of the genomes. No *P. aeruginosa* genomes were classified as nonpathogenic[299]. In an

adjacent field, machine learning has been applied to predict host tropism of bacteriophages from viral sequence with an AUC of >0.85 for nine bacterial genera[300].

Studies have begun to look at whether the genome of an infecting isolate can be used to predict patient outcomes. Recker et al. focused on patient mortality caused by two *S. aureus* clonal complexes (CC22 and CC30). By looking at each clonal complex individually they were able to limit the amount of genetic variation in their dataset, allowing them to consider any SNV in a gene (if nonsynonymous and not in a mobile genetic element) or intergenic region as a single feature. Using random forest models, they found that that predictive models based on these genetic features had AUCs of 0.75 (CC22) and 0.79 (CC30). This suggests that bacterial genetic variation, at least in closely related groups of isolates, may be moderately predictive of patient outcomes, but the authors did not further confirm this using an external test set[282]. A recent study conducted by Lapp et al. highlights the challenges of trying to predict clinical outcomes, finding that bacterial genomics, patient characteristics, and the combination of the two were all only weak predictors of whether a given *Klebsiella pneumoniae* isolate was identified as colonizing or infecting a patient[294]. Further studies are needed to elucidate the extent to which the bacterial genome (alone or in combination with patient factors) can be used to predict patient outcomes, as well as how this varies by infecting pathogen, type of infection, and outcome in question.

While generally using established machine learning frameworks[287,290,293], the studies I have described largely employed their own pipelines for data processing, analysis, and interpretation. This creates a barrier to entry to biologists with limited computational or machine learning experience. With that in mind, some groups are beginning to focus on making more universal and user-friendly machine learning applications specifically for problems related to

bacterial genomics. One such example this is in a recent publication by Lees et al., where the authors have added a machine learning pipeline to the existing bacterial genome-wide association study (GWAS) software pyseer. This pipeline builds elastic net models from genomic data (optionally weighted by population structure) that can be used to both predict phenotype in new samples and estimate the heritability of the phenotype in question (the extent to which it is controlled by the genetic information supplied). They applied their approach to number of previously published datasets with phenotypes including AMR and disease vs. colonization. The authors state that their approach can be used without coding experience, but it does still require comfort with a command line interface and ideally use of an external bioinformatic tools to generate the genomic feature set[301]. I expect future packages by Lees et al. or others will push the barrier to entry even lower, likely leading to large increases in the number of studies incorporating machine learning techniques into their analyses.

Introduction to the current work

Over the past decade, the field of bacteriology has undergone a sea change with rapid increases in the number whole genome sequences following advances in next-generation sequencing technology. This has included a large effort in the Hauser laboratory to sequence *P. aeruginosa* isolates, both from Northwestern Memorial Hospital and from other sites. However, the wealth of genomic information this has generated poses new challenges in both analysis and interpretation and has necessitated a revolution in the way that we think about bacterial research. To this end, I present two large-scale comparative genomics studies examining antimicrobial resistance and virulence in *P. aeruginosa*. In Chapter 2, I take a phylogenomic approach to uncover a prolonged epidemic of a highly resistant *P. aeruginosa* subclade at NMH and describe

a large, novel plasmid present in many of these strains. In Chapter 3, I show that machine learning techniques can be used to predict the virulence of *P. aeruginosa* isolates from genomic information.

CHAPTER 2

Identifying and characterizing a prolonged local epidemic of extensively drug-resistant *Pseudomonas aeruginosa* at Northwestern Memorial Hospital

Chapter-Specific Acknowledgements

The work described in this chapter was primarily conducted as part of Pincus et al. 2019[302]. This was a collaborative project, and the work conducted by my coauthors was essential to its completion. In particular, Dr. Kelly Bachta performed antimicrobial susceptibility testing to determine minimum inhibitory concentrations (MICs) of the described antibiotics for each isolate. Dr. Egon Ozer performed the complete genome assembly of PABL048 from the combined long- and short-read sequence data and performed some of the *in silico* sequence typing and draft genome assembly for this project. Additionally, I used multiple scripts written by Dr. Ozer to perform some of the genomic analyses in this study. Though not exhaustive, these included several scripts for generating multiple sequence alignments based on short reads and/or assembled genomes and subsequent filtering and masking steps as well as the script used for *in silico* sequence typing. Dr. Jonathan Allen's observation of a cluster of AMR genes present in a subset of genetically related isolates from Northwestern Memorial Hospital was the catalyst that started this project, and his input early in the project identified which strains possessed these AMR genes. Much of the short-read sequencing data used in this project was generated by various members of the Hauser laboratory prior to the project's initiation. I have further indicated in the methods associated with this project in Chapter 5 which analyses and experiments were performed by or with the help of other laboratory members and which analyses I performed using scripts written by Dr. Ozer. I generated all of the figures in this chapter.

Introduction

Pseudomonas aeruginosa is a major cause of serious nosocomial infections. Antimicrobial resistance (AMR) in *P. aeruginosa* is frequent and limits treatment options, which has led the Infectious Disease Society of America[142] and the World Health Organization[303] to list this bacterium as a priority pathogen for the development of new antimicrobials. Surveillance of highly drug-resistant *P. aeruginosa* is critical to better understand its epidemiology and limit its spread.

Multilocus sequence typing (MLST) has identified distinct patterns in the epidemiology of multidrug-resistant (MDR) and extensively drug-resistant (XDR) *P. aeruginosa* infections. While sporadic isolates may demonstrate high AMR, a large proportion of MDR/XDR infections are caused by a relatively small number of globally distributed sequence types (ST) termed "high-risk clones"[129,131,140,192,199,200]. Known high-risk clones such as ST235, ST111, and ST175 may possess a variety of resistance determinants, both horizontally acquired and mutational[129,132,168], suggesting that these clones' high potential for acquiring AMR plays a role in their survival and spread in human populations[131]. While these STs are relatively common, other high-risk clones also contribute to drug-resistant infections worldwide[129,131,186,199], and it is likely that additional high-risk clones have yet to be described.

In this study, we investigated clonal complex (CC) 446, containing major STs 446 and 298, as a potential emerging high-risk clone. We describe the global distribution of this lineage and the presence of highly resistant isolates at both our institution and others. In doing so, we identified the persistence of an XDR ST298 subclade (ST298*) possessing a large novel AMR plasmid at one academic medical center for at least 16 years.

Results

Geographic Distribution of CC446 Isolates

In the process of investigating *P. aeruginosa* strains from collections obtained at a single medical center (Northwestern Memorial Hospital – NMH), we noted an unusually large representation of isolates with the closely related ST298 and ST446 genotypes. BURST analysis identified these STs, which are single locus variants, as central members of a larger CC consisting of 20 STs. This CC was termed CC446 after the likely group founder (Figure 2.1).

We next used *in silico* MLST to screen six *P. aeruginosa* patient and healthcare environmental strain collections from Chicago, Boston, and Spain (a total of 1259 isolates) for CC446 isolates and identified 54 (Table 2.1). Additionally, we screened 2483 *P. aeruginosa* genomes previously deposited in the NCBI database to identify another 38 CC446 isolates (Table 2.1). In total, we identified 92 CC446 isolates (49 ST298 and 43 ST446, Table 2.2). All CC446 isolates in this study were either ST298 or ST446, suggesting that these are the dominant clinical STs in this clonal complex. Whole-genome sequences were available for each of these isolates, and several had been previously published[65,73,74,100,304]. We also found multiple instances of CC446 strains mentioned in the literature and the PubMLST database for which whole-genome sequences were not available[180,181,192,305-308]. These CC446 isolates were cultured from North America, South America, Europe, Asia, and Oceania, indicating that CC446 is globally distributed (Figure 2.2).

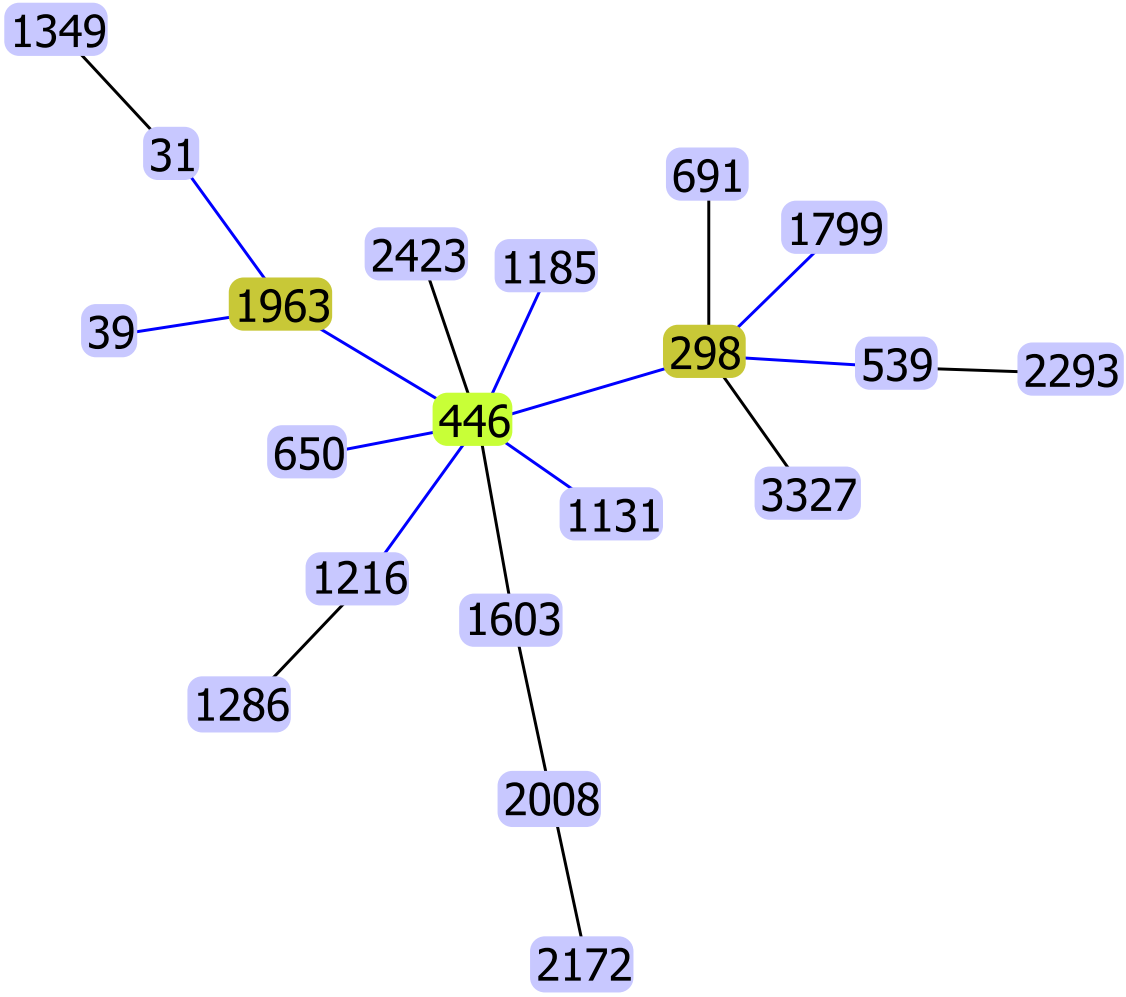


Figure 2.1 Global optimal eBURST diagram showing sequence types in CC446. Each sequence type is represented as a node with lines connecting single locus variants. ST446 (light green) was identified as the likely founder of the clonal complex because it possessed the largest number of single locus variants. Subgroup founders (ST298 and ST1963) are indicated in dark green.

Table 2.1 Sequenced CC446 Isolates and Collection Source

Collection	# Isolates	ST298 (#)	ST446 (#)	Total CC446 (#)	Location	Source	Years
PABL	100	9	2	11	NMH	Blood cultures	1999-2003
MolEpi	301	10	4	14	NMH	Microbiology and Molecular Epidemiology Lab	2002-2009
PA-NM	99	4	1	5	NMH	Patient samples	2013-2018
Hosp_Env	58	1	6	7	Chicago Metro Area	Health care facility environments (e.g. sinks)	2017-2018
BWH	100	2	1	3	Boston	Patient samples	2015-2016
PASP	601	2	12	14	Spain	Blood cultures	2008-2009
NCBI	2483	21	17	38	Various[a]	Publicly available genome sequences	-

[a]USA, Argentina, Belgium, Canada, Columbia, France, Germany, Netherlands, Pakistan, Portugal, Spain.

NMH: Northwestern Memorial Hospital, Chicago, USA.

Table 2.2 Whole Genome Sequenced CC446 *P. aeruginosa* Isolates Included in This Study

Name	Colection	Alternative Name	ST	Location	Year	BioSample Accession	SRA Accession	Genome Size (bp)	Contigs/Scaffolds (#)
PABL020	PABL	NA	298	NMH	2000	SAMN09831335	NA	6759755	509
PABL021	PABL	NA	298	NMH	2000	SAMN09831336	NA	6506522	512
PABL022	PABL	NA	298	NMH	2000	SAMN09831337	NA	6968177	152
PABL036	PABL	NA	298	NMH	2001	SAMN09831349	NA	7226411	375
PABL040	PABL	NA	*298*	NMH	NA	SAMN09831352	NA	6837974	431
PABL048	PABL	NA	298	NMH	2001	SAMN09831360	NA	7294576	2
PABL056	PABL	NA	298	NMH	2001	SAMN09831367	NA	7331427	663
PABL067	PABL	NA	298	NMH	2001	SAMN09831378	NA	6884951	576
PABL072	PABL	NA	446	NMH	2001	SAMN09831383	NA	6726822	165
PABL088	PABL	NA	298	NMH	2002	SAMN09831398	NA	7174621	350
PABL097	PABL	NA	446	NMH	2002	SAMN09831407	NA	6897023	246
PS1793	MolEpi	NA	298	NMH	NA	SAMN12162657	NA	7266255	119
PS1796	MolEpi	NA	298	NMH	NA	SAMN12162658	NA	7269149	138
PS1797	MolEpi	NA	298	NMH	NA	SAMN12162659	NA	7265792	160
PS1875	MolEpi	NA	298	NMH	2007	SAMN12162660	NA	7071643	227
PS1882	MolEpi	NA	298	NMH	2007	SAMN12162661	NA	7228492	223
PS1884	MolEpi	NA	446	NMH	2007	SAMN12162662	NA	6740329	124
PS1893	MolEpi	NA	298	NMH	2007	SAMN12162663	NA	7223965	252
PS1900	MolEpi	NA	298	NMH	2007	SAMN12162664	NA	7233292	243
PS1934	MolEpi	NA	298	NMH	2007	SAMN12162665	NA	7219953	254
PS1946	MolEpi	NA	446	NMH	2007	SAMN12162666	NA	6881977	165
PS1948	MolEpi	NA	446	NMH	2007	SAMN12162667	NA	6883295	171
PS1955	MolEpi	NA	298	NMH	2007	SAMN12162668	NA	6822019	234
PS1977	MolEpi	NA	446	NMH	2007	SAMN12162669	NA	6607435	123
PS2027	MolEpi	NA	298	NMH	2008	SAMN12162670	NA	6818562	234
PA-NM-015	PA-NM	NA	298	NMH	2014	SAMN12162671	NA	6661543	104
PA-NM-069	PA-NM	NA	298	NMH	2017	SAMN12162672	NA	7204989	229
PA-NM-074	PA-NM	NA	446	NMH	2017	SAMN12162673	NA	6735376	99
PA-NM-079	PA-NM	NA	298	NMH	2017	SAMN12162674	NA	7199132	217
PA-NM-088	PA-NM	NA	298	NMH	2017	SAMN12162675	NA	7201493	222
ENVO278	Hosp_Env	NA	446	Chicago, USA	2018	SAMN12162676	NA	6980029	330
ENVO281	Hosp_Env	NA	446	Chicago, USA	2018	SAMN12162677	NA	6837357	108
ENVO304	Hosp_Env	NA	446	Chicago, USA	2018	SAMN12162678	NA	6862184	117
OENV015	Hosp_Env	NA	298	Chicago, USA	2017	SAMN12162679	NA	6861252	137
OENV043	Hosp_Env	NA	446	Schaumburg, USA	2017	SAMN12162680	NA	6773683	101
OENV069	Hosp_Env	NA	446	Barrington, USA	2017	SAMN12162681	NA	6834888	288
OENV139	Hosp_Env	NA	446	Chicago, USA	2017	SAMN12162682	NA	6786588	108
BWH011	BWH	NA	298	Boston, USA	2015	SAMN12162683	NA	7075697	125
BWH031	BWH	NA	298	Boston, USA	2015	SAMN12162684	NA	7035660	101
BWH069	BWH	NA	446	Boston, USA	2016	SAMN12162685	NA	7005703	133
PASP010	PASP	NA	446	Spain	2008	SAMN12162686	NA	7164540	185
PASP063	PASP	NA	446	Spain	2008	SAMN12162687	NA	6870678	89
PASP107	PASP	NA	446	Spain	2008	SAMN12162688	NA	6837137	104
PASP118	PASP	NA	298	Spain	2008	SAMN12162689	NA	6981866	120
PASP145	PASP	NA	446	Spain	2008	SAMN12162690	NA	6980829	131
PASP163	PASP	NA	446	Spain	2008	SAMN12162691	NA	6929527	130
PASP170	PASP	NA	446	Spain	2008	SAMN12162692	NA	6859625	96
PASP174	PASP	NA	446	Spain	2008	SAMN12162693	NA	7072736	141
PASP199	PASP	NA	446	Spain	2008	SAMN12162694	NA	7115419	155
PASP363	PASP	NA	298	Spain	2008	SAMN12162695	NA	6906610	140
PASP368	PASP	NA	446	Spain	2008	SAMN12162696	NA	6977667	145
PASP375	PASP	NA	446	Spain	2009	SAMN12162697	NA	6785831	103
PASP418	PASP	NA	446	Spain	2009	SAMN12162698	NA	6813780	106
PASP614	PASP	NA	446	Spain	2009	SAMN12162699	NA	6784771	107
AXPE	NCBI	BL22	298	USA	NA	SAMN02360735	SRR1014184	6997218	12
JARI	NCBI	PA103	298	NA	NA	SAMN02951864	SRX2736379	6711305	262
JIEQ	NCBI	BWH060	298	NA	2013	SAMN02402444	SRX422978	6763011	18
JIEX	NCBI	BWH053	298	NA	2013	SAMN02402437	SRX422944	6840259	25
JTMS	NCBI	AZPAE15054	298	Bogota, Columbia	2012	SAMN03105751	NA	6647129	101
JTMZ	NCBI	AZPAE15047	298	Victoria, Argentina	2012	SAMN03105744	NA	6898282	123

Table 2.2 Continued

Name	Colection	Alternative Name	ST	Location	Year	BioSample Accession	SRA Accession	Genome Size (bp)	Contigs/Scaffolds (#)
JTND	NCBI	AZPAE15025	446	Madrid, Spain	2011	SAMN03105740	NA	6978967	109
JTNM	NCBI	AZPAE15034	298	Bilbao, Spain	2011	SAMN03105731	NA	6794620	92
JTPH	NCBI	AZPAE14987	298	Koln, Germany	2010	SAMN03105684	NA	6855155	121
JTTS	NCBI	AZPAE14870	298	Victoria, Argentina	2007	SAMN03105569	NA	6899104	112
JTYJ	NCBI	AZPAE14437	298	Canada	2010	SAMN03105448	NA	6714753	99
JTZF	NCBI	AZPAE13876	446	Portugal	2010	SAMN03105426	NA	6795623	153
JUMF	NCBI	953_PAER	446	Seattle, USA	2012-2013	SAMN03198173	SRX762871	7045281	899
JUNG	NCBI	928_PAER	298	Seattle, USA	2012-2013	SAMN03198146	SRX762844	6624059	890
JUNH	NCBI	927_PAER	298	Seattle, USA	2012-2013	SAMN03198145	SRX762843	6654556	474
JUZD	NCBI	637_PAER	298	Seattle, USA	2012-2013	SAMN03197837	SRX762535	6734429	1750
JVFW	NCBI	468_PAER	298	Seattle, USA	2012-2013	SAMN03197662	SRX762360	6803974	278
JVGC	NCBI	462_PAER	298	Seattle, USA	2012-2013	SAMN03197656	SRX762354	6898892	140
JVPD	NCBI	230_PAER	298	Seattle, USA	2012-2013	SAMN03197421	Not Usable	6919796	136
LLMB	NCBI	WH-SGI-V-07172	446	France	1992	SAMN04128510	SRX1437077	7091160	104
LLMY	NCBI	WH-SGI-V-07227	298	USA	1995	SAMN04128533	SRX1437100	7059764	124
LLNC	NCBI	WH-SGI-V-07231	446	USA	1995	SAMN04128537	SRX1437104	6610758	110
LLOJ	NCBI	WH-SGI-V-07385	446	France	1991	SAMN04128570	SRX1437137	7037854	168
LLPI	NCBI	WH-SGI-V-07421	298	USA	2005	SAMN04128595	SRX1437162	6874120	135
LLQA	NCBI	WH-SGI-V-07494	446	USA	2005	SAMN04128613	SRX1437180	6367706	137
LLRE	NCBI	WH-SGI-V-07633	298	USA	2005	SAMN04128643	SRX1437210	6942577	138
LLRF	NCBI	WH-SGI-V-07634	446	USA	2005	SAMN04128644	SRX1437211	6870291	156
LLSF	NCBI	WH-SGI-V-07685	298	USA	2005	SAMN04128670	SRX1437237	6895101	151
LLTF	NCBI	WH-SGI-V-07711	446	USA	2008	SAMN04128696	SRX1437263	7567509	305
LLTL	NCBI	WH-SGI-V-07251	446	Netherlands	1997	SAMN04128702	SRX1437269	6945431	125
LLUT	NCBI	WH-SGI-V-07297	446	Pakistan	1998	SAMN04128736	SRX1437303	6772578	132
MPVG	NCBI	CLB24232	298	NA	NA	SAMN05774262	SRX2410572	7087679	185
MPVJ	NCBI	CLB24412	446	NA	NA	SAMN05774265	SRX2410578	7094206	203
NMPT	NCBI	53014	446	Belgium	2013	SAMN07344900	NA	7050840	383
NMPU	NCBI	53012	446	Belgium	2013	SAMN07344899	NA	7042955	588
NMPV	NCBI	53011	446	Belgium	2012	SAMN07344898	NA	7005891	934
NMPW	NCBI	41437	446	Belgium	2010	SAMN07344897	NA	7100137	586
S04_90	NCBI	S04 90	446	Rotterdam, Netherlands	2013	SAMN03396926	SRX976879	7259150	2

NMH: Northwestern Memorial Hospital, Chicago, USA.

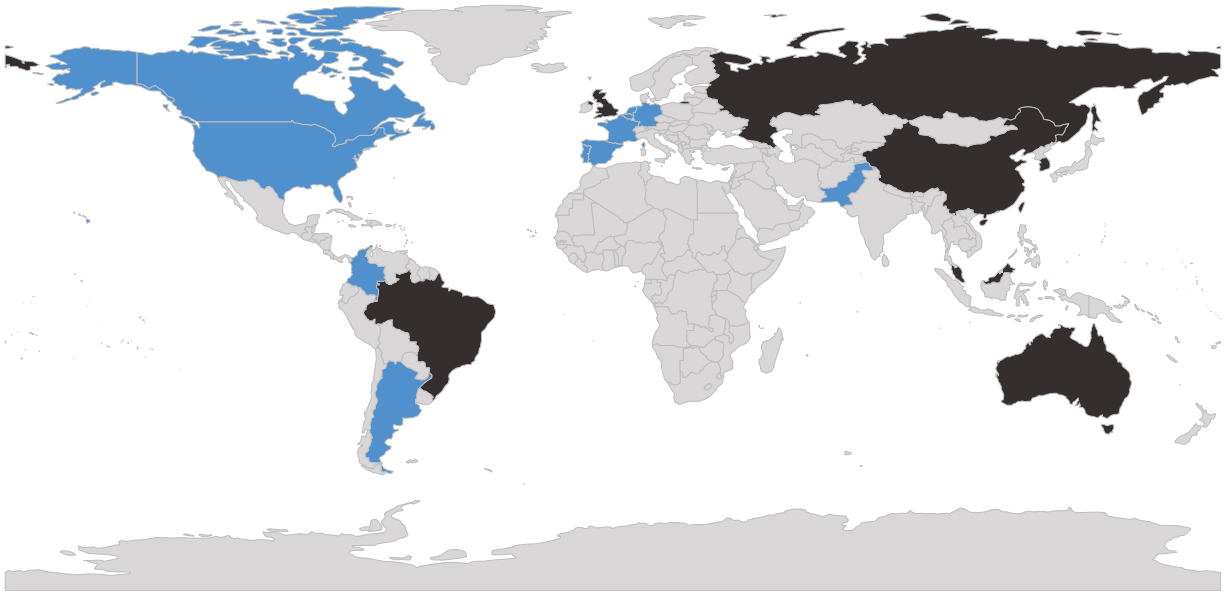


Figure 2.2 Global distribution of CC446. World map indicates countries where CC446 isolates have been detected. Countries with at least one isolate associated with a genome analyzed in this study are shaded blue. Countries in which a CC446 isolate has been reported (in the literature or PubMLST database[305]) but no genome was available are shaded in black.

Antimicrobial Resistance of CC446 Isolates

We had access to the 54 CC446 isolates from the Chicago, Boston, and Spain collections and performed microbroth-dilution antibiotic susceptibility testing on them (Table 2.3). Overall, 52% (28/54) of the isolates were MDR, of which 54% (15/28) were XDR. AMR was most prevalent in isolates collected at NMH, with 77% (23/30) of isolates MDR and 61% (14/23) of those XDR. In particular, almost all NMH ST298 isolates (21/23, 91%) were MDR, of which many (13/21, 62%) were XDR. Of the 14 Spanish CC446 isolates, four ST446 isolates were MDR, one of which was XDR. One of three CC446 isolates from Boston was MDR. In contrast to isolates collected from patient samples, none of the seven healthcare environmental CC446 isolates were MDR. Additionally, four XDR ST298 strains from NMH showed nonsusceptibility to recently developed β-lactam/β-lactamase inhibitor combinations ceftazidime-avibactam and ceftolozane-tazobactam on disk diffusion testing (Table 2.3). The high prevalence of MDR/XDR isolates in this study, coupled with the global distribution of CC446 (Figure 2.2) and previous reports of AMR in this clonal complex[65,74,100,192,307], support the classification of CC446 as an emerging high-risk clone.

Table 2.3 Antibacterial Susceptibility Testing of CC446 P. aeruginosa Included in This Study

Microbroth-dilution antimicrobial susceptibility testing and MIC determination performed by Dr. Kelly Bachta

					Minimum Inhibitory Concentration (µg/mL)								Zone Diameter (mm)			
Name	ST	Location	in1697 Presence	ST298* Subclade	Gent	Cep	Ctz	Pip/Tazo	Mero	Az	Cipro	Col	Ctz/Avi	Ceftol/Tazo	MDR	XDR
PABL020	298	NMH		+	8 - ns	8	4	16	16 - ns	16 - ns	64 - ns	1	21	25	+	
PABL021	298	NMH		+	16 - ns	4	8	8	16 - ns	32 - ns	64 - ns	0.5	23	24	+	
PABL036	298	NMH	±	+	>128 - ns	16 - ns	4	64 - ns	8 - ns	16 - ns	32 - ns	1	22	23	+	+
PABL040	298	NMH		+	4	8	4	16	4 - ns	16 - ns	32 - ns	1	24	25	+	
PABL048	298	NMH	+	+	>128 - ns	16 - ns	4	64 - ns	4 - ns	16 - ns	32 - ns	1	22	23	+	+
PABL056	298	NMH	+	+	>128 - ns	16 - ns	2	64 - ns	4 - ns	32 - ns	32 - ns	1	23	24	+	+
PABL067	298	NMH	±	+	>128 - ns	16 - ns	4	64 - ns	4 - ns	32 - ns	32 - ns	1	18 - ns	19 - ns	+	+
PABL088	298	NMH	+	+	>128 - ns	16 - ns	4	32- ns	16 - ns	32 - ns	4 - ns	0.5	25	26	+	+
PS1793	298	NMH	+	+	>128 - ns	16 - ns	64 - ns	64 - ns	16 - ns	16 - ns	32 - ns	1	16 - ns	19 - ns	+	+
PS1796	298	NMH	+	+	>128 - ns	16 - ns	64 - ns	64 - ns	16 - ns	16 - ns	32 - ns	1	17 - ns	18 - ns	+	+
PS1797	298	NMH	+	+	>128 - ns	32 - ns	64 - ns	128 - ns	16 - ns	8	32 - ns	1	17 - ns	18 - ns	+	+
PS1875	298	NMH	+	+	>128 - ns	8	2	64 - ns	8 - ns	8	16 - ns	1	21	21	+	
PS1882	298	NMH	+	+	>128 - ns	8	2	64 - ns	16 - ns	8	16 - ns	1	25	27	+	
PS1893	298	NMH	+	+	>128 - ns	8	4	64 - ns	16 - ns	16 - ns	32 - ns	1	23	24	+	+
PS1900	298	NMH	+	+	>128 - ns	4	2	32- ns	8 - ns	8	16 - ns	1	24	25	+	
PS1934	298	NMH	+	+	>128 - ns	16 - ns	2	128 - ns	16 - ns	16 - ns	8 - ns	0.5	27	29	+	+
PS1955	298	NMH		+	2	8	8	16	16 - ns	16 - ns	16 - ns	0.5	21	24	+	
PS2027	298	NMH		+	4	4	4	16	32 - ns	16 - ns	16 - ns	1	25	31	+	
PA-NM-069	298	NMH	+	+	>128 - ns	16 - ns	4	64 - ns	32 - ns	32 - ns	64 - ns	4 - ns	24	25	+	+
PA-NM-079	298	NMH	+	+	>128 - ns	16 - ns	4	64 - ns	32 - ns	32 - ns	32 - ns	1	23	23	+	+
PA-NM-088	298	NMH	+	+	>128 - ns	16 - ns	4	64 - ns	32 - ns	32 - ns	32 - ns	1	24	25	+	+
PABL022	298	NMH			2	4	1	4	4 - ns	4	<0.25	1	28	28		
PA-NM-015	298	NMH			2	2	2	4	0.5	4	<0.25	1	26	26		
OENV015	298	Chicago			4	4	2	8	8 - ns	4	<0.25	1	28	29		
BWH011	298	Boston			2	4	2	4	8 - ns	8	32 - ns	1	24	25		
BWH031	298	Boston			0.5	16 - ns	32 - ns	16	32 - ns	64 - ns	32 - ns	1	21	26	+	
PASP118	298	Spain			2	4	2	8	16 - ns	8	<0.25	1	24	27		
PASP363	298	Spain			2	1	1	4	0.5	4	<0.25	1	25	28		
PABL072	446	NMH			2	4	4	16	1	16 - ns	16 - ns	1	21	24		
PABL097	446	NMH			4	8	2	4	0.5	4	16 - ns	1	23	23		
PS1884	446	NMH			2	1	2	4	1	8	<0.25	1	24	23		
PS1946	446	NMH			4	8	8	32 - ns	16 - ns	8	32 - ns	1	24	22	+	
PS1948	446	NMH			8 - ns	8	16 - ns	32 - ns	16 - ns	8	32 - ns	1	23	22	+	+
PS1977	446	NMH			1	2	4	32 - ns	1	8	<0.25	1	24	24		
PA-NM-074	446	NMH			<0.25	4	4	8	2	16	<0.25	0.5	28	27		

Table 2.3 Continued

Name	ST	Location	in1697 Presence	ST298* Subclade	Minimum Inhibitory Concentration (µg/mL)								Zone Diameter (mm)		MDR	XDR
					Gent	Cep	Ctz	Pip/Tazo	Mero	Az	Cipro	Col	Ctz/Avi	Ceftol/Tazo		
OENV043	446	Chicago			2	4	2	8	0.5	16 - ns	<0.25	1	27	28		
OENV069	446	Chicago			4	2	2	4	0.5	8	<0.25	1	28	28		
OENV139	446	Chicago			2	4	2	8	1	8	<0.25	0.5	24	25		
ENVO278	446	Chicago			<0.25	4	4	4	2	16 - ns	<0.25	1	27	25		
ENVO281	446	Chicago			2	4	2	4	1	8	<0.25	1	25	26		
ENVO304	446	Chicago			4	2	1	4	0.5	4	<0.25	1	24	24		
BWH069	446	Boston			4	4	2	8	8 - ns	8	<0.25	1	22	23		
PASP010	446	Spain			2	2	16 - ns	16	0.5	8	<0.25	1	23	24		
PASP063	446	Spain			2	4	16 - ns	128 - ns	16 - ns	8	<0.25	1	21	21	+	
PASP107	446	Spain			2	1	1	4	1	4	<0.25	0.5	26	26		
PASP145	446	Spain			4	4	1	4	8 - ns	4	<0.25	1	24	24		
PASP163	446	Spain			2	1	1	4	1	4	<0.25	1	25	25		
PASP170	446	Spain			2	16 - ns	32 - ns	64 - ns	1	16 - ns	<0.25	1	27	26	+	
PASP174	446	Spain			2	2	2	16	0.5	8	<0.25	1	25	25		
PASP199	446	Spain			2	16 - ns	32 - ns	64 - ns	4	32 - ns	<0.25	1	25	25	+	
PASP368	446	Spain			4	8	1	4	0.5	4	0.5	1	23	22		
PASP375	446	Spain			2	2	1	8	1	4	<0.25	1	22	23		
PASP418	446	Spain			8 - ns	32 - ns	16 - ns	128 - ns	32 - ns	16 - ns	1	1	21	23	+	+
PASP614	446	Spain			2	4	4	16	1	16 - ns	0.5	1	21	22		

Gent: gentamicin, Cep: cefepime, Ctz: ceftazidime, Pip/Tazo: piperacillin-tazobactam, Mero: meropenem, Az: aztreonam, Cipro: ciprofloxacin, Col: Colistin, Ctz, Ctz/Avi: ceftazidime-avibactam, Ceftol/Tazo: ceftolozane-tazobactam.

NMH: Northwestern Memorial Hospital, Chicago, USA.

±: Heterogenous presence of in1697 in some colonies.

ns: non-susceptible (intermediate and resistant); Clinical Laboratory Standards Institute, MIC Interpretive Standards (µg/mL), 2018.

MDR: ns to at least one drug in >= 3 classes tested, XDR: ns to at least one drug in all but <=2 classes tested

Identification of AMR Integron in1697 in NMH ST298 Isolates

To determine the genetic basis for the high rates of AMR in ST298 isolates from NMH, we identified resistance genes from their whole genome sequences using the ResFinder database[309]. We identified a locus containing multiple AMR genes present in 16 MDR ST298 isolates from NMH. This locus was present in 76.2% (16/21) of MDR ST298 isolates from NMH, and 81.3% (13/16) of these isolates were XDR (Table 2.3). Characterization of this locus revealed it to be a novel class 1 integron designated in1697 (Figure 2.3). As is common for class 1 integrons, it consists of a 5' conserved segment (5'-CS) containing the *intl1* integrase gene and a promoter driving cassette expression, several resistance gene cassettes, and a 3' conserved segment (3'-CS) containing *sul1* (sulphonamide resistance) and *qacEΔ1* (quaternary ammonium compound, QAC, resistance)[98]. Gene cassettes in in1697 include the β-lactamase *blaOXA-10*, aminoglycoside resistance genes *aadB and aadA10e*, and QAC resistance gene *qacF*. Isolates with in1697 showed high levels of gentamicin resistance (>128 μg/mL) not seen among other CC446 isolates tested. These findings suggest that a novel integron, in1697, contributes to the antibiotic resistance of some CC446 isolates.

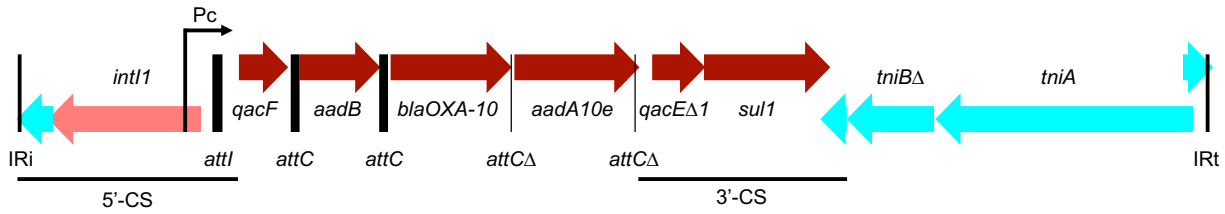


Figure 2.3. Diagram of AMR class I integron in1697. In1697 consists of a 5' conserved segment (5'-CS) with the integrase *intI1* and promoter Pc, AMR cassettes, and a 3' conserved segment (3'-CS) of *qacEΔ1* and *sul1*. Complete *attC* recombination sites were identified downstream of *qacF* and *aadB,* and truncated *attC* sites were identified downstream of *blaOXA-10* and *aad10Ae*. In1697 appears to be part of transposable-like element that includes a partial *tni* transposon operon and has as its borders IRi and IRt (25-bp imperfect [92% identity] inverted repeats).

Identification of a large AMR plasmid in NMH ST298 isolates

To investigate the genomic context of in1697, we performed long-read sequencing and complete genome construction for one in1697-positive isolate from NMH (PABL048). This yielded a 6,879,622 bp bacterial chromosome and revealed that in1697 is located on a large plasmid (414,954 bp) that we named pPABL048 (Figure 2.4). The plasmid pPABL048 contains 496 coding sequences, some of which were predicted to encode for antimicrobial/disinfectant resistance proteins, heavy metal resistance proteins, and chemotaxis proteins (Supplementary Table 2.1). Screening both the PABL048 chromosome and plasmid against the virulence factor database identified several predicted virulence factors on the plasmid with 3 related to Type IV pili and 1 potentially related to carbon storage regulation (Supplementary Table 2.2)[310]. Sequencing reads for all isolates containing in1697 showed substantial alignment to pPABL048 (generally >90% sequence coverage) with few SNVs (≤4), indicating that these isolates contain very similar plasmids. The exception is PS1875 (57.6% alignment), which is missing a large contiguous portion of the plasmid (Table 2.4 and Figure 2.5). In1697 was not found outside the context of pPABL048. We compared read alignments of in1697 containing isolates to the PABL048 chromosome and plasmid (excluding PS1875 and PABL048 itself). The median depth of plasmid alignments was on average 1.47 times the median depth of chromosome alignments, suggesting that pPABL048 is present at a low copy number (Table 2.5). In summary, pPABL048 is a large ST298-associated plasmid containing a novel AMR class 1 integron.

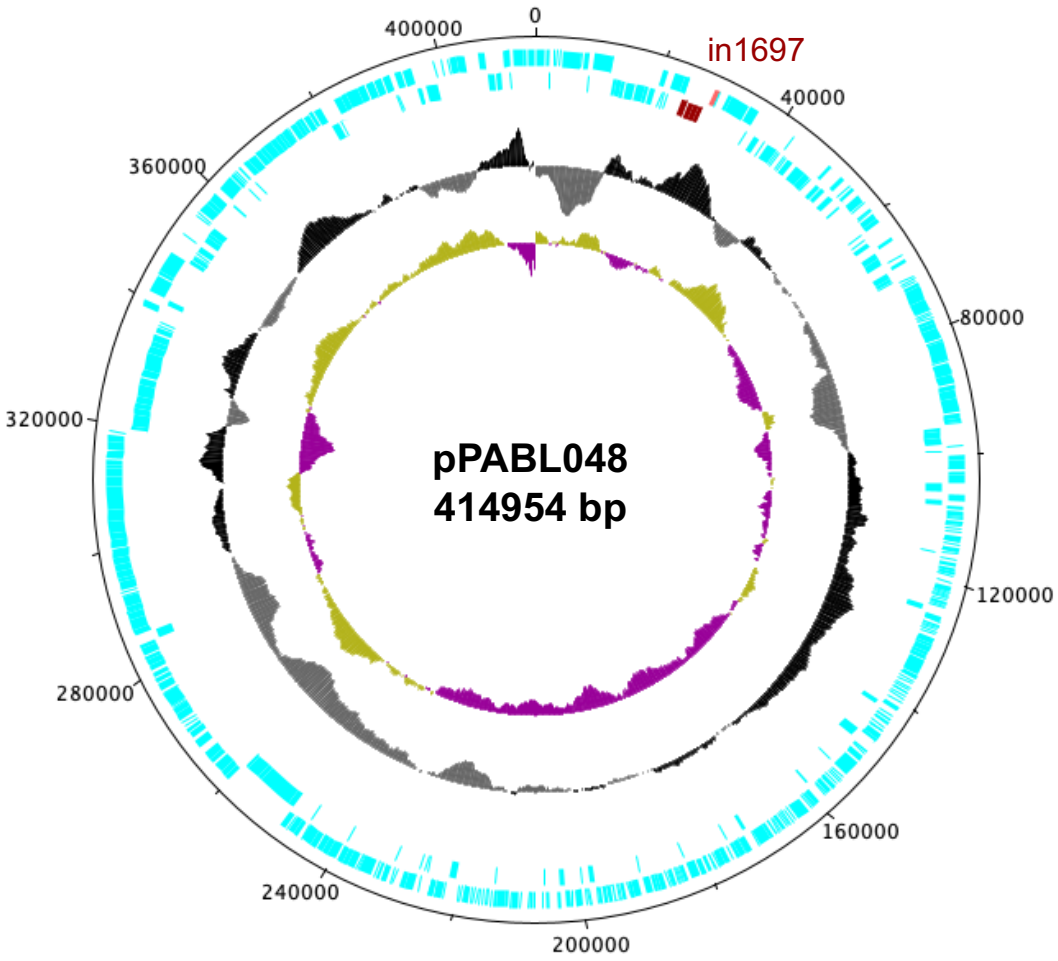


Figure 2.4 Diagram of the large AMR plasmid pPABL048. Rings (from in to out) show GC skew, GC%, coding sequences, and position in bp. The location of in1697 is highlighted in red.

Table 2.4 Alignment of CC446 Isolates to pPABL048

Name	Percent Alignment[a]	SNVs[b]	in1697
PABL020	15.6	296	
PABL021	0.8	0	
PABL022	5.7	113	
PABL036	99.4	2	+
PABL036-GentR	99.8	2	+
PABL036-GentS	4.1	94	
PABL040	4.0	64	
PABL048-c1	4.1	93	
PABL048-c2	4.1	93	
PABL056	99.4	3	+
PABL067	4.1	87	
PABL067-GentR	99.9	1	+
PABL067-GentS	4.1	93	
PABL072	4.7	422	
PABL088	94.5	3	+
PABL097	4.4	467	
PS1793	98.8	4	+
PS1796	98.7	4	+
PS1797	98.7	4	+
PS1875	57.6	1	+
PS1882	99.7	3	+
PS1884	6.6	395	
PS1893	99.7	0	+
PS1900	99.8	3	+
PS1934	99.7	1	+
PS1946	4.4	462	
PS1948	4.4	466	
PS1955	3.9	77	
PS1977	6.0	140	
PS2027	3.9	73	
PA-NM-015	5.3	600	
PA-NM-069	93.4	1	+
PA-NM-074	6.0	287	
PA-NM-079	93.4	2	+
PA-NM-088	93.4	1	+
ENVO278	6.2	318	
ENVO281	6.2	310	
ENVO304	8.7	2099	
OENV015	4.5	606	
OENV043	4.1	501	
OENV069	5.4	119	
OENV139	4.1	535	
BWH011	3.7	162	
BWH031	3.8	199	
BWH069	4.2	206	
PASP010	4.2	275	
PASP063	6.1	273	
PASP107	5.5	141	
PASP118	5.6	98	
PASP145	4.1	420	
PASP163	6.5	391	
PASP170	5.8	171	
PASP174	7.4	381	
PASP199	8.2	519	
PASP363	5.2	597	
PASP368	4.6	471	
PASP375	5.9	176	
PASP418	5.8	174	
PASP614	5.8	157	

Table 2.4 Continued

Name	Percent Alignment[a]	SNVs[b]	in1697
AXPE	5.5	52	
JARI	6.0	115	
JIEQ	3.2	92	
JIEX	3.3	109	
JUMF	7.0	935	
JUNG	3.0	24	
JUNH	2.8	35	
JUZD	5.4	22	
JVFW	5.4	43	
JVGC	5.4	40	
LLMB	4.6	501	
LLMY	6.0	129	
LLNC	5.6	112	
LLOJ	4.5	462	
LLPI	6.0	118	
LLQA	6.6	369	
LLRE	5.1	442	
LLRF	4.0	481	
LLSF	6.0	599	
LLTF	86.8	5311	
LLTL	4.4	440	
LLUT	4.3	360	
MPVG	82.4	3424	
MPVJ	5.5	406	
S04_90	6.5	64	
JTMS	5.4	786	
JTMZ	4.6	472	
JTND	3.8	463	
JTNM	5.6	205	
JTPH	5.6	205	
JTTS	4.6	472	
JTYJ	5.6	207	
JTZF	4.5	382	
JVPD	4.7	799	
NMPT	5.1	546	
NMPU	5.1	546	
NMPV	5.1	546	
NMPW	5.1	546	

[a]Percentage of total length of pPABL048 covered by aligned sequences. When reads were available genomes were aligned to pPABL048 with BWA using a minimum depth cutoff of 5 reads. When only contigs were available genomes were aligned with NUCmer.

[b]SNVs after filtering masked positions.

c: experimentally cured of pPABL048, GentR: derived from gentamicin resistant colony, GentS: derived from gentamicin sensitive colony.

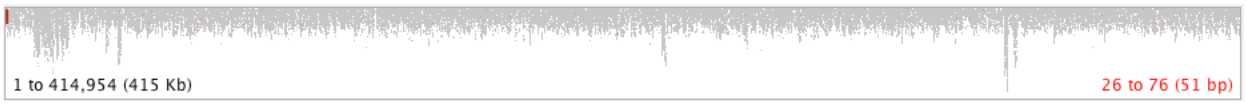
PABL020:

PABL021:

PABL022

PABL036:

PABL036-Gent$_R$:



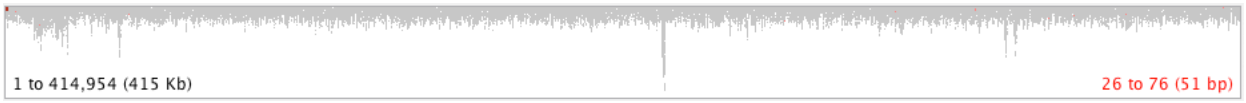
PABL036-Gent$_S$:

PABL040:

PABL048-c1:

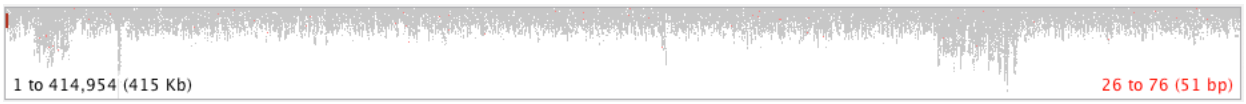
PABL048-c2:

PABL056:



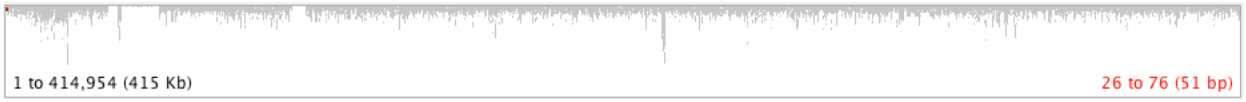
PABL067:

PABL067-Gent$_R$:

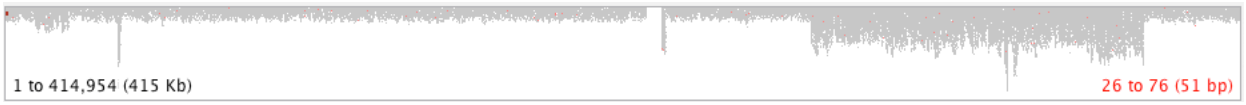


PABL067-Gent$_S$:

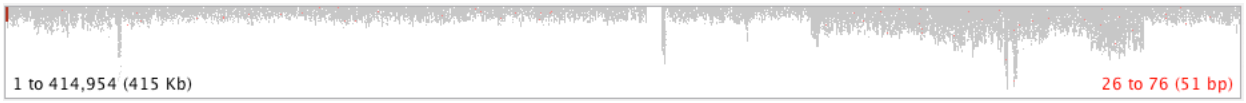
PABL088:



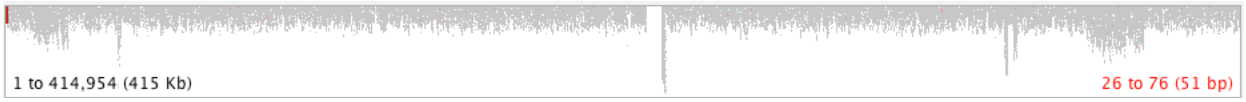
PS1793:



PS1796:

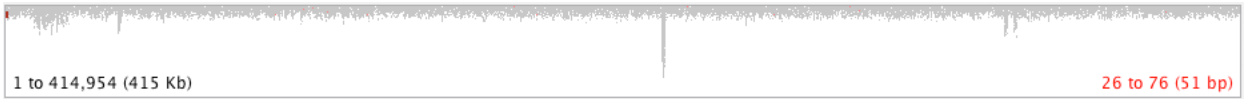


PS1797:

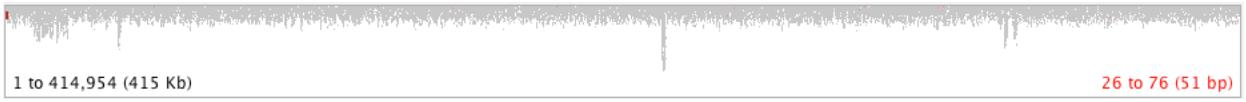


PS1875:

PS1882:

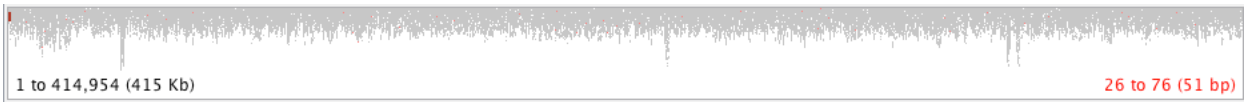


PS1893:



PS1900:

PS1934:



PS1955:

PS2027:

PA-NM-015:

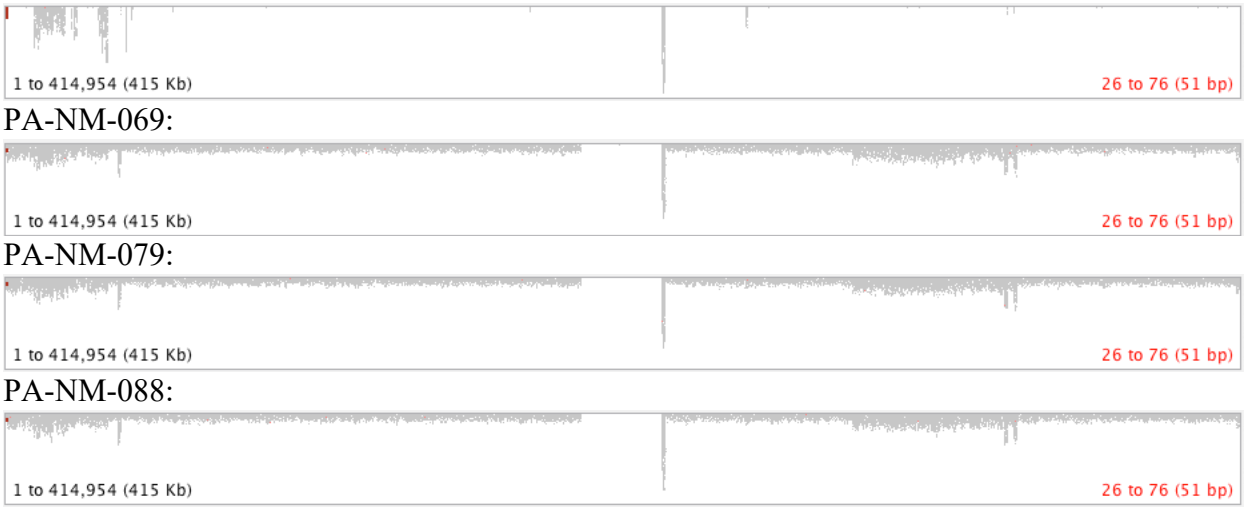


Figure 2.5 Visualization of NMH ST298 isolate read alignments to pPABL048. For strains with heterogenous plasmid presence (PABL036 and PABL067) alignments for the original reads, reads from sequencing a gentamicin resistant colony ($Gent_R$) and reads from sequencing a gentamicin sensitive colony ($Gent_S$) are shown. For PABL048, alignments of reads from two colonies experimentally cured of pPABL048 (c1 and c2) are shown.

Table 2.5 Comparison of Alignment Depth of in1697-Containing Isolates to the PABL048 Chromosome and Plasmid

Name	Percent Aligned Chromosome	Median Depth Chromosome	Percent Aligned Plasmid	Medium Depth Plasmid	Plasmid Depth / Chromosome Depth
PABL036	98.32	169	99.37	169	1.00
PABL056	98.34	136	99.40	172	1.26
PABL067-$Gent_R$	98.08	37	99.89	57	1.54
PABL088	97.96	115	94.50	142	1.23
PS1793	98.19	79	98.77	125	1.58
PS1796	97.93	33	98.72	34	1.03
PS1797	97.72	25	98.68	42	1.68
PS1882	98.67	53	99.66	71	1.34
PS1893	98.76	54	99.68	81	1.50
PS1900	98.82	60	99.78	84	1.40
PS1934	98.67	39	99.71	105	2.69
PA-NM-069	99.01	78	93.42	119	1.53
PA-NM-079	98.96	64	93.41	88	1.38
PA-NM-088	98.95	59	93.41	86	1.46

$Gent_R$: derived from gentamicin resistant colony

Two ST298 strains from NMH, PABL036 and PABL067, expressed a heterogeneous pattern of resistance. That is, some colonies from the same culture stock expressed high levels of gentamicin resistance while others did not. Whole genome sequencing confirmed that this variable pattern of resistance was associated with the presence or absence of the AMR plasmid pPABL048 (Table 2.4 and Figure 2.5). This discrepancy accounts for the lack of in1697 in our initial whole-genome sequence of PABL067. These findings indicate that pPABL048 can be spontaneously lost by some strains, which could potentially cause inaccurate antibiotic susceptibility results.

To explore the function of the plasmid, we generated plasmid-cured variants of PABL048 (PABL048-c1 and PABL048-c2, Table 2.4 and Figure 2.5) and tested the impact of pPABL048 on AMR. Isolates lacking the plasmid showed reduced MICs to gentamicin and piperacillin-tazobactam compared to their isogenic partners possessing the plasmid (Table 2.6), indicating that pPABL048 encodes for resistance to these antibiotics.

Table 2.6 Minimum Inhibitory Concentrations (MICs) of ST298* *P. aeruginosa* isolates with and without pPABL048

Microbroth-dilution antimicrobial susceptibility testing and MIC determination performed by Dr. Kelly Bachta

		MICs (µg/mL)							
Name	**pPABL048**	**Gent**	**Cep**	**Ctz**	**Pip/Tazo**	**Mero**	**Az**	**Cipro**	**Col**
PABL048	+	>128 - ns	8	2	64 – ns	8 - ns	32 - ns	32 - ns	0.5
PABL048-c1		4	4	4	16	4 – ns	16 - ns	16 - ns	0.5
PABL048-c2		8 – ns	4	4	16	4 - ns	16 - ns	16 - ns	0.5
PABL036-$Gent_R$	+	>128 - ns	8	4	32 - ns	8 - ns	16 - ns	32 - ns	0.5
PABL036-$Gent_S$		4	4	4	16	8 - ns	16 - ns	32 - ns	0.5
PABL067-$Gent_R$	+	>128 - ns	8	4	64 – ns	8 – ns	32 – ns	32 – ns	0.5
PABL067-$Gent_S$		4	4	4	16	4 - ns	16 - ns	32 - ns	0.5

Gent: gentamicin, Cep: cefepime, Ctz: ceftazidime, Pip/Tazo: piperacillin-tazobactam, Mero: meropenem, Az: aztreonam, Cipro: ciprofloxacin, Col: Colistin.

ns: non-susceptible (intermediate and resistant); Clinical Laboratory Standards Institute, MIC Interpretive Standards (µg/mL), 2018.

c: experimentally cured of pPABL048, $Gent_R$: derived from gentamicin resistant colony, $Gent_S$: derived from gentamicin sensitive colony.

Phylogenetic Analysis of CC446

To better understand the relationships between CC446 isolates included in this study, we constructed a recombination-corrected maximum likelihood phylogenetic tree based on core genome alignment to the PABL048 chromosome (Figure 2.6). We found that while ST298 and ST446 are closely related, they are phylogenetically distinct. The majority (21/30) of CC446 isolates from NMH clustered in a distinct ST298 subclade (designated ST298*), which was not seen in any of the other collections. Both pPABL048 and in1697 are exclusive to this subclade. ST298* isolates were collected between 2000 and 2017 (with pPABL048 first detected in 2001). While ST298* was only detected at NMH, ST298 and ST446 isolates outside of this subclade were also present at NMH. This suggests a prolonged local epidemic of ST298* had occurred in addition to the general circulation of other CC446 isolates. ST298* isolates showed high levels of AMR, while sporadic CC446 isolates from NMH were largely sensitive to antimicrobials, the exceptions being the ST446 isolates PS1946 (MDR) and PS1948 (XDR) (Table 2.3). Clustering the CC446 isolates using the hierBAPS algorithm[311] supported the definition of ST298* as a distinct subclade of ST298. It is notable that hierBAPS analysis also split ST446 into two subclades, a finding which was not investigated further (Figure 2.7).

We used Bayesian phylogenetic analysis to construct a time-scaled phylogenetic tree of 17 ST298* isolates with available collection dates (Figure 4). The most recent common ancestor for ST298* is estimated to have arisen in the year 1980 (mean 1980.9, 95% HPD interval 1973.8-1987.4). Based on this analysis, ST298* has been evolving at a rate of 1.80 (95% HPD interval 1.32-2.29) core genome SNVs/year. This is comparable to previous estimates in non-hypermutable *P. aeruginosa*[12,312].

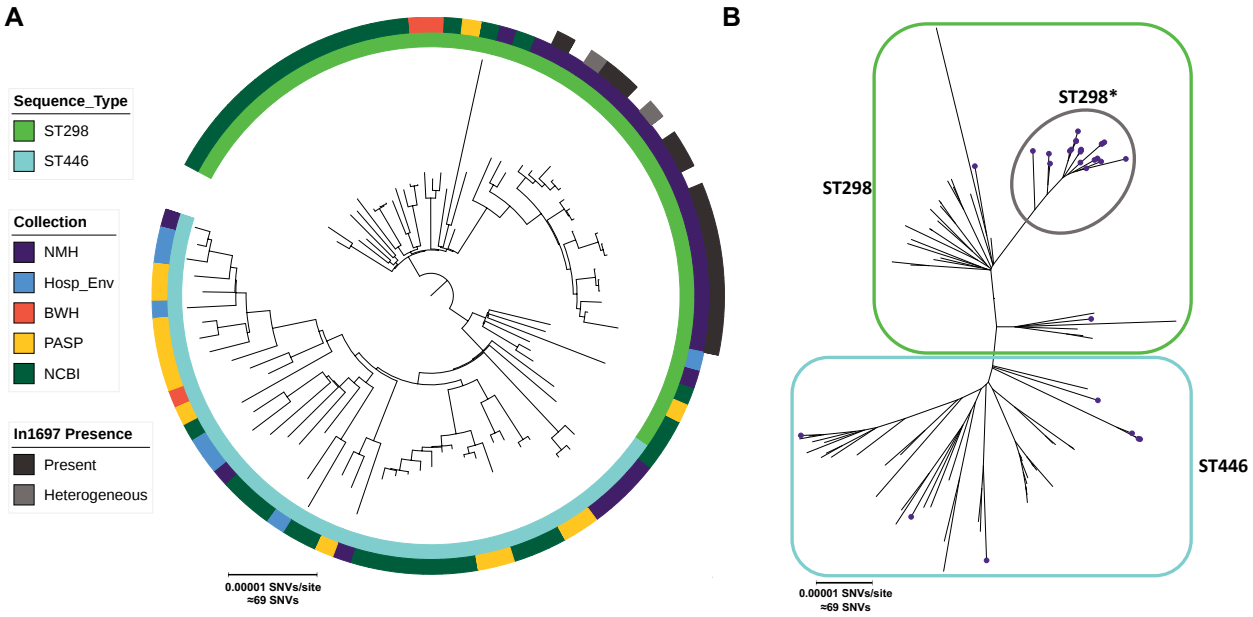


Figure 2.6. Recombination-corrected maximum likelihood phylogenetic tree of the CC446 isolates included in this study based on core genome alignment to the chromosome of PABL048. (A) Midpoint-rooted circular tree annotated (from inner to outer rings) with sequence type, collection of origin, and the presence of in1697. (B) Unrooted radial tree with sequence type and subclade indicated by blue (ST446), green (ST298), and grey (ST298*) outlines. Isolates collected from NMH are indicated with purple circles.

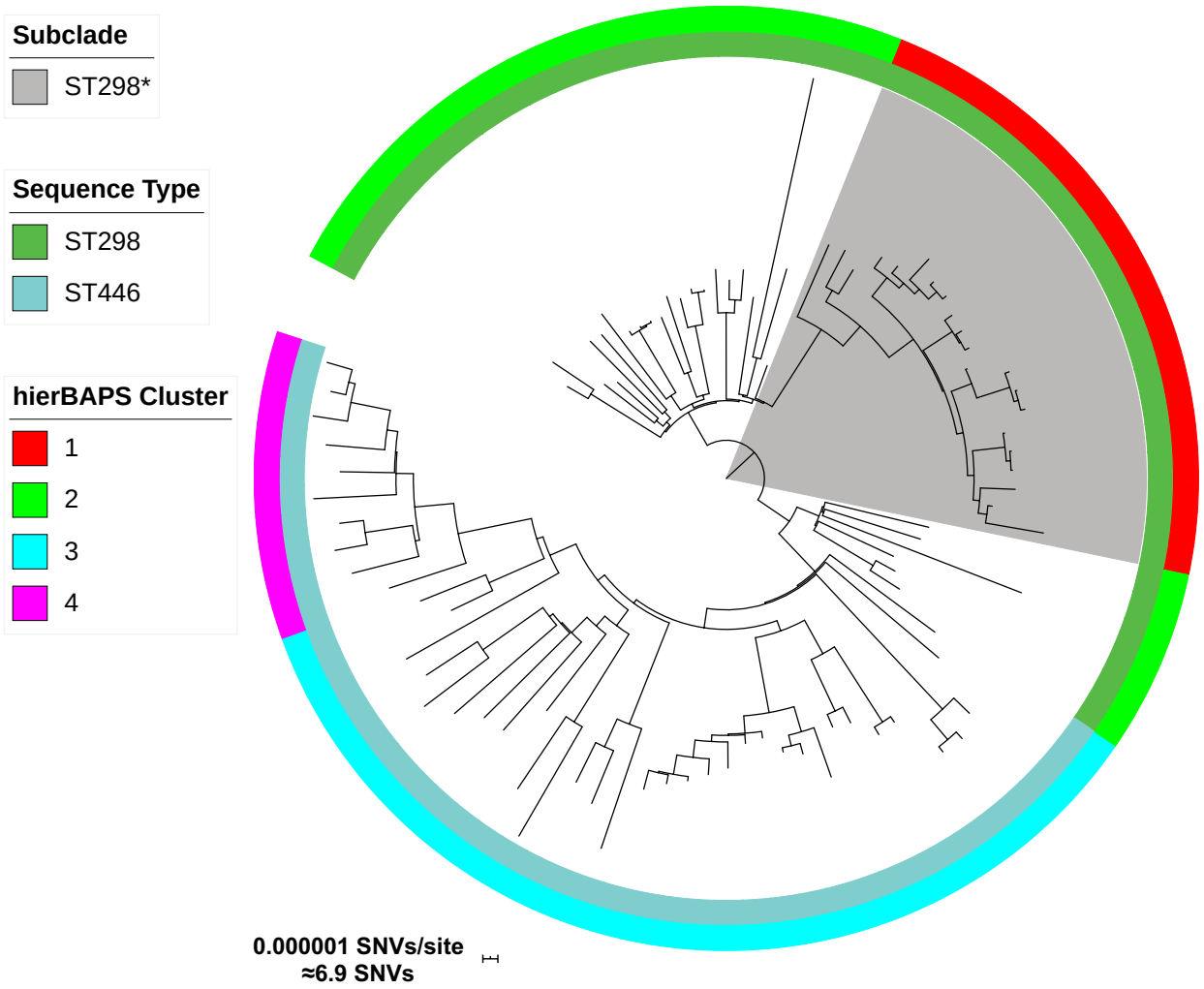


Figure 2.7 Genetic clustering with the hierBAPS algorithm agrees with the designation of the ST298* subclade. Mid-point rooted circular recombination-corrected maximum likelihood phylogenetic tree of the CC446 isolates included in this study. Annotations with sequence type, hierBAPS cluster assignment, and subclade (with ST298* isolates shaded grey). Only the first level of hierBAPS clustering is shown.

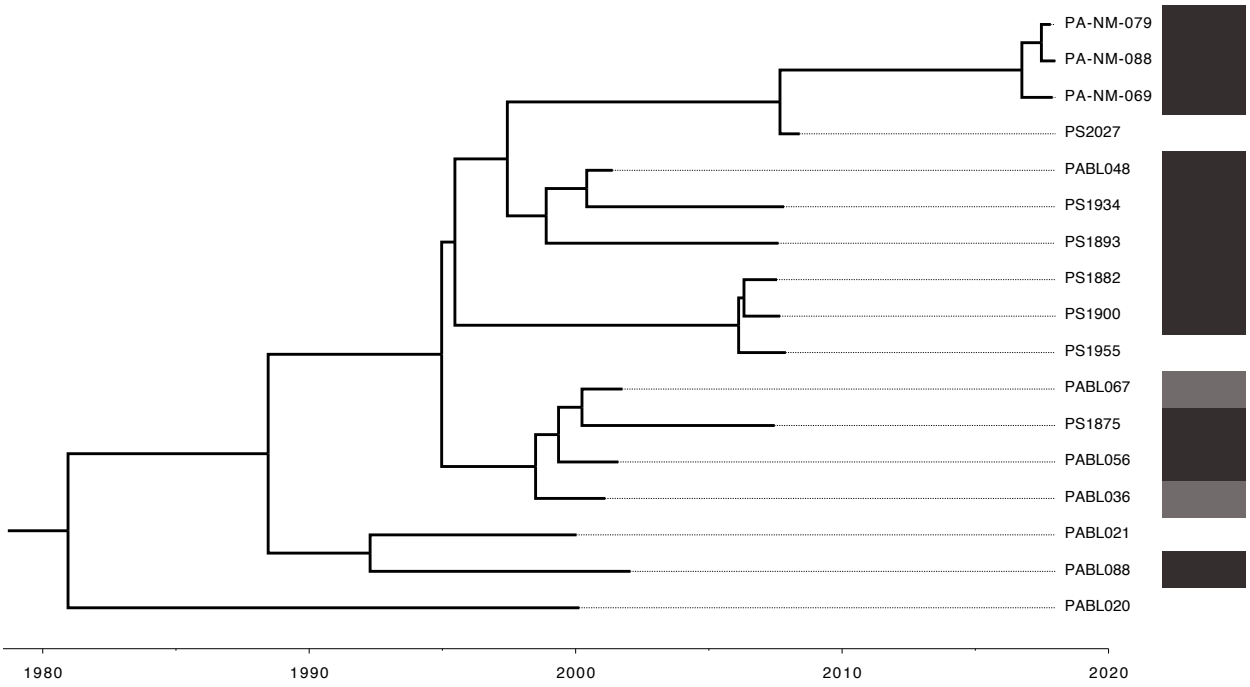


Figure 2.8. Time-scaled phylogenetic tree of ST298* isolates. Tips indicate date of isolation. Root is the estimated last common ancestor of these isolates (mean 1980.9, 95% HPD interval 1973.8-1987.4). The presence of in1697 is indicated by shaded bars on the right, with light grey indicating heterogenous presence in only some colonies for a given isolate.

*Mutational Resistance in ST298**

While pPABL048, likely through in1697, contributes to increased resistance to aminoglycosides and penicillins, acquired AMR genes do not explain the resistance of ST298* isolates to other antibacterials. We investigated whether mutations could explain these resistance patterns. PABL048 harbors a T83I substitution in GyrA and a S87L substitution in ParC that in combination confer high fluoroquinolone resistance[178]. PABL048 also possesses a 4 amino acid deletion (residues 12-15) and 3 single amino acid substitutions in NalC compared to the reference strain PAO1. While the impact of these mutations is unknown, inactivation of NalC increases resistance to multiple antibacterials through MexAB-OprM efflux pump overproduction[146]. Mutations impacting the porin OprD can play a role in carbapenem resistance[144]. ST298* isolates, all of which are meropenem non-susceptible, show multiple amino acid substitutions of unclear significance in OprD compared to PAO1. The ST298* isolates with the highest meropenem resistance (PS2027, PA-NM-069, PA-NM-079, PA-NM-088) show both amino acid deletions from residues 12-54 as well as amino acid substitutions in residues 2-10 (Figure 2.9). Ceftazidime resistance in the ST298* isolates PS1793, PS1796 and PS1797 is likely secondary to a deletion of amino acid residues 2-30 in AmpD, leading to AmpC overproduction[149] (Figure 2.10). These three isolates also share 2-amino-acid substitutions in the plasmid-borne OXA-10, which may confer extended spectrum β-lactamase activity. These substitutions include the G157D substitution previously seen in the extended spectrum OXA-10 variant OXA-14[187] as well as a F153S substitution (Figure 2.11). These findings suggest that ST298* isolates have accumulated mutations that confer antibiotic resistance.

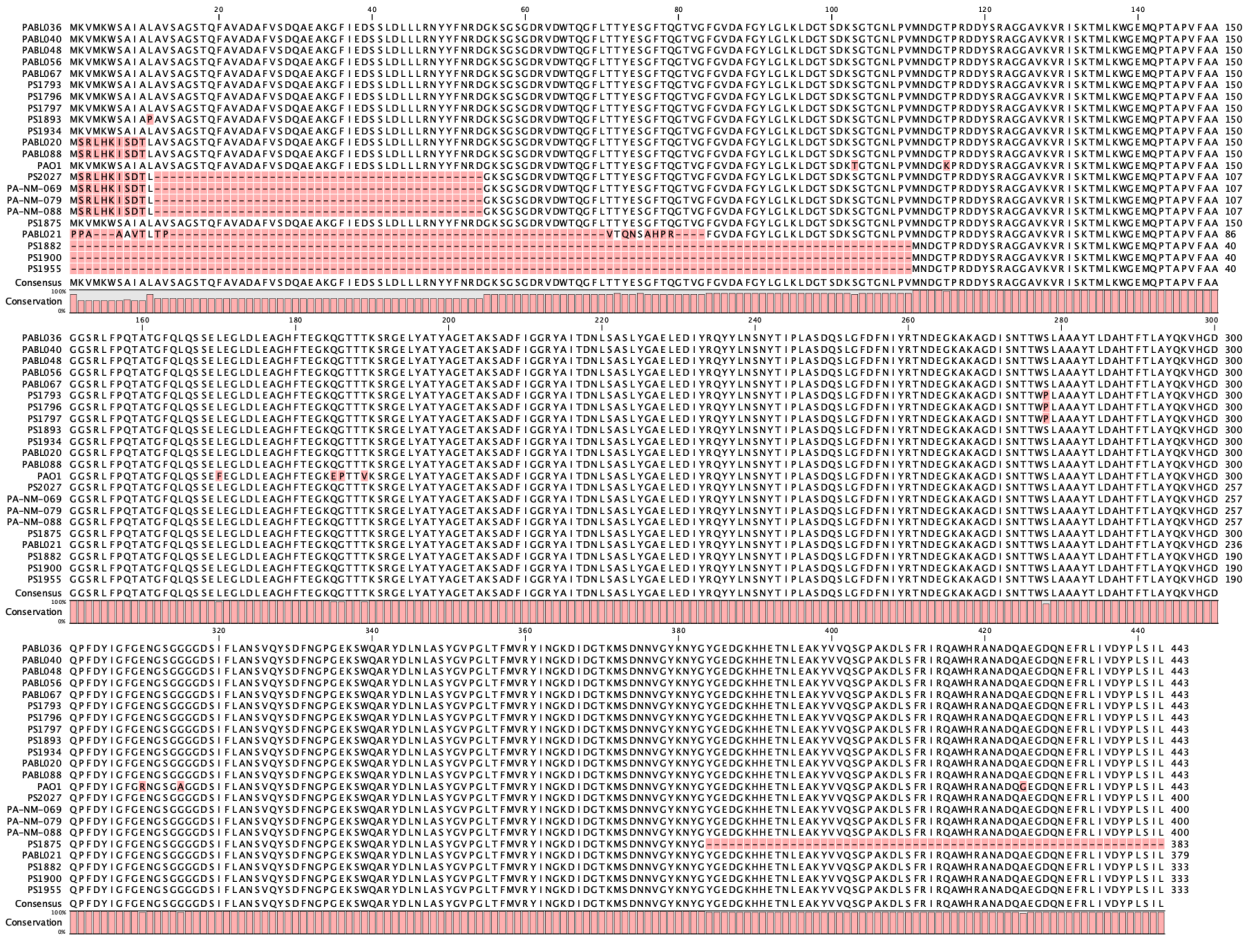


Figure 2.9 Multiple alignment of OprD protein sequences from ST298* isolates. Deviations from the consensus sequence are highlighted in pink. The sequence for PAO1 OprD is included as a reference.

Figure 2.10 Multiple alignment of AmpD protein sequences from ST298* isolates. Deviations from the consensus sequence are highlighted in pink. The sequence for PAO1 AmpD is included as a reference.

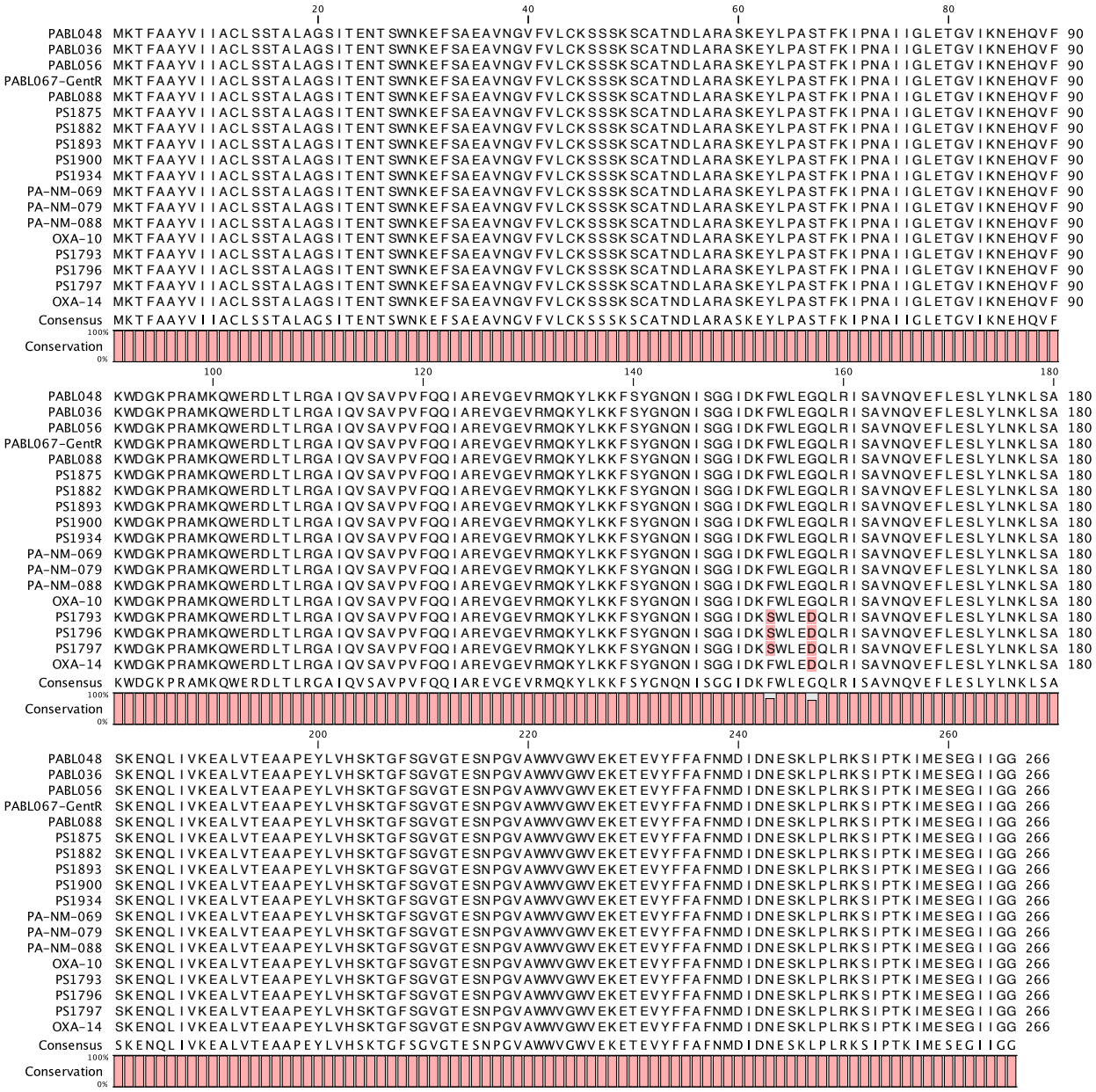


Figure 2.11 Multiple alignment of OXA-10 protein sequences from ST298* isolates possessing in1697, highlighting potential extended spectrum OXA-10 variants. Deviations from the consensus sequence are highlighted in pink. The sequence for OXA-10 and OXA-14 (a known extended spectrum variant) is included as a reference. For PABL067, sequence from a gentamicin resistant colony (GentR) is considered.

Comparative Genomics of pPABL048

While pPABL048 appears to be exclusive to ST298*, evidence of related plasmids can be seen in other isolates, including LLTF and MPVG from this study (Table 2.4). We identified 16 complete plasmids with substantial sequence alignment (≥70% coverage) to pPABL048 in multiple *Pseudomonas* species (Table 2.7), suggesting that pPABL048 is part of a family of large *Pseudomonas* genus plasmids. No similar plasmids were found in available sequences from non-*Pseudomonas* Gammaproteobacteria. We defined the "plasmid backbone" as sequence positions present in 16/17 of these plasmids (Figure 2.12A). While in1697 is not part of the backbone, other genetic features, such as replication and partitioning genes, a chemotaxis locus, putative pilus locus, and a tellurium resistance locus are common to these plasmids. The "backbone" replication protein gene common to these plasmids has not been characterized. Of note, other cases of integron-mediated AMR have been described in this family of plasmids[95,185,313]. To identify additional *Pseudomonas* isolates that may carry pPABL048-like plasmids, we screened publicly available draft genomes and identified 32 with >70% alignment to pPABL048 (Table 2.8). Phylogenetic analysis of all 63 pPABL048-like sequence alignments show that they do not appear to segregate by species (Figure 2.12B-C). Additionally, ST298* pPABL048 alignments form a distinct group, showing that pPABL048 itself has not been previously reported. These results that pPABL048 is a novel member of a family of large *Pseudomonas* genus plasmids.

Table 2.7 Previously Identified *Pseudomonas* Genus Plasmids With >70% Alignment to pPABL048

Name	Species	GenBank Accession	Size (bp)	Strain	Collection Date	Strain ST	Query Coverage[a]	Query Identity[a]	Percent Alignment[b]	SNVs[b]
pPABL048	*aeruginosa*	NA	414954	PABL048	2001	298	NA	NA	NA	NA
AR439_plasmid_unnamed2	*aeruginosa*	CP029096.1	437392	AR439	NA	179	92%	99%	91.6	5265
p12939-OXA	*aeruginosa*	MF344569.1	496436	NA	NA	NA	89%	99%	88.1	5873
pJB37	*aeruginosa*	KY494864.1	464804	FFUP_PS_37	NA	NA	87%	99%	86.3	8372
RW109_plasmid_1	*aeruginosa*	LT969519.1	555265	RW109	NA	111	85%	99%	84.6	2673
pOZ176	*aeruginosa*	KC543497.1	500839	PA96	2000	1129	85%	99%	84.1	7775
AR_0356_plasmid_unnamed2	*aeruginosa*	CP027170.1	438531	AR_0356	NA	1006	85%	99%	83.9	4080
AR441_plasmid_unnamed3	*aeruginosa*	CP029094.1	438529	AR441	NA	1006	85%	99%	83.9	4082
pBM413	*aeruginosa*	CP016215.1	423017	PA121617	2012	389	82%	99%	82.6	6511
p727-IMP	*aeruginosa*	MF344568	430173	NA	NA	NA	78%	99%	77.4	6146
pA681-IMP	*aeruginosa*	MF344570.1	397519	NA	NA	NA	73%	98%	73.8	6067
pR31014-IMP	*aeruginosa*	MF344571.1	374000	NA	NA	NA	70%	99%	70.2	5396
pRBL16	*citronellolis*	CP015879.1	370338	SJTE-3	2015	NA	82%	99%	82.8	6095
P19E3_plasmid_p1	*koreensis*	CP027478.1	467568	P19E3	2014	NA	85%	99%	85.5	5635
pTTS12	*putida*	CP009975.1	583900	S12	1989	NA	84%	98%	82.0	7776
pSY153-MDR	*putida*	KY883660.1	468170	SY153	2012	NA	81%	98%	81.3	6321
p12969-DIM	*putida*	KU130294.1	409102	12969	2013	NA	78%	98%	77.1	9631

[a]Based on BLASTn with pPABL048 as query.

[b]Based on NUCmer alignment of plasmid to pPABL048 reference.

Table 2.8 *Pseudomonas* Genus Genomes With >70% Alignment to pPABL048

Name	Species	Biosample Accession	Refseq Accession	Percent Alignment[b]	SNVs[b]
AZPAE14689	*aeruginosa*	SAMN03105468	GCF_000794545.1	94.28	5430
WH-SGI-V-07698	*aeruginosa*	SAMN04128683	GCF_001453725.1	92.84	4699
WH-SGI-V-07237	*aeruginosa*	SAMN04128543	GCF_001450005.1	92.75	4888
105777	*aeruginosa*	SAMN03076164	GCF_001560865.1	87.7	7875
WH-SGI-V-07253	*aeruginosa*	SAMN04128704	GCF_001452905.1	87.32	5736
WH-SGI-V-07711 (LLTF)[a]	*aeruginosa*	SAMN04128696	GCF_001452765.1	86.77	5859
AZPAE14838	*aeruginosa*	SAMN03105539	GCF_000794335.1	86.41	7898
AZPAE14863	*aeruginosa*	SAMN03105562	GCF_000790145.1	86.01	5576
GTC 10899	*monteilii*	SAMD00031681	GCF_001753835.1	85.69	5970
AZPAE14956	*aeruginosa*	SAMN03105654	GCF_000791685.1	85.38	8605
BWHPSA028	*aeruginosa*	SAMN02360700	GCF_000481145.1	84.65	2380
isolate 15.111a	*aeruginosa*	SAMEA3296128	GCF_001374055.1	84.63	2608
WH-SGI-V-07492	*aeruginosa*	SAMN04128611	GCF_001453185.1	84.31	5944
M140A	*aeruginosa*	SAMN04966044	GCF_001750425.1	84.19	6462
Stone 130	*aeruginosa*	SAMN01779569	GCF_000478465.2	84.19	7688
WCHP16	*sp.*	SAMN05415086	GCF_001695625.1	84.04	6975
PI1	*sp.*	SAMN03262487	GCF_000812405.1	83.69	8870
AZPAE14840	*aeruginosa*	SAMN03105541	GCF_000789555.1	83.63	7822
WH-SGI-V-07709	*aeruginosa*	SAMN04128694	GCF_001452755.1	83.09	8635
WH-SGI-V-07378	*aeruginosa*	SAMN04128563	GCF_001450265.1	82.89	4436
AZPAE14872	*aeruginosa*	SAMN03105571	GCF_000790355.1	82.85	8196
NBRC 111118	*sp.*	SAMD00031653	GCF_001320085.1	82.55	8084
AZPAE14871	*aeruginosa*	SAMN03105570	GCF_000790325.1	82.35	5936
PA13SY16	*aeruginosa*	SAMN04966046	GCF_001750225.1	82.1	6501
CLB24232 (MPVG)[a]	*aeruginosa*	SAMN05774262	GCF_001909485.1	81.88	3619
S12	*putida*	SAMN02470946	GCF_000287915.1	81.75	7777
WH-SGI-V-07300	*aeruginosa*	SAMN04128737	GCF_001454265.1	81.35	5850
GTC 10897	*monteilii*	SAMD00031680	GCF_001319945.1	81.07	7501
AZPAE14827	*aeruginosa*	SAMN03105528	GCF_000795365.1	80.62	4334
P179	*sp.*	SAMN01779567	GCF_000478485.2	80.2	9103
Isolate 10% 5	*aeruginosa*	SAMEA3296136	GCF_001374215.1	80.07	9236
WH-SGI-V-07165	*aeruginosa*	SAMN04128503	GCF_001449465.1	72.85	6947

[a]CC298 Pseudomonas aeruginosa genome included in this study, see Table 2.2.

[b]Based on NUCmer alignment of plasmid to pPABL048 reference.

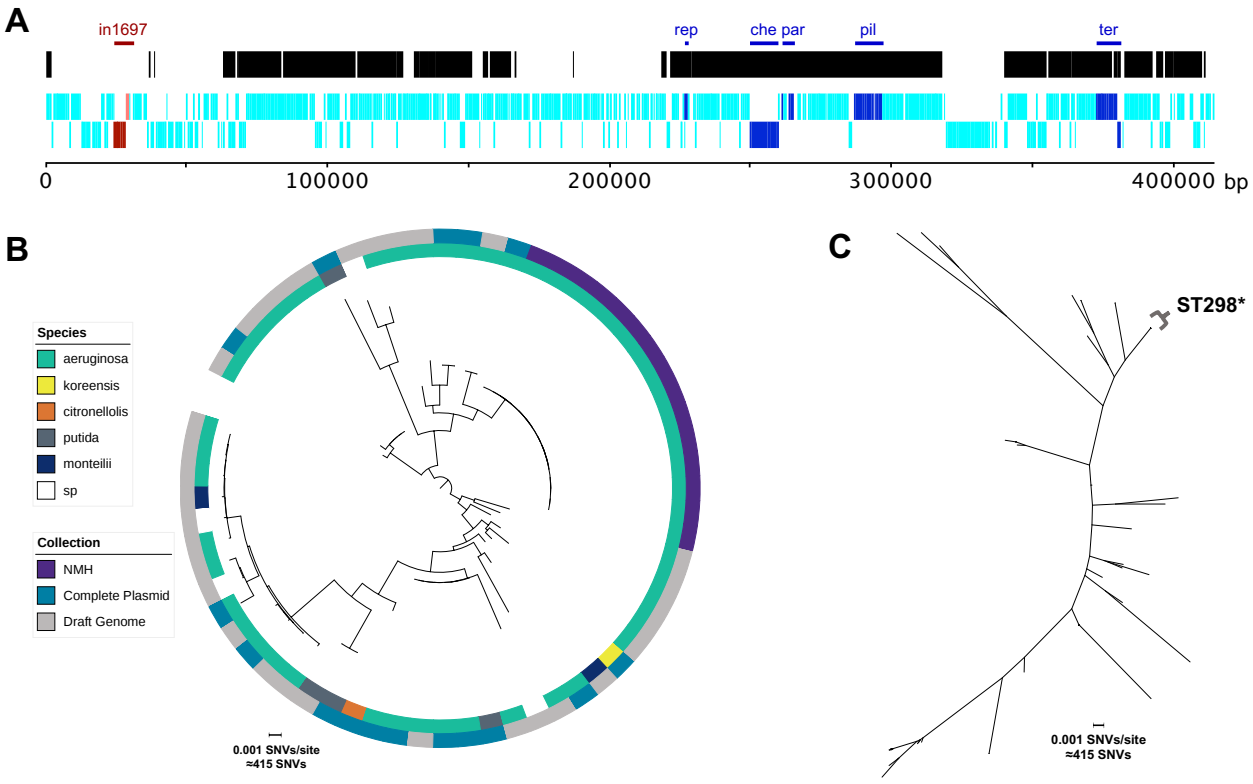


Figure 2.12 Comparative genomic analysis of pPABL048. (A) Linear diagram of pPABL084 showing coding sequences (light blue) and the plasmid backbone (black) defined as positions present in at least 16 of 17 similar plasmids. In1697 is indicated in red. Plasmid backbone features including putative replication (rep) and partitioning (par) genes, chemotaxis locus (che), putative pili locus (pil), and tellurium resistance locus (ter) are indicated in dark blue. (B) Midpoint-rooted circular and (C) unrooted radial maximum likelihood phylogenetic trees based on alignment of 63 *Pseudomonas* genus sequences to pPABL048. Sequences (ST298* read alignments, complete plasmids, and draft genomes) with >70% alignment to pPABL048 by length were included, and SNVs in positions present in 62/63 alignments (plasmid backbone) were considered. The circular tree is annotated with species (inner ring) and collection (outer ring). On the radial tree, pPABL048 alignments from ST298* isolates are indicated.

Discussion

In this study, we identified a subclade of CC446, termed ST298*, that is responsible for a prolonged epidemic of XDR *P. aeruginosa* infections at NMH from at least 2001 through 2017. Extensive antimicrobial resistance in ST298* was due in part to the presence of the large novel AMR plasmid pPABL048, but ST298* isolates lacking this plasmid were still universally MDR. The long-term persistence of ST298* *P. aeruginosa* is clinically significant, both from the standpoint of infection prevention at this institution and in highlighting the potential risk posed by CC446 at healthcare centers in general. Additionally, with its global distribution and multiple incidents of high AMR both from this study and other reports in the literature[65,74,100,192,307], we provide evidence that CC446 is an emerging high-risk lineage of *P. aeruginosa.*

While we were able to identify the XDR subclade ST298* at NMH and show that it has repeatedly caused highly AMR infections, we lack additional epidemiological data to link these cases. However, our findings suggest the existence of a persistent reservoir for ST298* isolates over the last two decades. We hypothesize that this reservoir could be within NMH itself, from a common source outside the hospital (e.g. a long-term acute care hospital), or more widespread throughout Chicago healthcare settings. It is notable that the estimated date of emergence of the last common ancestor for the ST298* subclade (1980, Figure 2.8) is nearly 20 years prior to the opening of the current NMH inpatient facility in 1999. It is important to note that only a limited number of isolates from Chicago came from sources outside of NMH, and we were unable to determine the extent to which ST298* has spread throughout the region. It will be critical to assess whether this lineage is unique to NMH or more widespread. As such, future work integrating both microbiological and epidemiological approaches is needed to identify the reservoir and geographic spread of ST298*.

The plasmid pPABL048 containing the AMR integron in1697 has contributed to the XDR phenotypes of ST298* isolates. While pPABL048 is unique to ST298*, it is part of a family of large *Pseudomonas* genus plasmids. The involvement of both pPABL048 and related plasmids in drug-resistant infections[95,185,313] highlights the clinical importance of this plasmid family. Further investigation is needed to determine the impact of the pPABL048 family of plasmids on bacterial phenotypes that could contribute to increased persistence or fitness, with a specific focus on predicted virulence factors that may affect adhesion, motility, and carbon storage (Supplementary Table 2.2).

Although recognized high-risk clones such as ST235, ST111, and ST175 are enriched for antibiotic resistance and cause a large proportion of MDR/XDR infections worldwide[129,140,192,199,200], relatively susceptible isolates from these STs also occur[65,130,199,200]. Additionally, the genetic basis for AMR in these STs are diverse[129,132], suggesting that the propensity to acquire antibiotic resistance is a hallmark of high-risk clones. Our findings show that CC446 has many of these same features. Although our study documents XDR ST298* isolates only at a single institution, CC446 organisms as a whole are responsible for clinically-significant infections worldwide (Figure 2.2). While not all CC446 isolates tested in this study were MDR/XDR, the frequency of these phenotypes unmasks the high potential for AMR within this clonal complex. Previous findings support this, with multiple geographically distinct reports of MDR CC446 isolates[65,74,100,192,307]. Mechanisms of resistance among these isolates are varied and include other AMR plasmids[74,100], extended spectrum β-lactamases (e.g. VIM-2)[74,180,181,192,306], and chromosomal mutations[73,74,100,307]. These findings are consistent with the assertion that CC446 represents an emerging high-risk clone with the potential to cause further MDR outbreaks.

CHAPTER 3

Using the *Pseudomonas aeruginosa* genome to predict virulence in a mouse model of bacteremia

Chapter-Specific Acknowledgements

The work described in this chapter was primarily conducted as part of Pincus et al., 2020[314]. I would like to acknowledge work conducted by my coauthors that was essential to the completion of this project. A large portion of the experiments testing the virulence of the PABL and SF isolates in mice were conducted before I started this project by Dr. Jonathan Allen, Dr. Egon Ozer, and other members of the Hauser laboratory. Much of the raw whole-genome sequence data used in this project was generated by various members of the Hauser laboratory prior the project's initiation. Dr. Egon Ozer performed complete genome assembly for the isolates PABL012, PABL017, PAC1, and PAC6 from the combined long and short read sequence data (along with PABL048 as described in Chapter 2). Additionally, Dr. Ozer wrote scripts that I used to filter draft genome assemblies for short and low-coverage contigs, generate alignments of short reads for all genomes to PAO1, and filter to core genome variant SNV sites. Draft genome assemblies and sequencing reads for the Early Pseudomonas Infection Control (EPIC) isolates were obtained from Dr. Maulin Soneji. Classification of these isolates as persistent or eradicated was obtained from Dr. Sumitra Mitra based on analyses performed by Dr. Maulin Soneji. I have noted in the methods associated with this project in Chapter 5 which experiments and analyses were performed by or with the help of other laboratory members and which analyses I performed using scripts written by Dr. Ozer. I generated all of the figures in this chapter.

Introduction

Pseudomonas aeruginosa is a ubiquitous gram-negative opportunistic pathogen that infects 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 community[2,49,142]. The *P. aeruginosa* species, however, is not monolithic. Instead, it shows a large degree of genomic diversity both through polymorphisms and differences in gene content[47,51,52]. 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 infections at increased risk for poor outcomes 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[47,48]. While only 10-15% of a typical strain's genome is accessory, when combined from all strains these sequences comprise the vast majority of the *P. aeruginosa* pangenome[48,51,54]. Variations in both the core and accessory genomes impact 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 *in vitro* virulence markers[208], and these strains have attenuated virulence in animal models of acute infection[27]. Genomic islands, major components of the accessory genome, are enriched for predicted virulence factors[315]. 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 models[24,75,76]. We recently identified, within the accessory genome, multiple novel virulence determinants in a mouse model of bacteremia[70]. Conversely, a study using a *Caenorhabditis elegans* model identified several *P. aeruginosa* accessory genes whose presence reduced virulence[31]. Further, the presence of active CRISPR systems was associated with increased virulence[31], supporting the hypothesis that many horizontally transferred elements are genetic parasites with respect to the host bacterium[63]. 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 prediction, however, is not 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[25,261], but some strains naturally lacking *exoU* are more virulent than those possessing the gene[70]. As virulence is a complex and combinatorial phenotype, the strategy taken to study it must be appropriately robust to that complexity.

In supervised machine learning, samples belonging to known classes are used to build a computational model which can then predict the class of new samples[273]. Supervised 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 approach has proven successful in a variety of species and using a variety of genomic features[66,274-279]. These studies 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 cytotoxicity[296], using variation in core genome loci to predict

patient mortality in specific *S. aureus* clones[282], and using predicted perturbations in protein coding sequences to classify *Salmonella* strains as causing either gastrointestinal or extraintestinal infections[280]. A support vector machine approach has been used to distinguish the transcriptomes of *P. aeruginosa* in human infection compared to in vitro growth[281]. However, to our knowledge there has been no study directly modeling *P. aeruginosa* pathogenicity from genomic content.

In this study, we utilize a supervised 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. The predictions appear 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 whether the *P. aeruginosa* genome can be used 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. We used two previously reported collections: 98 archived isolates from adults with bacteremia at Northwestern Memorial Hospital (NMH) in Chicago, USA[316] 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 Taiwan[317] (Table 3.1). These 115 isolates formed our training set. We performed whole genome sequencing for each of the isolates that had not been previously sequenced. Likewise, we supplemented

previously reported virulence data[70,317] with additional experiments (Supplementary Table 3.1) 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, we estimated a modified 50% lethal dose (mLD_{50}, termed as such because it includes pre-lethal illness) for each of the 115 *P. aeruginosa* isolates (Table 3.2). The isolates showed a median mLD_{50} of 6.9 $\log_{10}$ CFU but 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 isolates[70]. For the purpose of this study, we classified isolates with an estimated mLD_{50} below the median value for the group as "high virulence" and the remainder as "low virulence" (Figure 3.1). 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 3.2). 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*[51,52]. One distinct outlier isolate from the PA7-like clade was also present in the collection[51]. 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.

Table 3.1 *P. aeruginosa* Isolates Included in This Study

Name	Train/Test Set	Location	BioSample Accession[b]	Genome Size (bp)	Contigs (#)	Virulence (High: Rounded LD50 < Train Set Median)	Initial Report of Isolate	Initial Report of Sequencing	Report of Assembly Used in This Study
PABL001	Train	Chicago, USA	SAMN09831318	6856138	156	High	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL002	Train	Chicago, USA	SAMN09831319	6792893	126	High	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Bachta et al., *Nat Commun* (2020) / Allen et al., *PNAS* (2020)	This Study
PABL003	Train	Chicago, USA	SAMN09831320	6406489	212	High	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL004	Train	Chicago, USA	SAMN09831321	6480052	133	Low	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL006	Train	Chicago, USA	SAMN09831322	6812450	467	High	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL007	Train	Chicago, USA	SAMN09831323	6381929	105	High	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL009	Train	Chicago, USA	SAMN09831324	6819960	431	Low	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL010	Train	Chicago, USA	SAMN09831325	6857215	457	Low	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL011	Train	Chicago, USA	SAMN09831326	6626407	266	Low	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL012	Train	Chicago, USA	SAMN09831327	6584048	2	High	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Bachta et al., *Nat Commun* (2020) / Allen et al., *PNAS* (2020) / This Study	This Study
PABL013	Train	Chicago, USA	SAMN09831328	6768256	180	High	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL014	Train	Chicago, USA	SAMN09831329	6491067	288	Low	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL015	Train	Chicago, USA	SAMN09831330	6518289	318	Low	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL016	Train	Chicago, USA	SAMN09831331	6453620	132	High	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Bachta et al., *Nat Commun* (2020) / Allen et al., *PNAS* (2020)	This Study
PABL017	Train	Chicago, USA	SAMN09831332	6528721	1	High	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Bachta et al., *Nat Commun* (2020) / Allen et al., *PNAS* (2020) / This Study	This Study
PABL018	Train	Chicago, USA	SAMN09831333	6357732	336	Low	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL020	Train	Chicago, USA	SAMN09831335	6756276	495	Low	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Pincus et al., *CID* (2019)	This Study
PABL021	Train	Chicago, USA	SAMN09831336	6504687	503	Low	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Pincus et al., *CID* (2019)	This Study
PABL022	Train	Chicago, USA	SAMN09831337	6968177	152	High	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Pincus et al., *CID* (2019)	This Study
PABL023	Train	Chicago, USA	SAMN09831338	6415121	216	High	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL024	Train	Chicago, USA	SAMN09831339	6486560	297	High	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL026	Train	Chicago, USA	SAMN09831340	6541241	136	High	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL027	Train	Chicago, USA	SAMN09831341	6673341	318	Low	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL028	Train	Chicago, USA	SAMN09831342	6366972	106	Low	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Bachta et al., *Nat Commun* (2020) / Allen et al., *PNAS* (2020)	This Study
PABL029	Train	Chicago, USA	SAMN09831343	6963015	611	High	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL030	Train	Chicago, USA	SAMN09831344	6370252	184	High	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Bachta et al., *Nat Commun* (2020) / Allen et al., *PNAS* (2020)	This Study
PABL031	Train	Chicago, USA	SAMN09831345	6757528	170	High	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL032	Train	Chicago, USA	SAMN09831346	6515492	178	High	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL034	Train	Chicago, USA	SAMN09831347	6598709	133	High	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL035	Train	Chicago, USA	SAMN09831348	6782268	319	Low	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL036	Train	Chicago, USA	SAMN09831349	7226449	375	Low	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Pincus et al., *CID* (2019)	This Study

Table 3.1 Continued

Name	Train/Test Set	Location	BioSample Accession[b]	Genome Size (bp)	Contigs (#)	Virulence (High: Rounded LD50 < Train Set Median)	Initial Report of Isolate	Initial Report of Sequencing	Report of Assembly Used in This Study
PABL037	Train	Chicago, USA	SAMN09831350	6977344	140	Low	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL038	Train	Chicago, USA	SAMN09831351	6820361	234	Low	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL040	Train	Chicago, USA	SAMN09831352	6838534	432	Low	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Pincus et al., *CID* (2019)	This Study
PABL041	Train	Chicago, USA	SAMN09831353	6712192	219	High	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Bachta et al., *Nat Commun* (2020) / Allen et al., *PNAS* (2020)	This Study
PABL042	Train	Chicago, USA	SAMN09831354	6922755	208	Low	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL043	Train	Chicago, USA	SAMN09831355	6277849	271	Low	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL044	Train	Chicago, USA	SAMN09831356	6961975	392	Low	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL046	Train	Chicago, USA	SAMN09831358	6206873	87	Low	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Bachta et al., *Nat Commun* (2020) / Allen et al., *PNAS* (2020)	This Study
PABL047	Train	Chicago, USA	SAMN09831359	6585916	203	Low	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL048	Train	Chicago, USA	SAMN09831360	7294576	2	Low	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Pincus et al., *CID* (2019)	Pincus et al., *CID* (2019)
PABL049	Train	Chicago, USA	SAMN09831361	6861902	221	High	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Bachta et al., *Nat Commun* (2020) / Allen et al., *PNAS* (2020)	This Study
PABL051	Train	Chicago, USA	SAMN09831362	6227393	254	Low	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL052	Train	Chicago, USA	SAMN09831363	6609700	218	Low	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL053	Train	Chicago, USA	SAMN09831364	6814304	286	Low	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL054	Train	Chicago, USA	SAMN09831365	6972705	232	Low	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL055	Train	Chicago, USA	SAMN09831366	6779537	190	Low	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL056	Train	Chicago, USA	SAMN09831367	7222869	353	Low	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Pincus et al., *CID* (2019)	This Study
PABL057	Train	Chicago, USA	SAMN09831368	6428244	393	Low	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL058	Train	Chicago, USA	SAMN09831369	6910579	231	Low	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL059	Train	Chicago, USA	SAMN09831370	6761248	237	Low	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL060	Train	Chicago, USA	SAMN09831371	6473694	216	High	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL061	Train	Chicago, USA	SAMN09831372	6391296	104	High	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL062	Train	Chicago, USA	SAMN09831373	6889022	202	Low	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL063	Train	Chicago, USA	SAMN09831374	6863118	428	Low	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL064	Train	Chicago, USA	SAMN09831375	6912173	513	Low	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL065	Train	Chicago, USA	SAMN09831376	6695382	454	Low	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL066	Train	Chicago, USA	SAMN09831377	6791517	285	Low	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL067	Train	Chicago, USA	SAMN09831378	6802331	328	Low	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Pincus et al., *CID* (2019)	This Study
PABL068	Train	Chicago, USA	SAMN09831379	6396063	124	High	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL069	Train	Chicago, USA	SAMN09831380	6650002	159	Low	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL070	Train	Chicago, USA	SAMN09831381	6536641	194	High	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL071	Train	Chicago, USA	SAMN09831382	6554431	234	Low	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL072	Train	Chicago, USA	SAMN09831383	6726621	164	High	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Pincus et al., *CID* (2019)	This Study

Table 3.1 Continued

Name	Train/Test Set	Location	BioSample Accession[b]	Genome Size (bp)	Contigs (#)	Virulence (High: Rounded LD50 < Train Set Median)	Initial Report of Isolate	Initial Report of Sequencing	Report of Assembly Used in This Study
PABL073	Train	Chicago, USA	SAMN09831384	6351380	127	Low	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL074	Train	Chicago, USA	SAMN09831385	6754467	207	Low	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL075	Train	Chicago, USA	SAMN09831386	6756836	433	Low	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL076	Train	Chicago, USA	SAMN09831387	6429295	145	Low	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL077	Train	Chicago, USA	SAMN09831388	6395399	103	High	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL078	Train	Chicago, USA	SAMN09831389	6435111	211	High	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL079	Train	Chicago, USA	SAMN09831390	6568448	909	High	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL080	Train	Chicago, USA	SAMN09831391	6561442	249	Low	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL081	Train	Chicago, USA	SAMN09831392	6648976	480	Low	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL082	Train	Chicago, USA	SAMN09831393	6988878	190	Low	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL083	Train	Chicago, USA	SAMN09831394	6631960	121	High	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL084	Train	Chicago, USA	SAMN09831395	6722847	203	High	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL085	Train	Chicago, USA	SAMN09831396	7043013	269	High	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL086	Train	Chicago, USA	SAMN09831397	6936883	213	Low	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL088	Train	Chicago, USA	SAMN09831398	7153658	277	Low	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Pincus et al., *CID* (2019)	This Study
PABL089	Train	Chicago, USA	SAMN09831399	6761681	184	Low	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL090	Train	Chicago, USA	SAMN09831400	6834993	176	High	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL091	Train	Chicago, USA	SAMN09831401	6734477	345	Low	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL092	Train	Chicago, USA	SAMN09831402	6733564	205	Low	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL093	Train	Chicago, USA	SAMN09831403	6951684	395	High	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL094	Train	Chicago, USA	SAMN09831404	6986084	414	Low	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL095	Train	Chicago, USA	SAMN09831405	6524119	161	High	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Bachta et al., *Nat Commun* (2020) / Allen et al., *PNAS* (2020)	This Study
PABL096	Train	Chicago, USA	SAMN09831406	6673595	148	High	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL097	Train	Chicago, USA	SAMN09831407	6897174	246	High	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Pincus et al., *CID* (2019)	This Study
PABL098	Train	Chicago, USA	SAMN09831408	6473620	410	Low	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL100	Train	Chicago, USA	SAMN09831409	6935034	218	Low	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL101	Train	Chicago, USA	SAMN09831410	6756723	438	Low	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL102	Train	Chicago, USA	SAMN09831411	6424468	126	High	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL103	Train	Chicago, USA	SAMN09831412	6711993	192	Low	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL104	Train	Chicago, USA	SAMN09831413	6731743	176	Low	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL105	Train	Chicago, USA	SAMN09831414	6887663	122	Low	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL106	Train	Chicago, USA	SAMN09831415	6677390	202	High	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PABL107	Train	Chicago, USA	SAMN09831416	6615031	112	High	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Bachta et al., *Nat Commun* (2020) / Allen et al., *PNAS* (2020)	This Study

Table 3.1 Continued

Name	Train/Test Set	Location	BioSample Accession[b]	Genome Size (bp)	Contigs (#)	Virulence (High: Rounded LD50 < Train Set Median)	Initial Report of Isolate	Initial Report of Sequencing	Report of Assembly Used in This Study
PABL108	Train	Chicago, USA	SAMN09831417	6602827	202	High	Scheetz et al., *Diagn Microbiol Infect Dis* (2009)	Allen et al., *PNAS* (2020)	This Study
PAC1	Train	Taiwan	SAMN14970706	7605607	4	High	Chuang et al., *Gut* (2014)	This Study	This Study
PAC6	Train	Taiwan	SAMN14970707	6560636	1	High	Chuang et al., *Gut* (2014)	Bachta et al., *Nat Commun* (2020) / This Study	This Study
S10[a]	Train	Taiwan	SAMN14970708	6798627	143	High	Chuang et al., *Gut* (2014)	Bachta et al., *Nat Commun* (2020) / This Study	This Study
S11	Train	Taiwan	SAMN14970709	6446879	123	High	Chuang et al., *Gut* (2014)	This Study	This Study
S12	Train	Taiwan	SAMN14970710	6383022	99	High	Chuang et al., *Gut* (2014)	This Study	This Study
S13	Train	Taiwan	SAMN14970711	6736624	121	High	Chuang et al., *Gut* (2014)	This Study	This Study
S14	Train	Taiwan	SAMN14970712	7164601	180	High	Chuang et al., *Gut* (2014)	This Study	This Study
S15	Train	Taiwan	SAMN14970713	6341129	91	High	Chuang et al., *Gut* (2014)	This Study	This Study
S16	Train	Taiwan	SAMN14970714	7164497	178	Low	Chuang et al., *Gut* (2014)	This Study	This Study
S17	Train	Taiwan	SAMN14970715	7033494	143	High	Chuang et al., *Gut* (2014)	This Study	This Study
S2	Train	Taiwan	SAMN14970716	6312249	77	High	Chuang et al., *Gut* (2014)	This Study	This Study
S3	Train	Taiwan	SAMN14970717	6321387	117	High	Chuang et al., *Gut* (2014)	This Study	This Study
S4	Train	Taiwan	SAMN14970718	6525413	138	High	Chuang et al., *Gut* (2014)	This Study	This Study
S5	Train	Taiwan	SAMN14970719	6543702	106	High	Chuang et al., *Gut* (2014)	This Study	This Study
S7	Train	Taiwan	SAMN14970720	6175995	87	High	Chuang et al., *Gut* (2014)	This Study	This Study
S8	Train	Taiwan	SAMN14970721	6367209	103	High	Chuang et al., *Gut* (2014)	This Study	This Study
S9	Train	Taiwan	SAMN14970722	6427741	97	High	Chuang et al., *Gut* (2014)	This Study	This Study
PASP048	Test	Spain	SAMN14970723	6757144	110	High	Peña et al., *AAC* (2012)	This Study	This Study
PASP146	Test	Spain	SAMN14970724	6330026	59	High	Peña et al., *AAC* (2012)	This Study	This Study
PASP170	Test	Spain	SAMN12162692	6859625	96	Low	Peña et al., *AAC* (2012)	Pincus et al., *CID* (2019)	Pincus et al., *CID* (2019)
PASP198	Test	Spain	SAMN14970725	6279554	75	High	Peña et al., *AAC* (2012)	This Study	This Study
PASP204	Test	Spain	SAMN14970726	6798486	88	High	Peña et al., *AAC* (2012)	This Study	This Study
PASP208	Test	Spain	SAMN14970727	6855539	173	High	Peña et al., *AAC* (2012)	This Study	This Study
PASP251	Test	Spain	SAMN14970728	6875145	85	High	Peña et al., *AAC* (2012)	This Study	This Study
PASP269	Test	Spain	SAMN14970729	6411445	71	High	Peña et al., *AAC* (2012)	This Study	This Study
PASP309	Test	Spain	SAMN14970730	7020830	145	High	Peña et al., *AAC* (2012)	This Study	This Study
PASP315	Test	Spain	SAMN14970731	6910212	128	High	Peña et al., *AAC* (2012)	This Study	This Study
PASP352	Test	Spain	SAMN14970732	6712762	78	High	Peña et al., *AAC* (2012)	This Study	This Study
PASP388	Test	Spain	SAMN14970733	7012486	160	Low	Peña et al., *AAC* (2012)	This Study	This Study
PASP398	Test	Spain	SAMN14970734	6465306	99	Low	Peña et al., *AAC* (2012)	This Study	This Study
PASP415	Test	Spain	SAMN14970735	6074119	62	High	Peña et al., *AAC* (2012)	This Study	This Study
PASP450	Test	Spain	SAMN14970736	6994850	179	Low	Peña et al., *AAC* (2012)	This Study	This Study
PASP453	Test	Spain	SAMN14970737	6794060	162	High	Peña et al., *AAC* (2012)	This Study	This Study
PASP471	Test	Spain	SAMN14970738	6492447	94	High	Peña et al., *AAC* (2012)	This Study	This Study
PASP475	Test	Spain	SAMN14970739	6494538	114	High	Peña et al., *AAC* (2012)	This Study	This Study
PASP496	Test	Spain	SAMN14970740	6432527	82	Low	Peña et al., *AAC* (2012)	This Study	This Study
PASP499	Test	Spain	SAMN14970741	7121993	174	Low	Peña et al., *AAC* (2012)	This Study	This Study
PASP518	Test	Spain	SAMN14970742	7043592	264	Low	Peña et al., *AAC* (2012)	This Study	This Study
PASP527	Test	Spain	SAMN14970743	6818948	135	High	Peña et al., *AAC* (2012)	This Study	This Study
PASP612	Test	Spain	SAMN14970744	6962611	209	Low	Peña et al., *AAC* (2012)	This Study	This Study
PASP639	Test	Spain	SAMN14970745	7145660	144	Low	Peña et al., *AAC* (2012)	This Study	This Study
PASP657	Test	Spain	SAMN14970746	6836322	134	Low	Peña et al., *AAC* (2012)	This Study	This Study

[a]Isolated from ascites fluid. All other isolates in this study were from blood cultures

[b]For all isolates, the version of the assemblies used in this study are available at https://github.com/nathanpincus/PA_Virulence_Prediction

Table 3.2 Estimated mLD_{50} Values for Isolates Included in this Study

Name	Estimated LD50 (log_{10} CFU)	Standard Deviation	Rounded LD50 (log_{10} CFU)
PABL001	6.69	0.10	6.7
PABL002	6.36	0.28	6.4
PABL003	6.49	0.21	6.5
PABL004	7.48	4.63	7.5
PABL006	6.50	0.10	6.5
PABL007	6.45	0.34	6.5
PABL009	7.28	0.23	7.3
PABL010	7.43	12.40	7.4
PABL011	7.11	0.16	7.1
PABL012	5.85	0.12	5.8
PABL013	6.23	0.11	6.2
PABL014	7.27	8.60	7.3
PABL015	7.40	0.53	7.4
PABL016	6.30	0.06	6.3
PABL017	6.79	0.16	6.8
PABL018	7.14	23.98	7.1
PABL020	7.53	20.99	7.5
PABL021	7.44	13.65	7.4
PABL022	6.63	0.03	6.6
PABL023	6.63	0.03	6.6
PABL024	6.49	0.12	6.5
PABL026	6.56	0.24	6.6
PABL027	8.28	0.04	8.3
PABL028	7.52	3.15	7.5
PABL029	6.59	18.11	6.6
PABL030	6.84	0.22	6.8
PABL031	6.53	0.89	6.5
PABL032	6.71	0.05	6.7
PABL034	6.76	0.44	6.8
PABL035	6.95	6.24	6.9
PABL036	7.00	23.74	7.0
PABL037	6.99	0.12	7.0
PABL038	7.40	0.00	7.4
PABL040	7.19	4.23	7.2
PABL041	6.64	0.16	6.6
PABL042	7.31	0.14	7.3
PABL043	8.10	5.09	8.1
PABL044	7.20	22.56	7.2
PABL046	7.00	152.45	7.0
PABL047	7.04	14.15	7.0
PABL048	7.04	175.91	7.0
PABL049	6.11	0.14	6.1
PABL051	7.49	2.85	7.5
PABL052	7.69	0.13	7.7
PABL053	7.00	5.89	7.0
PABL054	6.92	0.08	6.9
PABL055	6.97	0.06	7.0
PABL056	7.66	6.86	7.7
PABL057	7.26	0.11	7.3
PABL058	7.60	0.12	7.6
PABL059	7.35	0.27	7.3
PABL060	6.78	20.20	6.8
PABL061	6.46	0.34	6.5
PABL062	7.52	0.29	7.5
PABL063	7.20	0.75	7.2
PABL064	7.43	1.34	7.4
PABL065	7.91	6.94	7.9
PABL066	7.67	2.17	7.7
PABL067	7.36	0.17	7.4
PABL068	6.53	0.01	6.5
PABL069	7.02	0.09	7.0
PABL070	6.51	0.05	6.5
PABL071	7.28	0.05	7.3
PABL072	6.70	0.33	6.7
PABL073	6.87	0.11	6.9
PABL074	7.30	0.30	7.3
PABL075	6.94	0.34	6.9
PABL076	6.94	0.23	6.9
PABL077	6.70	0.57	6.7
PABL078	6.77	0.60	6.8
PABL079	6.56	0.22	6.6

Table 3.2 Continued

Name	Estimated LD50 (log_{10} CFU)	Standard Deviation	Rounded LD50 (log_{10} CFU)
PABL080	7.12	0.48	7.1
PABL081	7.57	0.05	7.6
PABL082	7.84	0.11	7.8
PABL083	6.42	0.07	6.4
PABL084	6.58	9.88	6.6
PABL085	6.70	0.30	6.7
PABL086	7.15	6.18	7.2
PABL088	7.42	3.68	7.4
PABL089	7.32	0.08	7.3
PABL090	6.67	0.12	6.7
PABL091	7.15	0.39	7.1
PABL092	7.66	0.18	7.7
PABL093	6.31	1.32	6.3
PABL094	7.11	0.46	7.1
PABL095	6.68	12.02	6.7
PABL096	6.54	0.14	6.5
PABL097	6.64	0.16	6.6
PABL098	7.00	0.06	7.0
PABL100	6.88	0.35	6.9
PABL101	6.86	0.35	6.9
PABL102	6.25	0.79	6.2
PABL103	8.15	0.08	8.1
PABL104	7.93	0.14	7.9
PABL105	7.66	0.14	7.7
PABL106	6.45	0.45	6.4
PABL107	6.35	0.03	6.4
PABL108	6.83	0.21	6.8
PAC1	6.33	0.07	6.3
PAC6	6.10	0.06	6.1
S10	5.98	0.12	6.0
S11	6.17	0.11	6.2
S12	6.50	0.10	6.5
S13	6.27	1.90	6.3
S14	6.81	11.14	6.8
S15	6.36	2.34	6.4
S16	7.17	0.12	7.2
S17	6.81	0.07	6.8
S2	6.32	0.02	6.3
S3	6.05	0.13	6.1
S4	6.02	0.03	6.0
S5	5.97	0.09	6.0
S7	6.33	0.11	6.3
S8	6.14	0.13	6.1
S9	6.70	0.15	6.7
PASP048	6.54	0.14	6.5
PASP146	6.41	1.94	6.4
PASP170	7.40	0.15	7.4
PASP198	6.47	0.30	6.5
PASP204	6.31	5.95	6.3
PASP208	6.29	2.37	6.3
PASP251	6.32	0.22	6.3
PASP269	6.39	0.04	6.4
PASP309	6.68	0.10	6.7
PASP315	6.49	7.38	6.5
PASP352	6.64	0.30	6.6
PASP388	7.23	6.37	7.2
PASP398	7.32	0.12	7.3
PASP415	6.77	0.45	6.8
PASP450	7.86	0.26	7.9
PASP453	6.28	0.26	6.3
PASP471	6.31	3.40	6.3
PASP475	6.32	0.21	6.3
PASP496	7.43	6.69	7.4
PASP499	7.35	0.01	7.3
PASP518	7.04	0.21	7.0
PASP527	6.31	0.08	6.3
PASP612	7.11	0.18	7.1
PASP639	7.38	0.12	7.4
PASP657	7.83	0.06	7.8

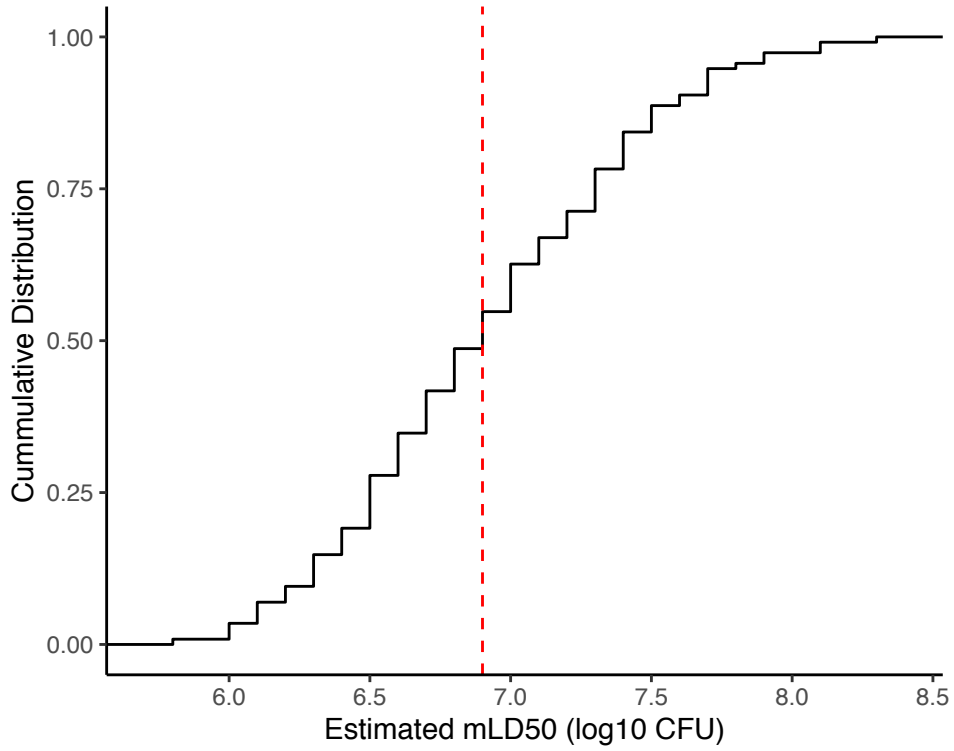


Figure 3.1 Cumulative distribution function of estimated mLD_{50} values for the 115 training isolates in a mouse model of bacteremia. Isolates with estimated mLD_{50} values less than the median value (red dashed line) were designated as high virulence, with the remainder designated as low virulence.

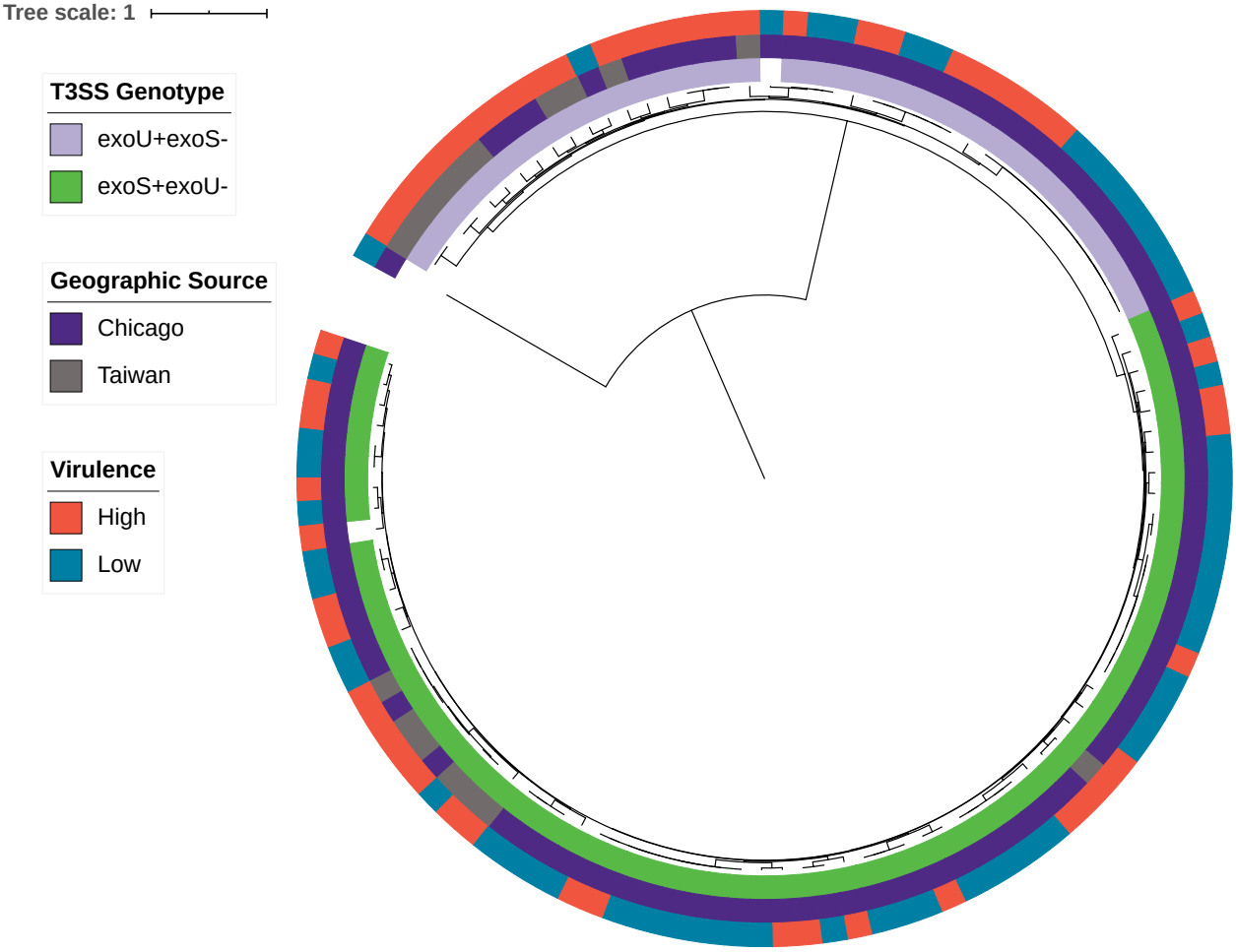


Figure 3.2 Core genome comparisons for the training set of 115 *P. aeruginosa* isolates. Mid-point rooted core genome phylogenetic tree of the 115 training isolates constructed from SNV loci present in at least 95% of genomes, annotated with T3SS genotype, geographic source, and virulence level.

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 others[48]. For the purpose of this study, noncontiguous accessory sequences were grouped and considered as a single AGE if they were perfectly correlated (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 3.2). 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 (Figure 3.3A). Consistent with previous findings[51], 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 3.3).

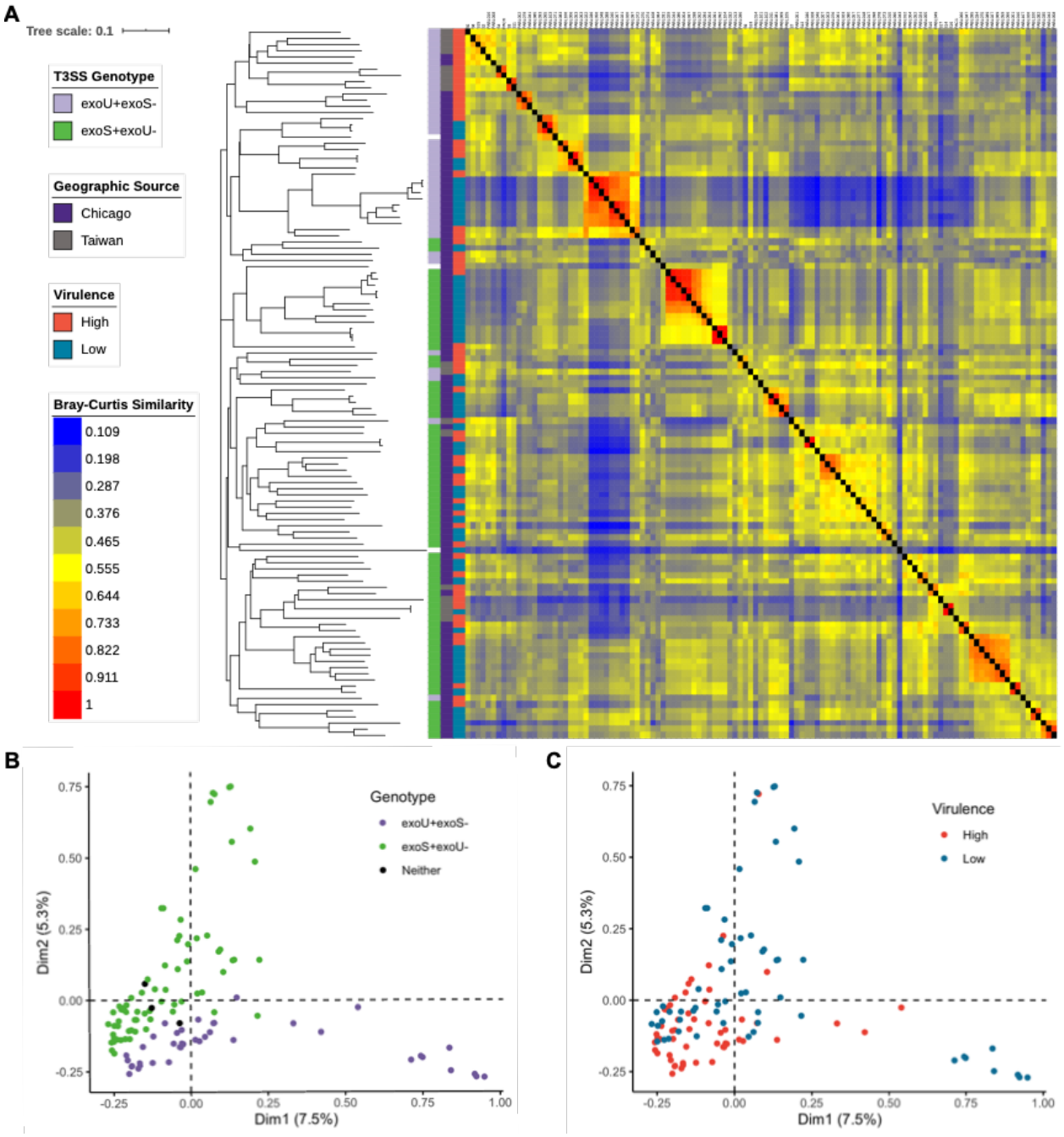

A
Tree scale: 0.1
T3SS Genotype
exoU+exoS-
exoS+exoU-
Geographic Source
Chicago
Taiwan
Virulence
High
Low
Bray-Curtis Similarity
0.109
0.198
0.287
0.376
0.465
0.555
0.644
0.733
0.822
0.911
1
B
Genotype
exoU+exoS-
exoS+exoU-
Neither
Dim2 (5.3%)
Dim1 (7.5%)
C
Virulence
High
Low
Dim2 (5.3%)
Dim1 (7.5%)

Figure 3.3 Accessory genome comparisons for the training set of 115 *P. aeruginosa* isolates. (A) 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, geographic source, virulence level, and the dissimilarity heatmap. A higher value indicates that two isolates have more similar accessory genomes. Multiple correspondence analysis (MCA) performed based on AGE presence/absence in the 115 training set isolates and annotated based on (B) T3SS genotype and (C) virulence level. The first two dimensions, and the percentage of variance they explain, are shown.

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[47,48,71] and includes multiple known virulence determinants[24,70,75], it would contain information predictive of strain virulence in mice. To test this hypothesis, we took a supervised machine learning approach (Figure 3.4). Through 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. During model construction, optimal hyperparameters were chosen using grid-search cross-validation. Here, all possible combinations of hyperparameters were tested through 10-fold cross-validation. The best-performing combination was then used to build a final model. Model performance was estimated using 10-fold nested cross-validation. In this process, grid-search cross-validation was performed within an outer cross-validation loop. For each training fold in this outer loop, a model was built through grid-search cross-validation, and its performance was tested against the cross-validation fold. Nested cross-validation does not return a final machine learning model but instead examines how multiple models 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 3.5). 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.

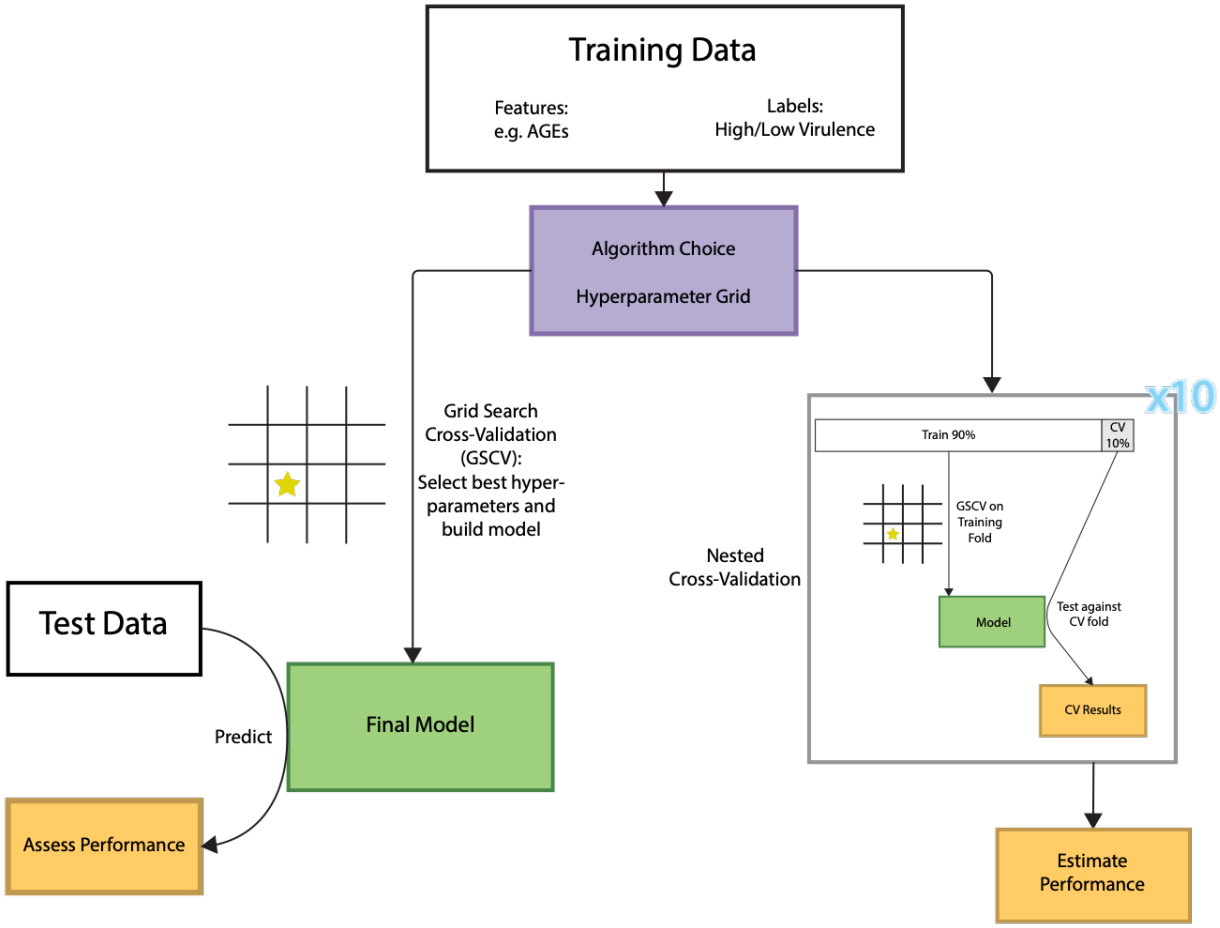


Figure 3.4 Overview of the machine learning pipeline. AGE: accessory genomic element, CV: cross-validation.

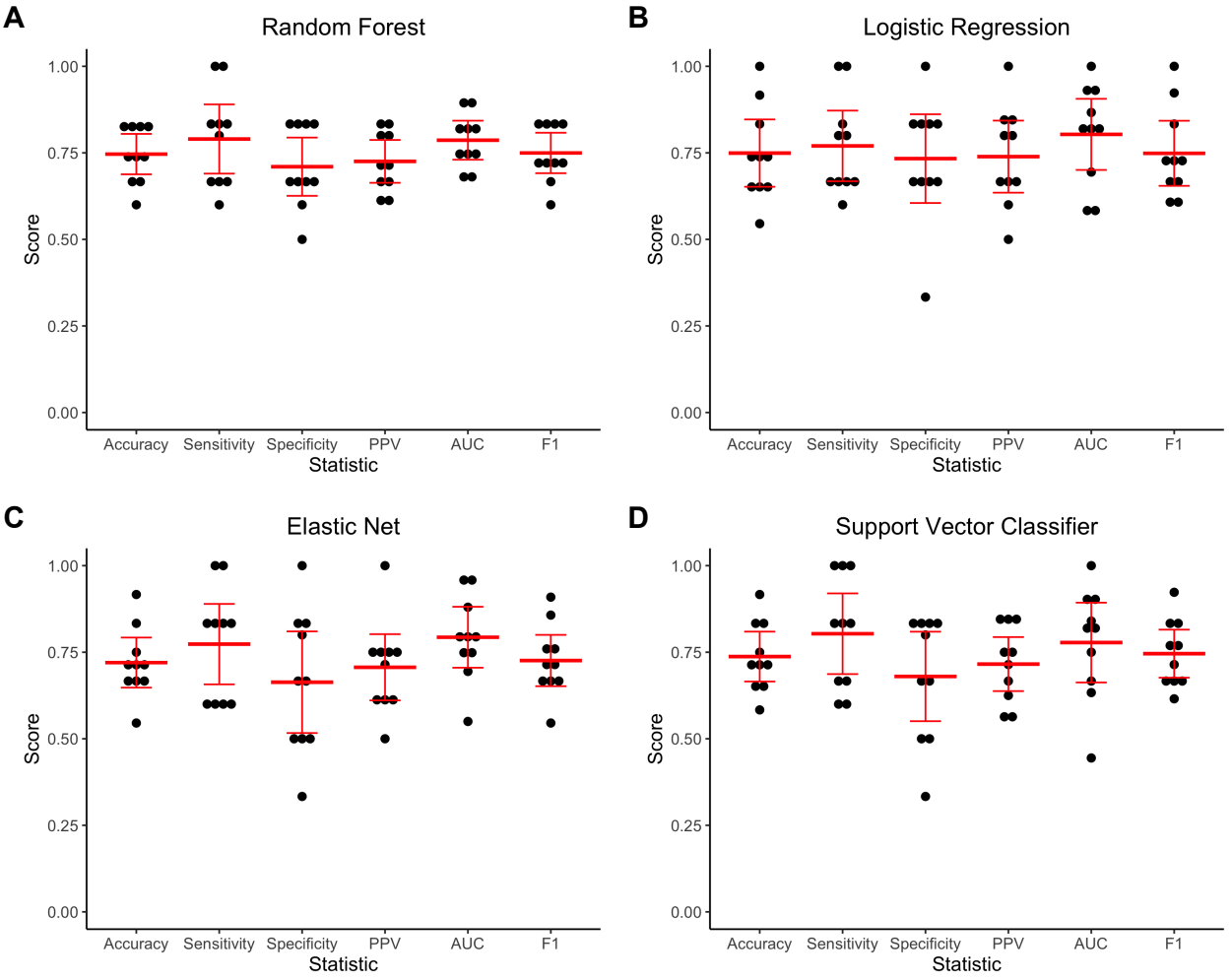


Figure 3.5 Nested 10-fold cross-validation performance of machine learning algorithms in predicting *P. aeruginosa* virulence in mice based on accessory genomic content. (A) Random forest, (B) L2-regularized logistic regression, (C) elastic net logistic regression, and (D) support vector classifier algorithms 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.

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 3.6A). 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 can only consider AGEs contained in the training set and cannot account for the impact of additional AGEs (or different patterns of AGE carriage) found when including new genetically distinct isolates. Learning curves for the other machine learning algorithms similarly showed that there is unlikely to be substantial improvement in cross-validation accuracy if we were to increase the number of training isolates (Figure 3.6B-D).

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 (Table 3.3), 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 (Table 3.3 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.

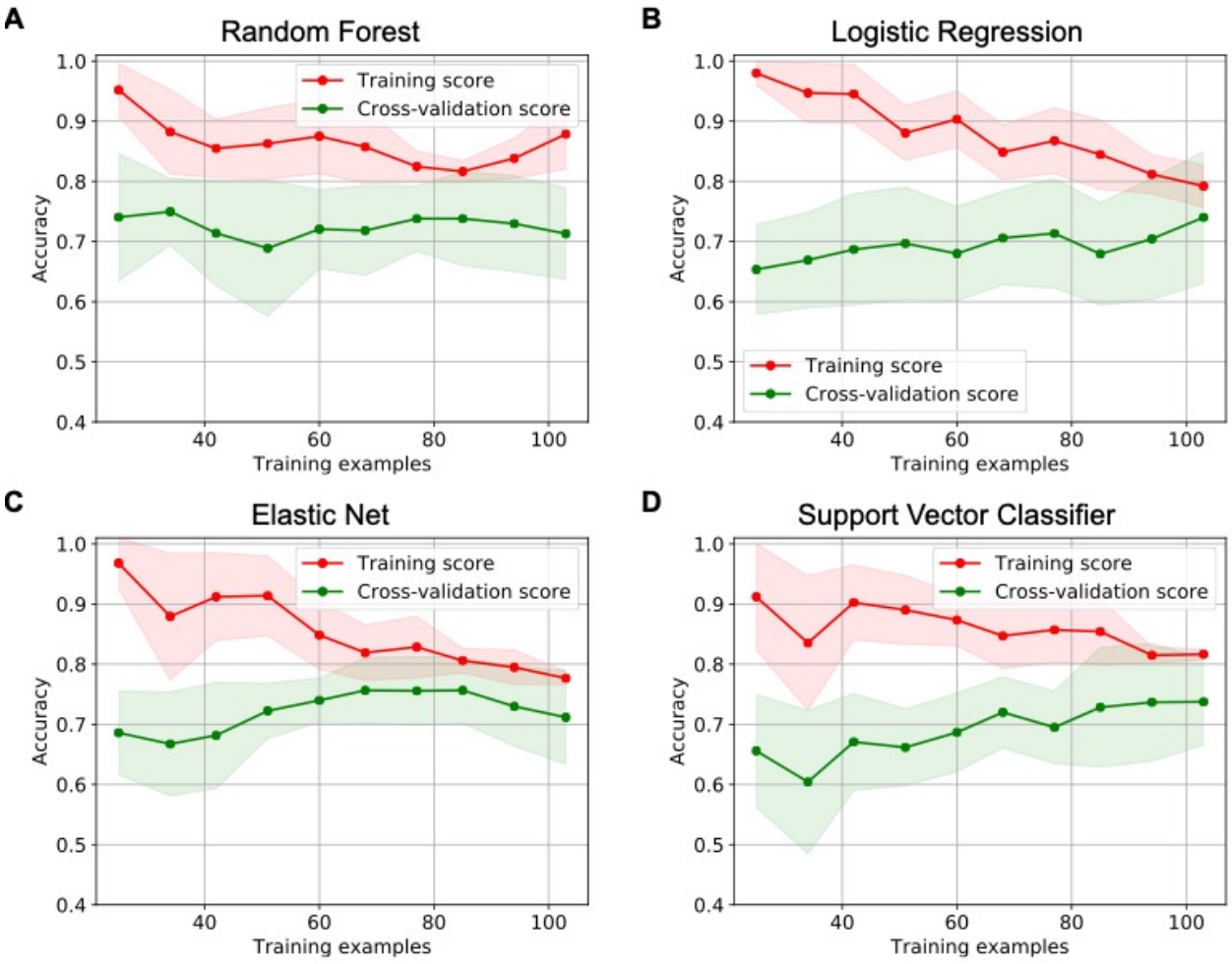


Figure 3.6 Learning curves showing change in mean training accuracy and cross-validation accuracy in predicting *P. aeruginosa* virulence as increasing numbers of isolates are used with different machine learning algorithms. (A) Random forest, (B) L2-regularized logistic regression, (C) elastic net logistic regression, and (D) support vector classifier algorithms were tested. Mean training accuracy (red line) and cross-validation accuracy (green line) are shown. Shading indicates the 95% confidence interval. Assessments at each number of training examples were through 10-fold nested cross-validation.

Table 3.3 Performance of the Accessory Genome Random Forest Model Against the Training Set of 115 P. aeruginosa Isolates

Out-of-bag Accuracy	Training Accuracy	Training Sensitivity	Training Specificity	Training PPV	Training AUC	Training F1
0.75	0.79	0.84	0.75	0.76	0.91	0.80

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 when a given feature is randomly permuted) for each AGE. To do this, we randomly permuted each AGE 100 times and the 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 3.7A). The vast majority of features (2,979/3,013; 98.9%) had no impact on out-of-bag accuracy when randomly permuted (Supplementary Table 3.2), indicating that the machine learning model based decisions on a genomic signature predictive of virulence level rather than by 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 3.7B). As such, it was not possible for a single AGE to have a large impact on the prediction of virulence.

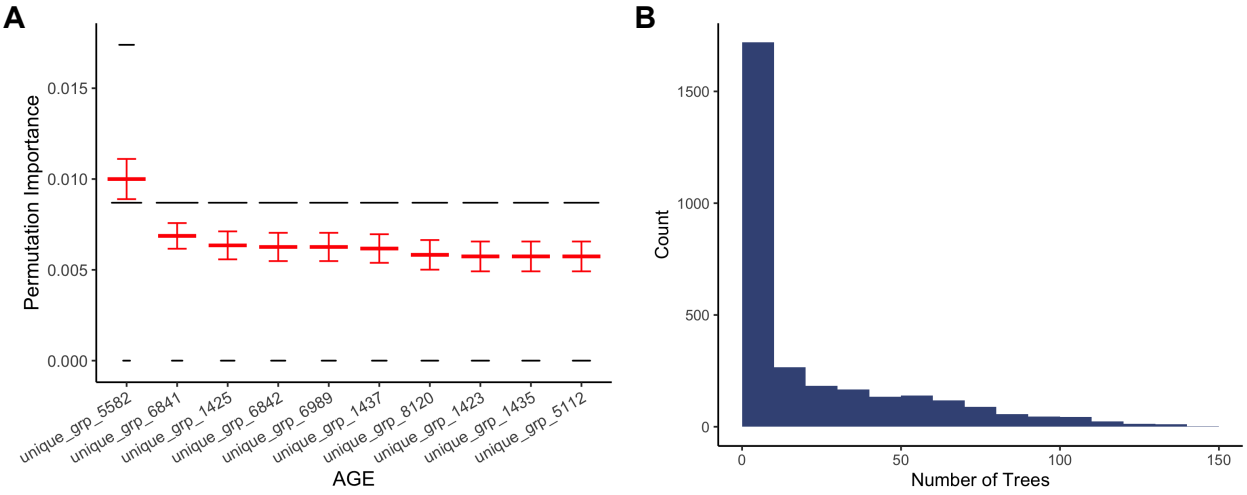


Figure 3.7 Evaluation of features in the random forest model predicting *P. aeruginosa* virulence based on accessory genomic content. (A) Out-of-bag permutation importance for the 10 most important AGEs in the random forest model, showing decrease in accuracy when these AGEs were randomly permuted. Permutation 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. (B) Histogram indicating how many trees within the random forest model contained each AGE (feature), out of a total of 10,000 trees.

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 (Figure 3.8A-C). We next tested dividing the training AGEs into 100 random subsets, finding the average mean nested cross-validation accuracy across all subsets to be still 0.67. Performance of many subsets did deteriorate at this level of data reduction (with 14 subsets having a mean accuracy < 0.6), indicating that in some cases the remaining AGEs lacked sufficient signal to be good predictors of virulence (Figure 3.8D). Together, these findings provide additional 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. Further, it is consistent with a recent finding that antimicrobial resistance in several species can be accurately predicted by only considering variation in a small subset of core genes (and excluding known resistance genes)[291].

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 virulence isolates (Table 3.4, Supplementary Table 3.3). Expanding this analysis to all AGEs with non-zero permutation importance showed that 32/34 were more prevalent in low virulence isolates (Supplementary Table 3.2). This is consistent with the finding that horizontally acquired genetic elements, major components of the accessory genome[47,63], can

incur a fitness cost on the host bacterium[63]. While some genomic islands encode virulence factors[315], many horizontally acquired elements can have a parasitic relationship with the bacterium[63]. 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. It is also important to consider that reduced virulence in our mouse model does not necessarily equate to an overall reduction in bacterial fitness. A fitness cost in the experimental condition we are measuring may be associated with increased fitness in other scenarios (such as environmental persistence).

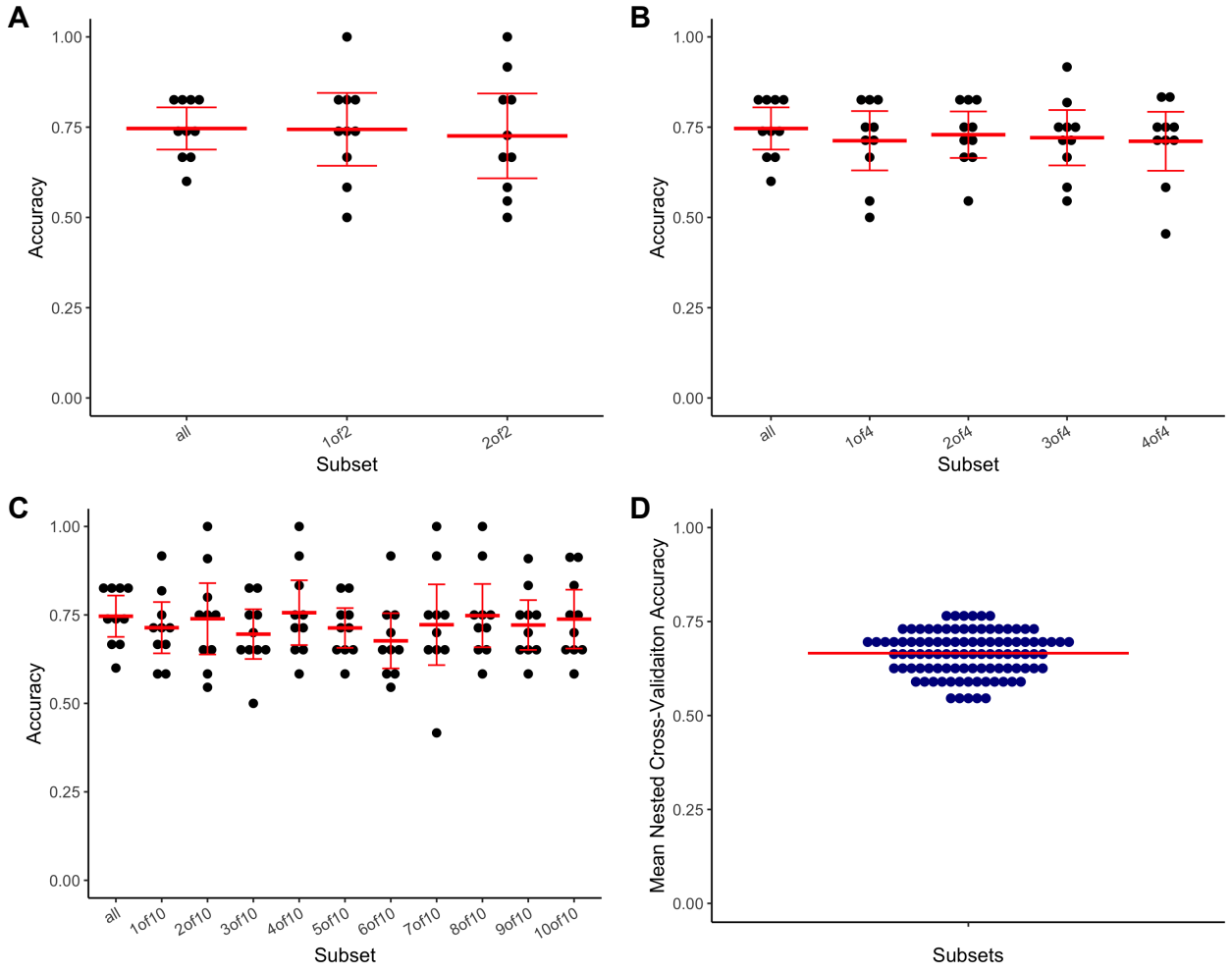

A
1.00
0.75
0.50
0.25
0.00
Accuracy
all
1of2
2of2
Subset
B
1.00
0.75
0.50
0.25
0.00
Accuracy
all
1of4
2of4
3of4
4of4
Subset
C
1.00
0.75
0.50
0.25
0.00
Accuracy
all
1of10
2of10
3of10
4of10
5of10
6of10
7of10
8of10
9of10
10of10
Subset
D
1.00
0.75
0.50
0.25
0.00
Mean Nested Cross-Validaiton Accuracy
Subsets

Figure 3.8 Nested 10-fold cross-validation accuracy of a random forest model in predicting *P. aeruginosa* virulence when trained on random subsets of accessory genomic features. The 3,013 AGEs in the training set were randomly split into (A) 2, (B) 4, and (C) 10 subsets and the accuracy of models trained using each of these subsets of features was estimated through nested cross-validation. The nested cross-validation accuracy obtained when all features are used for training (as in Figure 3.5A) is included for reference. For each subset, accuracy seen in each cross-validation fold are shown in black with the mean accuracy and 95% confidence interval indicated in red. The 3,013 AGEs in the training set were then split into (D) 100 subsets and the accuracy of models trained using each subset estimated through nested cross-validation. The mean nested cross-validation accuracy of each subset is shown in blue with the mean across all subsets indicated in red.

Table 3.4. 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

[a]Based on annotation of any ORF with at least 50 bp overlap with the AGE sequence when blasted against the Pseudomonas Genome Database[50]

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 2008-2009[6] and which we have whole genome sequenced (Table 3.1 and Figure 3.9). 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 $\log_{10}$ CFU) defined for the training set (Figure 3.10, Supplementary Table 3.1, and Table 3.2). The test set was more pathogenic on average than the training set, with 15/25 (60%) of isolates classified as high virulence. This means that a trivial model uniformly predicting high virulence would show an accuracy of 0.6, higher than the null accuracy (0.51) of the training set. However, as the model we are testing was trained on a dataset in which low virulence is the majority class (prevalence 0.51), we would not expect this to occur. We identified which of the 3,013 AGEs used as training features were present in each of the test isolates. 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 3.11A), a finding supported by multiple correspondence analysis (Figure 3.11B-D).

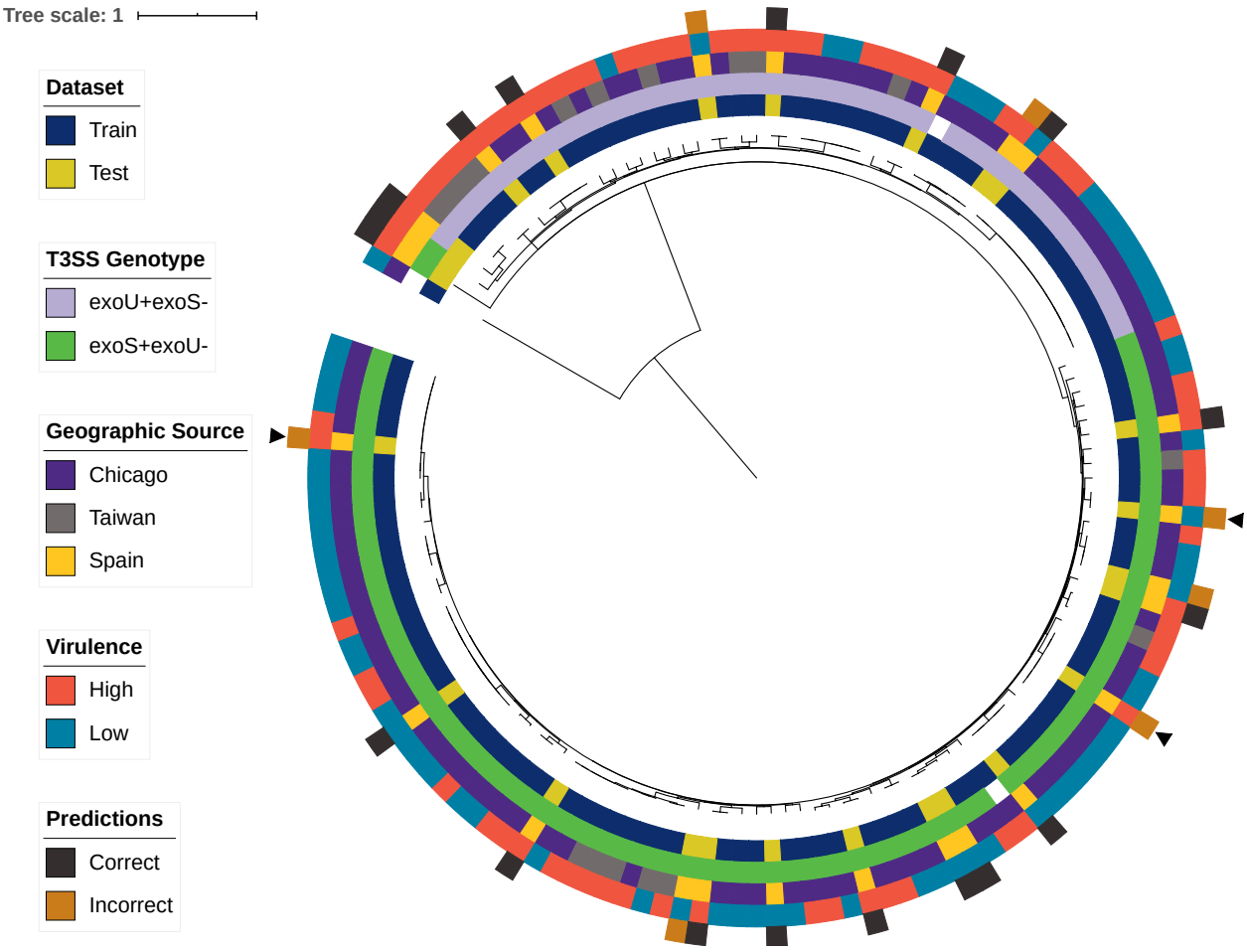


Figure 3.9 Core genome comparisons for the training set of all 140 *P. aeruginosa* isolates considered in this study. Mid-point rooted core genome phylogenetic tree of the 115 training isolates and 25 test isolates constructed from SNV loci present in at least 95% of genomes, annotated (from inner to outer rings) with dataset, T3SS genotype, geographic source, virulence level, and accuracy of prediction by the accessory genome random forest model for test set isolates. Arrowheads indicate examples of incorrectly classified test set strains whose closest core and accessory genomic neighbor(s) show a discordant virulence phenotype.

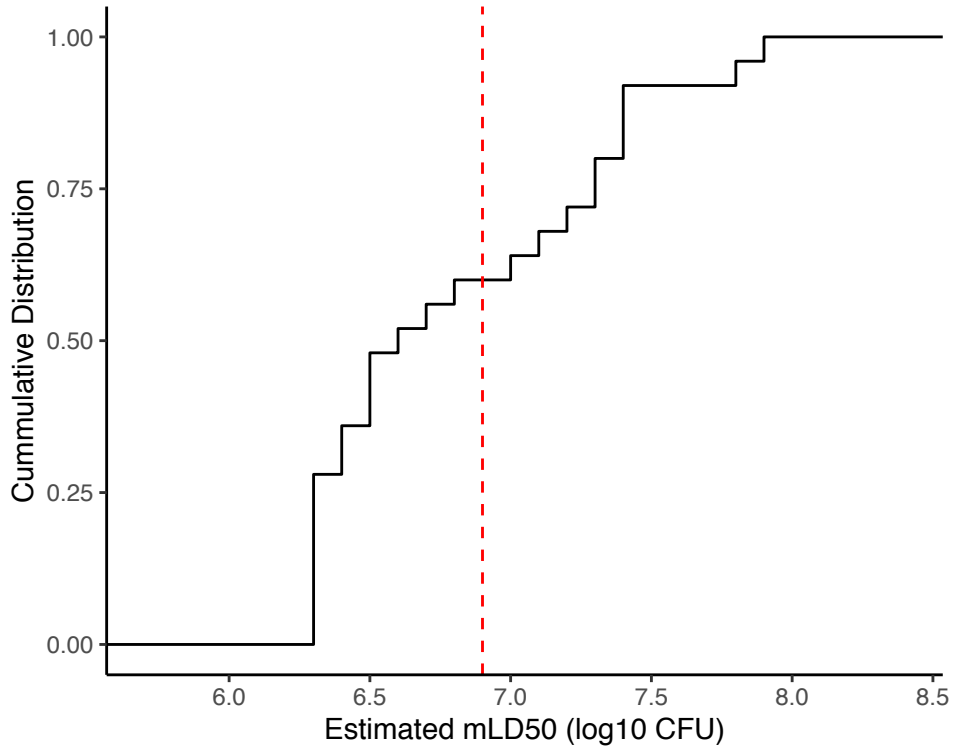
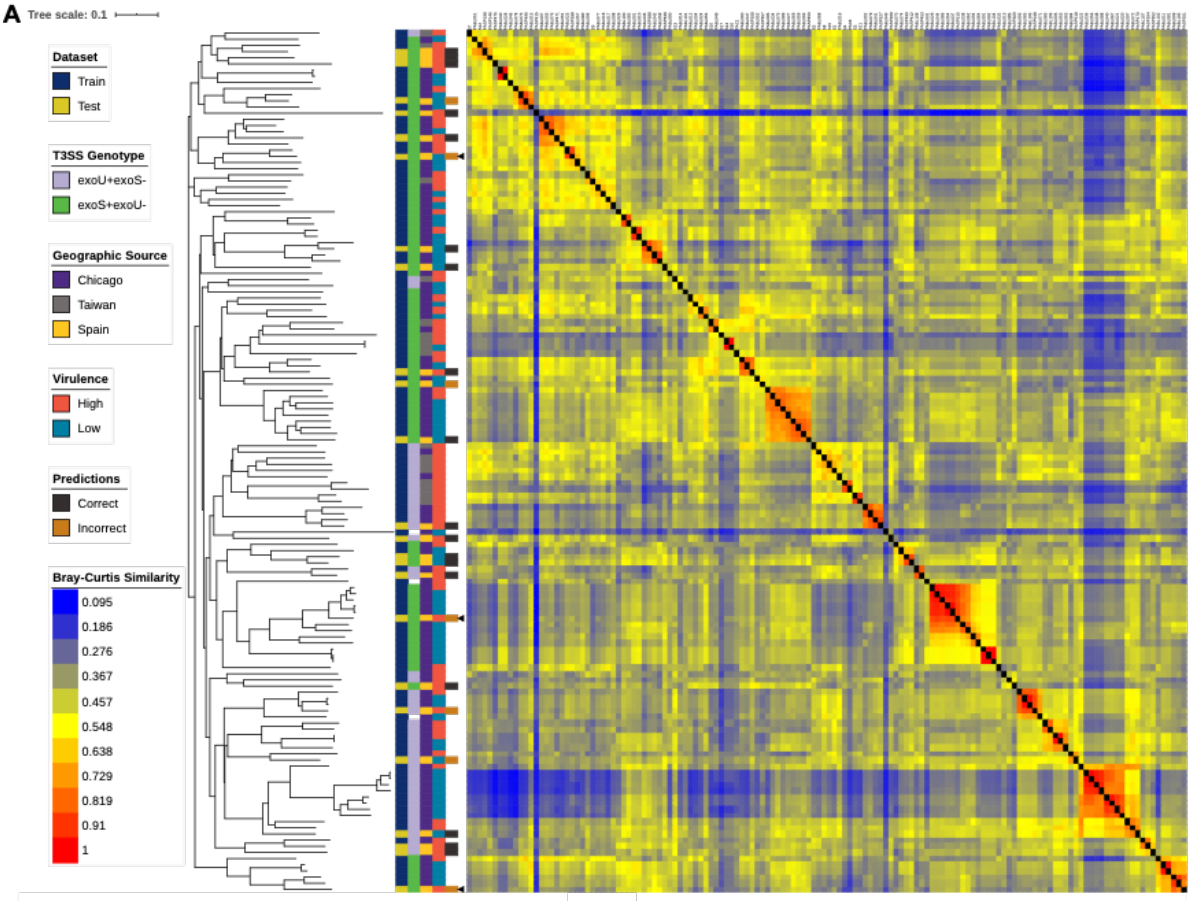
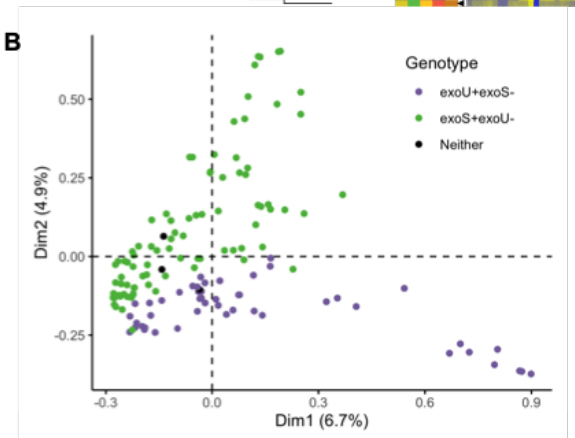
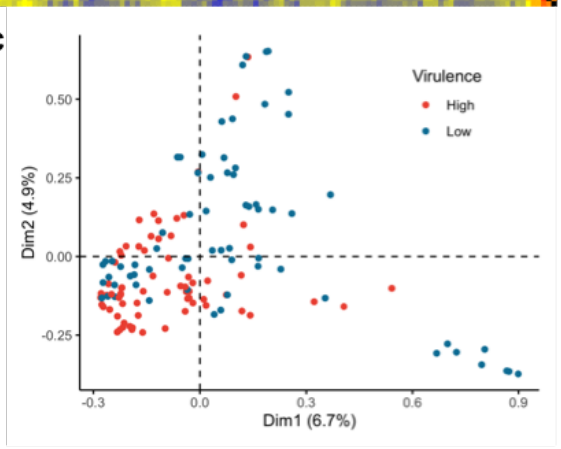


Figure 3.10 Cumulative distribution function of estimated mLD_{50} values for the 25 *P. aeruginosa* isolates making up the independent test set in a mouse model of bacteremia. Isolates with estimated mLD_{50} values less than the median estimated mLD_{50} of the training set (red dashed line) were designated as high virulence, with the remainder designated as low virulence.


A
Tree scale: 0.1
Dataset
Train
Test
T3SS Genotype
exoU+exoS-
exoS+exoU-
Geographic Source
Chicago
Taiwan
Spain
Virulence
High
Low
Predictions
Correct
Incorrect
Bray-Curtis Similarity
0.095
0.186
0.276
0.367
0.457
0.548
0.638
0.729
0.819
0.91
1


B
Genotype
exoU+exoS-
exoS+exoU-
Neither
Dim2 (4.9%)
Dim1 (6.7%)
0.50
0.25
0.00
-0.25
-0.3
0.0
0.3
0.6
0.9


C
Virulence
High
Low
Dim2 (4.9%)
Dim1 (6.7%)
0.50
0.25
0.00
-0.25
-0.3
0.0
0.3
0.6
0.9

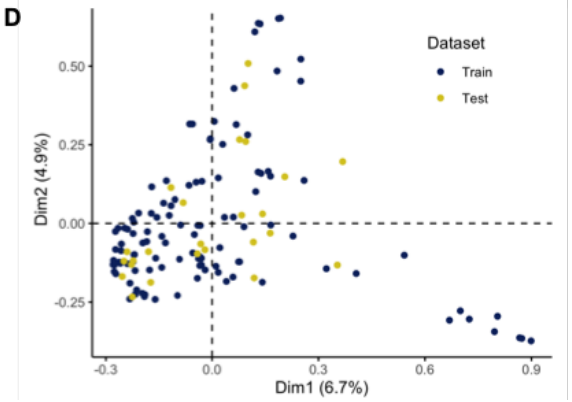

D
Dataset
Train
Test
Dim2 (4.9%)
Dim1 (6.7%)
0.50
0.25
0.00
-0.25
-0.3
0.0
0.3
0.6
0.9

Figure 3.11 Accessory genome comparisons for the training set of all 140 *P. aeruginosa* isolates considered in this study. (A) 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 dataset, T3SS genotype, geographic source, virulence level, accuracy of prediction by the accessory genome random forest model in test set isolates (arrowheads highlighting specific incorrectly classified test set strains as in Figure 3.9), and the dissimilarity heatmap. A higher value indicates that two isolates have more similar accessory genomes. Multiple correspondence analysis (MCA) performed on all 140 isolates in both the training and test sets, considering only the 3,013 AGEs defined from the training set and annotated based on (B) T3SS genotype, (C) virulence level, and (D) dataset. The first two dimensions, and the percentage of variance they explain, are shown.

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 3.5 and Figure 3.12A) 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 (Figures 3.9 and 3.11A). 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 randomly permuted the predicted virulence of the 25 test set isolates to model the null distribution of test set accuracies that we would expect if no link between accessory genome content and virulence existed in the test set. After one million permutations, an accuracy of at least 0.72 was found in 53,476 cases (one-sided $p = 0.053$) (Figure 3.12B). 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.

Table 3.5 Performance of Random Forest Models Trained Using Different Genomic Features Against the 25 Test Isolates

Feature Set	Accuracy	Sensitivity	Specificity	PPV	AUC	F1
AGEs	0.72	0.80	0.60	0.75	0.77	0.77
Core SNVs	0.72	0.67	0.80	0.83	0.69	0.74
8-mers	0.60	0.53	0.70	0.73	0.63	0.62
10-mers	0.68	0.73	0.60	0.73	0.72	0.73

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

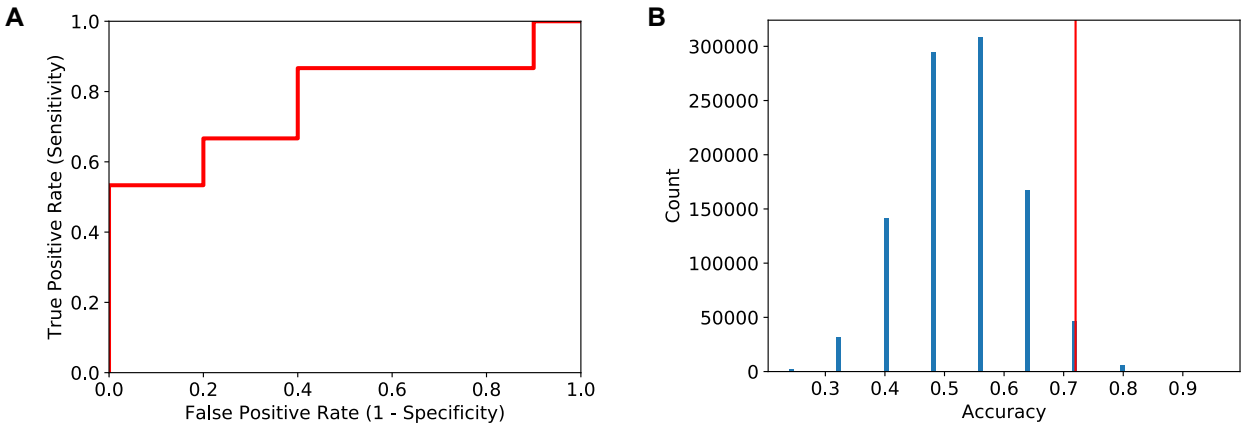


Figure 3.12 Performance of the random forest model trained on accessory genomic content in predicting virulence in an independent test set of 25 isolates. (A) Receiver operating characteristic curve for predictions of the 25 test set isolates using the random forest model (AUC = 0.77). (B) Permutation 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).

Addressing Model Limitations by Removing Isolates with Intermediate Levels of Virulence

While the models generated thus far showed that the accessory genome is predictive of *P. aeruginosa* virulence in mice, limitations inherent to our binary classification of virulence may have constrained their performance. The first lies in the resolution of the mLD_{50} 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 Table 3.1 and Table 3.2). Second, isolates with mLD_{50} 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 mLD_{50} values). This enforced a greater separation of isolates classified as high and low virulence (Figure 3.13A). 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 3.13B). The learning curve, however, showed a greater distance between the training and cross-validation scores (Figure 3.13C). 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.

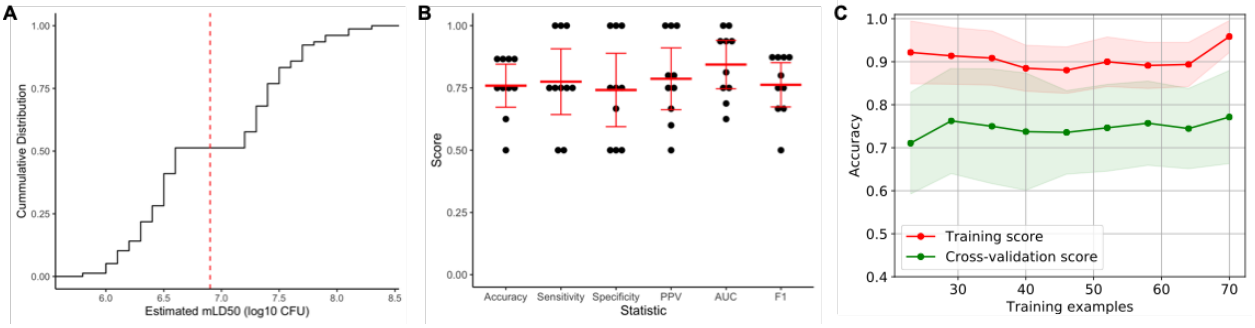


Figure 3.13 Performance of the random forest algorithm in predicting *P. aeruginosa* virulence from accessory genomic content when intermediate virulence isolates (middle 3rd of estimated mLD_{50} values) were removed. (A) Cumulative distribution function of estimated mLD_{50} values after removing intermediate virulence isolates. Isolates with estimated mLD_{50} 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.

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 changed the median estimated mLD_{50}, 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 (Figure 3.14A). 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 mLD_{50} cutoff of the 115 training isolates and 0.69 (95% CI 0.60-0.78) when using the median mLD_{50} cutoff of all 140 isolates (Figure 3.14C and E). It is notable that adding an additional 25 isolates to the training set (and considering the new AGEs in these isolates) did not result in an improvement in model performance. 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 (Figure 3.14D 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 (Figure 3.14B, G, and H).

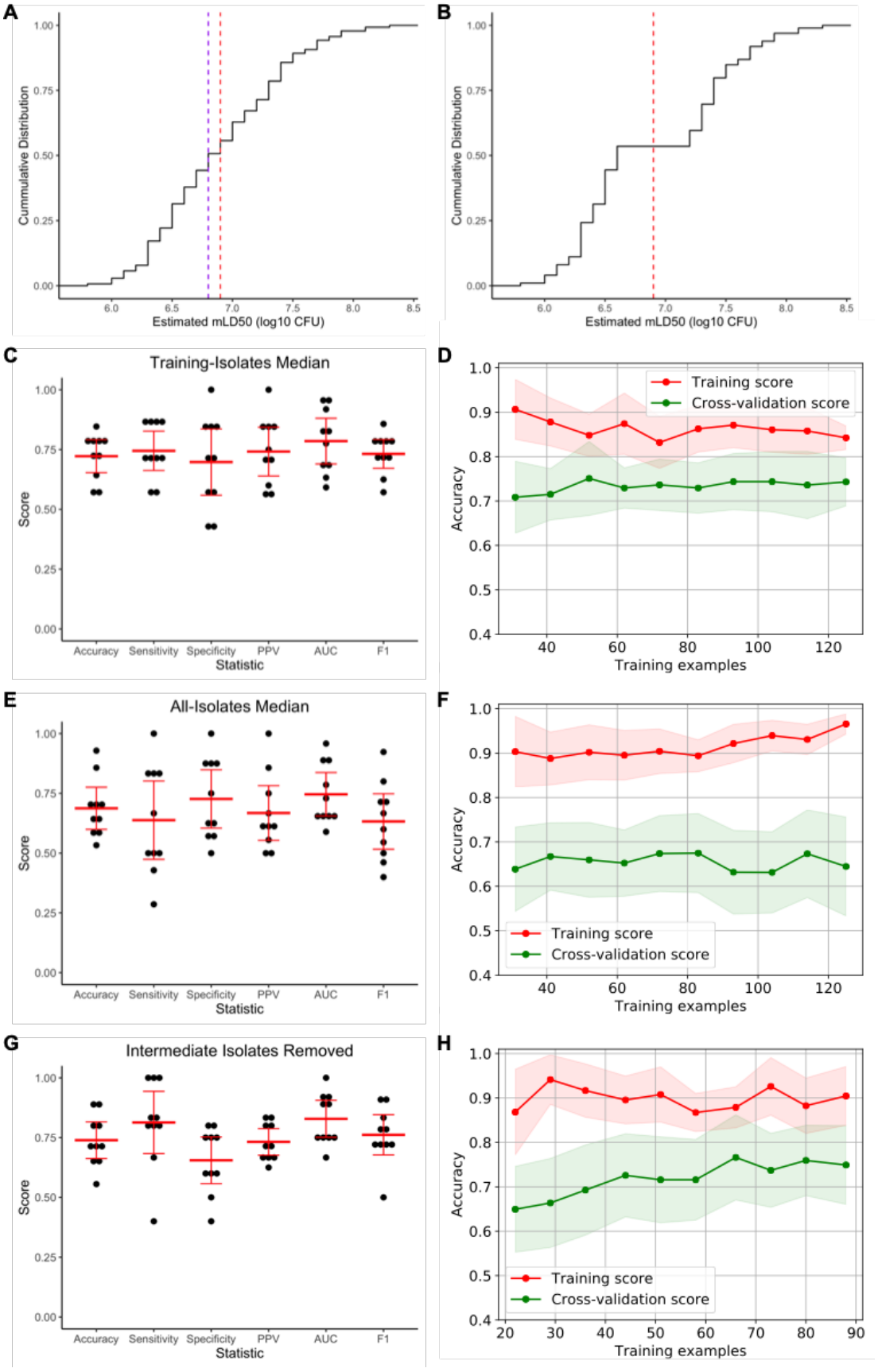

A
Cummulative Distribution
Estimated mLD50 (log10 CFU)
B
Cummulative Distribution
Estimated mLD50 (log10 CFU)
C
Training-Isolates Median
Score
Accuracy
Sensitivity
Specificity
PPV
AUC
F1
Statistic
D
Training score
Cross-validation score
Accuracy
Training examples
E
All-Isolates Median
Score
Accuracy
Sensitivity
Specificity
PPV
AUC
F1
Statistic
F
Training score
Cross-validation score
Accuracy
Training examples
G
Intermediate Isolates Removed
Score
Accuracy
Sensitivity
Specificity
PPV
AUC
F1
Statistic
H
Training score
Cross-validation score
Accuracy
Training examples

Figure 3.14 Performance of the random forest algorithm in predicting virulence from accessory genomic content when all 140 tested *P. aeruginosa* isolates were used to train the model. Cumulative distribution functions of estimated mLD_{50} values considering (A) all 140 tested isolates and (B) after removing intermediate virulence isolates. Isolates were designated as high or low virulence based on whether their estimated mLD_{50} was lower than the median value in the training isolates (red dashed line) or all isolates (purple dashed line). Nested cross-validation performance when defining high virulence based on the median estimated mLD_{50} in the (C) training isolates, (E) all tested isolates, and (G) after removing intermediate virulence isolates, 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. Learning curves showing change in mean training accuracy (red line) and cross-validation accuracy (green line) with increasing training set sizes when defining high virulence based on the median estimated mLD_{50} in the (D) training isolates, (F) all tested isolates, and (H) after removing intermediate virulence isolates. Shading indicates the 95% confidence interval. Assessments at each number of training examples were through 10-fold nested cross-validation.

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 reports[51] and by comparing core and accessory genome measures of strain relatedness (Figures 3.2 and 3.3). 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. Here we used one-hot encoding in our machine learning pipeline to convert SNVs from nucleotides 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. K-mers are defined by dividing the genome into overlapping sequences of length k. We considered k-mer lengths 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 3.15A). 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. This was identical to the test set accuracy seen for the accessory genomic model but with a lower sensitivity and higher specificity (Table 3.5). Despite its lower nested cross-validation accuracy, we 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 3.15B 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 3.15C). 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 (Table 3.5). 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.

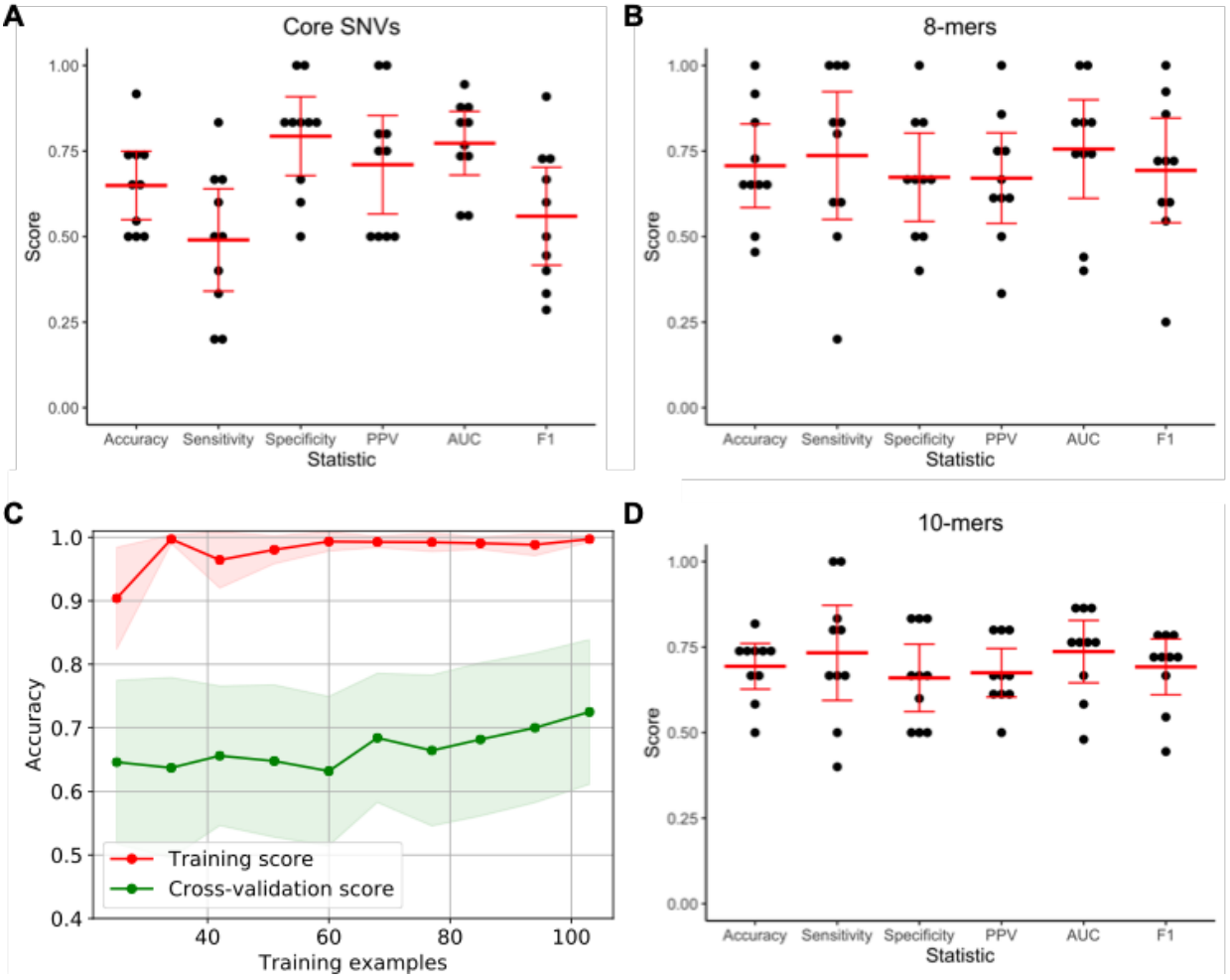

A
Core SNVs
Score
1.00
0.75
0.50
0.25
0.00
Accuracy
Sensitivity
Specificity
PPV
AUC
F1
Statistic
B
8-mers
Score
1.00
0.75
0.50
0.25
0.00
Accuracy
Sensitivity
Specificity
PPV
AUC
F1
Statistic
C
Accuracy
1.0
0.9
0.8
0.7
0.6
0.5
0.4
40
60
80
100
Training examples
Training score
Cross-validation score
D
10-mers
Score
1.00
0.75
0.50
0.25
0.00
Accuracy
Sensitivity
Specificity
PPV
AUC
F1
Statistic

Figure 3.15 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-fold nested cross-validation. Learning curves were not constructed when using core genome SNV or 10-mer counts as features for reasons of computational feasibility.

Simulating Scenarios Where There is No Relationship Between the Accessory Genome and Virulence and When a Single AGE Perfectly Predicts Virulence

In the course of conducting this study, two questions arose regarding how our machine learning approach would respond to different scenarios. We saw that there was signal in the accessory genome predictive of virulence class, but as a comparison we wanted to test how our machine learning approach would perform if there was no relationship between an isolate's accessory genome and our phenotype of interest, in this case virulence. We also saw that a random forest model of virulence appeared to be making predictions based on a diffuse genomic signal rather than the presence of specific virulence or anti-virulence factors. This raised the question of how our models would perform if virulence class was controlled by a single factor. We investigated both of these questions by simulating scenarios where virulence phenotype of the training set was randomly shuffled (breaking the link between accessory genomic content and phenotype) and estimated the performance of different machine learning algorithms with or without the addition of a perfectly predictive feature. To account for variation in results based on how the phenotype was permuted, this process repeated 10 times and the mean nested cross-validation accuracy from each replicate was compared (Figure 3.16).

When there was no signal between accessory genome content and phenotype mean nested cross-validation accuracy was approximately 0.5 (with mean nested cross-validation accuracy averaging 0.47-0.51 across the 10 seeds tested) regardless of algorithm choice (Figure 3.16). This serves as a negative control to provide more evidence that there is true signal in the accessory genome predictive of virulence. When an artificial feature is added that perfectly predicts phenotype, accuracy increases for all algorithms tested. However, the degree of improvement varied by algorithm. Notably, random forest showed the smallest improvement in

nested cross-validation accuracy (mean 0.65), while elastic net logistic regression (mean 1) and support vector classifier (mean 0.92) showed much higher performance (Figure 3.16). This can be attributed to differences in the way these algorithms learn from the training data. At each node of each decision tree in the random forest, only a subsample of potential features are tested. When the number of features is much larger than the number of samples, as is the case for the AGEs in our training set, many trees would never test a given feature. This limits the ability of that feature to influence model predictions. As the random forest algorithm performed similarly to all other algorithms in predicting virulence based on accessory genomic content (Figure 3.5), it is unlikely that there is a single AGE is highly predictive of phenotype in the *P. aeruginosa* isolates. The L1-regularization component of elastic net logistic regression can set the weight of uninformative features to 0 (effectively removing them from the model), which may explain why it was best able to learn from a single perfectly predictive feature. More investigation would be needed to test how different algorithms would react to the phenotype being dictated by the combination of a small number of features, but a similar pattern may arise in that scenario.

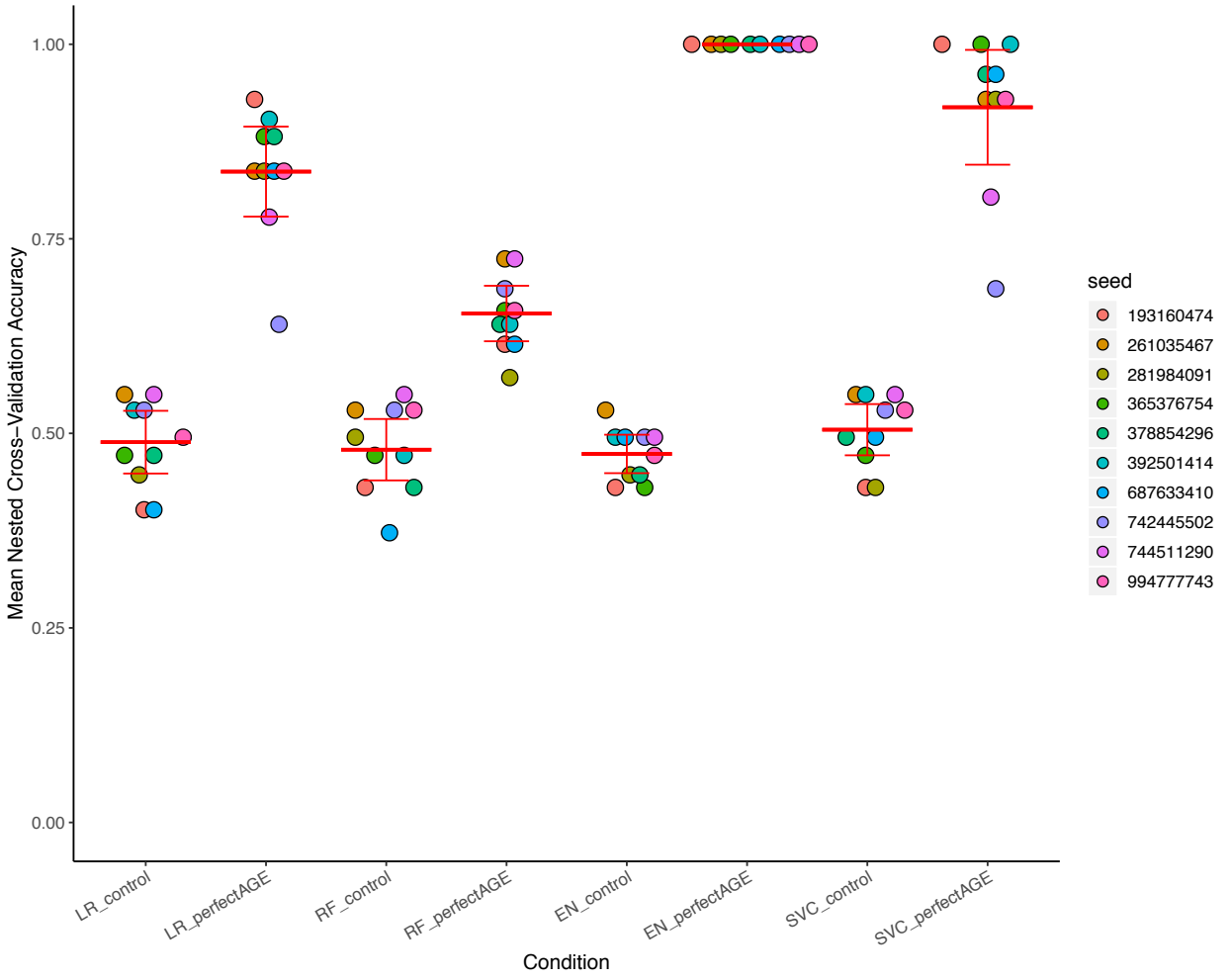


Figure 3.16 Mean nested cross-validation accuracy of machine learning algorithms in predicting a randomly permuted phenotype based on accessory genomic content and after adding an artificial perfectly predictive AGE. L2-regularized logistic regression (LR), random forest (RF), elastic net logistic regression (EN), and support vector classifier (SVC) algorithms were tested. Nested cross-validation accuracy was determined when using the 3,013 AGEs for the 115 training isolates to predict (control) and after adding an additional feature identical to the labels (perfectAGE). This process was repeated using 10 random seeds (indicated by color), with the mean and 95% confidence interval between seeds indicated in red.

Evaluating Machine Learning Models Predicting Persistence or Eradication from Early Cystic Fibrosis P. aeruginosa Isolates Based on Genomic Content

As a comparison to our work using the *P. aeruginosa* genome to predict an isolate's virulence in mice, we investigated whether the genome could be used to predict whether an isolate is persistent or eradicated in the lungs of cystic fibrosis patients. To do this, we investigated a collection of 207 early cystic fibrosis isolates collected as part of the Early Pseudomonas Infection Control (EPIC) Clinical Trial[318,319] for which we possess both whole-genome sequencing data and know whether the isolate was persistent or eradicated in the study patient. The majority (0.72) of these isolates were eradicated (Supplemental Table 3.4). A core genome phylogenetic tree showed that both persistent and eradicated isolates are genetically diverse (Figure 3.17). It is important to note that as this collection does not contain PA7-like outlier strains, the scale of this phylogenetic tree differs from trees considering the collection we used to examine the relationship between the *P. aeruginosa* genome and virulence in mice (e.g. Figure 3.2). We defined the accessory genome of each of the 207 cystic fibrosis isolates. This identified a total of 4156 AGEs, with a mean length of 3875 bp, median length of 756 bp, and forming a pan-accessory genome size of 16.1 Mb (Supplemental Table 3.5). Alignment of sequencing reads from these isolates to PAO1 identified 308,999 variant core genome SNV loci.

We used the machine learning approach described in Figure 3.4 to estimate how well the accessory genome is able to predict whether an isolate is persistent or eradicated. However, as accuracy can be a poor measure of model performance in unbalanced datasets, F1 score was used for hyperparameter selection during grid-search cross-validation and considered as the primary outcome metric. In nested cross-validation, the mean F1 score was low (0.26), showing that the accessory genome was a poor predictor of whether an isolate was persistent or eradicated.

Additionally, the mean accuracy (0.56) was lower than simply picking the majority class (0.72), as F1 score was maximized at the expense of accuracy (Figure 3.18A). Examination of the learning curve for the F1-tuned approach showed a large gap between the training F1 score and the cross-validation F1 score, suggesting that performance may improve with increasing sample size. However, as there is no upward trend in F1 score as the training size increases, an appreciable improvement may not be possible with a realistic number of samples (Figure 3.18B). When accuracy was instead used for hyperparameter selection, the optimal model exclusively classifies isolates as the majority class. This leads to a mean cross-validation accuracy of 0.72, a specificity of 1, and a sensitivity, PPV, and F1 score of 0 (Figure 3.18C). In the learning curve the training and cross-validation scores have already converged by the largest sample size tested, suggesting that there will be no further improvement in the model with increasing sample size (Figure 3.18D). Similarly, when core genome SNVs were used to train a random forest model, an F1-tuned approach performed poorly in nested cross-validation (mean F1 score 0.21), and an accuracy-tuned approach produced models that primarily pick the majority class (mean accuracy 0.71 and mean specificity 0.99) but never correctly identify a persistent isolate (mean sensitivity and F1 0) (Figure 3.19). Altogether, these findings show that in our dataset the *P. aeruginosa* genome is not predictive of whether an isolate is persistent or eradicated during early cystic fibrosis infection.

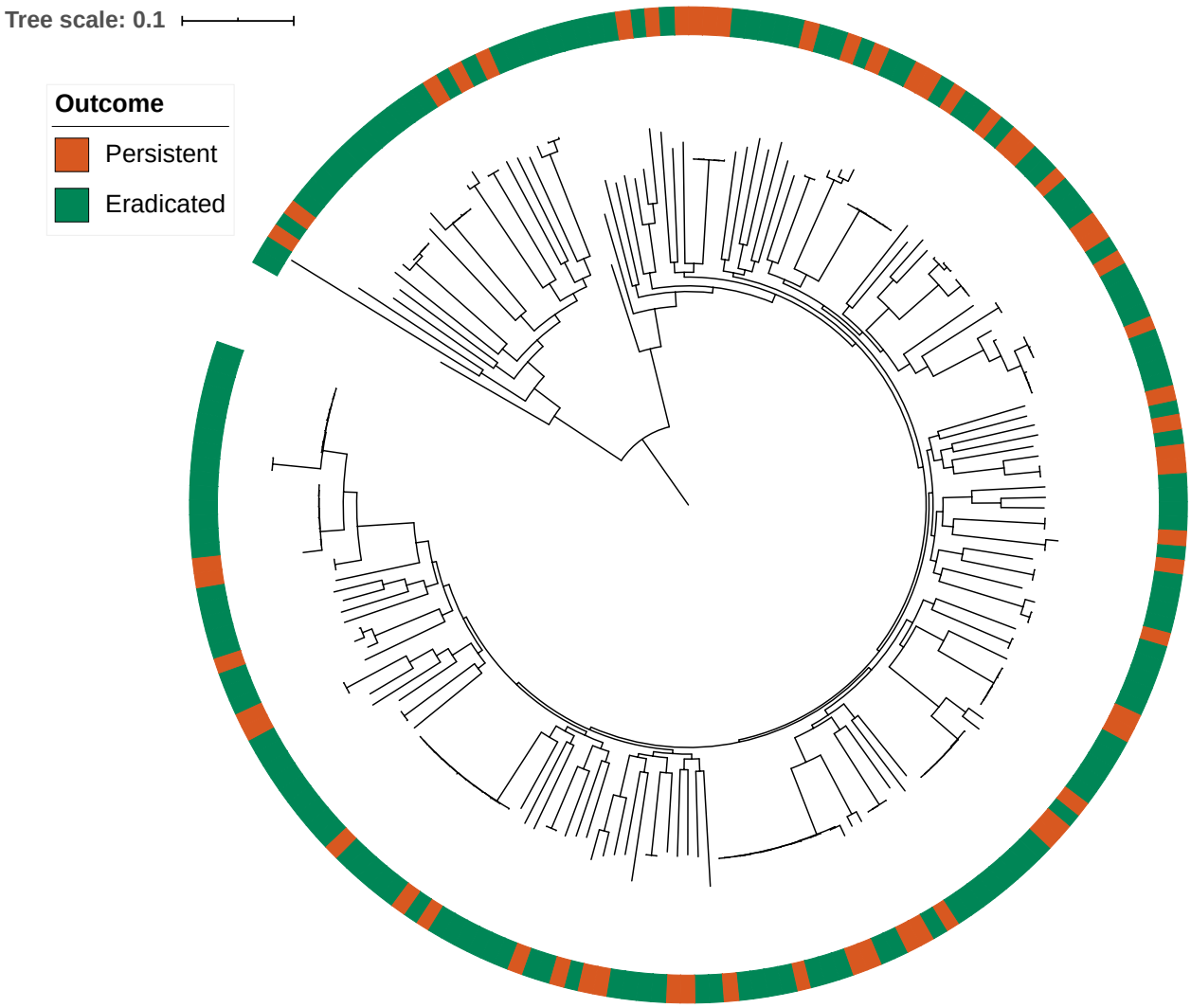


Figure 3.17 Core genome comparisons for the collection of 207 early cystic fibrosis *P. aeruginosa* isolates. Mid-point rooted core genome phylogenetic tree of the 115 training isolates constructed from SNV loci present in at least 95% of genomes, annotated by clinical outcome (persistent or eradicated).

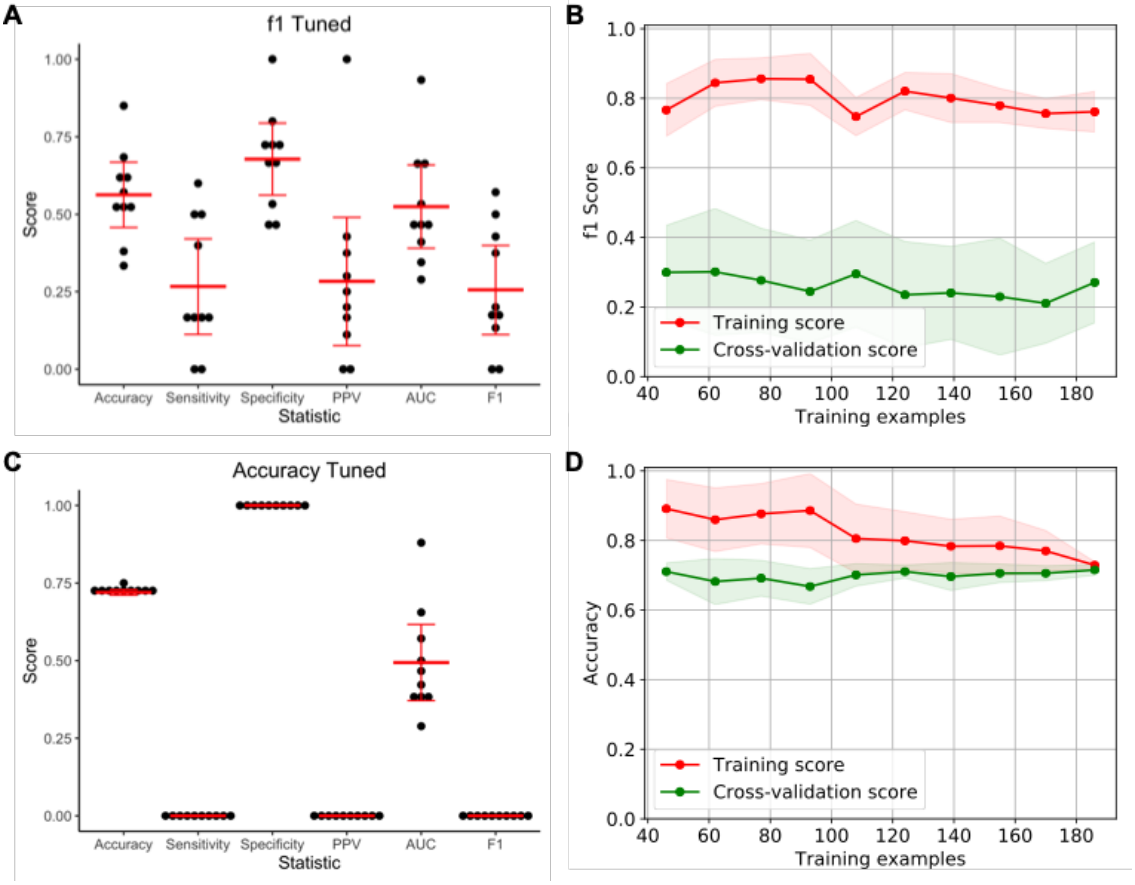


Figure 3.18 The accessory genome is not predictive of persistence or eradication in a collection of early cystic fibrosis isolates. Nested cross-validation performance when using F1 score (A) or accuracy (C) to select hyperparameters during grid-search cross-validation including accuracy, sensitivity, 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. Learning curves showing change in mean training accuracy (red line) and cross-validation accuracy (green line) with increasing training set sizes when using F1 score (B) or accuracy (D) to select hyperparameters during grid-search cross-validation. Shading indicates the 95% confidence interval. Assessments at each number of training examples were through 10-fold nested cross-validation.

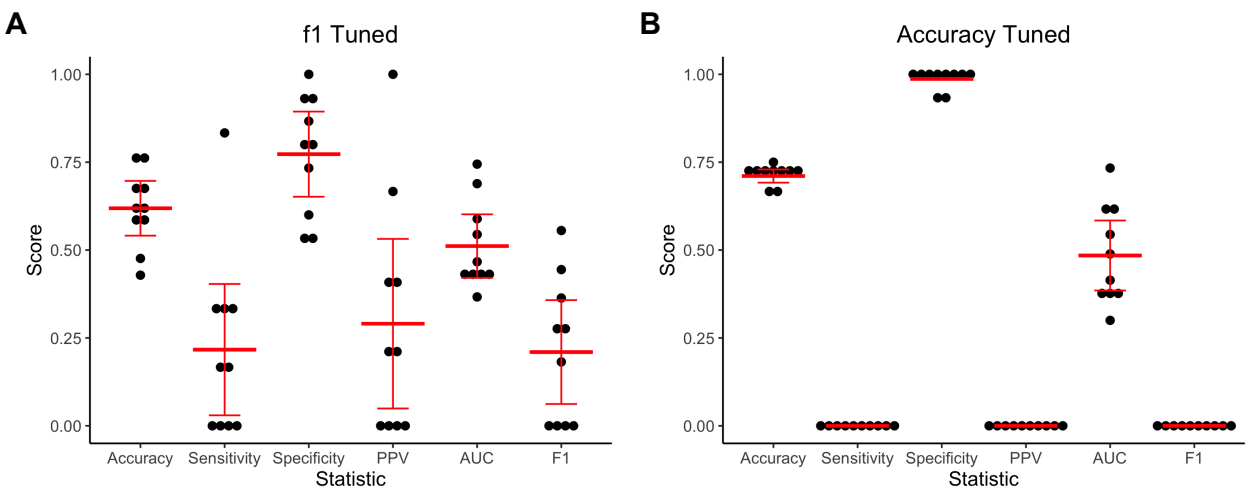


Figure 3.19 Core genome SNVs are not predictive of persistence or eradication in a collection of early cystic fibrosis isolates. Nested cross-validation performance when using F1 score (A) or accuracy (B) to select hyperparameters during grid-search cross-validation including accuracy, sensitivity, specificity, 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.

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. The genomic signature detected may approximate lineage, echoing the recent finding that genomic neighbors are highly predictive of antimicrobial resistance in *Streptococcus pneumoniae* and *Neisseria gonorrhoeae*[279]. Supporting this conclusion is the finding that individual AGEs showed low importance in random forest model predictions and that models could be built using only a random tenth of the total AGEs without a dramatic loss of performance. Further, some of the misclassified test set strains were virulence outliers relative to their phylogenetic neighbors. Still, information encoded in the genome is not necessarily simply phylogenetic. This was shown in recent study by Khaledi et al. using genomic

and transcriptomic features to predict antimicrobial resistance in *P. aeruginosa*. They tested the influence of phylogenetics on their resistance predictions through "block cross-validation", in which they enforced that training and cross-validation folds contained non-overlapping sequence types. This resulted in modest reductions in performance but showed that resistance could be predicted even when testing against phylogenetically distinct isolates[66]. Future studies should determine the extent to which *P. aeruginosa* virulence correlates with phylogenetic relationships.

While individual AGEs showed low importance in model predictions, 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 the presence of specific *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 model[31]. While certain AGEs enhance virulence[70], many AGEs (e.g. parasitic phages, plasmids, or ICEs) may decrease virulence through mechanisms such as dysregulation of regulatory networks, insertion into important genes, or imposition of an additional metabolic burden. The latter possibility could be assessed by examining the in vitro growth rate of the isolates included in this study and determining whether AGEs predictive of low virulence were associated with slower growth. In addition, it could be determined whether deletion of these AGEs resulted in an increased growth rate. This should be accompanied by a systematic investigation into the types of AGEs that are associated with low and high 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[140] 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 the model was 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 mLD_{50} values 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 mLD_{50} of 7.0 $\log_{10}$ CFU is near the cutoff of less than 6.9 $\log_{10}$ CFU for high virulence. This highlights inherent limitations of this study: that virulence exists on a continuum not neatly divided into binary classes and that the limited number of mice tested for each isolate creates uncertainty in the estimations of the mLD_{50} values. Both of these factors could decrease the accuracy of our models. To address these limitations, 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 3.13). On the other hand, as mentioned above some of the incorrect predictions in the test set were exceptions in virulence compared to closely related isolates. For example, PASP251 has an estimated mLD_{50} of 6.3 $\log_{10}$ CFU, while its nearest four phylogenetic neighbors all have estimated mLD_{50} values of greater than 7 $\log_{10}$ 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 3.6A) 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[66,274,275,277,279], opening the door for sequence analysis to supplement or replace traditional antimicrobial susceptibility testing if routine sequencing becomes commonplace. This study serves as a proof of concept that the *P. aeruginosa* genome can be used to predict its pathogenicity. This is notable because, with its relatively large and complex genome, *P. aeruginosa* has the potential to be more difficult to evaluate through machine learning modeling than other bacterial species. For example, in a study performed by Hyun et al., AMR prediction models for *P. aeruginosa* tended to have worse performance than those for *S. aureus* or *E. coli*[278]. We also considered a genetically diverse population, compared to the simpler scenario examined by some other studies which focused on variation within a specific clone[282,296]. 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* clones[282]. This could improve the sophistication of current diagnostics and allow clinicians to rapidly identify patients at highest risk for poor outcomes.

CHAPTER 4

Discussion

In this dissertation, I have described two projects examining antimicrobial resistance and virulence in *Pseudomonas aeruginosa*. One was a genomic epidemiology study in which we uncovered a clone that has caused a prolonged epidemic at our institution and characterized a large plasmid which contributed to its resistance. In the other, we used a machine learning approach to show that the *P. aeruginosa* genome can be used to predict virulence in a mouse model of infection. While these studies may seem distinct, they are tied together by a unifying thread of bacterial genomics. As bacterial whole genome sequencing becomes an increasingly practical tool in the clinical microbiology laboratory, I expect that the gap between these two areas of research will continue to narrow. Ultimately, the goal of both studies was to use the *P. aeruginosa* genome to better track and understand bacterial phenotypes. They employed many of the same techniques, including whole-genome sequencing and assembly, sequence alignment, and phylogenetics. In these chapters, I have laid the foundation for future work to further investigate the epidemic subclade ST298* and its resistance plasmid and to interrogate additional *P. aeruginosa* phenotypes via machine learning. In this section, I provide context to the results of these studies and propose future directions for additional research.

Prolonged epidemic of XDR ST298* *P. aeruginosa* at Northwestern Memorial Hospital

In Chapter 2, I describe the repeated isolation of a specific highly drug-resistant *P. aeruginosa* subclade (termed ST298*) for at least 16 years from patients at Northwestern Memorial Hospital (NMH), with time-scaled phylogenetic analysis suggesting that the ST298*

subclade was established substantially earlier. ST298* shows high levels of antimicrobial resistance (AMR), driving the high rate of drug resistance seen in CC446 isolates from NMH. This is in part due to the presence of a large AMR plasmid containing a novel class I integron, which in many cases pushed isolates from multidrug resistant (MDR) to extensively drug resistant (XDR). This plasmid is so far unique to ST298* but is a member of a family of large *Pseudomonas* genus plasmids. While ST298* *P. aeruginosa* has caused a prolonged local epidemic of drug resistant infections, CC446 more broadly has caused drug resistant infections around the globe.

The results of this study have both general and local significance. Broadly, they indicate the importance of CC446 as an emerging high-risk clone. Locally, this work uncovers the existence a source of highly drug resistant *P. aeruginosa* infections at our institution. Identifying, and hopefully eradicating, the reservoir of ST298* would be critical to prevent future infections and lower the burden of drug resistance in the hospital environment. While the AMR plasmid (pPABL048) and integron (in1697) have not been detected outside of the ST298* subclade thus far, it is possible it could facilitate the spread of resistance to other *Pseudomonas* isolates. Further, in1697 if found in only a small portion of pPABL048. Examining other roles pPABL048 plays in the physiology or pathogenicity of ST298* may help us better understand the biology this subclade, high-risk clones in general, and large *Pseudomonas* plasmids.

CC446 as an emerging global high-risk clone

One of the major motivations of this project was to show that CC446 has not only caused a local epidemic at NMH (in the form of ST298*) but that it also represents an emerging high-risk clone. The isolates included in our study and previous reports in the literature and PubMLST database show that CC446 has a global distribution, with cases found on five

continents[65,73,74,100,180,181,192,304-308]. Further, multiple studies (including our own) reported at least MDR (non-susceptibility to a drug in ≥3 classes) in one or more CC446 isolates[65,73,74,100,192,306,307]. Two other studies each noted carbapenem resistance and the VIM-2 metallo-β-lactamase in an ST298 isolate[180,181]. These characteristics (global spread and repeated drug-resistant infections) are the key indicators of a high-risk clone[129,131].

An important question regarding CC446's status as an emerging high-risk clone is whether it continues to cause outbreaks of drug-resistant infections globally. An updated literature search found an additional 12 studies reporting ST298 or ST446 isolates (either from patients or in the hospital environment) not captured in our original study[167,169,183,184,202,320-326]. These studies were published between September 2018 and June 2020, and identified CC446 isolates in the United States, Canada, China, Indonesia, Iran, Ireland, Myanmar, South Korea, Spain, Russia. Detection of CC446 in Indonesia, Iran, Ireland, and Myanmar was not noted in our original study, highlighting that extent of global spread for this clone is even greater than previously appreciated. Notably, nine of these studies report concerning AMR (either carbapenem nonsusceptibility or explicit multidrug resistance)[167,169,183,184,202,321,323-325]. A surveillance study of sites in five states led by the Centers for Disease Control and Prevention (CDC) highlights the importance and spread of CC446 in drug-resistant *P. aeruginosa* infections in the United States. They sequenced 128 of 129 carbapenem-resistant isolates submitted to the CDC as part of this study (106/129 of which were confirmed to be carbapenem resistant on repeat testing by the CDC). ST298 was the second most common ST identified with 10 isolates, second only to the global epidemic strain ST235 (14 isolates)[202].

As appreciated in Chapter 2, the CC446 isolates in these studies showed AMR through diverse mechanisms. Carbapenem resistance in a set of ST446 isolates from four Russian

hospitals was largely explainable (7/8 isolates) by an insertion sequence disrupting *oprD*[167]. Similarly, four carbapenem non-susceptible ST446 isolates from Oregon showed a premature stop codon leading to truncation of OprD[169]. On the other hand, a carbapenem resistant ST446 isolate from Spain harbored the metallo-β-lactamase IMP-8, while an isolate from Myanmar contained VIM-2. Both of these isolates also contained one or more acquired aminoglycoside resistance genes[184,324].

Even in cases where concerning drug resistance was not described, there were findings that may be relevant to CC446's global spread. Moghadam et al. found that ST446 was common in urine samples from prostate and bladder cancer patients at an Iranian hospital (11.8% of total *P. aeruginosa* isolates). These were not carbapenem resistant but still showed some AMR, with 3/8 resistant to ceftazidime and 5/8 resistant to gentamicin[326]. Moloney et al. described 4 ST298 *P. aeruginosa* isolates from two washbasin U-bends in a Dublin Dental University Hospital clinic and an additional isolate from a washbasin U-bend in a second nearby hospital[322]. The authors did not test for AMR in these isolates, but this study provides an example of ST298 *P. aeruginosa* establishing itself in a healthcare environmental reservoir.

Altogether, the evolving literature suggests that CC446 continues to be a global problem clade, with numerous new cases reported since the completion of our original study. This cements its status as an emerging, if not already established, high-risk clone. As such, further surveillance is needed to monitor for new or larger outbreaks and determine whether CC446 is increasing in clinical significance over time.

*Potential reservoirs for ST298**

We identified a total of 21 ST298* isolates (all of which were MDR and 13/21 of which were XDR) that were collected from NMH over the course of 16 years. This suggests the

existence of a stable reservoir from which ST298* *P. aeruginosa* periodically emerges to cause drug-resistant infections. Further, time-scaled phylogenetic analysis estimates that the last common ancestor of these isolates occurred in approximately 1980. This is relevant because the current NMH inpatient facility opened in 1999, suggesting that ST298* began diverging (and likely had acquired the AMR plasmid) prior to its opening. It is notable that this new facility is on the same campus as the prior hospital and is connected to buildings that predate this estimated last common ancestor (e.g. the Olson Pavilion, opened 1979). Our earliest ST298* isolate (PABL020) was collected in 2000, shortly after the opening of the new facility. Identifying the reservoir of ST298* is important from a public health perspective and is necessary to end this prolonged epidemic and prevent future highly drug-resistant infections caused by this subclade. As stated in Chapter 2, this reservoir could be within NMH itself (perhaps seeded by an already-diverging ST298* population), at an outside site from which patients are admitted to NMH (such as a long-term acute care hospital or skilled nursing facility), or more widely distributed in the Chicago metropolitan area. An important limitation in our current study is that we did not possess epidemiological data about the isolate collection beyond date of isolation. For example, we do not know the patient room or ward from which these isolates were collected and whether the affected patients were linked by common facilities that could serve as an infection source. This makes it impossible to pinpoint the reservoir of ST298* based on the evidence we have now. Future work will need to incorporate epidemiological data, including chart review of patients infected with ST298* isolates and targeted sampling of the healthcare environment. With that in mind, it is important discuss the potential reservoirs for ST298* to develop a framework for these future studies.

A recent study performed by Chng et al. at a tertiary care center in Singapore found that, for multiple pathogenic species (*Elizabethkingia anopheles*, *Staphylococcus aureus, Acinetobacter baumannii*), isolates collected from the hospital environment could be genetically linked to previously collected patient isolates from either that hospital or others in Singapore. Some of these linked patient isolates were collected over eight years prior[327]. This provides evidence healthcare settings can drive prolonged outbreaks like we see in ST298*. Further, it is possible that admissions to NMH or discharges to other facilities could drive transmission and spread the outbreak. This can be illustrated by a study of transmission in a regional *Klebsiella pneumoniae* outbreak performed by Snitkin et al., where a nursing home provided a link for transmission between an acute care hospital and a long-term acute care hosptital[328].

Water systems are a prime candidate for the ST298* reservoir and as such deserve increased attention. Maloney et al. cultured ST298 *P. aeruginosa* from washbasin U-bends[322]. This is consistent with our study, where 7 CC446 isolates (6 ST446 and 1 ST298) isolates were collected from sinks in healthcare facilities in the Chicago metropolitan area as part of the healthcare environmental collection. None of these were ST298* and all were susceptible to antibiotics, but they did not originate from NMH where all ST298* isolates thus far have been identified. While Olivia Pura, an undergraduate researcher in the Hauser laboratory, sampled one or more sinks at NMH while assembling the healthcare environmental collection, only one isolate (which was not CC446) was collected at this site. As such, it is possible that we have simply not sampled the right locations. It is well established that *P. aeruginosa* can colonize water systems, which can then serve as likely foci for hospital outbreaks[21-23]. It has been shown that *P. aeruginosa* can incorporate into biofilms on common plumbing materials under experimental conditions[329] and that *P. aeruginosa* is biofilm-associated in sink taps[23]. This

biofilm lifestyle may promote the formation of a stable and persistent reservoir. A particularly insightful example of the importance of water systems in infection control is that of a prolonged outbreak (2006-2016) of another opportunistic pathogen, *Sphingomonas koreensis*, at the NIH clinical center investigated by Johnson et al. They cultured water systems (e.g. water, sinks, plumbing components) in rooms where patients were identified as infected as well as other possible sources of exposure, identifying a number of environmental isolates genetically related to those causing infections. Because the phylogenetic structure of this cluster was broadly diverse and replaced sinks became culture-positive over time they were able to infer that the colonization occurred deep within the water system shortly after the NIH clinical center was constructed in 2004. Finally, they were able to prevent additional infections by increasing chlorination and hot water temperature at the clinical center, showing that this type of investigation is actionable[330].

While water systems are the most likely reservoir for ST298*, it is possible that colonization of other surfaces plays an important role. Chng et al. were able to detect multiple pathogenic bacteria (including *P. aeruginosa*) on surfaces in hospital rooms, including both sinks and dry surfaces. Further, they showed that clones of multiple species (particularly *E. anopheles*, *Serratia marcescens*, and *Staphylococcus haemolyticus*) were persistent in the hospital environment over a period of approximately 1.5 years[327]. The increased exposure of these surfaces to both decontamination procedures and desiccation make it less likely that they serve as the ultimate reservoir, but the disinfectant resistance genes on in1697 may promote transient colonization of surfaces by ST298* and provide opportunities for patient infection. Colonized or infected patients may also serve as a vector for the spread of ST298*.

As stated, future work is needed to identify the reservoir of ST298*. The Hauser and Ozer laboratories have continued to collect *P. aeruginosa* isolates from NMH. As these isolates are sequenced for other purposes, they should be screened for the occurrence of ST298* and the presence of pPABL048. This is important to determine whether this prolonged epidemic is still ongoing or whether the plasmid has disseminated. It will also be important to, where possible, acquire epidemiologic data on patients infected by ST298* isolates through retrospective chart review. This will determine if there are clusters of cases which share a floor or hospital room as well as other epidemiologic links (e.g. common procedures). Additionally, it could uncover whether the patients were admitted to the hospital from a common location, such as the same long-term acute care hospital. Recently collected ST298* isolates would be the most useful targets for chart review, as they may point to areas currently colonized by ST298*. After chart review, the best next step would be systematic environmental sampling following the approach Johnson et al. used to identify the reservoir of *S. koreensis* in the NIH Clinical Center[330]. Water systems from areas where patients could have been exposed (both patient rooms and common systems) should be sampled and cultured either in the hospital or if relevant in an outside site. If sinks or drains colonized by ST298* *P. aeruginosa* were detected, it could prompt sampling deeper in the water system to attempt to identify a root source.

As noted in Chapter 2 it is possible that ST298* is more widespread in healthcare settings throughout the region and has not been previously detected simply for a lack of screening. With that in mind, NMH should collaborate with other Chicago institutions to screen banked isolates for the presence of ST298*. To alleviate the cost of sequencing, initial screening could be done in a rapid manner through either MLST or targeted PCR of sequence specific to ST298* or pPABL048. Of particular importance would be Lurie Children's Hospital, which in 2012 moved

to a primary facility adjacent to NMH. Collaboration here could potentially reveal if this move was associated with the acquisition of ST298*.

Finally, in the process of investigating the potential reservoir of ST298* it would also be useful to examine the characteristics of the patients who were infected by these isolates. This subclade has clearly been successful in that it has caused repeated infections at NMH over an extended period of time. This may be due, at least in part, to its high level of AMR, but it is unknown if its mutational and acquired resistance mechanisms (or other genomic characteristics of ST298*) come at a cost to acute pathogenicity. Notably, all eight ST298* isolates included as part of the training set in Chapter 3 were classified as low virulence in the mouse model of bacteremia (Table 2.3 and Table 3.1). While virulence in mice does not necessarily equate to pathogenicity during human infection, it would be useful to know if patients who are infected by ST298* have more significant underlying disease or immune compromise than those infected by other *P. aeruginosa* strains.

Characterization of the large AMR plasmid pPABL048

The novel AMR plasmid pPABL048 contributes to the high level of resistance in ST298* through the presence of the class I integron in1697, with many in1697-containing isolates showing an XDR phenotype. However, this integron makes up only a fraction of pPABL048, which is 415,954 bp in size and contains 496 coding sequences. More study is needed to understand the role of pPABL048 in ST298* and to compare it to other related plasmids. The development of plasmid-cured PABL048 in this study provides an ideal tool to investigate plasmid-mediated phenotypes.

In this study, I described a novel large plasmid identified at NMH that is a member of a family of large *Pseudomonas* genus plasmids, many of which are associated with AMR[95,185,313].

Cazares et al. recently described two additional members of this plasmid family in clinical *P. aeruginosa* isolates from Thailand. Alignments detected these plasmids in four isolates from two distinct *P. aeruginosa* clades. Further, they performed comparative genomics analyses characterizing this plasmid family. The 13 complete plasmid sequences which they compared to their two novel plasmids were also considered in the plasmid comparative genomic analyses conducted in our study (Table 2.7)[331]. As such, their findings can provide insights relevant to pPABL048. Cazares et al. show that, in addition to being found in multiple *Pseudomonas* species, these plasmids have been detected in both Asia and Europe and from both clinical and environmental sources. They found that the plasmid family has a conserved core genome of approximately 261 genes, encoding for functions similar to what I described in the pPABL048 plasmid backbone (Figure 2.12). This core genome included genes encoding the conjugal transfer proteins TraG, TraB, and TraV[331]. In our annotations of pPABL048, a gene encoding TraB is present adjacent to a gene encoding an unspecified conjugal transfer protein, suggesting these conjugative proteins are also present on pPABL048 and raising the possibility that it could be transmitted to other *Pseudomonas* strains (Supplementary Table 2.1). Additionally, they found a diverse accessory genome among the 15 plasmids that included a large number of AMR genes[331].

Cazares et al. postulated that these plasmids may be in the IncP-2 incompatibility group due to previous experiments performed on pOZ176[331]. However, pOZ176 also contained additional putative replication and partitioning genes[95]. These additional genes were present in the related plasmid p12969-DIM (notably along with a IncQ2-type replication system), while the IncP-2 type replication and partitioning genes were not[185]. The pPABL048 replication and partitioning genes were part of the plasmid backbone (Figure 2.12A), and as noted by Sun et al.

in the p12969-DIM study[185], the backbone replication gene has not been characterized. This suggests that the IncP-2 incompatibility group is likely not a core characteristic of this plasmid family, but instead that at least two of these plasmids have acquired additional replication and partitioning systems from known groups. As with our analyses, Cazares et al. identified through alignment the likely presence of this plasmid family in a number of *Pseudomonas* genus draft genomes[331].

For the final experiment in their study, Cazares et al. showed the transfer of plasmids pOZ176 and p1 (whose sequences are included in our study) from their parent strains into the *Pseudomonas flourescens* strain SBW25. This could be due to the conjugal transfer genes present in the plasmid core genome described above. However, pOZ176 carries a operon containing a number of conjugal transfer genes, including *trbBCDEJLFGI* and several others[95], that are not present in pPABL048 or the other previously described plasmid p12969-DIM[185]. Conjugation has been attempted in the related plasmids p12969-DIM and pSY153-MDR without success[185,313]. In p12969-DIM this can be explained by the fact that these plasmids are likely *Pseudomonas* genus restricted, as the authors only tested conjugation into *Escherichia coli*[185]. However, pSY153-MDR was also unable to be conjugated into the *P. aeruginosa* strain PAO1[313]. As such, it is possible that conjugation is not a fundamental characteristic of this plasmid family. However, it is clear that they can at the very least acquire systems that allow for their transfer. The simultaneous presence of another plasmid with intact conjugation machinery may also be able to mediate transfer. The ability of pPABL048 to transfer via conjugation should be tested. Neither pOZ176 nor p1 conferred a fitness cost on *P. fluorescens* SBW25 in KB media in a competition assay. Further, a sizable proportion of the competitors acquired the plasmids,

showing the risk that conjugal transfer could play in the spread of AMR by this plasmid family[331].

Beyond conjugation, there are multiple other ways in which pPABL048 could impact the phenotypes of ST298* isolates that warrant investigation. Screening pPABL048 against the virulence factor database[310] identified three potential virulence factors related to type IV pili and the plasmid backbone includes both a chemotaxis locus and a putative pilus locus (Figure 2.12 and Supplementary Table 2.2). Type IV pili are known to play important roles in motility, adhesion, aggregation, and DNA binding (and in some species DNA uptake)[2,212,332,333]. With that in mind, PABL048 motility and cell adhesion (to cells or surfaces) should be tested in the presence or absence of pPABL048 using previously described assays[213,214,223]. As noted in Chapter 2, three ST298* isolates contain an uncharacterized variant of the plasmid-borne β-lactamase OXA-10 that may confer extended spectrum activity, particularly towards ceftazidime. Dr. Kelly Bachta and I are collaborating with the Center for Structural Genomics of Infectious Diseases to obtain the crystal structure of this novel OXA-10 variant, which can be compared to the existing structure of OXA-10[334] to determine whether there are any structural factors that may contribute to extended spectrum activity. Dr. Bachta is testing whether this variant confers ceftazidime resistance by expressing both OXA-10 and the novel variant in PAO1 and PA14 and observing impact on MICs. The impact of pPABL048 on fitness in animal models of infection, including bacteremia and pneumonia, should also be characterized.

The Hauser laboratory has now performed long-read MinION sequencing on four additional ST298* isolates: PABL020 (ST298* isolate with only 15.6% alignment to pPABL048), PS1793 (which contains the novel OXA-10 variant), PS1875 (which has in1697 but is missing a large contiguous segment of pPABL048), and PA-NM-088 (a recent ST298*

isolate). Assembling complete genomes for these isolates and comparing to the PABL048 chromosome and plasmid will provide insights into the evolution and divergence of pPABL048-like plasmids in ST298*.

In the course of working with PABL048 and its plasmid-cured variant, I encountered a phenotype associated with the carriage of pPABL048 that warrants further investigation. That is, late stationary phase cultures of plasmid-cured PABL048 appear to precipitate out of solution when left without shaking while the parental PABL048 strain remains suspended. This phenotype was also observed in other ST298* isolates, where isolates lacking the plasmid (PABL040 and PS2027) precipitated while an isolate possessing the plasmid (PA-NM-088) remained largely suspended. The exception was PABL020, which based on read alignment lacks the majority of pPABL048 sequence (Appendix I). The time-scaled phylogenetic tree shows that PABL020 is on the deepest branch within the ST298* subclade (Figure 2.8), and the isolate contains sequence aligning to 15.6% of the plasmid (Table 2.4). ST298 strains outside of this subclade did not precipitate (Appendix I). Systematic experiments are needed to fully characterize this phenotype to both better understand the conditions provoking it and ensure that it can be consistently replicated. Still, its observation raises interesting questions about the biology of pPABL048. It is somewhat unexpected that we see a precipitation phenotype upon plasmid loss, as plasmids are often associated with increased aggregation (including in cases where they encode pilus systems)[335-337]. It is possible that this phenotype is secondary to the adaption of ST298* to the carriage of pPABL048 and is unmasked upon its loss, as it is well known that carriage of plasmids can lead to compensatory mutations in the bacterial chromosome[63,338,339]. Additionally, the fact that several ST298* isolates lacking the plasmid show the same precipitation phenotype as plasmid-cured PABL048 supports the conclusion that

pPABL048 was most likely acquired a single time by the subclade and has subsequently been lost in several isolates. This is also supported by the presence of the plasmid on multiple branches of the ST298* time-scaled phylogenetic tree (Figure 2.8), and our observation that the plasmid can be lost in both natural (heterogenous presence in PABL036 and PABL067) and experimental (plasmid-cured PABL048) conditions. It is possible that an ancestor of PABL020 also contained this plasmid and that retention of a portion of it explains the lack of precipitation phenotype. Once this phenotype is better characterized, the region of the plasmid present in PABL020 would be a good place to start in understanding its mechanism. This phenotype also increases the necessity to observe the plasmid's impact on motility and adhesion.

A machine learning approach to predict *P. aeruginosa* virulence in mice from genomic data

In Chapter 3, I use machine learning to investigate whether the genome of *P. aeruginosa* is predictive of strain virulence in a mouse model of bacteremia. Models trained using accessory genomic information (in the form of AGE presence or absence) were able to predict level of virulence (high or low), with a mean nested cross-validation accuracy ranging from 0.72 to 0.75 depending on the algorithm used. This demonstrates that there is signal in the *P. aeruginosa* accessory genome predictive of virulence. We confirmed this finding using an independent test set, where a random forest model trained on accessory genomic information predicted virulence with an accuracy of 0.72. Further, we showed that core genome information (in the form of core genome SNVs) and whole-genome information (in the form of k-mer counts) can also be used to predict virulence.

While individual factors increasing or decreasing the virulence of *P. aeruginosa* have been well described, we show that the genome as a whole can be used, with moderate

performance, to predict an isolate's virulence. Further study into how these genome-based models make their predictions would improve our understanding of what dictates variation in pathogenicity between *P. aeruginosa* isolates. Importantly, our accessory genome random forest model made predictions based on a diffuse genomic signature rather than the presence or absence of individual AGEs. This raises questions regarding what strategies would be most effective in identifying novel virulence or anti-virulence factors from genomic data. Unlike virulence in mice, the *P. aeruginosa* genome was not able to predict persistence or eradication in early cystic fibrosis isolates. Comparing these two scenarios may provide insights into the type of phenotypes that can be predicted using genomic information alone. Finally, the machine learning methodology employed in this study can be used as a framework for future studies investigating the relationship between the bacterial genome and diverse phenotypes. This includes both simpler phenotypes (such as cytotoxicity or growth rate) where models may be able to predict with higher accuracy or identify causal features, and more actionable phenotypes (such as patient morality) where model predictions could serve as a guide for clinical practice and identify high-risk infections.

The P. aeruginosa genome is predictive of virulence in mice

Our results showing that the *P. aeruginosa* genome possesses signal predictive of virulence in a mouse model helps move us beyond the established understanding the individual factors can increase or decrease pathogenicity[24,25,31,70,75,76,208,261] and towards a more global understanding of intraspecific variation in *P. aeruginosa* virulence. Still, it is not clear from the current study how our models use this genomic information to predict virulence. In Chapter 3, we discussed that the accessory genome random forest model appeared to be basing its predictions on a diffuse genomic signature, and that this signal perhaps resembled phylogenetic

structure. Studies focusing on AMR prediction have shown that both simply identifying an isolate's closest genomic neighbors in a labeled database can accurately predict resistance level and that models can be highly predictive even when phylogeny is corrected for by ensuring that a test isolate's closest neighbors were not used to train the model[66,279]. While seemingly conflicting, these results are not necessarily inconsistent. For example, a clade may have acquired an accessory gene or chromosomal mutation conferring resistance recently in evolutionary history. This would cause closely related isolates to show a similar phenotype even if it was not conserved deeper in the phylogenetic tree, and the same causal features may also be present on different branches. As noted in Chapter 3, the simplest way to test whether phylogenetic structure is playing a large role in our virulence predictions may be through a "block cross-validation" strategy similar to that utilized by Khaledi et al. Here, they grouped isolates into blocks by ST and assigned blocks to the training and cross-validation folds rather than individual isolates[66]. Grouping isolates by ST may not be the best way to rule out the influence of phylogenetics on model predictions as isolates from different sequence types can still be very closely related if they are in the same clonal complex (e.g. ST298 and ST446). Other options would be to expand to clonal complexes or use an algorithm such as hierBAPs[340] to cluster isolates into subpopulations before dividing into blocks. It would be possible to perform this analysis in a stepwise fashion, seeing how model performance changes as increasingly large portions of the phylogenetic tree are grouped into blocks. At the highest level, we could ask the question of whether a model built on isolates from the *exoS*+ clade shows any predictive power against isolates from the *exoU*+ clade. A recent study by Nguyen et al. took a different approach, instead correcting for phylogeny during the model building process by putting greater weight on

isolates from rare clades (and conversely less weight on any given isolate from a common clade)[291].

Our ability to predict virulence was moderate. Using a random forest approach, models built using accessory genomic features, core genome SNVs, whole-genome 8-mers, and whole genome 10-mers showed mean nested cross-validation accuracies of 0.75, 0.65, 0.69, and 0.71 respectively (Figures 3.5 and 3.15). These approaches performed similarly when a final model was used to predict the virulence of the test set, with accuracies of 0.72, 0.72, 0.6, and 0.68 respectively (Table 3.5). As discussed in Chapter 3, limitations in our study related to how isolates were assigned a virulence class may have decreased the performance of our models. Briefly, virulence is continuous and not neatly split into binary classes and error in our mLD_{50} estimations could lead to isolates being assigned incorrectly. Additionally, our sample size was limited by the large number of mice required to perform these experiments, which did not lend itself easily to a high-throughput assay. This is particularly apparent in our test set, which contained only 25 isolates. Learning curve analysis suggests that performance would not necessarily improve with more samples, at least for the accessory genome models, but that may at least in part be secondary to the difficulties in assigning isolates to a clear virulence class. That our machine learning approach performed as well as observed given these limitations makes our findings all the more impressive. Still, the limitations we encountered in this study will be important to consider when designing future machine learning experiments.

In this study, we used AGE presence or absence, core genome SNVs, and whole-genome k-mers as model features, but these are not the only ways to represent genomic information. Other types of genomic features may allow for the same information to be presented in a more compact form, potentially reducing both model complexity (and through this computational

demands) and overfitting. For example, considering only nonsynonymous SNVs would reduce the size of our core genome SNVs feature set while focusing on the features that would be most likely to have functional impacts. A drawback here is that this may lead to the exclusion of important SNVs in noncoding regions. Additionally, there are graph-based methods to condense k-mers into more interpretable features termed "unitigs"[341]. As noted in Chapter 1, an alternative to using AGE presence or absence as accessory genomic features would be to consider gene or protein families[67,68]. This may result in features which are easier to interpret but may also result in the loss of important accessory sequences that are either domains within a gene or are part of intergenic regions. The type of genomic information used should be thoughtfully considered when designing future machine learning studies.

A strength of our study is that we explicitly examined how model performance was impacted by our sample size through learning curve analyses. This allowed us to conclude, for example, that the ability of our accessory genome models to predict virulence was unlikely to substantially improve with additional samples (Figure 3.6) but that there was more room for improvement if moderate virulence isolates were excluded (Figure 3.13C), suggesting that the challenge of assigning isolates to a given virulence class was a limiting factor. Similar analyses have been conducted in other machine learning studies. For example, Nguyen et al. evaluated changes in the accuracy of models predicting AMR in *Salmonella* with increasing training set size[275]. However, this practice is not currently standard among studies using machine learning to predict bacterial phenotypes. This leads to unanswered questions regarding how much genomic information is needed to predict different phenotypes in examined bacterial populations, which could have important implications in both one's understanding of the results (e.g. the extent

performance was limited by insufficient training data) and the design of future machine learning studies.

Model predictions are not based on individual virulence or anti-virulence factors

In our random forest model of virulence trained using accessory genomic content, all AGEs included showed low permutation importance, with the highest ranked AGE lowering out-of-bag accuracy by only 1% on average when randomly permuted. In fact, almost all AGEs (2,979/3,013) could be permuted without any change in model accuracy (Supplementary Table 3.2). We further showed that we could randomly subset the AGE feature set down to at least a tenth of its initial size with little loss of performance (Figure 3.8). Together, these findings suggest that there is a high degree of redundancy in the accessory genomic feature set. In our simulation analyses, we found that the random forest algorithm was least sensitive to the presence of a single causative feature (Figure 3.16), but since our accessory genome models showed equivalent nested cross-validation performance regardless of algorithm choice (Figure 3.5) it is unlikely that examining models built using other algorithms would have uncovered individual highly predictive AGEs. As we discussed in Chapter 3, our model appears to be learning a diffuse genomic signature which may, at least in part, approximate phylogenetic structure. This may be a reason why we are able to predict virulence to a similar extent when using different types of genomic features (AGEs, core genome SNVs, k-mers). The core and accessory genomes are not independent of each other, which can be seen by the finding that the two major *P. aeruginosa* clades defined from a core genome phylogenetic tree also largely cluster by accessory genome content[51].

Our conclusions should also serve as a cautionary note when interpreting the results of other machine learning studies. If we had not considered permutation importance and instead

used a different metric of feature importance, such as Gini importance (the default for random forest in scikit-learn[287]) or ranking features by coefficient in the regularized logistic regression model, it may have been less obvious that individual AGEs played a very small role in model predictions. If we simply had a list of the AGEs that were "most important", it may have led us to overestimate how large a role they play in virulence. Similarly, if we had included a feature selection step to reduce the number of features used to train the final model, each of the retained features would have a greater impact on model predictions. This could be similarly misleading, as they could not be disentangled from highly correlated features that were excluded. Features responsible for causing a phenotype are not necessary to build a highly predictive model in a system as complicated as the bacterial genome[279,291]. In studies modeling AMR, the detection of genes or variants known to play a role in resistance is often used as evidence that the model in question can identify resistance factors[276,278,342]. While this proves that these models can identify true resistance elements, it does not mean that novel features identified through these approaches are necessarily going to impact AMR. In a study modeling patient mortality in two *S. aureus* clones, Recker et al. highlighted a number of loci as predictive of mortality in their models. They selected one of these, the known virulence factor *capA*, for further analysis. They identified a SNV is this gene resulting in defective capsule production, and therefore susceptibility to killing by neutrophils, in isolates from six patients who survived infection[282]. While this is a promising result, it cannot prove that defective capsule production contributed to the survival of the infected patients or guarantee that any of the other predictive loci are involved in pathogenesis. Any genomic features identified as predictive in these modeling approaches are hypotheses. Microbiologic studies are required to prove that they are actually playing a role in that phenotype.

One of our original goals when embarking on this project was to see if machine learning could be used not only to predict virulence but also to identify novel virulence factors. A recent project in the Hauser laboratory, described in Allen et al. 2020, performed a pan-accessory genome wide screen to identify novel AGEs that contribute to virulence in *P. aeruginosa*[70]. In this study, candidate AGEs were identified that both correlated with virulence and showed homology to known virulence factors (in the MvirDB[343] and Effective[344] databases). This approach was successful, and for 11/15 tested AGEs virulence in mice was attenuated when that AGE was deleted[70]. This approach was not unbiased, and by its design could only identify AGEs which resembled known virulence factors. It was thought that our machine learning approach may allow us to identify novel virulence (or anti-virulence) determinants in a less constrained manner. We did observe an association of the most important AGEs with low virulence. All of the top 10 AGEs, and 32/34 of the AGEs with non-zero permutation importance, were more prevalent in low virulence isolates (Table 3.4 and Supplementary Table 3.2). As stated in Chapter 3, this should prompt further investigation into whether AGEs have a tendency to impose fitness costs that could lead to reduced virulence. However, for the reasons described above it is clear that we cannot use the models developed in this study to identify novel virulence factors. This highlights the importance of problem formulation when designing a computational study. The approach best suited to predicting a phenotype may not be the same approach that is best suited to identifying individual features that are causal of that phenotype. As stated above, powerful predictors can be developed even when causative features are not considered[279,291].

An alternative approach to identifying genomic features causing increased or decreased virulence would be through a bacterial genome-wide association study (GWAS), a technique that

has been a major focus of development in the past several years. GWAS is a well-established technique in eukaryotic genomics (particularly humans), but the presence of accessory genomes and extensive horizontal gene transfer creates challenges not seen in higher eukaryotes[345]. An important factor in bacterial GWAS approaches is control for the effects of population structure, which attempts to reduce the effect of spurious correlations between genotype and phenotype that are simply the result of this structure[345-347]. This may allow for unbiased searches while still limiting the number of false positives (features that are associated with, but not causative of, the phenotype). Two main ways in which this population structure is accounted for is through linear mixed models[346,348] and through phylogenetic tree-based methods[347,349]. Some of the tools developed for bacterial GWAS, such as treeWAS, have been designed to examine associations between both core genome SNVs and accessory genes with the phenotype, though not necessarily at the same time[347]. Bacterial GWAS has been used to explore factors associated with antimicrobial resistance[341,346,347], invasive disease vs. carriage in *Neisseria meningitidis*[347], pyomyositis vs. carriage in *S. aureus*[350], and gastric cancer from *Helicobacter pylori* infection[351]. As with the machine learning approaches, any hits identified through these screens are hypotheses and would need to be tested to determine whether they indeed play a causal role in the phenotype in question.

Comparing predictive models of P. aeruginosa virulence in mice and persistence in cystic fibrosis patients

As discussed, there was signal in the *P. aeruginosa* genome predictive of virulence, regardless of whether accessory genome AGEs, core genome SNVs, or whole-genome k-mers were used as features in the machine learning models. This, however, was not the case when we attempted to predict persistence in isolates from the EPIC Clinical Trial. When tuning our

models based on F1 score to account for class imbalance, the resulting F1 scores in nested cross-validation were quite low (0.26 when using AGEs and 0.21 when using core genome SNVs). When we instead tuned our models based on accuracy, the optimal solution was to simply call isolates eradicated (exclusively when training on AGEs and near-exclusively when training on core genome SNVs) (Figures 3.18 and 3.19).

From one perspective, the negative finding that the *P. aeruginosa* genome is not predictive of persistence or eradication in early cystic fibrosis raises our confidence that there is true genomic signal predictive of *P. aeruginosa* virulence in mice. It shows that, while the ability of genomic information to predict virulence in mice was moderate, we cannot expect to see the same response for all phenotypes. This is supported by the findings of our simulation analyses, where randomly shuffling which isolates were classified as high or low virulence resulted in nested cross-validation accuracies near 0.5 (Figure 3.16). This is what one would expect to find in the case where the features possess no predictive signal. The performance we observed when tuning models based on accuracy also illustrates the danger of relying on accuracy as the only outcome metric, especially in cases with class imbalance in the training dataset. For the results of a machine learning study to be meaningful it is essential that an appropriate outcome metric be used (both during model training and performance evaluation), and a critical eye for this must be taken when interpreting the literature.

The question remains why we were unable to predict persistence in early cystic fibrosis isolates. It is well known that cystic fibrosis patients can be persistently colonized by a single *P. aeruginosa* clone[11,12,352], with one study showing through sequencing the carriage of a single clone (which diversified into subpopulations) over 32 years[11]. It is possible that whether a given strain is able to establish long-term colonization is largely stochastic or determined by factors

other than the bacterium itself. There could be far more eradicated isolates that were never detected, simply because they had already been cleared (or not yet acquired) at the time of sampling. Similarly, it is possible that isolates we classified as eradicated could persist (at least for a period of time) under different scenarios, and that isolates we classified as persistent may not ultimately establish prolonged colonization. This possibility is apparent in the phylogenetic tree of this isolate collection (Figure 3.17), where closely related isolates can be seen belonging to both classes. A question this raises is whether any *P. aeruginosa* isolate is then able to colonize a cystic fibrosis patient, either by chance or if given the right environment. This would be inconsistent with the finding that certain *P. aeruginosa* clones have caused large or prolonged epidemics in cystic fibrosis patients. The most famous of these is LES (the Liverpool Epidemic Strain), which has spread out of the United Kingdom, and another example is the DK1 clone in Denmark[11,129,133,353,354]. Additionally, isolates possessing the gene *exoS* appear to be overrepresented in *P. aeruginosa* from cystic fibrosis patients compared to what is seen in acute infection[121,128]. Isolates from the natural environment appear to also be enriched for *exoS* presence[16], so this predilection could be in part due to exposure and increased susceptibility of cystic fibrosis patients to *P. aeruginosa* in general. That would not, however, account for the success of epidemic strains in the cystic fibrosis population.

The definition of persistence used in this study (detection of an isolate closely related to the first isolate for that patient at a later study date) could have exacerbated the stochasticity described above and contributed to the poor performance of our machine learning models. The EPIC clinical trial followed patients for only 18 months[318], and an isolate did not have to be detected throughout the entire study period to be called persistent. As such, it would not be impossible that an isolate we defined as eradicated actually colonized a patient for the same

amount of time as another isolate we defined as persistent. A stricter definition of persistence (e.g. carriage of a given clone for a minimum of 5 years) may reveal that there is a greater importance of bacterial factors in chronic colonization of cystic fibrosis patients. Unfortunately, this is outside of the scope of the data we have from the EPIC Clinical Trial.

As stated above, another possibility is that patient or environmental factors largely dictate whether a given *Pseudomonas* strain is able to successfully colonize the cystic fibrosis lung. Even if the bacterial genome plays a role in persistence, these confounding factors may mask their impact. Our mouse experiments were performed with an inbred mouse strain in a controlled environment. While this does not remove all sources of experimental variance, this is far less than what one would expect in clinical study. For example, ages of patients in the EPIC trial ranged from 1-12 years[318]. Additionally, patients were assigned to one of four antibiotic treatment arms. While these treatment arms showed no significant impact on study outcomes[318], they could have still had an impact on the persistence of individual isolates. The degree of lung pathology in different patients, or differences in their microbiomes or environmental exposures, may also be important factors in whether a given strain is able to persist. These potential confounding factors should be considered when designing future studies evaluating the relationship between bacterial genomics and phenotype, particularly when studying clinical collections and patient outcomes.

Machine learning as a tool to interrogate P. aeruginosa phenotypes

In this project we have established a machine learning approach to predict *P. aeruginosa* virulence in mice from genomic features. The framework laid out in this study can be applied to a variety of scenarios investigating the relationship between the *P. aeruginosa* genome and diverse phenotypes. For example, we have already used this approach to show that the *P.*

aeruginosa genome is not predictive of persistence in early cystic fibrosis isolates, at least in the conditions evaluated in the EPIC Clinical Trial. Moving forward, I see two main avenues for future studies. The first would be to focus on phenotypes which are simpler or more amenable to high-throughput screening than mouse virulence, which could avoid some of the limitations of our present study. The second would be to focus on clinical outcomes, such as patient mortality in acute infection. This would be more challenging, but the models developed could be clinically actionable and help understand how large a role bacterial factors play in patient outcomes.

As previously discussed, practical limitations in our mouse model create uncertainty in our virulence estimates, and virulence in mice was not easily separable into binary classes. Further, virulence is a complex phenotype likely dictated by the combinatorial effects of numerous genetic factors. Examining simpler phenotypes may allow us to avoid at least some of these limitations. This could result in higher model performance and a greater potential to identify predictive features. A first step here would be to examine in vitro growth rate. As noted in Chapter 3, almost all AGEs with non-zero permutation importance in our accessory genome model of virulence were more prevalent in low-virulence isolates. We posited that this might be due to fitness costs imposed by the accessory genome. Examining whether the accessory genome can be used to predict growth rate, and if any specific AGEs are predictive of increased or decreased growth rate, would help us understand the extent to which the accessory genome is a metabolic burden. Additional in vitro assays that could serve as the basis for machine learning studies include cytotoxicity, adhesion, or resistance to killing by immune cells. In many of these scenarios there may still be issues with continuous phenotypes that are not easily divided into discrete classes. If the data cluster into clear groups, it may be simple to frame a given phenotype as a classification problem. On the other hand, with sufficiently precise data and a large enough

sample size, it may be possible to take a regression strategy rather than a classification strategy for machine learning. This has been done with AMR prediction, predicting MICs rather than resistance vs. susceptibility[275,292].

With simpler phenotypes and larger datasets, it may be possible identify predictive features that play a potential role in the phenotype in question. This would be consistent with studies predicting AMR, where known resistance determinants are often highly ranked in their models[275,278,342]. These predictive features would form testable hypothesis for further microbiologic experiments. As described previously, machine learning studies could be supplemented with bacterial GWAS if a main goal is to identify potential causal elements.

Other options would be to examine virulence in simpler model organisms, such as *C. elegans*[27,29,31,32], *G. mellonella*[29,30], *D. melanogaster*[27-29], or even plants[24,30]. These infection models would scale easier than mice, which would allow us to both screen a larger number of *P. aeruginosa* isolates and estimate virulence with more precision. If the same strains were tested in different model organisms, it would be possible to assess if virulence in one organism can be used to predict virulence in another. Features predictive of virulence in multiple infection models would be ideal targets for further analysis.

Applying machine learning to clinical phenotypes would be more challenging, as highlighted by the poor performance of the *P. aeruginosa* genome in predicting persistence in early cystic fibrosis and mixed results in the literature. This, however, only increases the necessity for additional studies to move the field forward. When modeling patient mortality in specific *S. auerus* clones, Recker et al. found that genomic features could predict mortality with an area under the receiver operating characteristic curve (AUC) of 0.75 for CC22 and 0.79 for CC30[282]. A recent study by Lapp et al. found that the ability of the *K. pneumoniae* genome to

differentiate between isolates causing infection or simply colonization was relatively poor, with AUC values showing an inter-quartile range of 0.55-0.61[294]. If bacterial factors are not predictive of patient outcomes, it is still an important finding and would suggest that future studies should focus on patient, environmental, or treatment factors that influence disease outcomes.

In machine learning studies examining the predictive power the bacterial genome in clinical phenotypes, it would also be important to consider the impact of patient factors (such as age or comorbidities). This would allow us to compare whether bacterial or patient factors are more predictive of outcomes, and whether there is any benefit of considering them together. For example, it may be that the combination of a particularly virulent bacterial isolate with an especially vulnerable (e.g. immunocompromised) host creates the highest risk for poor outcomes. Another possibility would be that genomic determinants of bacterial virulence are important to cause disease in an otherwise healthy patient but are dispensable in one who is already severely compromised. In their study of mortality caused by *S. aureus* clones, Recker et al. found that for CC22, bacterial genotype (AUC 0.75) was more predictive than clinical features (AUC 0.66) and there appeared to be additional benefit of using these together along with bacterial phenotypic features (AUC 0.84). In the other clone, CC30, bacterial genotype (AUC 0.79) was more predictive than clinical features (AUC 0.5) and little benefit was observed when combining these together with bacterial phenotype (AUC 0.81)[282]. As they only report single AUC values for each condition (based on out-of-bag performance in their random forest models)[282], it is unclear whether these differences in performance are significant. A cross-validation approach may be better suited to compare the effectiveness of different feature sets. On the other hand, Lapp et al. found that patient and bacterial features resulted in similarly weak predictors[294]. To our

knowledge, no study thus far has compared the influence of bacterial genomics and clinical factors on patient outcomes in *P. aeruginosa* infections.

The highest priority clinical phenotype for our machine learning approach would be patient mortality during acute infection. A model predictive of patient mortality would be clinically actionable by identifying patients at increased risk for poor outcomes. Previous studies have suggested that the presence of *exoU* may be associated with quicker time-to-death[140] and that an MDR phenotype or inappropriate therapy (due to drug resistance) can be associated with worse outcomes[140,141], but no study to date has examined whether the genome as a whole is predictive of patient mortality. It would be difficult to conduct clinical studies prospectively. The best approach, at least to start, would then be retrospective analysis of existing studies with banked or already-sequenced isolates. Another question that could be explored through our machine learning approach would be whether there is a genomic signature that differentiates clinical and environmental *P. aeruginosa* isolates. Investigating this could resemble a previous study that used the pangenomes of *Salmonella enterica* and *E. coli* to predict host source (human, avian, bovine, or swine)[283] and could take advantage of public databases to analyze a large number of genomes. It would also be useful to revisit persistence in cystic fibrosis patients. For the EPIC Clinical Trial, it would be possible to determine whether the incorporation of clinical data improves the performance of our machine learning models. With a new dataset, we could examine whether bacterial factors become important when using a stricter definition of persistence that requires chronic carriage.

CHAPTER 5

Materials and Methods

Materials and Methods used in Chapter 2. Identifying and characterizing a prolonged local epidemic of extensively drug-resistant *Pseudomonas aeruginosa* at Northwestern Memorial Hospital

Bacterial Isolates

Several collections of *P. aeruginosa* isolates available in the Hauser laboratory were evaluated (Table 2.1). These include 3 cohorts of isolates collected from patients and clinical settings at Northwestern Memorial Hospital (NMH) in Chicago: 100 bloodstream isolates collected 1999-2003 ("PABL")[316], 301 isolates from clinical specimens and hospital environments collected 2002-2009 ("MolEpi"), and 99 isolates from patient samples collected 2013-2018 ("PA-NM"). Other patient isolates screened included 601 bloodstream isolates collected from 10 public hospitals in Spain between 2008-2009[6] and 100 isolates from patient samples collected at Brigham and Women's Hospital (BWH) in Boston between 2015-2016. Also included were 58 *P. aeruginosa* isolates collected from multiple healthcare facility environments (e.g. sinks) in the Chicago metropolitan area between 2017-2018. While these isolates were not collected from patients, they are healthcare associated and could be of human origin or serve as a reservoir for potential infections. CC446 isolates were identified from these collections through post-sequencing *in silico* MLST using allele sequences and MLST profiles listed in the PubMLST database[305]. Both Dr. Egon Ozer and I performed *in silico* MLST analyses using a script written by Dr. Egon Ozer.

In addition to the isolates described above, 2483 *P. aeruginosa* genomes previously deposited in the NCBI database (accessed October 26, 2017, deposited 2006-2017) were screened using *in silico* MLST to identify CC446 isolates. Dr. Egon Ozer performed this screen. The 38 CC446 isolates identified in this screen were included in the genomic analyses performed in this study.

Antimicrobial Resistance Determination

Minimum inhibitory concentrations (MICs) against 8 antibacterial agents from 7 antipseudomonal classes were determined for each CC446 isolate included in this study that was available for testing using microbroth dilution. MICs were determined in triplicate using the broth microdilution protocol by Wiegand, et al.[355] and are reported in Table 2.3 and Table 2.6. The following antibiotics were prepared from commercially available sources and were used to assess MICs: gentamicin, cefepime, ceftazidime, piperacillin-tazobactam, meropenem, aztreonam, ciprofloxacin, and colistin. Where discordant values were obtained, the median was used. Dr. Kelly Bachta performed antimicrobial susceptibility testing and determined MICs for the tested CC446 isolates. Isolates were classified as susceptible or non-susceptible (intermediate and resistant) to each antibiotic based on 2018 Clinical and Laboratory Standards Institute (CLSI) breakpoints[310]. An isolate was classified as multidrug resistant (MDR) if it was non-susceptible to at least one antibiotic from ≥3 classes tested and classified as extensively drug resistant (XDR) if non-susceptible to at least one antibiotic from ≥5 classes tested (susceptible to antibiotics tested from ≤2 classes)[134].

Susceptibility to ceftazidime-avibactam and ceftolozane-tazobactam was assessed through Kirby-Bauer disk diffusion testing using HardyDisk AST disks (Hardy Diagnostics).

Isolates were classified as susceptible or non-susceptible based on 2018 CLSI breakpoints[310]. I performed this experiment with Dr. Kelly Bachta.

BURST Analysis

To investigate the relationships between ST298, ST446, and other related sequence types, BURST analysis was performed using the goeBURST algorithm[356] as implemented in PHYLOViZ (v2.0)[357], and the resulting clonal complex containing these sequence types was identified. All *Pseudomonas aeruginosa* sequence types listed in the PubMLST database (accessed August 12, 2019) were considered[305].

Whole Genome Sequencing

To construct a complete genome sequence for PABL048, long-read sequencing was performed on a PacBio RS II machine at the University of Maryland Institute for Genome Sciences. PacBio raw data were corrected and assembled using HGAP assembler (SMRT Analysis 2.3.0), Canu assembler (v1.2)[358], and Celera assembler (v8.2)[359]. The assemblers were run using default settings. Resulting contigs were combined and circularized using Circlator (v1.5.1)[360]. The final assembly was polished using Quiver (SMRT Analysis 2.3.0). Indel errors were corrected with Pilon (v1.21)[361] using 100-bp paired-end reads generated on an Illumina HiSeq 2000 system, with an average read coverage of 190-fold. The genome was annotated through the NCBI Prokaryotic Genome Annotation Pipeline[362] and has been deposited to GenBank with the accession numbers CP039293.1 (chromosome) and CP039294.1 (plasmid). I prepared genomic DNA of PABL048. LB broth overnight culture (4 mL) was resuspended in 2 mL PBS. Genomic DNA was extracted using a Promega Maxwell Cell DNA Purification Kit, pooling the eluants of 3 extractions using 400 uL resuspended culture each. Library preparation and sequencing were then performed by the University of Maryland Institute for Genome

Sciences. Dr. Egon Ozer performed the complete genome assembly and deposited the genome to NCBI.

Potential virulence factors present within the PABL048 chromosome and plasmid were identified using the VFanalyzer pipeline to screen against the virulence factor database (VFDB)[310].

CC446 isolates were whole genome sequenced using Illumina HiSeq and MiSeq platforms. Sequencing was performed at Northwestern University Feinberg School of Medicine and at the University of Maryland Institute for Genome Sciences. Preparation of libraries for sequencing was performed by various members of the Hauser laboratory, including myself. Sequencing reads were trimmed using Trimmomatic (v0.36)[363] to remove low-quality bases and assembled into draft genome contigs using SPAdes (v3.9.1)[364]. Contigs shorter than 200 bp were filtered out. Both Dr. Egon Ozer and I performed draft genome assemblies. I additionally performed short-read sequencing to investigate heterogenous plasmid presence in PABL036 and PABL067 and to confirm plasmid curing in PABL048. Genomic DNA was extracted from LB broth overnight cultures using a Promega Maxwell Cell DNA Purification Kit. Libraries were prepared using a Nextera XT kit and run using a MiSeq Reagent Kit v3 (600 cycle) on an Illumina MiSeq instrument to yield 2 x 300 bp paired-end reads.

Sequence Alignment

All CC446 genomes were aligned to the complete genome sequence of PABL048, with separate alignments to the PABL048 chromosome and pPABL048 plasmid. Both the chromosome and plasmid were indexed with BWA (v0.7.15) using the command “bwa index” and samtools (v0.1.19-44428cd)[365] using the command “samtools faidx”. I then determined

regions of the chromosome and plasmid that are repetitive for later masking using a custom script written by Dr. Egon Ozer which blasts 500 bp fragments of the sequence against itself.

For isolates with reads available (Table 2.2), read-trimming was performed with Trimmomatic (v0.36)[363] to remove low quality bases, and alignment was then performed using BWA (v0.7.15) with the BWA-MEM algorithm[366]. Alignments were then sorted and indexed using samtools (v0.1.19-44428cd)[365] with commands "samtools sort" and "samtools index". Single nucleotide variants (SNVs) relative to the reference were called using the mpileup function of samtools (v0.1.19-44428cd)[365] with the following settings: -E (recalculate extended BAQ), -M 0 (cap mapping quality at 0), -Q 25 (skip bases with BAQ less than 25), -q 30 (skip alignments with mapQ less than 30), -m 2 (minimum gapped reads for indel candidates of 2), -D (output per-sample DP in binary call format [BCF]), -S (output per-sample strand bias P-value in BCF), and -g (generate BCF output). The resulting bcf file was viewed using bcftools (v0.1.19-44428cd) and I generated the final fasta-format alignment using a custom script written by Dr. Egon Ozer as follows. SNVs were filtered if they failed to meet 1 or more of the following criteria: minimum SNV quality score of 200, minimum read consensus of 75%, minimum of 5 reads covering the SNV position, maximum of 3 times the median read depth of the total alignment, minimum of 1 read in either direction covering the SNV position, homozygous under the diploid model, and not within a repetitive region as determined by BLAST alignment of fragments of the reference sequence against itself. Any positions in the reference sequence with SNVs that passed the above filters were changed to the SNV base. Positions with SNVs that did not pass the above filters were changed to a missing base character. Non-SNV positions with coverage of fewer than 5 reads were also changed to a missing base character. For 13 NCBI genomes usable reads were not available, so draft genome contigs were aligned to PABL048

using NUCmer (v3.1)[367] with SNVs within repetitive regions masked (replaced with "N"). I performed the NUCmer alignments using a custom script written by Dr. Ozer.

Alignments of reads from the NMH CC446 isolates to pPABL048 were visualized with Tablet (v1.19.09.03)[368] using the sorted BAM files generated above.

For subsequent phylogenetic analysis, alignments of all CC446 isolates to the PABL048 chromosome were concatenated into a single fasta file. The core genome was defined as all non-missing and non-filtered positions present in 91 (98%) of the 92 genomes. Bases in all non-core positions were replaced by the corresponding base in the PABL048 reference. I performed core genome filtering using a custom script written by Dr. Egon Ozer.

Phylogenetic Analysis

A maximum likelihood phylogenetic tree was constructed based on core genome alignments to PABL048 using RAxML (v8.2.11)[369]. Tree construction was performed using a gamma model of rate heterogeneity (-m GTRGAMMA) with 1000 rapid bootstraps (-f a -N 1000) to assess support. The phylogenetic tree was corrected for the impact of recombination using ClonalFrameML (v1.11-3-g4f12f23) with default settings[370]. The recombination-corrected phylogenetic tree was visualized and annotated using iTOL[371]. I then masked predicted recombinant regions in the core genome alignment using a custom script written by Dr. Egon Ozer.

Clustering of CC446 isolates was performed using the hierBAPS algorithm as implemented in the rhierbaps (v1.1.2) package in R (v3.6.0)[311,340,372]. The recombination-filtered core genome alignment was used as input, the maximum number of populations (n.pops) set to 20, and the maximum depth of clustering (max.depth) set to 2.

To model the evolution of the ST298* subclade, a time-scaled phylogenetic tree was constructed for bacteria from this subclade with known isolation dates (Supplementary Tables 1 and 2). Recombination-filtered core genome alignments of these isolates to the PABL048 chromosome were extracted, and the recombination-corrected maximum likelihood tree was pruned to contain only these isolates using the ape package (v5.1) in R[373,374]. These, along with isolation dates, were used as input for Bayesian analysis. I generated the input file for this analysis from the recombination-filtered alignments, pruned tree, and isolation dates using a custom script written by Dr. Egon Ozer (which I modified to accept a yyyy-MM-dd date format). Bayesian analysis was performed using BEAST (v2.5.1) with a gamma site model, strict clock rate, Yule tree prior, and chain length of 100 million, sampling every 1000 states[375]. Sampled states were analyzed with Tracer (v1.7.1) to determine the clock rate and last common ancestor date for ST298*, considering the first 10% of states as burn-in[376]. SNVs/year were determined by multiplying estimated clock rate (SNVs per site per year) by PABL048 chromosome size. To construct the final time-scaled tree, TreeAnnotator (v2.5.1) was used to form a maximum clade credibility tree from the sample trees with node heights as common ancestor heights, using the first 10% of trees as burn-in. The time-scaled tree was visualized using FigTree (v1.4.4).

Determination of Heterogenous Plasmid Presence in ST298 and Curing Plasmid from PABL048*

ST298 isolates from NMH were screened for heterogenous resistance to gentamicin by patching individual colonies onto LB agar supplemented with gentamicin (50 μg/mL). Gentamicin-resistant and -sensitive colonies of PABL036 and PABL067 were selected for further analysis. These underwent MIC testing and whole genome sequencing as described

above. Plasmid presence and chromosomal SNVs were determined by read alignment to the PABL048 complete genome.

To cure pPABL048 from PABL048, we used a combination of sodium dodecyl sulfate (SDS) and elevated temperature[377]. Colonies were inoculated into 5 mL LB with 2% SDS, cultured at 42ºC for approximately 24 hours, and plated on LB agar with irgasan (5 µg/mL). Colonies were then screened for loss of gentamicin resistance on LB agar supplemented with gentamicin (50 µg/mL) and irgasan (5 µg/mL) as a marker for pPABL048 presence. Loss of pPABL048 was confirmed through whole genome sequencing and alignment as described above.

Characterization of in1697, pPABL048, and Plasmid Comparative Genomics

The AMR integron in1697 was identified through detection of several AMR genes in multiple NMH ST298 isolates using the ResFinder database[309]. Dr. Jonathan Allen identified this set of AMR genes and which isolates possessed them (with the exception of PABL067, an isolate with heterogenous presence of the AMR plasmid where these genes were lacking in our initial sequence, see above). I then characterized them as being part of a novel class 1 integron. The sequence of this locus was referenced against the PABL048 complete genome to determine its genomic context. In1697 was further characterized through sequence alignment of translated coding sequences to the NCBI non-redundant protein database and through the INTEGRALL integron database[378], through which it was assigned the unique name in1697.

Plasmids similar to pPABL048 were identified using BLASTn, separately screening *P. aeruginosa*, non-*aeruginosa Pseudomonas*, and non-*Pseudomonas* Gammaproteobacteria sequences in the NCBI nucleotide database (nr/nt). This identified 16 plasmids with a minimum of query coverage of 70% (Table 2.7). SPINE (v0.3) was used to determine the plasmid backbone of pPABL048 based on sequences conserved in 16/17 complete plasmids analyzed[48].

To identify other isolates which harbor plasmids similar to pPABL048, 3133 *Pseudomonas* genus draft genomes cataloged by the *Pseudomonas* Genome Database (accessed January 2019)[50] were aligned to pPABL048 using NUCmer as described above and screened for genome sequences with >70% alignment by length (Table 2.8). A 98% "core" sequence alignment to pPABL048 (considering all non-missing and non-filtered positions in 62/63 sequences) was determined. A maximum likelihood phylogenetic tree was constructed to show relationships between these plasmids using RAxML (GTRGamma model, 1000 rapid bootstraps)[369].

Mutational Resistance Analysis

To examine the role of mutational resistance in the observed AMR phenotype, a panel of PABL048 genes was screened for mutations known to confer resistance in *P. aeruginosa*[145]. In cases where resistance is imparted through specific gain-of-function mutations, translated coding sequences were screened for previously reported alleles known to be involved in resistance. In cases where resistance is conferred from loss-of-function mutations (e.g. gene disruption), translated coding sequences were compared to that of PAO1 as a reference to assess for gross changes in the amino acid sequence. The genomes of ceftazidime-resistant ST298* isolates PS1793, PS1796, and PS1797 were similarly screened to investigate mechanisms of ceftazidime resistance. Protein sequences for OprD, AmpD, and OXA-10 were extracted for isolates in the ST298* subclade and multiple sequence alignment was performed using CLC Sequence Viewer (v8.0) with default parameters. For OprD and AmpD, the PAO1 protein sequence was incuded as a reference. For OXA-10, the sequence of OXA-10 and known variant OXA-14 were included as a reference. For PABL067, OXA-10 protein sequence was extracted from the assembly of a gentamicin resistant colony.

Materials and Methods used in Chapter 3. Using the *Pseudomonas aeruginosa* genome to predict virulence in a mouse model of bacteremia

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* bacteremia[316] 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 2003-2008[317] 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 2008-2009[6] (Table 3.1).

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 described[317]. Isolates were plated from freezer stocks onto lysogeny broth (LB) agar, and single colonies were inoculated into MINS broth[379] and grown overnight at 37 ºC. Overnight cultures were then subcultured in fresh MINS broth for approximately 3 hours at 37 ºC. 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 3.1). 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., 2020[70]. Some experiments with the Taiwan isolates PAC1 and PAC6 were performed as part of Chuang et al., 2014[317]. Many of these experiments (in particular all conducted prior to 2017) were conducted by other members of the Hauser laboratory prior to the initiation of this project, including by Dr. Jonathan Allen and Dr. Egon Ozer.

A modified 50% lethal dose (mLD_{50}) for each isolate was estimated from the above experiments using the drc package (v3.0-1)[380] in R (v3.6.1)[372]. One outlier experiment for strain S2, which caused 20% mortality at a dose of ~7.2 $\log_{10}$ CFU, was excluded as doses of ~6.3 and ~6.8 $\log_{10}$ CFU caused 80% and 100% mortality, respectively, in other experiments. Percent mortality as a function of dose (in units of $\log_{10}$ CFU) was modeled using a two-parameter log-logistic function and binomial data type. Experiments were weighted by number of mice. These models were used to estimate the mLD_{50} for each isolate, which was then rounded to the nearest tenth (Table 3.2). Cumulative distribution functions were constructed in R to examine the distribution of virulence in the isolates. Isolates with rounded mLD_{50} 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. Much of this sequencing was performed before the initiation of this project and facilitated by various members of the Hauser laboratory.

Genomic DNA was extracted from LB broth overnight cultures using Promega Maxwell Cell DNA Purification Kits. Library preparation was performed by the University of Maryland or by members of the Hauser laboratory. Sequencing was performed by the University of Maryland (HiSeq) or in the Hauser laboratory (MiSeq). I performed library preparation for 17/25 of the Spanish isolates included in the test set using Nextera XT library preparation kits.

Reads were trimmed using Trimmomatic (v0.36)[363], 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)[364] with the careful and automatic read coverage cutoff options. I further filtered draft genomes to remove contigs shorter than 200 bp, with less than 5-fold mean read coverage, or with alignment to phiX, using a custom script written by Dr. Egon Ozer. 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[70,246,302]. 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 by the University of Maryland Institute for Genome Sciences. Raw data were assembled using the HGAP assembler (SMRT Analysis v2.3.0), Canu assembler (v1.2)[358], and Celera assembler (v8.2)[359], all using default settings. Contigs were combined and circularized using Circlator (v1.5.1)[360]. Assemblies were polished using Quiver (SMRT Analysis v2.3.0). Indel errors were corrected using Pilon (v1.21)[361] using paired-end reads generated on Illumina HiSeq or MiSeq platforms. Dr. Egon Ozer performed the hybrid assembly for these

isolates. The complete genome for PABL048 was generated as part of the work conducted in Chapter 2 and its associated sequencing and assembly are described in detail in the methods associated with that chapter.

The initial report of each isolate, its sequencing, and the assembly used in this study are listed in Table 3.1. For all isolates, the version of the genome assemblies used in this study are available on GitHub (https://github.com/nathanpincus/PA_Virulence_Prediction).

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 trees[381]. The phylogenetic trees were annotated and plots generated using iTOL (v4)[371].

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)[48,69]. Spine was used with Prokka[382]-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"[69,70]. 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)" using a custom R script. 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. 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.

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 using a custom R script. Screened length was used to no penalize unique groups where some subelements were too small to be output by clustAGE.

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 ≥200 bp were used as input for subelement_to_tree.pl rather than the default subelements, necessitating the creation of custom input files. The neighbor joining tree and associated heatmap of Bray-Curtis

dissimilarities were annotated and visualized with iTOL (v4)[371]. 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 was 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 perfumed considering which of the training isolate AGEs were identified all 140 isolates. MCA was performed in R (v3.6.1)[372] using the FactoMineR (v2.3)[383] package ("MCA" function) and visualized using the factoextra (v1.0.6) package ("fviz_mca_ind" function).

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 described in the methods associated with Chapter 2. Briefly, reads were trimmed with Trimmomatic (v0.36)[363] and aligned to PAO1 with BWA (v0.7.15)[366]. 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 and SNV positions present in fewer than 95% of genomes were filtered. I then removed invariant sites using a custom script written by Dr. Egon Ozer, yielding a final 95% core variant SNV site alignment.

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. I extracted the sequence present at these variant

positions in the PAO1 alignments for each of the 25 test set isolates to create a SNV feature set corresponding to that used in the training set using a custom python script.

K-mer Counts

K-mer counts (using either 8 or 10 bp k-mers) were determined for each genome using KMC3 (v3.0.0)[384]. All k-mers occurring at least once in each genome's fasta file were identified using the kmc application (k-mer size of 8 or 10, multi-fasta input format, include k-mers occurring at least once, maximum k-mer count of 1677215), and a count file was generated using the kmc_dump application. KMC3 run settings were modeled off of Nguyen et al, 2019[275]. 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 using a custom python script. This served as k-mer feature set in subsequent machine learning analyses.

To define the k-mers feature set present in the 25 test set isolates, k-mer counts were determined using KMC3 as above. A custom python script was then used to create a dataframe of counts for 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)[287] in Python (v3.6.9). The general workflow for the machine learning pipeline is described in Figure 3.4. 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 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)[372] with the tidyverse library suite (v1.2.1)[385].

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.

For the case of the final random forest model trained on AGE presence/absence in the 115 training isolate, training performance was measured by predicting virulence of the training set and comparing to the true values. Additionally, the number of component decision trees was

Code used for machine learning analyses to predict virulence from genomic data, including details on hyperparameters used during grid-search cross-validation, and for plotting the results are available on GitHub. Input data for these analyses (including all AGE, core SNV, and k-mer feature sets) are also available on GitHub (https://github.com/nathanpincus/PA_Virulence_Prediction).

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 "oob_importances" function in the rfpimp (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. The mean permutation importance was then calculated for each AGE and the 10 AGEs with the highest mean permutation importance were plotted using a custom R script (Supplementary Table 3.2 and Figure 3.7A).

The putative annotation of the top 10 AGEs identified by permutation importance was determined by blast search of subelement sequences against the Pseudomonas Genome

Database[50] 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 was 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 calculate 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 compared to the true values 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.

Code and input data used for these analyses are available on GitHub (https://github.com/nathanpincus/PA_Virulence_Prediction).

Simulating the Performance of Accessory Genome Models When Phenotype is Randomly Permuted and a Perfectly Predictive AGE is Added

The accessory genome feature set of 3,013 AGEs for the 115 training isolates was used as the starting feature set for the purpose of the simulations. Labels were assigned to be equivalent

to the proportion of low (59) and high (56) virulent isolates seen in this study and randomly shuffled. Nested cross-validation was conducted to estimate generalization performance models trained using these AGE features and shuffled labels with the random forest, L2-regularized logistic regression, elastic net logistic regression, or support vector classifier algorithms as described above. Additionally, an additional perfectly predictive feature (identical to the labels) was added to the AGE feature set and nested cross-validation was repeated to observe the extent to which this improved model accuracy. As performance could vary depending on how the labels were randomly assigned, this process was repeated using 10 random seeds. The mean nested cross-validation accuracy (with and without the perfectly predictive AGE) for each seed was plotted for each algorithm, along with the mean and 95% confidence interval between seeds, using R (v3.6.1)[372] with the tidyverse library suite (v1.2.1)[385].

Predicting Persistence or Eradication in a Collection of Early Cystic Fibrosis P. aeruginosa Isolates from Genomic Features

A set of 207 *P. aeruginosa* isolates collected from early infection in cystic fibrosis patients was considered. These isolates were originally collected as part of the Early Pseudomonas Infection Control (EPIC) program[318,319]. The first isolate from each of 207 patients was considered and classified as persistent if another isolate with <1000 SNVs was collected from the same patient at a later study visit. Both draft genome assemblies and paired-end sequencing reads for each isolate were obtained from Dr. Maulin Soneji, who also performed SNV comparisons to call isolates as persistent or eradicated. The classification of each isolate as persistent or eradicated was obtained from Dr. Sumitra Mitra.

To analyze population structure of these early cystic fibrosis isolates, a 95% core genome phylogenetic tree was constructed from the draft genomes using kSNP (v3.0.21)[381] as described above and visualized and annotated with iTOL (v4)[371].

Both accessory and core genome feature sets were defined as described above. For the accessory genome feature sets, AGEs (unique groups of subelements ≥200 bp) were considered. For the core genome feature set, variant SNV positions based on 95% core genome read alignment to PAO1 were considered.

Machine learning analysis was performed using the random forest algorithm and generalization performance estimated using nested cross-validation as described above with the follow modification. In order to account for class imbalance (between eradicted and persistant isoaltes) in the dataset an additional hyperparameter , “class_weight”, was added with the options “balanced” (scale the cost of misclassifying an isolate during model training by the prevalence of its class), “balanced_subsample” (as with balanced but independently for each decision tree in the random forest), and “None” (the default, no weighting by class prevalence). In addition, as accuracy can be a poor or misleading performance metric in unbalanced datasets, hyperparameter selection during grid-search cross-validation was performed using F1 score as the scoring metric. This was compared to results from using accuracy as the scoring metric during grid-search cross-validation. For the accessory genome feature set, learning curves were constructed for both the F1-tuned and accuracy-tuned approaches, with the change in the relevant scoring metric as the training sample size increases plotted.

REFERENCES

1 Palleroni, N. J. in *Bergey's Manual of Systematics of Archaea and Bacteria* (eds W. B. Whitman *et al.*) (John Wiley & Sons, Inc., in association with Bergey's Manual Trust, 2015).

2 Gellatly, S. L. & Hancock, R. E. W. *Pseudomonas aeruginosa* : new insights into pathogenesis and host defenses. *Pathog. Dis.* **67**, 159-173, doi:10.1111/2049-632X.12033 (2013).

3 Rafael Araos, E. D. A. in *Mandell, Douglas, and Bennett's Principles and Practice of Infectious Diseases* (ed Raphael Dolin John E. Bennett, Martin J. Blaser) Ch. 219, 2686-2699.e2683 (Elsevier, Inc., 2020).

4 Sadikot, R. T., Blackwell, T. S., Christman, J. W. & Prince, A. S. Pathogen–host interactions in *Pseudomonas aeruginosa* pneumonia. *Am. J. Respir. Crit. Care Med.* **171**, 1209-1223, doi:10.1164/rccm.200408-1044SO (2005).

5 Weiner, L. M., Webb, A. K., Limbago, B., Dudeck, M. A., Patel, J., Kallen, A. J., Edwards, J. R. & Sievert, D. M. Antimicrobial-resistant pathogens associated with healthcare-associated infections: Summary of data reported to the National Healthcare Safety Network at the Centers for Disease Control and Prevention, 2011–2014. *Infect. Control Hosp. Epidemiol.* **37**, 1288-1301, doi:10.1017/ice.2016.174 (2016).

6 Peña, C., Suarez, C., Gozalo, M., Murillas, J., Almirante, B., Pomar, V., Aguilar, M., Granados, A., Calbo, E., Rodríguez-Baño, J., Rodríguez, F., Tubau, F., Martínez-Martínez, L., Oliver, A. & Diseases, f. t. S. N. f. R. i. I. Prospective multicenter study of the impact of carbapenem resistance on mortality in *Pseudomonas aeruginosa* bloodstream infections. *Antimicrob. Agents Chemother.* **56**, 1265-1272, doi:10.1128/aac.05991-11 (2012).

7 Thaden, J. T., Park, L. P., Maskarinec, S. A., Ruffin, F., Fowler, V. G. & van Duin, D. Increased mortality associated with bloodstream infections caused by *Pseudomonas aeruginosa* as compared to other bacteria: results of a 13-year prospective cohort study. *Antimicrob. Agents Chemother.*, doi:10.1128/aac.02671-16 (2017).

8 Micek, S. T., Wunderink, R. G., Kollef, M. H., Chen, C., Rello, J., Chastre, J., Antonelli, M., Welte, T., Clair, B., Ostermann, H., Calbo, E., Torres, A., Menichetti, F., Schramm, G. E. & Menon, V. An international multicenter retrospective study of *Pseudomonas aeruginosa* nosocomial pneumonia: impact of multidrug resistance. *Crit. Care* **19**, 219, doi:10.1186/s13054-015-0926-5 (2015).

9 Planquette, B., Timsit, J.-F., Misset, B. Y., Schwebel, C., Azoulay, E., Adrie, C., Vesin, A., Jamali, S., Zahar, J.-R., Allaouchiche, B., Souweine, B., Darmon, M., Dumenil, A.-S., Goldgran-Toledano, D., Mourvillier, B. H. & Bédos, J.-P. *Pseudomonas aeruginosa* ventilator-associated pneumonia. Predictive factors of treatment failure. *Am. J. Respir. Crit. Care Med.* **188**, 69-76, doi:10.1164/rccm.201210-1897OC (2013).

10 Martínez-Solano, L., Macia, M. D., Fajardo, A., Oliver, A. & Martinez, J. L. Chronic *Pseudomonas aeruginosa* infection in chronic obstructive pulmonary disease. *Clin. Infect. Dis.* **47**, 1526-1533, doi:10.1086/593186 (2008).

11 Markussen, T., Marvig, R. L., Gómez-Lozano, M., Aanæs, K., Burleigh, A. E., Høiby, N., Johansen, H. K., Molin, S. & Jelsbak, L. Environmental heterogeneity drives within-

host diversification and evolution of *Pseudomonas aeruginosa*. *mBio* **5**, doi:10.1128/mBio.01592-14 (2014).

12 Marvig, R. L., Johansen, H. K., Molin, S. & Jelsbak, L. Genome analysis of a transmissible lineage of *Pseudomonas aeruginosa* reveals pathoadaptive mutations and distinct evolutionary paths of hypermutators. *PLoS Genet.* **9**, e1003741, doi:10.1371/journal.pgen.1003741 (2013).

13 Pamukcu, A., Bush, A. & Buchdahl, R. Effects of *Pseudomonas aeruginosa* colonization on lung function and anthropometric variables in children with cystic fibrosis. *Pediatr. Pulmonol.* **19**, 10-15, doi:10.1002/ppul.1950190103 (1995).

14 Nixon, G. M., Armstrong, D. S., Carzino, R., Carlin, J. B., Olinsky, A., Robertson, C. F. & Grimwood, K. Clinical outcome after early *Pseudomonas aeruginosa* infection in cystic fibrosis. *J. Pediatr.* **138**, 699-704, doi:https://doi.org/10.1067/mpd.2001.112897 (2001).

15 Pellett, S., Bigley, D. V. & Grimes, D. J. Distribution of *Pseudomonas aeruginosa* in a riverine ecosystem. *Appl. Environ. Microbiol.* **45**, 328-332 (1983).

16 Rutherford, V., Yom, K., Ozer, E. A., Pura, O., Hughes, A., Murphy, K. R., Cudzilo, L., Mitchell, D. & Hauser, A. R. Environmental reservoirs for *exoS*+ and *exoU*+ strains of *Pseudomonas aeruginosa*. *Environ. Microbiol. Rep.* **10**, 485-492, doi:10.1111/1758-2229.12653 (2018).

17 Schroth, M. N., Cho, J. J., Green, S. K., Kominos, S. D. & Publishing, M. S. Epidemiology of *Pseudomonas aeruginosa* in agricultural areas. *J. Med. Microbiol.* **67**, 1191-1201, doi:https://doi.org/10.1099/jmm.0.000758 (2018).

18 Green, S. K., Schroth, M. N., Cho, J. J., Kominos, S. D. & Vitanza-Jack, V. B. Agricultural plants and soil as a reservoir for *Pseudomonas aeruginosa*. *Appl. Microbiol.* **28**, 987 (1974).

19 Hong, J. H., Kim, J., Choi, O. K., Cho, K.-S. & Ryu, H. W. Characterization of a diesel-degrading bacterium, *Pseudomonas aeruginosa* IU5, isolated from oil-contaminated soil in Korea. *World J. Microbiol. Biotechnol.* **21**, 381-384, doi:10.1007/s11274-004-3630-1 (2005).

20 Saikia, R. R., Deka, S., Deka, M. & Banat, I. M. Isolation of biosurfactant-producing *Pseudomonas aeruginosa* RS29 from oil-contaminated soil and evaluation of different nitrogen sources in biosurfactant production. *Ann. Microbiol.* **62**, 753-763, doi:10.1007/s13213-011-0315-5 (2012).

21 Crivaro, V., Di Popolo, A., Caprio, A., Lambiase, A., Di Resta, M., Borriello, T., Scarcella, A., Triassi, M. & Zarrilli, R. *Pseudomonas aeruginosa* in a neonatal intensive care unit: molecular epidemiology and infection control measures. *BMC Infect. Dis.* **9**, 70, doi:10.1186/1471-2334-9-70 (2009).

22 Reuter, S., Sigge, A., Wiedeck, H. & Trautmann, M. Analysis of transmission pathways of *Pseudomonas aeruginosa* between patients and tap water outlets. *Crit. Care Med.* **30** (2002).

23 Walker, J. T., Jhutty, A., Parks, S., Willis, C., Copley, V., Turton, J. F., Hoffman, P. N. & Bennett, A. M. Investigation of healthcare-acquired infections associated with *Pseudomonas aeruginosa* biofilms in taps in neonatal units in Northern Ireland. *J. Hosp. Infect.* **86**, 16-23, doi:https://doi.org/10.1016/j.jhin.2013.10.003 (2014).

24 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. *J. Bacteriol.* **190**, 7130-7140, doi:10.1128/jb.00785-08 (2008).

25 Shaver, C. M. & Hauser, A. R. Relative contributions of *Pseudomonas aeruginosa* ExoU, ExoS, and ExoT to virulence in the lung. *Infect. Immun.* **72**, 6969-6977, doi:10.1128/iai.72.12.6969-6977.2004 (2004).

26 Growcott, E. J., Coulthard, A., Amison, R., Hardaker, E. L., Saxena, V., Malt, L., Jones, P., Grevot, A., Poll, C., Osborne, C. & Banner, K. H. Characterisation of a refined rat model of respiratory infection with *Pseudomonas aeruginosa* and the effect of ciprofloxacin. *J. Cyst. Fibros.* **10**, 166-174, doi:https://doi.org/10.1016/j.jcf.2010.12.007 (2011).

27 Lorè, N. I., Cigana, C., De Fino, I., Riva, C., Juhas, M., Schwager, S., Eberl, L. & Bragonzi, A. Cystic fibrosis-niche adaptation of *Pseudomonas aeruginosa* reduces virulence in multiple infection hosts. *PLoS One* **7**, e35648, doi:10.1371/journal.pone.0035648 (2012).

28 Kim, S.-H., Park, S.-Y., Heo, Y.-J. & Cho, Y.-H. *Drosophila melanogaster*-based screening for multihost virulence factors of *Pseudomonas aeruginosa* PA14 and identification of a virulence-attenuating factor, HudA. *Infect. Immun.* **76**, 4152-4162, doi:10.1128/iai.01637-07 (2008).

29 Dubern, J.-F., Cigana, C., De Simone, M., Lazenby, J., Juhas, M., Schwager, S., Bianconi, I., Döring, G., Eberl, L., Williams, P., Bragonzi, A. & Cámara, M. Integrated whole-genome screening for *Pseudomonas aeruginosa* virulence genes using multiple disease models reveals that pathogenicity is host specific. *Environ. Microbiol.* **17**, 4379-4393, doi:10.1111/1462-2920.12863 (2015).

30 Hilker, R., Munder, A., Klockgether, J., Losada, P. M., Chouvarine, P., Cramer, N., Davenport, C. F., Dethlefsen, S., Fischer, S., Peng, H., Schonfelder, T., Turk, O., Wiehlmann, L., Wolbeling, F., Gulbins, E., Goesmann, A. & Tummler, B. Interclonal gradient of virulence in the *Pseudomonas aeruginosa* pangenome from disease and environment. *Environ. Microbiol.* **17**, 29-46, doi:10.1111/1462-2920.12606 (2015).

31 Vasquez-Rifo, A., Veksler-Lublinsky, I., Cheng, Z., Ausubel, F. M. & Ambros, V. The *Pseudomonas aeruginosa* accessory genome elements influence virulence towards *Caenorhabditis elegans*. *Genome Biol.* **20**, 270, doi:10.1186/s13059-019-1890-1 (2019).

32 Sánchez-Diener, I., Zamorano, L., Peña, C., Ocampo-Sosa, A., Cabot, G., Gómez-Zorrilla, S., Almirante, B., Aguilar, M., Granados, A., Calbo, E., Rodriguez-Baño, J., Rodríguez-López, F., Tubau, F., Martínez-Martínez, L., Navas, A. & Oliver, A. Weighting the impact of virulence on the outcome of *Pseudomonas aeruginosa* bloodstream infections. *Clin. Microbiol. Infect.* **26**, 351-357, doi:10.1016/j.cmi.2019.06.034 (2020).

33 Alibaud, L., Köhler, T., Coudray, A., Prigent-Combaret, C., Bergeret, E., Perrin, J., Benghezal, M., Reimmann, C., Gauthier, Y., Van Delden, C., Attree, I., Fauvarque, M.-O. & Cosson, P. *Pseudomonas aeruginosa* virulence genes identified in a *Dictyostelium* host model. *Cell. Microbiol.* **10**, 729-740, doi:10.1111/j.1462-5822.2007.01080.x (2008).

34 Chellaiah, E. R. Cadmium (heavy metals) bioremediation by *Pseudomonas aeruginosa*: a minireview. *Appl. Water Sci.* **8**, 154, doi:10.1007/s13201-018-0796-5 (2018).

35 Ganguli, A. & Tripathi, A. Bioremediation of toxic chromium from electroplating effluent by chromate-reducing *Pseudomonas aeruginosa* A2Chr in two bioreactors. *Appl. Microbiol. Biotechnol.* **58**, 416-420, doi:10.1007/s00253-001-0871-x (2002).

36 Sun, S., Wang, Y., Zang, T., Wei, J., Wu, H., Wei, C., Qiu, G. & Li, F. A biosurfactant-producing *Pseudomonas aeruginosa* S5 isolated from coking wastewater and its application for bioremediation of polycyclic aromatic hydrocarbons. *Bioresour. Technol.* **281**, 421-428, doi:10.1016/j.biortech.2019.02.087 (2019).

37 Maier, R. M. & Soberón-Chávez, G. *Pseudomonas aeruginosa* rhamnolipids: biosynthesis and potential applications. *Appl. Microbiol. Biotechnol.* **54**, 625-633, doi:10.1007/s002530000443 (2000).

38 Hauser, A. R. The type III secretion system of *Pseudomonas aeruginosa*: infection by injection. *Nat. Rev. Microbiol.* **7**, 654-665, doi:10.1038/nrmicro2199 (2009).

39 Matz, C., Moreno, A. M., Alhede, M., Manefield, M., Hauser, A. R., Givskov, M. & Kjelleberg, S. *Pseudomonas aeruginosa* uses type III secretion system to kill biofilm-associated amoebae. *ISME J.* **2**, 843-852, doi:10.1038/ismej.2008.47 (2008).

40 Segata, N., Haake, S. K., Mannon, P., Lemon, K. P., Waldron, L., Gevers, D., Huttenhower, C. & Izard, J. Composition of the adult digestive tract bacterial microbiome based on seven mouth surfaces, tonsils, throat and stool samples. *Genome Biol.* **13**, R42, doi:10.1186/gb-2012-13-6-r42 (2012).

41 Blanc, D. S., Francioli, P. & Zanetti, G. Molecular epidemiology of *Pseudomonas aeruginosa* in the intensive care units - a review. *Open Microbiol. J.* **1**, 8-11, doi:10.2174/1874285800701010008 (2007).

42 Cohen, R., Babushkin, F., Cohen, S., Afraimov, M., Shapiro, M., Uda, M., Khabra, E., Adler, A., Ben Ami, R. & Paikin, S. A prospective survey *of Pseudomonas aeruginosa* colonization and infection in the intensive care unit. *Antimicrob Resist Infect Control* **6**, 7-7, doi:10.1186/s13756-016-0167-7 (2017).

43 Kusters, J. G., van Vliet, A. H. M. & Kuipers, E. J. Pathogenesis of *Helicobacter pylori* infection. *Clin. Microbiol. Rev.* **19**, 449-490, doi:10.1128/CMR.00054-05 (2006).

44 Quillin, S. J. & Seifert, H. S. *Neisseria gonorrhoeae* host adaptation and pathogenesis. *Nat. Rev. Microbiol.* **16**, 226-240, doi:10.1038/nrmicro.2017.169 (2018).

45 Gagneux, S. Ecology and evolution of *Mycobacterium tuberculosis*. *Nat. Rev. Microbiol.* **16**, 202-213, doi:10.1038/nrmicro.2018.8 (2018).

46 Stover, C. K., Pham, X. Q., Erwin, A. L., Mizoguchi, S. D., Warrener, P., Hickey, M. J., Brinkman, F. S. L., Hufnagle, W. O., Kowalik, D. J., Lagrou, M., Garber, R. L., Goltry, L., Tolentino, E., Westbrock-Wadman, S., Yuan, Y., Brody, L. L., Coulter, S. N., Folger, K. R., Kas, A., Larbig, K., Lim, R., Smith, K., Spencer, D., Wong, G. K. S., Wu, Z., Paulsen, I. T., Reizer, J., Saier, M. H., Hancock, R. E. W., Lory, S. & Olson, M. V. Complete genome sequence of *Pseudomonas aeruginosa* PAO1, an opportunistic pathogen. *Nature* **406**, 959-964, doi:http://www.nature.com/nature/journal/v406/n6799/suppinfo/406959a0_S1.html (2000).

47 Kung, V. L., Ozer, E. A. & Hauser, A. R. The accessory genome of *Pseudomonas aeruginosa*. *Microbiol. Mol. Biol. Rev.* **74**, 621-641, doi:10.1128/MMBR.00027-10 (2010).

48 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).

49 Silby, M. W., Winstanley, C., Godfrey, S. A. C., Levy, S. B. & Jackson, R. W. *Pseudomonas* genomes: diverse and adaptable. *FEMS Microbiol. Rev.* **35**, 652-680, doi:10.1111/j.1574-6976.2011.00269.x (2011).

50 Winsor, G. L., Griffiths, E. J., Lo, R., Dhillon, B. K., Shay, J. A. & Brinkman, F. S. L. Enhanced annotations and features for comparing thousands of *Pseudomonas* genomes in the *Pseudomonas* genome database. *Nucleic Acids Res.* **44**, D646-D653, doi:10.1093/nar/gkv1227 (2016).

51 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).

52 Freschi, L., Vincent, A. T., Jeukens, J., Emond-Rheault, J.-G., Kukavica-Ibrulj, I., Dupont, M.-J., Charette, S. J., Boyle, B. & Levesque, R. C. 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).

53 Valot, B., Guyeux, C., Rolland, J. Y., Mazouzi, K., Bertrand, X. & Hocquet, D. What It Takes to Be a Pseudomonas aeruginosa? The Core Genome of the Opportunistic Pathogen Updated. *PLoS One* **10**, e0126468, doi:10.1371/journal.pone.0126468 (2015).

54 Mosquera-Rendón, J., Rada-Bravo, A. M., Cárdenas-Brito, S., Corredor, M., Restrepo-Pineda, E. & Benítez-Páez, A. Pangenome-wide and molecular evolution analyses of the *Pseudomonas aeruginosa* species. *BMC Genomics* **17**, 45, doi:10.1186/s12864-016-2364-4 (2016).

55 Skurnik, D., Roux, D., Aschard, H., Cattoir, V., Yoder-Himes, D., Lory, S. & Pier, G. B. A comprehensive analysis of in vitro and in vivo genetic fitness of *Pseudomonas aeruginosa* using high-throughput sequencing of transposon libraries. *PLoS Pathog.* **9**, e1003582, doi:10.1371/journal.ppat.1003582 (2013).

56 Turner, K. H., Wessel, A. K., Palmer, G. C., Murray, J. L. & Whiteley, M. Essential genome of *Pseudomonas aeruginosa* in cystic fibrosis sputum. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 4110-4115, doi:10.1073/pnas.1419677112 (2015).

57 Turner, K. H., Everett, J., Trivedi, U., Rumbaugh, K. P. & Whiteley, M. Requirements for *Pseudomonas aeruginosa* acute burn and chronic surgical wound infection. *PLoS Genet.* **10**, e1004518, doi:10.1371/journal.pgen.1004518 (2014).

58 Poulsen, B. E., Yang, R., Clatworthy, A. E., White, T., Osmulski, S. J., Li, L., Penaranda, C., Lander, E. S., Shoresh, N. & Hung, D. T. Defining the core essential genome of *Pseudomonas aeruginosa*. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 10072, doi:10.1073/pnas.1900570116 (2019).

59 Juhas, M. *Pseudomonas aeruginosa* essentials: an update on investigation of essential genes. *Microbiology* **161**, 2053-2060, doi:https://doi.org/10.1099/mic.0.000161 (2015).

60 Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., Angiuoli, S. V., Crabtree, J., Jones, A. L., Durkin, A. S., DeBoy, R. T., Davidsen, T. M., Mora, M., Scarselli, M., Margarit y Ros, I., Peterson, J. D., Hauser, C. R., Sundaram, J. P., Nelson, W. C., Madupu, R., Brinkac, L. M., Dodson, R. J., Rosovitz, M. J., Sullivan,

S. A., Daugherty, S. C., Haft, D. H., Selengut, J., Gwinn, M. L., Zhou, L., Zafar, N., Khouri, H., Radune, D., Dimitrov, G., Watkins, K., O'Connor, K. J. B., Smith, S., Utterback, T. R., White, O., Rubens, C. E., Grandi, G., Madoff, L. C., Kasper, D. L., Telford, J. L., Wessels, M. R., Rappuoli, R. & Fraser, C. M. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial "pan-genome". *Proc. Natl. Acad. Sci. U. S. A.* **102**, 13950-13955, doi:10.1073/pnas.0506758102 (2005).

61 Tettelin, H., Riley, D., Cattuto, C. & Medini, D. Comparative genomics: the bacterial pan-genome. *Curr. Opin. Microbiol.* **11**, 472-477, doi:10.1016/j.mib.2008.09.006 (2008).

62 Medini, D., Donati, C., Tettelin, H., Masignani, V. & Rappuoli, R. The microbial pan-genome. *Curr. Opin. Genet. Dev.* **15**, 589-594, doi:http://dx.doi.org/10.1016/j.gde.2005.09.006 (2005).

63 Brockhurst, M. A., Harrison, E., Hall, J. P. J., Richards, T., McNally, A. & MacLean, C. The ecology and evolution of pangenomes. *Curr. Biol.* **29**, R1094-R1103, doi:https://doi.org/10.1016/j.cub.2019.08.012 (2019).

64 Rouli, L., Merhej, V., Fournier, P. E. & Raoult, D. The bacterial pangenome as a new tool for analysing pathogenic bacteria. *New Microbes New Infect.* **7**, 72-85, doi:10.1016/j.nmni.2015.06.005 (2015).

65 van Belkum, A., Soriaga, L. B., LaFave, M. C., Akella, S., Veyrieras, J.-B., Barbu, E. M., Shortridge, D., Blanc, B., Hannum, G., Zambardi, G., Miller, K., Enright, M. C., Mugnier, N., Brami, D., Schicklin, S., Felderman, M., Schwartz, A. S., Richardson, T. H., Peterson, T. C., Hubby, B. & Cady, K. C. Phylogenetic distribution of CRISPR-Cas systems in antibiotic-resistant *Pseudomonas aeruginosa*. *mBio* **6** (2015).

66 Khaledi, A., Weimann, A., Schniederjans, M., Asgari, E., Kuo, T.-H., Oliver, A., Cabot, G., Kola, A., Gastmeier, P., Hogardt, M., Jonas, D., Mofrad, M. R. K., Bremges, A., McHardy, A. C. & Häussler, S. Predicting antimicrobial resistance in *Pseudomonas aeruginosa* with machine learning-enabled molecular diagnostics. *EMBO Mol. Med.* **12**, e10264, doi:10.15252/emmm.201910264 (2020).

67 Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150-3152, doi:10.1093/bioinformatics/bts565 (2012).

68 Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T. G., Fookes, M., Falush, D., Keane, J. A. & Parkhill, J. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691-3693, doi:10.1093/bioinformatics/btv421 (2015).

69 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).

70 Allen, J. P., Ozer, E. A., Minasov, G., Shuvalova, L., Kiryukhina, O., Satchell, K. J. F. & Hauser, A. R. A comparative genomics approach identifies contact-dependent growth inhibition as a virulence determinant. *Proc. Natl. Acad. Sci. U. S. A.*, 201919198, doi:10.1073/pnas.1919198117 (2020).

71 Mathee, K., Narasimhan, G., Valdes, C., Qiu, X., Matewish, J. M., Koehrsen, M., Rokas, A., Yandava, C. N., Engels, R., Zeng, E., Olavarietta, R., Doud, M., Smith, R. S., Montgomery, P., White, J. R., Godfrey, P. A., Kodira, C., Birren, B., Galagan, J. E. &

Lory, S. Dynamics of *Pseudomonas aeruginosa* genome evolution. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 3100-3105, doi:10.1073/pnas.0711982105 (2008).
72 Tatusov, R. L., Koonin, E. V. & Lipman, D. J. A genomic perspective on protein families. *Science* **278**, 631, doi:10.1126/science.278.5338.631 (1997).
73 Kos, V. N., Déraspe, M., McLaughlin, R. E., Whiteaker, J. D., Roy, P. H., Alm, R. A., Corbeil, J. & Gardner, H. The resistome of *Pseudomonas aeruginosa* in relationship to phenotypic susceptibility. *Antimicrob. Agents Chemother.* **59**, 427-436, doi:10.1128/AAC.03954-14 (2015).
74 van der Zee, A., Kraak, W. B., Burggraaf, A., Goessens, W. H. F., Pirovano, W., Ossewaarde, J. M. & Tommassen, J. Spread of carbapenem resistance by transposition and conjugation among *Pseudomonas aeruginosa*. *Front. Microbiol.* **9**, doi:10.3389/fmicb.2018.02057 (2018).
75 He, J., Baldini, R. L., Déziel, E., Saucier, M., Zhang, Q., Liberati, N. T., Lee, D., Urbach, J., Goodman, H. M. & Rahme, L. G. The broad host range pathogen *Pseudomonas aeruginosa* strain PA14 carries two pathogenicity islands harboring plant and animal virulence genes. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 2530-2535, doi:10.1073/pnas.0304622101 (2004).
76 Harrison, E. M., Carter, M. E. K., Luck, S., Ou, H.-Y., He, X., Deng, Z., O'Callaghan, C., Kadioglu, A. & Rajakumar, K. Pathogenicity islands PAPI-1 and PAPI-2 contribute individually and synergistically to the virulence of *Pseudomonas aeruginosa* strain PA14. *Infect. Immun.* **78**, 1437-1446, doi:10.1128/IAI.00621-09 (2010).
77 Kulasekara, B. R., Kulasekara, H. D., Wolfgang, M. C., Stevens, L., Frank, D. W. & Lory, S. Acquisition and evolution of the *exoU* locus in *Pseudomonas aeruginosa*. *J. Bacteriol.* **188**, 4037-4050, doi:10.1128/JB.02000-05 (2006).
78 Reboud, E., Elsen, S., Bouillot, S., Golovkine, G., Basso, P., Jeannot, K., Attrée, I. & Huber, P. Phenotype and toxicity of the recently discovered *exlA*-positive *Pseudomonas aeruginosa* strains collected worldwide. *Environ. Microbiol.* **18**, 3425-3439, doi:10.1111/1462-2920.13262 (2016).
79 Juhas, M., van der Meer, J. R., Gaillard, M., Harding, R. M., Hood, D. W. & Crook, D. W. Genomic islands: tools of bacterial horizontal gene transfer and evolution. *FEMS Microbiol. Rev.* **33**, 376-393, doi:10.1111/j.1574-6976.2008.00136.x (2009).
80 Dobrindt, U., Hochhut, B., Hentschel, U. & Hacker, J. Genomic islands in pathogenic and environmental microorganisms. *Nat. Rev. Microbiol.* **2**, 414-424, doi:10.1038/nrmicro884 (2004).
81 Klockgether, J., Reva, O., Larbig, K. & Tümmler, B. Sequence analysis of the mobile genome island pKLC102 of *Pseudomonas aeruginosa* C. *J. Bacteriol.* **186**, 518, doi:10.1128/JB.186.2.518-534.2004 (2004).
82 Winstanley, C., Langille, M. G. I., Fothergill, J. L., Kukavica-Ibrulj, I., Paradis-Bleau, C., Sanschagrin, F., Thomson, N. R., Winsor, G. L., Quail, M. A., Lennard, N., Bignell, A., Clarke, L., Seeger, K., Saunders, D., Harris, D., Parkhill, J., Hancock, R. E. W., Brinkman, F. S. L. & Levesque, R. C. Newly introduced genomic prophage islands are critical determinants of in vivo competitiveness in the Liverpool Epidemic Strain of *Pseudomonas aeruginosa*. *Genome Res.* **19**, 12-23, doi:10.1101/gr.086082.108 (2009).
83 Nakayama, K., Takashima, K., Ishihara, H., Shinomiya, T., Kageyama, M., Kanaya, S., Ohnishi, M., Murata, T., Mori, H. & Hayashi, T. The R-type pyocin of *Pseudomonas*

aeruginosa is related to P2 phage, and the F-type is related to lambda phage. *Mol. Microbiol.* **38**, 213-231, doi:10.1046/j.1365-2958.2000.02135.x (2000).

84 Pohl, S., Klockgether, J., Eckweiler, D., Khaledi, A., Schniederjans, M., Chouvarine, P., Tümmler, B. & Häussler, S. The extensive set of accessory *Pseudomonas aeruginosa* genomic components. *FEMS Microbiol. Lett.* **356**, 235-241, doi:10.1111/1574-6968.12445 (2014).

85 Johnson, C. M. & Grossman, A. D. Integrative and conjugative elements (ICEs): what they do and how they work. *Annu. Rev. Genet.* **49**, 577-601, doi:10.1146/annurev-genet-112414-055018 (2015).

86 Qiu, X., Gurkar, A. U. & Lory, S. Interstrain transfer of the large pathogenicity island (PAPI-1) of *Pseudomonas aeruginosa*. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 19830-19835, doi:10.1073/pnas.0606810104 (2006).

87 Cury, J., Touchon, M. & Rocha, E. P. C. Integrative and conjugative elements and their hosts: composition, distribution and organization. *Nucleic Acids Res.* **45**, 8943-8956, doi:10.1093/nar/gkx607 (2017).

88 Botelho, J., Grosso, F. & Peixe, L. Unravelling the genome of a *Pseudomonas aeruginosa* isolate belonging to the high-risk clone ST235 reveals an integrative conjugative element housing a $bla_{\text{GES-6}}$ carbapenemase. *J. Antimicrob. Chemother.* **73**, 77-83, doi:10.1093/jac/dkx337 (2017).

89 Ding, Y., Teo, J. W. P., Drautz-Moses, D. I., Schuster, S. C., Givskov, M. & Yang, L. Acquisition of resistance to carbapenem and macrolide-mediated quorum sensing inhibition by *Pseudomonas aeruginosa* via $\text{ICE}_{\text{Tn4371}}$6385. *Commun. Biol.* **1**, 57-57, doi:10.1038/s42003-018-0064-0 (2018).

90 Budzik, J. M., Rosche, W. A., Rietsch, A. & O'Toole, G. A. Isolation and characterization of a generalized transducing phage for *Pseudomonas aeruginosa* strains PAO1 and PA14. *J. Bacteriol.* **186**, 3270-3273, doi:10.1128/jb.186.10.3270-3273.2004 (2004).

91 Cavenagh, M. M. & Miller, R. V. Specialized transduction of *Pseudomonas aeruginosa* PAO by bacteriophage D3. *J. Bacteriol.* **165**, 448, doi:10.1128/jb.165.2.448-452.1986 (1986).

92 Rice, S. A., Tan, C. H., Mikkelsen, P. J., Kung, V., Woo, J., Tay, M., Hauser, A., McDougald, D., Webb, J. S. & Kjelleberg, S. The biofilm life cycle and virulence of *Pseudomonas aeruginosa* are dependent on a filamentous prophage. *ISME J* **3**, 271-282, doi:10.1038/ismej.2008.109 (2009).

93 Hayashi, T., Baba, T., Matsumoto, H. & Terawaki, Y. Phage-conversion of cytotoxin production in *Pseudomonas aeruginosa*. *Mol. Microbiol.* **4**, 1703-1709, doi:10.1111/j.1365-2958.1990.tb00547.x (1990).

94 Partridge, S. R., Kwong, S. M., Firth, N. & Jensen, S. O. Mobile genetic elements associated with antimicrobial resistance. *Clin. Microbiol. Rev.* **31** (2018).

95 Xiong, J., Alexander, D. C., Ma, J. H., Déraspe, M., Low, D. E., Jamieson, F. B. & Roy, P. H. Complete sequence of pOZ176, a 500-kilobase IncP-2 plasmid encoding IMP-9-mediated carbapenem resistance, from outbreak isolate *Pseudomonas aeruginosa* 96. *Antimicrob. Agents Chemother.* **57**, 3775-3782, doi:10.1128/aac.00423-13 (2013).

96 Naas, T., Bonnin, R. A., Cuzon, G., Villegas, M.-V. & Nordmann, P. Complete sequence of two KPC-harbouring plasmids from *Pseudomonas aeruginosa*. *J. Antimicrob. Chemother.* **68**, 1757-1762, doi:10.1093/jac/dkt094 (2013).

97 Abril, D., Marquez-Ortiz, R. A., Castro-Cardozo, B., Moncayo-Ortiz, J. I., Olarte Escobar, N. M., Corredor Rozo, Z. L., Reyes, N., Tovar, C., Sánchez, H. F., Castellanos, J., Guaca-González, Y. M., Llanos-Uribe, C. E., Vanegas Gómez, N. & Escobar-Pérez, J. Genome plasticity favours double chromosomal Tn4401b-bla_{KPC-2} transposon insertion in the *Pseudomonas aeruginosa* ST235 clone. *BMC Microbiol.* **19**, 45, doi:10.1186/s12866-019-1418-6 (2019).

98 Partridge, S. R., Tsafnat, G., Coiera, E. & Iredell, J. R. Gene cassettes and cassette arrays in mobile resistance integrons. *FEMS Microbiol. Rev.* **33**, 757-784, doi:10.1111/j.1574-6976.2009.00175.x (2009).

99 Mazel, D. Integrons: agents of bacterial evolution. *Nat. Rev. Microbiol.* **4**, 608, doi:10.1038/nrmicro1462 (2006).

100 Bonnin, R. A., Bogaerts, P., Girlich, D., Huang, T.-D., Dortet, L., Glupczynski, Y. & Naas, T. Molecular characterization of OXA-198 carbapenemase-producing *Pseudomonas aeruginosa* clinical isolates. *Antimicrob. Agents Chemother.* **62**, doi:10.1128/aac.02496-17 (2018).

101 Liapis, E., Bour, M., Triponney, P., Jové, T., Zahar, J.-R., Valot, B., Jeannot, K. & Plésiat, P. Identification of diverse integron and plasmid structures carrying a novel carbapenemase among *Pseudomonas* species. *Front. Microbiol.* **10**, doi:10.3389/fmicb.2019.00404 (2019).

102 Xu, Z., Li, L., Shirtliff, M. E., Alam, M. J., Yamasaki, S. & Shi, L. Occurrence and characteristics of class 1 and 2 integrons in *Pseudomonas aeruginosa* isolates from patients in southern China. *J. Clin. Microbiol.* **47**, 230, doi:10.1128/JCM.02027-08 (2009).

103 Yamamoto, M., Matsumura, Y., Gomi, R., Matsuda, T., Nakano, S., Nagao, M., Tanaka, M. & Ichiyama, S. Molecular analysis of a bla_{IMP-1} harboring class 3 integron in multidrug-resistant *Pseudomonas fulva*. *Antimicrob. Agents Chemother.* **62**, e00701-00718, doi:10.1128/aac.00701-18 (2018).

104 Shibata, N., Doi, Y., Yamane, K., Yagi, T., Kurokawa, H., Shibayama, K., Kato, H., Kai, K. & Arakawa, Y. PCR typing of genetic determinants for metallo-β-lactamases and integrases carried by gram-negative bacteria isolated in Japan, with focus on the class 3 integron. *J. Clin. Microbiol.* **41**, 5407-5413, doi:10.1128/jcm.41.12.5407-5413.2003 (2003).

105 Mazel, D., Dychinco, B., Webb, V. A. & Davies, J. A distinctive class of integron in the *Vibrio cholerae* genome. *Science* **280**, 605, doi:10.1126/science.280.5363.605 (1998).

106 Rowe-Magnus, D. A., Guerout, A.-M., Ploncard, P., Dychinco, B., Davies, J. & Mazel, D. The evolutionary history of chromosomal super-integrons provides an ancestry for multiresistant integrons. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 652, doi:10.1073/pnas.98.2.652 (2001).

107 Chakrabarty, A. M. Plasmids in *Pseudomonas*. *Annu. Rev. Genet.* **10**, 7-30, doi:10.1146/annurev.ge.10.120176.000255 (1976).

108 Jacoby, G. A. Properties of R plasmids determining gentamicin resistance by acetylation in *Pseudomonas aeruginosa*. *Antimicrob. Agents Chemother.* **6**, 239, doi:10.1128/AAC.6.3.239 (1974).
109 Stanisich, V. A. Interaction between an R factor and an element conferring resistance to mercuric ions in *Pseudomonas aeruginosa*. *Mol. Gen. Genet.* **128**, 201-212, doi:10.1007/BF00267109 (1974).
110 Galetti, R., Andrade, L. N., Chandler, M., Varani, A. d. M. & Darini, A. L. C. New small plasmid harboring bla_{KPC-2} in *Pseudomonas aeruginosa*. *Antimicrob. Agents Chemother.* **60**, 3211-3214, doi:10.1128/AAC.00247-16 (2016).
111 Xiong, J., Hynes, M. F., Ye, H., Chen, H., Yang, Y., M'Zali, F. & Hawkey, P. M. bla_{IMP-9} and its association with large plasmids carried by *Pseudomonas aeruginosa* isolates from the People's Republic of China. *Antimicrob. Agents Chemother.* **50**, 355-358, doi:10.1128/aac.50.1.355-358.2006 (2006).
112 Poirel, L., Naas, T., Nicolas, D., Collet, L., Bellais, S., Cavallo, J.-D. & Nordmann, P. Characterization of VIM-2, a carbapenem-hydrolyzing metallo-β-lactamase and its plasmid- and integron-borne gene from a *Pseudomonas aeruginosa* clinical isolate in France. *Antimicrob. Agents Chemother.* **44**, 891, doi:10.1128/AAC.44.4.891-897.2000 (2000).
113 Li, Z., Cai, Z., Cai, Z., Zhang, Y., Fu, T., Jin, Y., Cheng, Z., Jin, S., Wu, W., Yang, L. & Bai, F. Molecular genetic analysis of an XDR *Pseudomonas aeruginosa* ST664 clone carrying multiple conjugal plasmids. *J. Antimicrob. Chemother.* **75**, 1443-1452, doi:10.1093/jac/dkaa063 (2020).
114 Orlek, A., Stoesser, N., Anjum, M. F., Doumith, M., Ellington, M. J., Peto, T., Crook, D., Woodford, N., Walker, A. S., Phan, H. & Sheppard, A. E. Plasmid classification in an era of whole-genome sequencing: application in studies of antibiotic resistance epidemiology. *Front. Microbiol.* **8**, doi:10.3389/fmicb.2017.00182 (2017).
115 Smith, E. E., Sims, E. H., Spencer, D. H., Kaul, R. & Olson, M. V. Evidence for diversifying selection at the pyoverdine locus of *Pseudomonas aeruginosa*. *J. Bacteriol.* **187**, 2138, doi:10.1128/JB.187.6.2138-2147.2005 (2005).
116 Raymond, C. K., Sims, E. H., Kas, A., Spencer, D. H., Kutyavin, T. V., Ivey, R. G., Zhou, Y., Kaul, R., Clendenning, J. B. & Olson, M. V. Genetic variation at the O-antigen biosynthetic locus in *Pseudomonas aeruginosa*. *J. Bacteriol.* **184**, 3614, doi:10.1128/JB.184.13.3614-3622.2002 (2002).
117 Kus, J. V., Tullis, E., Cvitkovitch, D. G. & Burrows, L. L. Significant differences in type IV pilin allele distribution among *Pseudomonas aeruginosa* isolates from cystic fibrosis (CF) versus non-CF patients. *Microbiology* **150**, 1315-1326, doi:https://doi.org/10.1099/mic.0.26822-0 (2004).
118 Allen, J. P. & Hauser, A. R. Diversity of contact-dependent growth inhibition systems of *Pseudomonas aeruginosa*. *J. Bacteriol.* **201**, e00776-00718, doi:10.1128/JB.00776-18 (2019).
119 Ernst, R. K., D′Argenio, D. A., Ichikawa, J. K., Bangera, M. G., Selgrade, S., Burns, J. L., Hiatt, P., McCoy, K., Brittnacher, M., Kas, A., Spencer, D. H., Olson, M. V., Ramsey, B. W., Lory, S. & Miller, S. I. Genome mosaicism is conserved but not unique in *Pseudomonas aeruginosa* isolates from the airways of young children with cystic

fibrosis. *Environ. Microbiol.* **5**, 1341-1349, doi:10.1111/j.1462-2920.2003.00518.x (2003).

120 Hocquet, D., Petitjean, M., Rohmer, L., Valot, B., Kulasekara, H. D., Bedel, E., Bertrand, X., Plésiat, P., Köhler, T., Pantel, A., Jacobs, M. A., Hoffman, L. R. & Miller, S. I. Pyomelanin-producing *Pseudomonas aeruginosa* selected during chronic infections have a large chromosomal deletion which confers resistance to pyocins. *Environ. Microbiol.* **18**, 3482-3493, doi:10.1111/1462-2920.13336 (2016).

121 Feltman, H., Schulert, G., Khan, S., Jain, M., Peterson, L. & Hauser, A. R. Prevalence of type III secretion genes in clinical and environmental isolates of *Pseudomonas aeruginosa*. *Microbiology* **147**, 2659-2669, doi:https://doi.org/10.1099/00221287-147-10-2659 (2001).

122 Garey, K. W., Vo, Q. P., Larocco, M. T., Gentry, L. O. & Tam, V. H. Prevalence of type III secretion protein exoenzymes and antimicrobial susceptibility patterns from bloodstream isolates of patients with *Pseudomonas aeruginosa* bacteremia. *J. Chemother.* **20**, 714-720, doi:10.1179/joc.2008.20.6.714 (2008).

123 Roy, P. H., Tetu, S. G., Larouche, A., Elbourne, L., Tremblay, S., Ren, Q., Dodson, R., Harkins, D., Shay, R., Watkins, K., Mahamoud, Y. & Paulsen, I. T. Complete genome sequence of the multiresistant taxonomic outlier *Pseudomonas aeruginosa* PA7. *PLoS One* **5**, e8842, doi:10.1371/journal.pone.0008842 (2010).

124 Huber, P., Basso, P., Reboud, E. & Attrée, I. *Pseudomonas aeruginosa* renews its virulence factors. *Environ. Microbiol. Rep.* **8**, 564-571, doi:10.1111/1758-2229.12443 (2016).

125 Morita, Y., Tomida, J. & Kawamura, Y. Primary mechanisms mediating aminoglycoside resistance in the multidrug-resistant *Pseudomonas aeruginosa* clinical isolate PA7. *Microbiology* **158**, 1071-1083, doi:https://doi.org/10.1099/mic.0.054320-0 (2012).

126 Kinene, T., Wainaina, J., Maina, S. & Boykin, L. M. in *Encyclopedia of Evolutionary Biology* (ed Richard M. Kliman) 489-493 (Academic Press, 2016).

127 Curran, B., Jonas, D., Grundmann, H., Pitt, T. & Dowson, C. G. Development of a multilocus sequence typing scheme for the opportunistic pathogen *Pseudomonas aeruginosa*. *J. Clin. Microbiol.* **42**, 5644, doi:10.1128/JCM.42.12.5644-5649.2004 (2004).

128 Pirnay, J.-P., Bilocq, F., Pot, B., Cornelis, P., Zizi, M., Van Eldere, J., Deschaght, P., Vaneechoutte, M., Jennes, S., Pitt, T. & De Vos, D. *Pseudomonas aeruginosa* population structure revisited. *PLoS One* **4**, e7740, doi:10.1371/journal.pone.0007740 (2009).

129 Oliver, A., Mulet, X., López-Causapé, C. & Juan, C. The increasing threat of *Pseudomonas aeruginosa* high-risk clones. *Drug Resist. Updat.* **21-22**, 41-59, doi:https://doi.org/10.1016/j.drup.2015.08.002 (2015).

130 Maatallah, M., Cheriaa, J., Backhrouf, A., Iversen, A., Grundmann, H., Do, T., Lanotte, P., Mastouri, M., Elghmati, M. S., Rojo, F., Mejdi, S. & Giske, C. G. Population structure of *Pseudomonas aeruginosa* from five Mediterranean countries: evidence for frequent recombination and epidemic occurrence of CC235. *PLoS One* **6**, e25617, doi:10.1371/journal.pone.0025617 (2011).

131 Woodford, N., Turton, J. F. & Livermore, D. M. Multiresistant Gram-negative bacteria: the role of high-risk clones in the dissemination of antibiotic resistance. *FEMS Microbiol. Rev.* **35**, 736-755, doi:10.1111/j.1574-6976.2011.00268.x (2011).

132 Treepong, P., Kos, V. N., Guyeux, C., Blanc, D. S., Bertrand, X., Valot, B. & Hocquet, D. Global emergence of the widespread *Pseudomonas aeruginosa* ST235 clone. *Clin. Microbiol. Infect.* **24**, 258-266, doi:https://doi.org/10.1016/j.cmi.2017.06.018 (2018).

133 Scott, F. W. & Pitt, T. L. Identification and characterization of transmissible *Pseudomonas aeruginosa* strains in cystic fibrosis patients in England and Wales. *J. Med. Microbiol.* **53**, 609-615, doi:https://doi.org/10.1099/jmm.0.45620-0 (2004).

134 Magiorakos, A. P., Srinivasan, A., Carey, R. B., Carmeli, Y., Falagas, M. E., Giske, C. G., Harbarth, S., Hindler, J. F., Kahlmeter, G., Olsson-Liljequist, B., Paterson, D. L., Rice, L. B., Stelling, J., Struelens, M. J., Vatopoulos, A., Weber, J. T. & Monnet, D. L. Multidrug-resistant, extensively drug-resistant and pandrug-resistant bacteria: an international expert proposal for interim standard definitions for acquired resistance. *Clin. Microbiol. Infect.* **18**, 268-281, doi:https://doi.org/10.1111/j.1469-0691.2011.03570.x (2012).

135 Xiong, J., Déraspe, M., Iqbal, N., Krajden, S., Chapman, W., Dewar, K. & Roy, P. H. Complete genome of a panresistant *Pseudomonas aeruginosa* strain, isolated from a patient with respiratory failure in a Canadian community hospital. *Genome Announc.* **5**, e00458-00417, doi:10.1128/genomeA.00458-17 (2017).

136 Molinaro, M., Morelli, P., De Gregori, M., De Gregori, S., Giardini, I., Tordato, F., Monzillo, V., Pocaterra, D. & Casari, E. Efficacy of intraventricular amikacin treatment in pan-resistant *Pseudomonas aeruginosa* postsurgical meningitis. *Infect. Drug Resist.* **11**, 1369-1372, doi:10.2147/idr.S169271 (2018).

137 Dobbin, C., Maley, M., Harkness, J., Benn, R., Malouf, M., Glanville, A. & Bye, P. The impact of pan-resistant bacterial pathogens on survival after lung transplantation in cystic fibrosis: results from a single large referral centre. *J. Hosp. Infect.* **56**, 277-282, doi:https://doi.org/10.1016/j.jhin.2004.01.003 (2004).

138 Winkler, M. L., Papp-Wallace, K. M., Hujer, A. M., Domitrovic, T. N., Hujer, K. M., Hurless, K. N., Tuohy, M., Hall, G. & Bonomo, R. A. Unexpected challenges in treating multidrug-resistant gram-negative bacteria: resistance to ceftazidime-avibactam in archived isolates of *Pseudomonas aeruginosa*. *Antimicrob. Agents Chemother.* **59**, 1020-1029, doi:10.1128/aac.04238-14 (2015).

139 Tam, V. H., Rogers, C. A., Chang, K.-T., Weston, J. S., Caeiro, J.-P. & Garey, K. W. Impact of multidrug-resistant *Pseudomonas aeruginosa* bacteremia on patient outcomes. *Antimicrob. Agents Chemother.* **54**, 3717-3722, doi:10.1128/aac.00207-10 (2010).

140 Peña, C., Cabot, G., Gómez-Zorrilla, S., Zamorano, L., Ocampo-Sosa, A., Murillas, J., Almirante, B., Pomar, V., Aguilar, M., Granados, A., Calbo, E., Rodríguez-Baño, J., Rodríguez-López, F., Tubau, F., Martínez-Martínez, L. & Oliver, A. Influence of virulence genotype and resistance profile in the mortality of *Pseudomonas aeruginosa* bloodstream infections. *Clin. Infect. Dis.* **60**, 539-548, doi:10.1093/cid/ciu866 (2015).

141 Morata, L., Cobos-Trigueros, N., Martínez, J. A., Soriano, Á., Almela, M., Marco, F., Sterzik, H., Núñez, R., Hernández, C. & Mensa, J. Influence of multidrug resistance and appropriate empirical therapy on the 30-day mortality rate of *Pseudomonas aeruginosa* bacteremia. *Antimicrob. Agents Chemother.* **56**, 4833-4837, doi:10.1128/aac.00750-12 (2012).

142 Talbot, G. H., Bradley, J., Edwards, J. E., Gilbert, D., Scheld, M. & Bartlett, J. G. Bad bugs need drugs: an update on the development pipeline from the Antimicrobial

Availability Task Force of the Infectious Diseases Society of America. *Clin. Infect. Dis.* **42**, 657-668, doi:10.1086/499819 (2006).
143 CDC. *Antibiotic resistance threats in the United States, 2019*. (U.S. Department of Health and Human Services, CDC, 2019).
144 Lister, P. D., Wolter, D. J. & Hanson, N. D. Antibacterial-resistant *Pseudomonas aeruginosa*: clinical impact and complex regulation of chromosomally encoded resistance mechanisms. *Clin. Microbiol. Rev.* **22**, 582, doi:10.1128/CMR.00040-09 (2009).
145 López-Causapé, C., Cabot, G., Del Barrio-Tofiño, E. & Oliver, A. The versatile mutational resistome of *Pseudomonas aeruginosa*. *Front. Microbiol.* **9**, 685-685, doi:10.3389/fmicb.2018.00685 (2018).
146 Li, X.-Z., Plésiat, P. & Nikaido, H. The challenge of efflux-mediated antibiotic resistance in Gram-negative bacteria. *Clin. Microbiol. Rev.* **28**, 337, doi:10.1128/CMR.00117-14 (2015).
147 Botelho, J., Grosso, F. & Peixe, L. Antibiotic resistance in *Pseudomonas aeruginosa* – mechanisms, epidemiology and evolution. *Drug Resist. Updat.* **44**, 100640, doi:https://doi.org/10.1016/j.drup.2019.07.002 (2019).
148 Sanders, C. C., Jr. & Sanders, W. E., Jr. Type I β-Lactamases of gram-negative bacteria: interactions with β-Lactam antibiotics. *J. Infect. Dis.* **154**, 792-800, doi:10.1093/infdis/154.5.792 (1986).
149 Bagge, N., Ciofu, O., Hentzer, M., Campbell, J. I. A., Givskov, M. & Høiby, N. Constitutive High Expression of Chromosomal β-Lactamase in *Pseudomonas aeruginosa* Caused by a New Insertion Sequence IS*1669* Located in *ampD*. *Antimicrob. Agents Chemother.* **46**, 3406, doi:10.1128/AAC.46.11.3406-3411.2002 (2002).
150 Girlich, D., Naas, T. & Nordmann, P. Biochemical characterization of the naturally occurring oxacillinase OXA-50 of *Pseudomonas aeruginosa*. *Antimicrob. Agents Chemother.* **48**, 2043-2048, doi:10.1128/AAC.48.6.2043-2048.2004 (2004).
151 White, P. A., Stokes, H. W., Bunny, K. L. & Hall, R. M. Characterisation of a chloramphenicol acetyltransferase determinant found in the chromosome of *Pseudomonas aeruginosa*. *FEMS Microbiol. Lett.* **175**, 27-35, doi:10.1111/j.1574-6968.1999.tb13598.x (1999).
152 Hächler, H., Santanam, P. & Kayser, F. H. Sequence and characterization of a novel chromosomal aminoglycoside phosphotransferase gene, aph (3')-IIb, in *Pseudomonas aeruginosa*. *Antimicrob. Agents Chemother.* **40**, 1254, doi:10.1128/AAC.40.5.1254 (1996).
153 Poole, K. Aminoglycoside resistance in *Pseudomonas aeruginosa*. *Antimicrob. Agents Chemother.* **49**, 479, doi:10.1128/AAC.49.2.479-487.2005 (2005).
154 Ito, R., Mustapha, M. M., Tomich, A. D., Callaghan, J. D., McElheny, C. L., Mettus, R. T., Shanks, R. M. Q., Sluis-Cremer, N. & Doi, Y. Widespread fosfomycin resistance in gram-negative bacteria attributable to the chromosomal *fosA* gene. *mBio* **8**, e00749-00717, doi:10.1128/mBio.00749-17 (2017).
155 Pang, Z., Raudonis, R., Glick, B. R., Lin, T.-J. & Cheng, Z. Antibiotic resistance in *Pseudomonas aeruginosa*: mechanisms and alternative therapeutic strategies. *Biotechnol. Adv.* **37**, 177-192, doi:https://doi.org/10.1016/j.biotechadv.2018.11.013 (2019).
156 Yoshimura, F. & Nikaido, H. Permeability of *Pseudomonas aeruginosa* outer membrane to hydrophilic solutes. *J. Bacteriol.* **152**, 636 (1982).

157 Masuda, N., Gotoh, N., Ishii, C., Sakagawa, E., Ohya, S. & Nishino, T. Interplay between chromosomal β-lactamase and the MexAB-OprM efflux system in intrinsic resistance to β-lactams in *Pseudomonas aeruginosa*. *Antimicrob. Agents Chemother.* **43**, 400-402, doi:10.1128/aac.43.2.400 (1999).

158 Li, X. Z., Nikaido, H. & Poole, K. Role of mexA-mexB-oprM in antibiotic efflux in *Pseudomonas aeruginosa*. *Antimicrob. Agents Chemother.* **39**, 1948, doi:10.1128/AAC.39.9.1948 (1995).

159 Juan, C., Maciá, M. D., Gutiérrez, O., Vidal, C., Pérez, J. L. & Oliver, A. Molecular mechanisms of β-lactam resistance mediated by AmpC hyperproduction in *Pseudomonas aeruginosa* clinical strains. *Antimicrob. Agents Chemother.* **49**, 4733, doi:10.1128/AAC.49.11.4733-4738.2005 (2005).

160 Juan, C., Moyá, B., Pérez, J. L. & Oliver, A. Stepwise upregulation of the *Pseudomonas aeruginosa* chromosomal cephalosporinase conferring high-level β-lactam resistance involves three AmpD homologues. *Antimicrob. Agents Chemother.* **50**, 1780, doi:10.1128/AAC.50.5.1780-1787.2006 (2006).

161 Langaee, T. Y., Gagnon, L. & Huletsky, A. Inactivation of the *ampD* gene in *Pseudomonas aeruginosa* leads to moderate-basal-level and hyperinducible AmpC β-lactamase expression. *Antimicrob. Agents Chemother.* **44**, 583, doi:10.1128/AAC.44.3.583-589.2000 (2000).

162 Schmidtke, A. J. & Hanson, N. D. Role of *ampD* homologs in overproduction of AmpC in clinical isolates of *Pseudomonas aeruginosa*. *Antimicrob. Agents Chemother.* **52**, 3922-3927, doi:10.1128/aac.00341-08 (2008).

163 Berrazeg, M., Jeannot, K., Ntsogo Enguéné, V. Y., Broutin, I., Loeffert, S., Fournier, D. & Plésiat, P. Mutations in β-lactamase AmpC increase resistance of *Pseudomonas aeruginosa* isolates to antipseudomonal cephalosporins. *Antimicrob. Agents Chemother.* **59**, 6248, doi:10.1128/AAC.00825-15 (2015).

164 Trias, J. & Nikaido, H. Outer membrane protein D2 catalyzes facilitated diffusion of carbapenems and penems through the outer membrane of *Pseudomonas aeruginosa*. *Antimicrob. Agents Chemother.* **34**, 52-57, doi:10.1128/aac.34.1.52 (1990).

165 Li, H., Luo, Y.-F., Williams, B. J., Blackwell, T. S. & Xie, C.-M. Structure and function of OprD protein in *Pseudomonas aeruginosa*: from antibiotic resistance to novel therapies. *International journal of medical microbiology : IJMM* **302**, 63-68, doi:10.1016/j.ijmm.2011.10.001 (2012).

166 Köhler, T., Michea-Hamzehpour, M., Epp, S. F. & Pechere, J.-C. Carbapenem activities against *Pseudomonas aeruginosa* respective contributions of OprD and efflux systems. *Antimicrob. Agents Chemother.* **43**, 424, doi:10.1128/AAC.43.2.424 (1999).

167 Bocharova, Y., Savinova, T., Lazareva, A., Polikarpova, S., Gordinskaya, N., Mayanskiy, N. & Chebotar, I. Genotypes, carbapenemase carriage, integron diversity and oprD alterations among carbapenem-resistant *Pseudomonas aeruginosa* from Russia. *Int. J. Antimicrob. Agents* **55**, 105899, doi:https://doi.org/10.1016/j.ijantimicag.2020.105899 (2020).

168 Cabot, G., López-Causapé, C., Ocampo-Sosa, A. A., Sommer, L. M., Domínguez, M. Á., Zamorano, L., Juan, C., Tubau, F., Rodríguez, C., Moyà, B., Peña, C., Martínez-Martínez, L., Plesiat, P. & Oliver, A. Deciphering the resistome of the widespread *Pseudomonas aeruginosa* sequence type 175 international high-risk clone through whole-

genome sequencing. *Antimicrob. Agents Chemother.* **60**, 7415, doi:10.1128/AAC.01720-16 (2016).

169 Hakki, M., Humphries, R. M., Hemarajata, P., Tallman, G. B., Shields, R. K., Mettus, R. T., Doi, Y. & Lewis, J. S., II. Fluoroquinolone prophylaxis selects for meropenem-nonsusceptible *Pseudomonas aeruginosa* in patients with hematologic malignancies and hematopoietic cell transplant recipients. *Clin. Infect. Dis.* **68**, 2045-2052, doi:10.1093/cid/ciy825 (2018).

170 Richardot, C., Plésiat, P., Fournier, D., Monlezun, L., Broutin, I. & Llanes, C. Carbapenem resistance in cystic fibrosis strains of *Pseudomonas aeruginosa* as a result of amino acid substitutions in porin OprD. *Int. J. Antimicrob. Agents* **45**, 529-532, doi:https://doi.org/10.1016/j.ijantimicag.2014.12.029 (2015).

171 Ochs, M. M., McCusker, M. P., Bains, M. & Hancock, R. E. W. Negative regulation of the *Pseudomonas aeruginosa* outer membrane porin OprD selective for imipenem and basic amino acids. *Antimicrob. Agents Chemother.* **43**, 1085-1090, doi:10.1128/aac.43.5.1085 (1999).

172 Srikumar, R., Paul, C. J. & Poole, K. Influence of mutations in the *mexR* repressor gene on expression of the MexA-MexB-OprM multidrug efflux system *Pseudomonas aeruginosa*. *J. Bacteriol.* **182**, 1410, doi:10.1128/JB.182.5.1410-1414.2000 (2000).

173 Llanes, C., Hocquet, D., Vogne, C., Benali-Baitich, D., Neuwirth, C. & Plésiat, P. Clinical strains of *Pseudomonas aeruginosa* overproducing MexAB-OprM and MexXY efflux pumps simultaneously. *Antimicrob. Agents Chemother.* **48**, 1797-1802, doi:10.1128/aac.48.5.1797-1802.2004 (2004).

174 del Barrio-Tofiño, E., López-Causapé, C., Cabot, G., Rivera, A., Benito, N., Segura, C., Montero, M. M., Sorlí, L., Tubau, F., Gómez-Zorrilla, S., Tormo, N., Durá-Navarro, R., Viedma, E., Resino-Foz, E., Fernández-Martínez, M., González-Rico, C., Alejo-Cancho, I., Martínez, J. A., Labayru-Echverria, C., Dueñas, C., Ayestarán, I., Zamorano, L., Martinez-Martinez, L., Horcajada, J. P. & Oliver, A. Genomics and susceptibility profiles of extensively drug-resistant *Pseudomonas aeruginosa* isolates from Spain. *Antimicrob. Agents Chemother.* **61**, e01589-01517, doi:10.1128/AAC.01589-17 (2017).

175 Higgins, P. G., Fluit, A. C., Milatovic, D., Verhoef, J. & Schmitz, F. J. Mutations in GyrA, ParC, MexR and NfxB in clinical isolates of *Pseudomonas aeruginosa*. *Int. J. Antimicrob. Agents* **21**, 409-413, doi:https://doi.org/10.1016/S0924-8579(03)00009-8 (2003).

176 Yoshida, H., Bogaki, M., Nakamura, M. & Nakamura, S. Quinolone resistance-determining region in the DNA gyrase *gyrA* gene of *Escherichia coli*. *Antimicrob. Agents Chemother.* **34**, 1271, doi:10.1128/AAC.34.6.1271 (1990).

177 Georgiou, M., Muñoz, R., Román, F., Cantón, R., Gómez-Lus, R., Campos, J. & De La Campa, A. G. Ciprofloxacin-resistant *Haemophilus influenzae* strains possess mutations in analogous positions of GyrA and ParC. *Antimicrob. Agents Chemother.* **40**, 1741-1744, doi:10.1128/AAC.40.7.1741 (1996).

178 Bruchmann, S., Dötsch, A., Nouri, B., Chaberny, I. F. & Häussler, S. Quantitative contributions of target alteration and decreased drug accumulation to *Pseudomonas aeruginosa* fluoroquinolone resistance. *Antimicrob. Agents Chemother.* **57**, 1361, doi:10.1128/AAC.01581-12 (2013).

179 Akasaka, T., Tanaka, M., Yamaguchi, A. & Sato, K. Type II topoisomerase mutations in fluoroquinolone-resistant clinical strains of *Pseudomonas aeruginosa* isolated in 1998 and 1999: role of target enzyme in mechanism of fluoroquinolone resistance. *Antimicrob. Agents Chemother.* **45**, 2263-2268, doi:10.1128/AAC.45.8.2263-2268.2001 (2001).

180 Chen, Y., Sun, M., Wang, M., Lu, Y. & Yan, Z. Dissemination of IMP-6-producing *Pseudomonas aeruginosa* ST244 in multiple cities in China. *Eur. J. Clin. Microbiol. Infect. Dis.* **33**, 1181-1187, doi:10.1007/s10096-014-2063-5 (2014).

181 Kim, M. J., Bae, I. K., Jeong, S. H., Kim, S. H., Song, J. H., Choi, J. Y., Yoon, S. S., Thamlikitkul, V., Hsueh, P.-R., Yasin, R. M., Lalitha, M. K. & Lee, K. Dissemination of metallo-β-lactamase-producing *Pseudomonas aeruginosa* of sequence type 235 in Asian countries. *J. Antimicrob. Chemother.* **68**, 2820-2824, doi:10.1093/jac/dkt269 (2013).

182 Mano, Y., Saga, T., Ishii, Y., Yoshizumi, A., Bonomo, R. A., Yamaguchi, K. & Tateda, K. Molecular analysis of the integrons of metallo-β-lactamase-producing *Pseudomonas aeruginosa* isolates collected by nationwide surveillance programs across Japan. *BMC Microbiol.* **15**, 41, doi:10.1186/s12866-015-0378-8 (2015).

183 Pelegrin, A. C., Saharman, Y. R., Griffon, A., Palmieri, M., Mirande, C., Karuniawati, A., Sedono, R., Aditianingsih, D., Goessens, W. H. F., van Belkum, A., Verbrugh, H. A., Klaassen, C. H. W. & Severin, J. A. High-risk international clones of carbapenem-nonsusceptible *Pseudomonas aeruginosa* endemic to Indonesian intensive care units: Impact of a multifaceted infection control intervention analyzed at the genomic level. *mBio* **10**, e02384-02319, doi:10.1128/mBio.02384-19 (2019).

184 Pérez-Vázquez, M., Sola-Campoy, P. J., Zurita, Á. M., Ávila, A., Gómez-Bertomeu, F., Solís, S., López-Urrutia, L., Gónzalez-Barberá, E. M., Cercenado, E., Bautista, V., Lara, N., Aracil, B., Oliver, A., Campos, J., Oteo-Iglesias, J. & Spanish Antibiotic Resistance Surveillance Program collaborating, G. Carbapenemase-producing *Pseudomonas aeruginosa* in Spain: interregional dissemination of the high risk-clones ST175 and ST244 carrying $bla_{\text{VIM-2}}$, $bla_{\text{VIM-1}}$, $bla_{\text{IMP-8}}$, $bla_{\text{VIM-20}}$ and $bla_{\text{KPC-2}}$. *Int. J. Antimicrob. Agents*, 106026, doi:10.1016/j.ijantimicag.2020.106026 (2020).

185 Sun, F., Zhou, D., Wang, Q., Feng, J., Feng, W., Luo, W., Liu, Y., Qiu, X., Yin, Z. & Xia, P. Genetic characterization of a novel *blaDIM-2*-carrying megaplasmid p12969-DIM from clinical *Pseudomonas putida*. *J. Antimicrob. Chemother.* **71**, 909-912, doi:10.1093/jac/dkv426 (2016).

186 Wright, L. L., Turton, J. F., Livermore, D. M., Hopkins, K. L. & Woodford, N. Dominance of international 'high-risk clones' among metallo-β-lactamase-producing *Pseudomonas aeruginosa* in the UK. *J. Antimicrob. Chemother.* **70**, 103-110, doi:10.1093/jac/dku339 (2014).

187 Danel, F., Hall, L. M., Gur, D. & Livermore, D. M. OXA-14, another extended-spectrum variant of OXA-10 (PSE-2) beta-lactamase from *Pseudomonas aeruginosa*. *Antimicrob. Agents Chemother.* **39**, 1881, doi:10.1128/AAC.39.8.1881 (1995).

188 Toleman, M. A., Rolston, K., Jones, R. N. & Walsh, T. R. Molecular and biochemical characterization of OXA-45, an extended-spectrum class 2d′ β-lactamase in *Pseudomonas aeruginosa*. *Antimicrob. Agents Chemother.* **47**, 2859, doi:10.1128/AAC.47.9.2859-2863.2003 (2003).

189 Danel, F., Hall, L. M., Gur, D. & Livermore, D. M. OXA-15, an extended-spectrum variant of OXA-2 beta-lactamase, isolated from a *Pseudomonas aeruginosa* strain. *Antimicrob. Agents Chemother.* **41**, 785, doi:10.1128/AAC.41.4.785 (1997).

190 Hall, L. M., Livermore, D. M., Gur, D., Akova, M. & Akalin, H. E. OXA-11, an extended-spectrum variant of OXA-10 (PSE-2) beta-lactamase from *Pseudomonas aeruginosa*. *Antimicrob. Agents Chemother.* **37**, 1637, doi:10.1128/AAC.37.8.1637 (1993).

191 El Garch, F., Bogaerts, P., Bebrone, C., Galleni, M. & Glupczynski, Y. OXA-198, an Acquired Carbapenem-Hydrolyzing Class D β-Lactamase from *Pseudomonas aeruginosa*. *Antimicrob. Agents Chemother.* **55**, 4828-4833, doi:10.1128/aac.00522-11 (2011).

192 Correa, A., del Campo, R., Perenguez, M., Blanco, V. M., Rodríguez-Baños, M., Perez, F., Maya, J. J., Rojas, L., Cantón, R., Arias, C. A. & Villegas, M. V. Dissemination of high-risk clones of extensively drug-resistant *Pseudomonas aeruginosa* in Colombia. *Antimicrob. Agents Chemother.* **59**, 2421, doi:10.1128/AAC.03926-14 (2015).

193 Paulsen, I. T., Littlejohn, T. G., Rådström, P., Sundström, L., Sköld, O., Swedberg, G. & Skurray, R. A. The 3' conserved segment of integrons contains a gene associated with multidrug resistance to antiseptics and disinfectants. *Antimicrob. Agents Chemother.* **37**, 761, doi:10.1128/AAC.37.4.761 (1993).

194 Paulsen, I. T., Brown, M. H. & Skurray, R. A. Proton-dependent multidrug efflux systems. *Microbiol. Rev.* **60**, 575-608 (1996).

195 Belotti, P. T., Thabet, L., Laffargue, A., André, C., Coulange-Mayonnove, L., Arpin, C., Messadi, A., M'Zali, F., Quentin, C. & Dubois, V. Description of an original integron encompassing $bla_{\text{VIM-2}}$, *qnrVC1* and genes encoding bacterial group II intron proteins in *Pseudomonas aeruginosa*. *J. Antimicrob. Chemother.* **70**, 2237-2240, doi:10.1093/jac/dkv103 (2015).

196 Liu, J., Yang, L., Chen, D., Peters, B. M., Li, L., Li, B., Xu, Z. & Shirtliff, M. E. Complete sequence of pBM413, a novel multidrug resistance megaplasmid carrying *qnrVC6* and $bla_{\text{IMP-45}}$ from *Pseudomonas aeruginosa*. *Int. J. Antimicrob. Agents* **51**, 145-150, doi:https://doi.org/10.1016/j.ijantimicag.2017.09.008 (2018).

197 Snesrud, E., Maybank, R., Kwak, Y. I., Jones, A. R., Hinkle, M. K. & McGann, P. Chromosomally encoded *mcr-5* in colistin-nonsusceptible *Pseudomonas aeruginosa*. *Antimicrob. Agents Chemother.* **62**, e00679-00618, doi:10.1128/aac.00679-18 (2018).

198 Caselli, E., D'Accolti, M., Soffritti, I., Piffanelli, M. & Mazzacane, S. Spread of *mcr-1*-driven colistin resistance on hospital surfaces, Italy. *Emerg. Infect. Dis.* **24**, 1752-1753, doi:10.3201/eid2409.171386 (2018).

199 Slekovec, C., Robert, J., van der Mee-Marquet, N., Berthelot, P., Rogues, A.-M., Derouin, V., Cholley, P., Thouverez, M., Hocquet, D. & Bertrand, X. Molecular epidemiology of *Pseudomonas aeruginosa* isolated from infected ICU patients: a French multicenter 2012–2013 study. *Eur. J. Clin. Microbiol. Infect. Dis.* **38**, 921-926, doi:10.1007/s10096-019-03519-w (2019).

200 Gomila, M., del Carmen Gallegos, M., Fernández-Baca, V., Pareja, A., Pascual, M., Díaz-Antolín, P., García-Valdés, E. & Lalucat, J. Genetic diversity of clinical *Pseudomonas aeruginosa* isolates in a public hospital in Spain. *BMC Microbiol.* **13**, 138, doi:10.1186/1471-2180-13-138 (2013).

201 Cholley, P., Thouverez, M., Hocquet, D., van der Mee-Marquet, N., Talon, D. & Bertrand, X. Most multidrug-resistant *Pseudomonas aeruginosa* isolates from hospitals in eastern France belong to a few clonal types. *J. Clin. Microbiol.* **49**, 2578-2583, doi:10.1128/jcm.00102-11 (2011).

202 Walters, M. S., Grass, J. E., Bulens, S. N., Hancock, E. B., Phipps, E. C., Muleta, D., Mounsey, J., Kainer, M. A., Concannon, C., Dumyati, G., Bower, C., Jacob, J., Cassidy, P. M., Beldavs, Z., Culbreath, K., Phillips, W. E., Jr., Hardy, D. J., Vargas, R. L., Oethinger, M., Ansari, U., Stanton, R., Albrecht, V., Halpin, A. L., Karlsson, M., Rasheed, J. K. & Kallen, A. Carbapenem-resistant *Pseudomonas aeruginosa* at US Emerging Infections Program sites, 2015. *Emerg. Infect. Dis.* **25**, 1281-1288, doi:10.3201/eid2507.181200 (2019).

203 Hattemer, A., Hauser, A., Diaz, M., Scheetz, M., Shah, N., Allen, J. P., Porhomayon, J. & El-Solh, A. A. Bacterial and clinical characteristics of health care- and community-acquired bloodstream infections due to *Pseudomonas aeruginosa*. *Antimicrob. Agents Chemother.* **57**, 3969-3975, doi:10.1128/aac.02467-12 (2013).

204 Aiello, D., Williams, J. D., Majgier-Baranowska, H., Patel, I., Peet, N. P., Huang, J., Lory, S., Bowlin, T. L. & Moir, D. T. Discovery and characterization of inhibitors of *Pseudomonas aeruginosa* type III secretion. *Antimicrob. Agents Chemother.* **54**, 1988, doi:10.1128/AAC.01598-09 (2010).

205 Anantharajah, A., Mingeot-Leclercq, M.-P. & Van Bambeke, F. Targeting the type three secretion system in *Pseudomonas aeruginosa*. *Trends Pharmacol. Sci.* **37**, 734-749, doi:https://doi.org/10.1016/j.tips.2016.05.011 (2016).

206 Lee, J.-H., Kim, Y.-G., Cho, M. H., Kim, J.-A. & Lee, J. 7-fluoroindole as an antivirulence compound against *Pseudomonas aeruginosa*. *FEMS Microbiol. Lett.* **329**, 36-44, doi:10.1111/j.1574-6968.2012.02500.x (2012).

207 Lee, D. G., Urbach, J. M., Wu, G., Liberati, N. T., Feinbaum, R. L., Miyata, S., Diggins, L. T., He, J., Saucier, M., Deziel, E., Friedman, L., Li, L., Grills, G., Montgomery, K., Kucherlapati, R., Rahme, L. G. & Ausubel, F. M. Genomic analysis reveals that *Pseudomonas aeruginosa* virulence is combinatorial. *Genome Biol.* **7**, R90, doi:10.1186/gb-2006-7-10-r90 (2006).

208 LaFayette, S. L., Houle, D., Beaudoin, T., Wojewodka, G., Radzioch, D., Hoffman, L. R., Burns, J. L., Dandekar, A. A., Smalley, N. E., Chandler, J. R., Zlosnik, J. E., Speert, D. P., Bernier, J., Matouk, E., Brochiero, E., Rousseau, S. & Nguyen, D. Cystic fibrosis–adapted *Pseudomonas aeruginosa* quorum sensing *lasR* mutants cause hyperinflammatory responses. *Sci. Adv.* **1**, e1500199, doi:10.1126/sciadv.1500199 (2015).

209 Bragonzi, A., Paroni, M., Nonis, A., Cramer, N., Montanari, S., Rejman, J., Serio, C. D., Döring, G. & Tümmler, B. *Pseudomonas aeruginosa* microevolution during cystic fibrosis lung infection establishes clones with adapted virulence. *Am. J. Respir. Crit. Care Med.* **180**, 138-145, doi:10.1164/rccm.200812-1943OC (2009).

210 Jain, M., Ramirez, D., Seshadri, R., Cullina, J. F., Powers, C. A., Schulert, G. S., Bar-Meir, M., Sullivan, C. L., McColley, S. A. & Hauser, A. R. Type III secretion phenotypes of *Pseudomonas aeruginosa* strains change during infection of individuals with cystic fibrosis. *J. Clin. Microbiol.* **42**, 5229-5237, doi:10.1128/jcm.42.11.5229-5237.2004 (2004).

211 Jain, M., Bar-Meir, M., McColley, S., Cullina, J., Potter, E., Powers, C., Prickett, M., Seshadri, R., Jovanovic, B., Petrocheilou, A., King, J. D. & Hauser, A. R. Evolution of *Pseudomonas aeruginosa* type III secretion in cystic fibrosis: a paradigm of chronic infection. *Transl. Res.* **152**, 257-264, doi:https://doi.org/10.1016/j.trsl.2008.10.003 (2008).

212 Burrows, L. L. *Pseudomonas aeruginosa* twitching motility: Type IV pili in action. *Annu. Rev. Microbiol.* **66**, 493-520, doi:10.1146/annurev-micro-092611-150055 (2012).

213 Giltner, C. L., Van Schaik, E. J., Audette, G. F., Kao, D., Hodges, R. S., Hassett, D. J. & Irvin, R. T. The *Pseudomonas aeruginosa* type IV pilin receptor binding domain functions as an adhesin for both biotic and abiotic surfaces. *Mol. Microbiol.* **59**, 1083-1096, doi:10.1111/j.1365-2958.2005.05002.x (2006).

214 Doig, P., Todd, T., Sastry, P. A., Lee, K. K., Hodges, R. S., Paranchych, W. & Irvin, R. T. Role of pili in adhesion of *Pseudomonas aeruginosa* to human respiratory epithelial cells. *Infect. Immun.* **56**, 1641 (1988).

215 O'Toole, G. A. & Kolter, R. Flagellar and twitching motility are necessary for *Pseudomonas aeruginosa* biofilm development. *Mol. Microbiol.* **30**, 295-304, doi:10.1046/j.1365-2958.1998.01062.x (1998).

216 Chiang, P. & Burrows, L. L. Biofilm formation by hyperpiliated mutants of *Pseudomonas aeruginosa*. *J. Bacteriol.* **185**, 2374-2378, doi:10.1128/jb.185.7.2374-2378.2003 (2003).

217 Persat, A., Inclan, Y. F., Engel, J. N., Stone, H. A. & Gitai, Z. Type IV pili mechanochemically regulate virulence factors in *Pseudomonas aeruginosa*. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 7563-7568, doi:10.1073/pnas.1502025112 (2015).

218 Siryaporn, A., Kuchma, S. L., O'Toole, G. A. & Gitai, Z. Surface attachment induces *Pseudomonas aeruginosa* virulence. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 16860-16865, doi:10.1073/pnas.1415712111 (2014).

219 Tang, H., Kays, M. & Prince, A. Role of *Pseudomonas aeruginosa* pili in acute pulmonary infection. *Infect. Immun.* **63**, 1278-1285 (1995).

220 Zolfaghar, I., Evans, D. J. & Fleiszig, S. M. J. Twitching motility contributes to the tole of pili in corneal infection caused by *Pseudomonas aeruginosa*. *Infect. Immun.* **71**, 5389-5393, doi:10.1128/iai.71.9.5389-5393.2003 (2003).

221 Kazmierczak, B. I., Schniederberend, M. & Jain, R. Cross-regulation of *Pseudomonas* motility systems: the intimate relationship between flagella, pili and virulence. *Curr. Opin. Microbiol.* **28**, 78-82, doi:10.1016/j.mib.2015.07.017 (2015).

222 Bruzaud, J., Tarrade, J., Coudreuse, A., Canette, A., Herry, J.-M., Taffin de Givenchy, E., Darmanin, T., Guittard, F., Guilbaud, M. & Bellon-Fontaine, M.-N. Flagella but not type IV pili are involved in the initial adhesion of *Pseudomonas aeruginosa* PAO1 to hydrophobic or superhydrophobic surfaces. *Colloids Surf. B Biointerfaces* **131**, 59-66, doi:https://doi.org/10.1016/j.colsurfb.2015.04.036 (2015).

223 Rashid, M. H. & Kornberg, A. Inorganic polyphosphate is needed for swimming, swarming, and twitching motilities of *Pseudomonas aeruginosa*. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 4885-4890 (2000).

224 Feldman, M., Bryan, R., Rajan, S., Scheffler, L., Brunnert, S., Tang, H. & Prince, A. Role of flagella in pathogenesis of *Pseudomonas aeruginosa* pulmonary infection. *Infect. Immun.* **66**, 43-51, doi:10.1128/iai.66.1.43-51.1998 (1998).

225 Arora, S. K., Neely, A. N., Blair, B., Lory, S. & Ramphal, R. Role of motility and flagellin glycosylation in the pathogenesis of *Pseudomonas aeruginosa* burn wound infections. *Infect. Immun.* **73**, 4395-4398, doi:10.1128/iai.73.7.4395-4398.2005 (2005).

226 Arora, S. K., Wolfgang, M. C., Lory, S. & Ramphal, R. Sequence polymorphism in the glycosylation island and flagellins of *Pseudomonas aeruginosa*. *J. Bacteriol.* **186**, 2115-2122, doi:10.1128/jb.186.7.2115-2122.2004 (2004).

227 Lau, G. W., Hassett, D. J., Ran, H. & Kong, F. The role of pyocyanin in *Pseudomonas aeruginosa* infection. *Trends Mol. Med.* **10**, 599-606, doi:https://doi.org/10.1016/j.molmed.2004.10.002 (2004).

228 Lau, G. W., Ran, H., Kong, F., Hassett, D. J. & Mavrodi, D. *Pseudomonas aeruginosa* pyocyanin Is critical for lung infection in mice. *Infect. Immun.* **72**, 4275-4278, doi:10.1128/iai.72.7.4275-4278.2004 (2004).

229 Bleves, S., Viarre, V., Salacha, R., Michel, G. P. F., Filloux, A. & Voulhoux, R. Protein secretion systems in *Pseudomonas aeruginosa*: A wealth of pathogenic weapons. *Int. J. Med. Microbiol.* **300**, 534-543, doi:http://doi.org/10.1016/j.ijmm.2010.08.005 (2010).

230 Braun, P., de Groot, A., Bitter, W. & Tommassen, J. Secretion of elastinolytic enzymes and their propeptides by *Pseudomonas aeruginosa*. *J. Bacteriol.* **180**, 3467-3469, doi:10.1128/jb.180.13.3467-3469.1998 (1998).

231 Lu, H. M., Mizushima, S. & Lory, S. A periplasmic intermediate in the extracellular secretion pathway of *Pseudomonas aeruginosa* exotoxin A. *J. Bacteriol.* **175**, 7463-7467, doi:10.1128/jb.175.22.7463-7467.1993 (1993).

232 Michalska, M. & Wolf, P. *Pseudomonas* Exotoxin A: optimized by evolution for effective killing. *Front. Microbiol.* **6**, 963-963, doi:10.3389/fmicb.2015.00963 (2015).

233 Jyot, J., Balloy, V., Jouvion, G., Verma, A., Touqui, L., Huerre, M., Chignard, M. & Ramphal, R. Type II secretion system of *Pseudomonas aeruginosa*: In vivo evidence of a significant role in death due to lung infection. *J. Infect. Dis.* **203**, 1369-1377, doi:10.1093/infdis/jir045 (2011).

234 Miyazaki, S., Matsumoto, T., Tateda, K., Ohno, A. & Yamaguchi, K. Role of exotoxin A in inducing severe *Pseudomonas aeruginosa* infections in mice. *J. Med. Microbiol.* **43**, 169-175, doi:https://doi.org/10.1099/00222615-43-3-169 (1995).

235 Ball, G., Durand, É., Lazdunski, A. & Filloux, A. A novel type II secretion system in *Pseudomonas aeruginosa*. *Mol. Microbiol.* **43**, 475-485, doi:10.1046/j.1365-2958.2002.02759.x (2002).

236 Ball, G., Viarre, V., Garvis, S., Voulhoux, R. & Filloux, A. Type II-dependent secretion of a Pseudomonas aeruginosa DING protein. *Res. Microbiol.* **163**, 457-469, doi:https://doi.org/10.1016/j.resmic.2012.07.007 (2012).

237 Cadoret, F., Ball, G., Douzi, B. & Voulhoux, R. Txc, a new type II secretion system of *Pseudomonas aeruginosa* strain PA7, is regulated by the TtsS/TtsR two-component system and directs specific secretion of the CbpE chitin-binding protein. *J. Bacteriol.* **196**, 2376, doi:10.1128/JB.01563-14 (2014).

238 Chen, L., Zou, Y., She, P. & Wu, Y. Composition, function, and regulation of T6SS in *Pseudomonas aeruginosa*. *Microbiol. Res.* **172**, 19-25, doi:https://doi.org/10.1016/j.micres.2015.01.004 (2015).

239 Sana, T. G., Soscia, C., Tonglet, C. M., Garvis, S. & Bleves, S. Divergent control of two type VI secretion systems by RpoN in *Pseudomonas aeruginosa*. *PLoS One* **8**, e76030, doi:10.1371/journal.pone.0076030 (2013).

240 Li, Y., Chen, L., Zhang, P., Bhagirath, A. Y. & Duan, K. ClpV3 of the H3-type VI secretion system (H3-T6SS) affects multiple virulence factors in *Pseudomonas aeruginosa*. *Front. Microbiol.* **11**, doi:10.3389/fmicb.2020.01096 (2020).

241 Han, Y., Wang, T., Chen, G., Pu, Q., Liu, Q., Zhang, Y., Xu, L., Wu, M. & Liang, H. A *Pseudomonas aeruginosa* type VI secretion system regulated by CueR facilitates copper acquisition. *PLoS Pathog.* **15**, e1008198, doi:10.1371/journal.ppat.1008198 (2019).

242 Lesic, B., Starkey, M., He, J., Hazan, R. & Rahme, L. G. Quorum sensing differentially regulates *Pseudomonas aeruginosa* type VI secretion locus I and homologous loci II and III, which are required for pathogenesis. *Microbiology* **155**, 2845-2855, doi:10.1099/mic.0.029082-0 (2009).

243 Diepold, A. & Armitage, J. P. Type III secretion systems: the bacterial flagellum and the injectisome. *Phil. Trans. R. Soc. B* **370**, 20150020, doi:doi:10.1098/rstb.2015.0020 (2015).

244 Vallis, A. J., Yahr, T. L., Barbieri, J. T. & Frank, D. W. Regulation of ExoS production and secretion by *Pseudomonas aeruginosa* in response to tissue culture conditions. *Infect. Immun.* **67**, 914, doi:10.1128/IAI.67.2.914-920.1999 (1999).

245 Rietsch, A. & Mekalanos, J. J. Metabolic regulation of type III secretion gene expression in *Pseudomonas aeruginosa*. *Mol. Microbiol.* **59**, 807-820, doi:10.1111/j.1365-2958.2005.04990.x (2006).

246 Bachta, K. E. R., Allen, J. P., Cheung, B. H., Chiu, C.-H. & Hauser, A. R. Systemic infection facilitates transmission of *Pseudomonas aeruginosa* in mice. *Nat. Commun.* **11**, 543, doi:10.1038/s41467-020-14363-4 (2020).

247 Miyata, S., Casey, M., Frank, D. W., Ausubel, F. M. & Drenkard, E. Use of the *Galleria mellonella* caterpillar as a model host to study the role of the type III secretion system in *Pseudomonas aeruginosa* pathogenesis. *Infect. Immun.* **71**, 2404, doi:10.1128/IAI.71.5.2404-2413.2003 (2003).

248 Berube, B. J., Murphy, K. R., Torhan, M. C., Bowlin, N. O., Williams, J. D., Bowlin, T. L., Moir, D. T. & Hauser, A. R. Impact of type III secretion effectors and of phenoxyacetamide inhibitors of type III secretion on abscess formation in a mouse model of *Pseudomonas aeruginosa* infection. *Antimicrob. Agents Chemother.* **61**, e01202-01217, doi:10.1128/aac.01202-17 (2017).

249 Smith, R. S., Wolfgang, M. C. & Lory, S. An adenylate cyclase-controlled signaling network regulates *Pseudomonas aeruginosa* virulence in a mouse model of acute pneumonia. *Infect. Immun.* **72**, 1677, doi:10.1128/IAI.72.3.1677-1684.2004 (2004).

250 Krall, R., Schmidt, G., Aktories, K. & Barbieri, J. T. *Pseudomonas aeruginosa* ExoT is a Rho GTPase-activating protein. *Infect. Immun.* **68**, 6066, doi:10.1128/IAI.68.10.6066-6068.2000 (2000).

251 Sun, J. & Barbieri, J. T. *Pseudomonas aeruginosa* ExoT ADP-ribosylates CT10 regulator of kinase (Crk) proteins. *J. Biol. Chem.* **278**, 32794-32800, doi:10.1074/jbc.M304290200 (2003).

252 Shafikhani, S. H., Morales, C. & Engel, J. The *Pseudomonas aeruginosa* type III secreted toxin ExoT is necessary and sufficient to induce apoptosis in epithelial cells. *Cell. Microbiol.* **10**, 994-1007, doi:10.1111/j.1462-5822.2007.01102.x (2008).

253 Geiser, T. K., Kazmierczak, B. I., Garrity-Ryan, L. K., Matthay, M. A. & Engel, J. N. *Pseudomonas aeruginosa* ExoT inhibits in vitro lung epithelial wound repair. *Cell. Microbiol.* **3**, 223-236, doi:10.1046/j.1462-5822.2001.00107.x (2001).

254 Yahr, T. L., Vallis, A. J., Hancock, M. K., Barbieri, J. T. & Frank, D. W. ExoY, an adenylate cyclase secreted by the *Pseudomonas aeruginosa* type III system. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 13899-13904, doi:10.1073/pnas.95.23.13899 (1998).

255 Hritonenko, V., Mun, J. J., Tam, C., Simon, N. C., Barbieri, J. T., Evans, D. J. & Fleiszig, S. M. J. Adenylate cyclase activity of *Pseudomonas aeruginosa* ExoY can mediate bleb-niche formation in epithelial cells and contributes to virulence. *Microb. Pathog.* **51**, 305-312, doi:https://doi.org/10.1016/j.micpath.2011.08.001 (2011).

256 Prasain, N., Alvarez, D. F., Frank, D. W. & Stevens, T. *exoY* increases *Pseudomonas aeruginosa* virulence. *FASEB J.* **22**, 928.926-928.926, doi:10.1096/fasebj.22.1_supplement.928.6 (2008).

257 Mikkelsen, H., Ball, G., Giraud, C. & Filloux, A. Expression of *Pseudomonas aeruginosa* CupD fimbrial genes Is antagonistically controlled by RcsB and the EAL-containing PvrR response regulators. *PLoS One* **4**, e6018, doi:10.1371/journal.pone.0006018 (2009).

258 Mikkelsen, H., Hui, K., Barraud, N. & Filloux, A. The pathogenicity island encoded PvrSR/RcsCB regulatory network controls biofilm formation and dispersal in *Pseudomonas aeruginosa* PA14. *Mol. Microbiol.* **89**, 450-463, doi:10.1111/mmi.12287 (2013).

259 Ferrara, S., Falcone, M., Macchi, R., Bragonzi, A., Girelli, D., Cariani, L., Cigana, C. & Bertoni, G. The PAPI-1 pathogenicity island-encoded small RNA PesA influences *Pseudomonas aeruginosa* virulence and modulates pyocin S3 production. *PLoS One* **12**, e0180386, doi:10.1371/journal.pone.0180386 (2017).

260 Sato, H. & Frank, D. W. ExoU is a potent intracellular phospholipase. *Mol. Microbiol.* **53**, 1279-1290, doi:10.1111/j.1365-2958.2004.04194.x (2004).

261 Hauser, A. R., Kang, P. J. & Engel, J. N. PepA, a secreted protein of *Pseudomonas aeruginosa*, is necessary for cytotoxicity and virulence. *Mol. Microbiol.* **27**, 807-818, doi:10.1046/j.1365-2958.1998.00727.x (1998).

262 Sato, H., Frank, D. W., Hillard, C. J., Feix, J. B., Pankhaniya, R. R., Moriyama, K., Finck-Barbançon, V., Buchaklian, A., Lei, M., Long, R. M., Wiener-Kronish, J. & Sawa, T. The mechanism of action of the *Pseudomonas aeruginosa*-encoded type III cytotoxin, ExoU. *EMBO J.* **22**, 2959-2969, doi:10.1093/emboj/cdg290 (2003).

263 Phillips, R. M., Six, D. A., Dennis, E. A. & Ghosh, P. In vivo phospholipase activity of the *Pseudomonas aeruginosa* cytotoxin ExoU and protection of mammalian cells with phospholipase A2 inhibitors. *J. Biol. Chem.* **278**, 41326-41332, doi:10.1074/jbc.M302472200 (2003).

264 Finck-Barbançon, V., Goranson, J., Zhu, L., Sawa, T., Wiener-Kronish, J. P., Fleiszig, S. M. J., Wu, C., Mende-Mueller, L. & Frank, D. W. ExoU expression by *Pseudomonas aeruginosa* correlates with acute cytotoxicity and epithelial injury. *Mol. Microbiol.* **25**, 547-557, doi:10.1046/j.1365-2958.1997.4891851.x (1997).

265 Howell, H. A., Logan, L. K. & Hauser, A. R. Type III secretion of ExoU Is critical during early *Pseudomonas aeruginosa* pneumonia. *mBio* **4**, doi:10.1128/mBio.00032-13 (2013).
266 Wolfgang, M. C., Kulasekara, B. R., Liang, X., Boyd, D., Wu, K., Yang, Q., Miyada, C. G. & Lory, S. Conservation of genome content and virulence determinants among clinical and environmental isolates of *Pseudomonas aeruginosa*. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 8484-8489, doi:10.1073/pnas.0832438100 (2003).
267 Yahr, T. L., Barbieri, J. T. & Frank, D. W. Genetic relationship between the 53- and 49-kilodalton forms of exoenzyme S from *Pseudomonas aeruginosa*. *J. Bacteriol.* **178**, 1412-1419, doi:10.1128/jb.178.5.1412-1419.1996 (1996).
268 Pederson, K. J., Vallis, A. J., Aktories, K., Frank, D. W. & Barbieri, J. T. The amino-terminal domain of *Pseudomonas aeruginosa* ExoS disrupts actin filaments via small-molecular-weight GTP-binding proteins. *Mol. Microbiol.* **32**, 393-401, doi:10.1046/j.1365-2958.1999.01359.x (1999).
269 Knight, D. A., Finck-Barbançon, V., Kulich, S. M. & Barbieri, J. T. Functional domains of *Pseudomonas aeruginosa* exoenzyme S. *Infect. Immun.* **63**, 3182-3186 (1995).
270 Elsen, S., Huber, P., Bouillot, S., Couté, Y., Fournier, P., Dubois, Y., Timsit, J.-F., Maurin, M. & Attrée, I. A type III secretion negative clinical strain of *Pseudomonas aeruginosa* employs a two-partner secreted exolysin to induce hemorrhagic pneumonia. *Cell Host Microbe* **15**, 164-176, doi:https://doi.org/10.1016/j.chom.2014.01.003 (2014).
271 Basso, P., Ragno, M., Elsen, S., Reboud, E., Golovkine, G., Bouillot, S., Huber, P., Lory, S., Faudry, E. & Attrée, I. *Pseudomonas aeruginosa* pore-forming exolysin and type IV pili cooperate to induce host cell lysis. *mBio* **8**, e02250-02216, doi:10.1128/mBio.02250-16 (2017).
272 Basso, P., Wallet, P., Elsen, S., Soleilhac, E., Henry, T., Faudry, E. & Attrée, I. Multiple *Pseudomonas* species secrete exolysin-like toxins and provoke Caspase-1-dependent macrophage death. *Environ. Microbiol.* **19**, 4045-4064, doi:10.1111/1462-2920.13841 (2017).
273 Libbrecht, M. W. & Noble, W. S. Machine learning applications in genetics and genomics. *Nat. Rev. Genet.* **16**, 321-332, doi:10.1038/nrg3920 (2015).
274 Davis, J. J., Boisvert, S., Brettin, T., Kenyon, R. W., Mao, C., Olson, R., Overbeek, R., Santerre, J., Shukla, M., Wattam, A. R., Will, R., Xia, F. & Stevens, R. Antimicrobial resistance prediction in PATRIC and RAST. *Sci. Rep.* **6**, 27930, doi:10.1038/srep27930 (2016).
275 Nguyen, M., Long, S. W., McDermott, P. F., Olsen, R. J., Olson, R., Stevens, R. L., Tyson, G. H., Zhao, S. & Davis, J. J. Using machine learning to predict antimicrobial MICs and associated genomic features for nontyphoidal *Salmonella*. *J. Clin. Microbiol.* **57**, e01260-01218, doi:10.1128/JCM.01260-18 (2019).
276 Kavvas, E. S., Catoiu, E., Mih, N., Yurkovich, J. T., Seif, Y., Dillon, N., Heckmann, D., Anand, A., Yang, L., Nizet, V., Monk, J. M. & Palsson, B. O. Machine learning and structural analysis of *Mycobacterium tuberculosis* pan-genome identifies genetic signatures of antibiotic resistance. *Nat. Commun.* **9**, 4306, doi:10.1038/s41467-018-06634-y (2018).
277 Chen, M. L., Doddi, A., Royer, J., Freschi, L., Schito, M., Ezewudo, M., Kohane, I. S., Beam, A. & Farhat, M. Beyond multidrug resistance: Leveraging rare variants with

machine and statistical learning models in *Mycobacterium tuberculosis* resistance prediction. *EBioMedicine* **43**, 356-369, doi:10.1016/j.ebiom.2019.04.016 (2019).

278 Hyun, J. C., Kavvas, E. S., Monk, J. M. & Palsson, B. O. Machine learning with random subspace ensembles identifies antimicrobial resistance determinants from pan-genomes of three pathogens. *PLoS Comput. Biol.* **16**, e1007608, doi:10.1371/journal.pcbi.1007608 (2020).

279 Břinda, K., Callendrello, A., Ma, K. C., MacFadden, D. R., Charalampous, T., Lee, R. S., Cowley, L., Wadsworth, C. B., Grad, Y. H., Kucherov, G., O'Grady, J., Baym, M. & Hanage, W. P. Rapid inference of antibiotic resistance and susceptibility by genomic neighbour typing. *Nat. Microbiol.*, doi:10.1038/s41564-019-0656-6 (2020).

280 Wheeler, N. E., Gardner, P. P. & Barquist, L. Machine learning identifies signatures of host adaptation in the bacterial pathogen *Salmonella enterica*. *PLoS Genet.* **14**, e1007333, doi:10.1371/journal.pgen.1007333 (2018).

281 Cornforth, D. M., Dees, J. L., Ibberson, C. B., Huse, H. K., Mathiesen, I. H., Kirketerp-Møller, K., Wolcott, R. D., Rumbaugh, K. P., Bjarnsholt, T. & Whiteley, M. *Pseudomonas aeruginosa* transcriptome during human infection. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E5125, doi:10.1073/pnas.1717525115 (2018).

282 Recker, M., Laabei, M., Toleman, M. S., Reuter, S., Saunderson, R. B., Blane, B., Török, M. E., Ouadi, K., Stevens, E., Yokoyama, M., Steventon, J., Thompson, L., Milne, G., Bayliss, S., Bacon, L., Peacock, S. J. & Massey, R. C. Clonal differences in *Staphylococcus aureus* bacteraemia-associated mortality. *Nat. Microbiol.* **2**, 1381-1388, doi:10.1038/s41564-017-0001-x (2017).

283 Lupolova, N., Dallman, T. J., Holden, N. J. & Gally, D. L. Patchy promiscuity: machine learning applied to predict the host specificity of *Salmonella enterica* and *Escherichia coli*. *Microb. Genom.* **3**, e000135-e000135, doi:10.1099/mgen.0.000135 (2017).

284 Müller, A. C. & Guido, S. *Introduction to Machine Learning with Python: A Guide for Data Scientists*. (O'Reilly Media, Inc., 2016).

285 Baştanlar, Y. & Özuysal, M. in *miRNomics: MicroRNA Biology and Computational Analysis* Vol. 1107 *Methods in Molecular Biology* (eds Malik Yousef & Jens Allmer) Ch. 7, 105-128 (Humana Press, 2014).

286 Eraslan, G., Avsec, Ž., Gagneur, J. & Theis, F. J. Deep learning: new computational modelling techniques for genomics. *Nat. Rev. Genet.* **20**, 389-403, doi:10.1038/s41576-019-0122-6 (2019).

287 Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825-2830 (2011).

288 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 Comput. Biol.* **14**, e1006258-e1006258, doi:10.1371/journal.pcbi.1006258 (2018).

289 Naidenov, B., Lim, A., Willyerd, K., Torres, N. J., Johnson, W. L., Hwang, H. J., Hoyt, P., Gustafson, J. E. & Chen, C. Pan-genomic and polymorphic driven prediction of antibiotic resistance in *Elizabethkingia*. *Front. Microbiol.* **10**, 1446-1446, doi:10.3389/fmicb.2019.01446 (2019).

290 Chen, T. & Guestrin, C. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (Association for Computing Machinery, San Francisco, California, USA, 2016).

291 Nguyen, M., Olson, R., Shukla, M., VanOeffelen, M. & Davis, J. J. Predicting antimicrobial resistance using conserved genes. *bioRxiv*, 2020.2004.2029.068254, doi:10.1101/2020.04.29.068254 (2020).

292 Nguyen, M., Brettin, T., Long, S. W., Musser, J. M., Olsen, R. J., Olson, R., Shukla, M., Stevens, R. L., Xia, F., Yoo, H. & Davis, J. J. Developing an in silico minimum inhibitory concentration panel test for *Klebsiella pneumoniae*. *Sci. Rep.* **8**, 421, doi:10.1038/s41598-017-18972-w (2018).

293 Kuhn, M. Building predictive models in R using the caret package. *J. Stat. Softw.* **28**, 1-26, doi:10.18637/jss.v028.i05 (2008).

294 Lapp, Z., Han, J., Wiens, J., Goldstein, E. J. C., Lautenbach, E. & Snitkin, E. Machine learning models to identify patient and microbial genetic factors associated with carbapenem-resistant Klebsiella pneumoniae infection. *medRxiv*, 2020.2007.2006.20147306, doi:10.1101/2020.07.06.20147306 (2020).

295 Liaw, A. & Wiener, M. Classification and regression by randomForest. *R news* **2**, 18-22 (2002).

296 Laabei, M., Recker, M., Rudkin, J. K., Aldeljawi, M., Gulay, Z., Sloan, T. J., Williams, P., Endres, J. L., Bayles, K. W., Fey, P. D., Yajjala, V. K., Widhelm, T., Hawkins, E., Lewis, K., Parfett, S., Scowen, L., Peacock, S. J., Holden, M., Wilson, D., Read, T. D., van den Elsen, J., Priest, N. K., Feil, E. J., Hurst, L. D., Josefsson, E. & Massey, R. C. Predicting the virulence of MRSA from its genome sequence. *Genome Res.* **24**, 839-849, doi:10.1101/gr.165415.113 (2014).

297 Lupolova, N., Dallman, T. J., Matthews, L., Bono, J. L. & Gally, D. L. Support vector machine applied to predict the zoonotic potential of *E. coli* O157 cattle isolates. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 11312, doi:10.1073/pnas.1606567113 (2016).

298 Andreatta, M., Nielsen, M., Møller Aarestrup, F. & Lund, O. In silico prediction of human pathogenicity in the γ-Proteobacteria. *PLoS One* **5**, e13680, doi:10.1371/journal.pone.0013680 (2010).

299 Barash, E., Sal-Man, N., Sabato, S. & Ziv-Ukelson, M. BacPaCS—Bacterial Pathogenicity Classification via Sparse-SVM. *Bioinformatics* **35**, 2001-2008, doi:10.1093/bioinformatics/bty928 (2018).

300 Zhang, M., Yang, L., Ren, J., Ahlgren, N. A., Fuhrman, J. A. & Sun, F. Prediction of virus-host infectious association by supervised learning methods. *BMC Bioinformatics* **18**, 60, doi:10.1186/s12859-017-1473-7 (2017).

301 Lees, J. A., Mai, T. T., Galardini, M., Wheeler, N. E., Horsfield, S. T., Parkhill, J. & Corander, J. Improved prediction of bacterial genotype-phenotype associations using interpretable pangenome-spanning regressions. *mBio* **11**, e01344-01320, doi:10.1128/mBio.01344-20 (2020).

302 Pincus, N. B., Bachta, K. E. R., Ozer, E. A., Allen, J. P., Pura, O. N., Qi, C., Rhodes, N. J., Marty, F. M., Pandit, A., Mekalanos, J. J., Oliver, A. & Hauser, A. R. 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).

303 Tacconelli, E. & Magrini, N. Global priority list of antibiotic-resistant bacteria to guide research, discovery, and development of new antibiotics. *World Health Organization* (2017).

304 Roach, D. J., Burton, J. N., Lee, C., Stackhouse, B., Butler-Wu, S. M., Cookson, B. T., Shendure, J. & Salipante, S. J. A year of infection in the intensive care unit: prospective whole genome wequencing of bacterial clinical isolates reveals cryptic transmissions and novel microbiota. *PLoS Genet.* **11**, e1005413, doi:10.1371/journal.pgen.1005413 (2015).

305 Jolley, K. A., Bray, J. E. & Maiden, M. C. J. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. *Wellcome Open Res.* **3**, 124-124, doi:10.12688/wellcomeopenres.14826.1 (2018).

306 Lee, J.-Y., Peck, K. R. & Ko, K. S. Selective advantages of two major clones of carbapenem-resistant *Pseudomonas aeruginosa* isolates (CC235 and CC641) from Korea: antimicrobial resistance, virulence and biofilm-forming activity. *J. Med. Microbiol.* **62**, 1015-1024, doi:doi:10.1099/jmm.0.055426-0 (2013).

307 Domitrovic, T. N., Hujer, A. M., Perez, F., Marshall, S. H., Hujer, K. M., Woc-Colburn, L. E., Parta, M. & Bonomo, R. A. Multidrug resistant *Pseudomonas aeruginosa* causing prosthetic valve endocarditis: a genetic-based chronicle of evolving antibiotic resistance. *Open Forum Infect. Dis.* **3**, ofw188-ofw188, doi:10.1093/ofid/ofw188 (2016).

308 Avetisian, L. R., Voronina, O. L., Chernukha, M., Kunda, M. S., Gabrielian, N. I., Lunin, V. G. & Shaginian, I. A. [Persistence of *Pseudomonas aeruginosa* strains in patients of Federal Scientific Center of Transplantology and Artificial Organs]. *Zhurnal mikrobiologii, epidemiologii, i immunobiologii*, 99-104 (2012).

309 Zankari, E., Hasman, H., Cosentino, S., Vestergaard, M., Rasmussen, S., Lund, O., Aarestrup, F. M. & Larsen, M. V. Identification of acquired antimicrobial resistance genes. *J. Antimicrob. Chemother.* **67**, 2640-2644, doi:10.1093/jac/dks261 (2012).

310 Liu, B., Zheng, D., Jin, Q., Chen, L. & Yang, J. VFDB 2019: a comparative pathogenomic platform with an interactive web interface. *Nucleic Acids Res.* **47**, D687-D692, doi:10.1093/nar/gky1080 (2018).

311 Tonkin-Hill, G., Lees, J. A., Bentley, S. D., Frost, S. D. W. & Corander, J. RhierBAPS: An R implementation of the population clustering algorithm hierBAPS. *Wellcome Open Res.* **3**, 93-93, doi:10.12688/wellcomeopenres.14694.1 (2018).

312 Duchêne, S., Holt, K. E., Weill, F.-X., Le Hello, S., Hawkey, J., Edwards, D. J., Fourment, M. & Holmes, E. C. Genome-scale rates of evolutionary change in bacteria. *Microb. Genom.* **2**, e000094-e000094, doi:10.1099/mgen.0.000094 (2016).

313 Yuan, M., Chen, H., Zhu, X., Feng, J., Zhan, Z., Zhang, D., Chen, X., Zhao, X., Lu, J., Xu, J., Zhou, D. & Li, J. pSY153-MDR, a p12969-DIM-related mega plasmid carrying *bla(IMP-45)* and *armA*, from clinical *Pseudomonas putida*. *Oncotarget* **8**, 68439-68447, doi:10.18632/oncotarget.19496 (2017).

314 Pincus, N. B., Ozer, E. A., Allen, J. P., Nguyen, M., Davis, J. J., Winter, D. R., Chuang, C.-H., Chiu, C.-H., Zamorano, L., Oliver, A. & Hauser, A. R. A genome-based model to predict the virulence of *Pseudomonas aeruginosa* isolates. *mBio* **11**, e01527-01520, doi:10.1128/mBio.01527-20 (2020).

315 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).

316 Scheetz, M. H., Hoffman, M., Bolon, M. K., Schulert, G., Estrellado, W., Baraboutis, I. G., Sriram, P., Dinh, M., Owens, L. K. & Hauser, A. R. Morbidity associated with *Pseudomonas aeruginosa* bloodstream infections. *Diagn. Microbiol. Infect. Dis.* **64**, 311-319, doi:http://dx.doi.org/10.1016/j.diagmicrobio.2009.02.006 (2009).

317 Chuang, C.-H., Wang, Y.-H., Chang, H.-J., Chen, H.-L., Huang, Y.-C., Lin, T.-Y., Ozer, E. A., Allen, J. P., Hauser, A. R. & Chiu, C.-H. Shanghai fever: a distinct *Pseudomonas aeruginosa* enteric disease. *Gut* **63**, 736-743, doi:10.1136/gutjnl-2013-304786 (2014).

318 Treggiari, M. M., Retsch-Bogart, G., Mayer-Hamblett, N., Khan, U., Kulich, M., Kronmal, R., Williams, J., Hiatt, P., Gibson, R. L., Spencer, T., Orenstein, D., Chatfield, B. A., Froh, D. K., Burns, J. L., Rosenfeld, M., Ramsey, B. W. & Investigators, f. t. E. P. I. C. Comparative Efficacy and Safety of 4 Randomized Regimens to Treat Early Pseudomonas aeruginosa Infection in Children With Cystic Fibrosis. *Arch. Pediatr. Adolesc. Med.* **165**, 847-856, doi:10.1001/archpediatrics.2011.136 (2011).

319 Treggiari, M. M., Rosenfeld, M., Mayer-Hamblett, N., Retsch-Bogart, G., Gibson, R. L., Williams, J., Emerson, J., Kronmal, R. A., Ramsey, B. W. & Group, E. S. Early anti-pseudomonal acquisition in young patients with cystic fibrosis: rationale and design of the EPIC clinical trial and observational study'. *Contemp. Clin. Trials* **30**, 256-268, doi:10.1016/j.cct.2009.01.003 (2009).

320 Feng, W., Huang, Q., Wang, Y., Yuan, Q., Li, X., Xia, P. & Sun, F. Changes in the resistance and epidemiological characteristics of Pseudomonas aeruginosa during a ten-year period. *J. Microbiol. Immunol. Infect.*, doi:https://doi.org/10.1016/j.jmii.2019.08.017 (2019).

321 McCracken, M. G., Adam, H. J., Blondeau, J. M., Walkty, A. J., Karlowsky, J. A., Hoban, D. J., Zhanel, G. G., Mulvey, M. R., Canadian Antimicrobial Resistance, A. & Canward. Characterization of carbapenem-resistant and XDR *Pseudomonas aeruginosa* in Canada: results of the CANWARD 2007–16 study. *J. Antimicrob. Chemother.* **74**, iv32-iv38, doi:10.1093/jac/dkz285 (2019).

322 Moloney, E. M., Deasy, E. C., Swan, J. S., Brennan, G. I., O'Donnell, M. J. & Coleman, D. C. Whole-genome sequencing identifies highly related Pseudomonas aeruginosa strains in multiple washbasin U-bends at several locations in one hospital: evidence for trafficking of potential pathogens via wastewater pipes. *J. Hosp. Infect.* **104**, 484-491, doi:https://doi.org/10.1016/j.jhin.2019.11.005 (2020).

323 Saharman, Y. R., Pelegrin, A. C., Karuniawati, A., Sedono, R., Aditianingsih, D., Goessens, W. H. F., Klaassen, C. H. W., van Belkum, A., Mirande, C., Verbrugh, H. A. & Severin, J. A. Epidemiology and characterisation of carbapenem-non-susceptible *Pseudomonas aeruginosa* in a large intensive care unit in Jakarta, Indonesia. *Int. J. Antimicrob. Agents* **54**, 655-660, doi:https://doi.org/10.1016/j.ijantimicag.2019.08.003 (2019).

324 Tada, T., Hishinuma, T., Watanabe, S., Uchida, H., Tohya, M., Kuwahara-Arai, K., Mya, S., Zan, K. N., Kirikae, T. & Tin, H. H. Molecular characterization of multidrug-resistant *Pseudomonas aeruginosa* isolates in hospitals in Myanmar. *Antimicrob. Agents Chemother.* **63**, e02397-02318, doi:10.1128/AAC.02397-18 (2019).

325 Yoon, E.-J., Kim, D., Lee, H., Lee, H. S., Shin, J. H., Park, Y. S., Kim, Y. A., Shin, J. H., Shin, K. S., Uh, Y. & Jeong, S. H. Mortality dynamics of *Pseudomonas aeruginosa* bloodstream infections and the influence of defective OprD on mortality: prospective

observational study. *J. Antimicrob. Chemother.* **74**, 2774-2783, doi:10.1093/jac/dkz245 (2019).

326 Moghadam, S. O., Afshar, D., Nowroozi, M. R., Behnamfar, A. & Farzin, A. Molecular Epidemiology of Carbapenemase-Producing Pseudomonas aeruginosa Isolated from an Iranian University Hospital: Evidence for Spread of High-Risk Clones. *Infect. Drug Resist.* **13**, 1583-1592 (2020).

327 Chng, K. R., Li, C., Bertrand, D., Ng, A. H. Q., Kwah, J. S., Low, H. M., Tong, C., Natrajan, M., Zhang, M. H., Xu, L., Ko, K. K. K., Ho, E. X. P., Av-Shalom, T. V., Teo, J. W. P., Khor, C. C., Danko, D., Bezdan, D., Afshinnekoo, E., Ahsanuddin, S., Bhattacharya, C., Butler, D. J., Chng, K. R., De Filippis, F., Hecht, J., Kahles, A., Karasikov, M., Kyrpides, N. C., Leung, M. H. Y., Meleshko, D., Mustafa, H., Mutai, B., Neches, R. Y., Ng, A., Nieto-Caballero, M., Nikolayeva, O., Nikolayeva, T., Png, E., Sanchez, J. L., Shaaban, H., Sierra, M. A., Tong, X., Young, B., Alicea, J., Bhattacharyya, M., Blekhman, R., Castro-Nallar, E., Cañas, A. M., Chatziefthimiou, A. D., Crawford, R. W., Deng, Y., Desnues, C., Dias-Neto, E., Donnellan, D., Dybwad, M., Elhaik, E., Ercolini, D., Frolova, A., Graf, A. B., Green, D. C., Hajirasouliha, I., Hernandez, M., Iraola, G., Jang, S., Jones, A., Kelly, F. J., Knights, K., Łabaj, P. P., Lee, P. K. H., Shawn, L., Ljungdahl, P., Lyons, A., Mason-Buck, G., McGrath, K., Mongodin, E. F., Moraes, M. O., Nagarajan, N., Noushmehr, H., Oliveira, M., Ossowski, S., Osuolale, O. O., Özcan, O., Paez-Espino, D., Rascovan, N., Richard, H., Rätsch, G., Schriml, L. M., Semmler, T., Sezerman, O. U., Shi, L., Song, L. H., Suzuki, H., Court, D. S., Thomas, D., Tighe, S. W., Udekwu, K. I., Ugalde, J. A., Valentine, B., Vassilev, D. I., Vayndorf, E., Velavan, T. P., Zambrano, M. M., Zhu, J., Zhu, S., Mason, C. E., Chen, S. L., Mason, C. E., Ng, O. T., Marimuthu, K., Ang, B., Nagarajan, N. & Consortium, M. S. U. B. Cartography of opportunistic pathogens and antibiotic resistance genes in a tertiary hospital environment. *Nat. Med.* **26**, 941-951, doi:10.1038/s41591-020-0894-4 (2020).

328 Snitkin, E. S., Won, S., Pirani, A., Lapp, Z., Weinstein, R. A., Lolans, K. & Hayden, M. K. Integrated genomic and interfacility patient-transfer data reveal the transmission pathways of multidrug-resistant *Klebsiella pneumoniae* in a regional outbreak. *Sci. Transl. Med.* **9**, eaan0093, doi:10.1126/scitranslmed.aan0093 (2017).

329 Moritz, M. M., Flemming, H.-C. & Wingender, J. Integration of *Pseudomonas aeruginosa* and *Legionella pneumophila* in drinking water biofilms grown on domestic plumbing materials. *Int. J. Hyg. Environ. Health* **213**, 190-197, doi:https://doi.org/10.1016/j.ijheh.2010.05.003 (2010).

330 Johnson, R. C., Deming, C., Conlan, S., Zellmer, C. J., Michelin, A. V., Lee-Lin, S., Thomas, P. J., Park, M., Weingarten, R. A., Less, J., Dekker, J. P., Frank, K. M., Musser, K. A., McQuiston, J. R., Henderson, D. K., Lau, A. F., Palmore, T. N. & Segre, J. A. Investigation of a cluster of *Sphingomonas koreensis* infections. *N. Engl. J. Med.* **379**, 2529-2539, doi:10.1056/NEJMoa1803238 (2018).

331 Cazares, A., Moore, M. P., Hall, J. P. J., Wright, L. L., Grimes, M., Emond-Rhéault, J.-G., Pongchaikul, P., Santanirand, P., Levesque, R. C., Fothergill, J. L. & Winstanley, C. A megaplasmid family driving dissemination of multidrug resistance in *Pseudomonas*. *Nat. Commun.* **11**, 1370, doi:10.1038/s41467-020-15081-7 (2020).

332 Craig, L., Pique, M. E. & Tainer, J. A. Type IV pilus structure and bacterial pathogenicity. *Nat. Rev. Microbiol.* **2**, 363-378, doi:10.1038/nrmicro885 (2004).

333 van Schaik, E. J., Giltner, C. L., Audette, G. F., Keizer, D. W., Bautista, D. L., Slupsky, C. M., Sykes, B. D. & Irvin, R. T. DNA binding: a novel function of *Pseudomonas aeruginosa* type IV pili. *J. Bacteriol.* **187**, 1455, doi:10.1128/JB.187.4.1455-1464.2005 (2005).

334 Paetzel, M., Danel, F., de Castro, L., Mosimann, S. C., Page, M. G. P. & Strynadka, N. C. J. Crystal structure of the class D β-lactamase OXA-10. *Nat. Struct. Biol.* **7**, 918 (2000).

335 Yoshida, T., Furuya, N., Ishikura, M., Isobe, T., Haino-Fukushima, K., Ogawa, T. & Komano, T. Purification and characterization of thin pili of IncI1 plasmids ColIb-P9 and R64: formation of PilV-specific cell aggregates by type IV pili. *J. Bacteriol.* **180**, 2842, doi:10.1128/JB.180.11.2842-2848.1998 (1998).

336 Mather, M. W. & Fee, J. A. Plasmid-associated aggregation in *Thermus thermophilus* HB8. *Plasmid* **24**, 45-56, doi:https://doi.org/10.1016/0147-619X(90)90024-7 (1990).

337 Jensen, G. B., Wilcks, A., Petersen, S. S., Damgaard, J., Baum, J. A. & Andrup, L. The genetic basis of the aggregation system in *Bacillus thuringiensis subsp. israelensis* is located on the large conjugative plasmid pXO16. *J. Bacteriol.* **177**, 2914, doi:10.1128/jb.177.10.2914-2917.1995 (1995).

338 Harrison, E., Guymer, D., Spiers, Andrew J., Paterson, S. & Brockhurst, Michael A. Parallel compensatory evolution stabilizes plasmids across the parasitism-mutualism continuum. *Curr. Biol.* **25**, 2034-2039, doi:https://doi.org/10.1016/j.cub.2015.06.024 (2015).

339 Loftie-Eaton, W., Bashford, K., Quinn, H., Dong, K., Millstein, J., Hunter, S., Thomason, M. K., Merrikh, H., Ponciano, J. M. & Top, E. M. Compensatory mutations improve general permissiveness to antibiotic resistance plasmids. *Nat. Ecol. Evol.* **1**, 1354-1363, doi:10.1038/s41559-017-0243-2 (2017).

340 Cheng, L., Connor, T. R., Sirén, J., Aanensen, D. M. & Corander, J. Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. *Mol. Biol. Evol.* **30**, 1224-1228, doi:10.1093/molbev/mst028 (2013).

341 Jaillard, M., Lima, L., Tournoud, M., Mahé, P., van Belkum, A., Lacroix, V. & Jacob, L. A fast and agnostic method for bacterial genome-wide association studies: Bridging the gap between k-mers and genetic events. *PLoS Genet.* **14**, e1007758, doi:10.1371/journal.pgen.1007758 (2018).

342 Shi, J., Yan, Y., Links, M. G., Li, L., Dillon, J.-A. R., Horsch, M. & Kusalik, A. Antimicrobial resistance genetic factor identification from whole-genome sequence data using deep feature selection. *BMC Bioinformatics* **20**, 535, doi:10.1186/s12859-019-3054-4 (2019).

343 Zhou, C. E., Smith, J., Lam, M., Zemla, A., Dyer, M. D. & Slezak, T. MvirDB--a microbial database of protein toxins, virulence factors and antibiotic resistance genes for bio-defence applications. *Nucleic Acids Res.* **35**, D391-D394, doi:10.1093/nar/gkl791 (2007).

344 Jehl, M.-A., Arnold, R. & Rattei, T. Effective--a database of predicted secreted bacterial proteins. *Nucleic Acids Res.* **39**, D591-D595, doi:10.1093/nar/gkq1154 (2011).

345 Power, R. A., Parkhill, J. & de Oliveira, T. Microbial genome-wide association studies: lessons from human GWAS. *Nat. Rev. Genet.* **18**, 41-50, doi:10.1038/nrg.2016.132 (2017).

346 Earle, S. G., Wu, C.-H., Charlesworth, J., Stoesser, N., Gordon, N. C., Walker, T. M., Spencer, C. C. A., Iqbal, Z., Clifton, D. A., Hopkins, K. L., Woodford, N., Smith, E. G., Ismail, N., Llewelyn, M. J., Peto, T. E., Crook, D. W., McVean, G., Walker, A. S. & Wilson, D. J. Identifying lineage effects when controlling for population structure improves power in bacterial association studies. *Nat. Microbiol.* **1**, 16041, doi:10.1038/nmicrobiol.2016.41 (2016).

347 Collins, C. & Didelot, X. A phylogenetic method to perform genome-wide association studies in microbes that accounts for population structure and recombination. *PLoS Comput. Biol.* **14**, e1005958, doi:10.1371/journal.pcbi.1005958 (2018).

348 Lees, J. A., Galardini, M., Bentley, S. D., Weiser, J. N. & Corander, J. pyseer: a comprehensive tool for microbial pangenome-wide association studies. *Bioinformatics* **34**, 4310-4312, doi:10.1093/bioinformatics/bty539 (2018).

349 Saund, K. & Snitkin, E. S. hogwash: three methods for genome-wide association studies in bacteria. *bioRxiv*, 2020.2004.2019.048421, doi:10.1101/2020.04.19.048421 (2020).

350 Young, B. C., Earle, S. G., Soeng, S., Sar, P., Kumar, V., Hor, S., Sar, V., Bousfield, R., Sanderson, N. D., Barker, L., Stoesser, N., Emary, K. R. W., Parry, C. M., Nickerson, E. K., Turner, P., Bowden, R., Crook, D. W., Wyllie, D. H., Day, N. P. J., Wilson, D. J. & Moore, C. E. Panton–Valentine leucocidin is the key determinant of *Staphylococcus aureus* pyomyositis in a bacterial GWAS. *eLife* **8**, e42486, doi:10.7554/eLife.42486 (2019).

351 Berthenet, E., Yahara, K., Thorell, K., Pascoe, B., Meric, G., Mikhail, J. M., Engstrand, L., Enroth, H., Burette, A., Megraud, F., Varon, C., Atherton, J. C., Smith, S., Wilkinson, T. S., Hitchings, M. D., Falush, D. & Sheppard, S. K. A GWAS on *Helicobacter pylori* strains points to genetic variants associated with gastric cancer risk. *BMC Biol.* **16**, 84, doi:10.1186/s12915-018-0550-3 (2018).

352 Marvig, R. L., Sommer, L. M., Molin, S. & Johansen, H. K. Convergent evolution and adaptation of *Pseudomonas aeruginosa* within patients with cystic fibrosis. *Nat. Genet.* **47**, 57-64, doi:10.1038/ng.3148 http://www.nature.com/ng/journal/v47/n1/abs/ng.3148.html#supplementary-information (2015).

353 Cheng, K., Smyth, R. L., Govan, J. R. W., Doherty, C., Winstanley, C., Denning, N., Heaf, D. P., van Saene, H. & Hart, C. A. Spread of β-lactam-resistant *Pseudomonas aeruginosa* in a cystic fibrosis clinic. *The Lancet* **348**, 639-642, doi:https://doi.org/10.1016/S0140-6736(96)05169-0 (1996).

354 Jelsbak, L., Johansen, H. K., Frost, A.-L., Thøgersen, R., Thomsen, L. E., Ciofu, O., Yang, L., Haagensen, J. A. J., Høiby, N. & Molin, S. Molecular epidemiology and dynamics of *Pseudomonas aeruginosa* populations in lungs of cystic fibrosis patients. *Infect. Immun.* **75**, 2214, doi:10.1128/IAI.01282-06 (2007).

355 Wiegand, I., Hilpert, K. & Hancock, R. E. W. Agar and broth dilution methods to determine the minimal inhibitory concentration (MIC) of antimicrobial substances. *Nat. Protoc.* **3**, 163, doi:10.1038/nprot.2007.521 (2008).

356 Francisco, A. P., Bugalho, M., Ramirez, M. & Carriço, J. A. Global optimal eBURST analysis of multilocus typing data using a graphic matroid approach. *BMC Bioinformatics* **10**, 152, doi:10.1186/1471-2105-10-152 (2009).

357 Nascimento, M., Sousa, A., Ramirez, M., Francisco, A. P., Carriço, J. A. & Vaz, C. PHYLOViZ 2.0: providing scalable data integration and visualization for multiple

phylogenetic inference methods. *Bioinformatics* **33**, 128-129, doi:10.1093/bioinformatics/btw582 (2016).

358 Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H. & Phillippy, A. M. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27**, 722-736, doi:10.1101/gr.215087.116 (2017).

359 Myers, E. W., Sutton, G. G., Delcher, A. L., Dew, I. M., Fasulo, D. P., Flanigan, M. J., Kravitz, S. A., Mobarry, C. M., Reinert, K. H. J., Remington, K. A., Anson, E. L., Bolanos, R. A., Chou, H.-H., Jordan, C. M., Halpern, A. L., Lonardi, S., Beasley, E. M., Brandon, R. C., Chen, L., Dunn, P. J., Lai, Z., Liang, Y., Nusskern, D. R., Zhan, M., Zhang, Q., Zheng, X., Rubin, G. M., Adams, M. D. & Venter, J. C. A whole-genome assembly of *Drosophila*. *Science* **287**, 2196, doi:10.1126/science.287.5461.2196 (2000).

360 Hunt, M., Silva, N. D., Otto, T. D., Parkhill, J., Keane, J. A. & Harris, S. R. Circlator: automated circularization of genome assemblies using long sequencing reads. *Genome Biol.* **16**, 294, doi:10.1186/s13059-015-0849-0 (2015).

361 Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S. K. & Earl, A. M. Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, e112963, doi:10.1371/journal.pone.0112963 (2014).

362 Tatusova, T., DiCuccio, M., Badretdin, A., Chetvernin, V., Nawrocki, E. P., Zaslavsky, L., Lomsadze, A., Pruitt, K. D., Borodovsky, M. & Ostell, J. NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Res.* **44**, 6614-6624, doi:10.1093/nar/gkw569 (2016).

363 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).

364 Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A. & Pevzner, P. A. 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).

365 Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987-2993, doi:10.1093/bioinformatics/btr509 (2011).

366 Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*, 1303.3997 (2013).

367 Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C. & Salzberg, S. L. Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12, doi:10.1186/gb-2004-5-2-r12 (2004).

368 Milne, I., Stephen, G., Bayer, M., Cock, P. J. A., Pritchard, L., Cardle, L., Shaw, P. D. & Marshall, D. Using Tablet for visual exploration of second-generation sequencing data. *Brief. Bioinform.* **14**, 193-202, doi:10.1093/bib/bbs012 (2012).

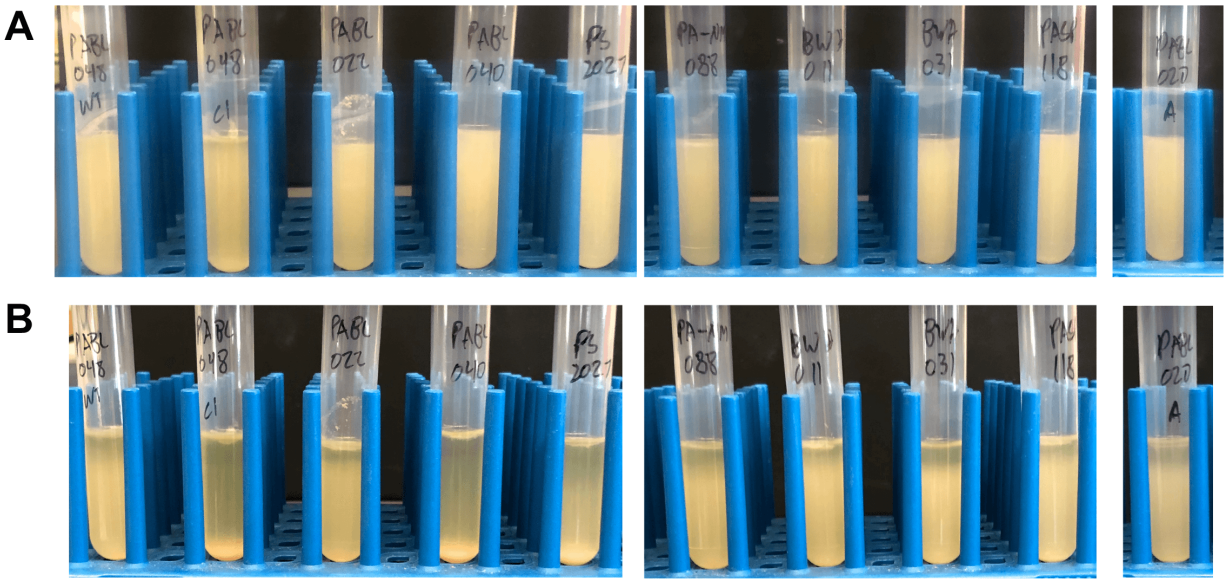
369 Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312-1313, doi:10.1093/bioinformatics/btu033 (2014).

370 Didelot, X. & Wilson, D. J. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. *PLoS Comput. Biol.* **11**, e1004041, doi:10.1371/journal.pcbi.1004041 (2015).

371 Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* **47**, W256-W259, doi:10.1093/nar/gkz239 (2019).
372 R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing, Vienna, Austria, 2019).
373 Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**, 289-290, doi:10.1093/bioinformatics/btg412 (2004).
374 R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing, Vienna, Austria, 2016).
375 Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., Suchard, M. A., Rambaut, A. & Drummond, A. J. BEAST 2: a software platform for bayesian evolutionary analysis. *PLoS Comput. Biol.* **10**, e1003537, doi:10.1371/journal.pcbi.1003537 (2014).
376 Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. Posterior summarization in bayesian phylogenetics using Tracer 1.7. *Syst. Biol.* **67**, 901-904, doi:10.1093/sysbio/syy032 (2018).
377 Trevors, J. T. Plasmid curing in bacteria. *FEMS Microbiol. Lett.* **32**, 149-157, doi:10.1111/j.1574-6968.1986.tb01189.x (1986).
378 Moura, A., Correia, A., Pereira, C., Henriques, I., Soares, M. & Leitão, N. INTEGRALL: a database and search engine for integrons, integrases and gene cassettes. *Bioinformatics* **25**, 1096-1098, doi:10.1093/bioinformatics/btp105 (2009).
379 Nicas, T. I. & Iglewski, B. H. Isolation and characterization of transposon-induced mutants of *Pseudomonas aeruginosa* deficient in production of exoenzyme S. *Infect. Immun.* **45**, 470 (1984).
380 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).
381 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).
382 Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068-2069, doi:10.1093/bioinformatics/btu153 (2014).
383 Lê, S., Josse, J. & Husson, F. FactoMineR: An R package for multivariate analysis. *J. Stat. Softw.* **25**, 18, doi:10.18637/jss.v025.i01 (2008).
384 Kokot, M., Długosz, M. & Deorowicz, S. KMC 3: counting and manipulating k-mer statistics. *Bioinformatics* **33**, 2759-2761, doi:10.1093/bioinformatics/btx304 (2017).
385 Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K. & Yutani, H. Welcome to the Tidyverse. *J. Open Source Softw.* **4**, 1686 (2019).

Appendix I

Precipitation phenotype in some ST298* isolates



Isolates tested include (from left to right) PABL048 (ST298*, plasmid+), PABL048-c1 (ST298*, plasmid-), PABL022 (ST298, non-ST298*), PABL040 (ST298*, plasmid-), PS2027 (ST298*, plasmid-), PA-NM-088 (ST298*, plasmid+), BWH011 (ST298, non-ST298*), BWH031 (ST298, non-ST298*), PASP118 (ST298, non-ST298*), and PABL020 (ST298*, plasmid-). Isolates were inoculated from LB agar into 5 mL LB broth and incubated overnight with shaking at 37 ºC. Cultures were photographed after removal from the incubator (A) and after leaving at room temperature without shaking for approximately 50 hours (B). ST298* isolates containing the plasmid (PABL048 and PA-NM) remained largely suspended in the media, while ST298* isolates lacking the plasmid (PABL048-c1, PABL040, and PS2027) precipitated resulting in an appreciable pellet. ST298* isolate PABL020 did not precipitate even though it does not possess sequence aligning to the majority of pPABL048. The non-ST298* isolates tested remained largely suspended in the media.