Identifying genetic determinants of complex phenotypes from whole genome sequence data and very small training sets

George S. Long[1], Mohammed Hussen[1], Jonathan Dench[1], Stéphane Aris-Brosou,[1,2⋆]

[1]Department of Biology, University of Ottawa, Ottawa, Ontario, Canada

[2]Department of Mathematics and Statistics, University of Ottawa, Ottawa, Ontario, Canada

⋆Corresponding author

Email: sarisbro@uottawa.ca (SAB)

Abstract

Machine learning (ML) encompasses numerous algorithms that aim at discovering complex patterns between elements within large data using limited prior assumptions or modeling. However, some scientific disciplines still produce small data sets: in particular, empirical studies that try to find the mutations responsible for complex phenotypes are often limited to very small sample sizes (n), while scanning a large number of amino acid sites (p) in a proteome. To date, little is known on how ML performs in this type of so-called "large p, small n" problem. To address this question, we evaluated the performance of two general ML classifiers, adaptive boosting (AB) and random forest, on two data sets. To assess the impact of proteome size, we contrasted a small (viral) genome with a larger (bacterial) one. To analyze large proteomes, we further developed a chunking algorithm, and introduce a repeated random forest (RRF) algorithm that stabilizes model predictions. With the influenza data, we were able to rediscover amino acid sites experimentally implicated in three different complex phenotypes (infectivity, transmissibility, and pathogenicity). Results for the larger proteome, pertaining to three types of drug resistance (Ciprofloxacin, Ceftazidime, and Gentamicin), were more nuanced, with RRF making more sensible predictions, with smaller errors rates, than AB. Furthermore, we show that chunking improved runtimes by an order of magnitude and may increase sensitivity of the predictions. Altogether, we demonstrate that ML algorithms can be used to identify genetic determinants in small proteomes (viruses), even with small numbers of individuals. We further show that even if the size of bacterial proteomes pushes AB to its limits in the context of small n, RRF may deserve more scrutiny, which should be facilitated by the plummeting costs of sequencing and, more critically, by phenotyping large cohorts of individuals.

Keywords: influenza virus; *Pseudomonas aeruginosa*; machine learning; genome-wide association study; drug resistance

Author Summary

Finding the genetic determinants of a phenotype is typically performed by testing for an association between a particular allele and a trait, carrying out the testing over a large number of loci in a large cohort of individuals, itself divided into two subsets of individuals: those who have the trait (cases), and those who do not (controls). However, recruiting large cohorts can be problematic in some experimental fields, while using genotypic information rather than complete genomes can miss some mutations. To address these issues, we implemented two machine learning (ML) algorithms, tweaked for analyzing large genomes and providing stable results. The analysis of a small viral genome, for which genetic determinants of three phenotypes are already known, showed that our approach can rediscover known mutations, almost irrespective of the ML algorithm used. However, the analysis of a larger bacterial genome, for which genetic determinants of three phenotypes are unknown, suggested that the simpler of our modified algorithms performed better, returning more sensitive predictions with lower error rates. This work demonstrates the feasibility of finding genetic determinants of complex phenotypes based on a small number of complete genomes.

Introduction

An overarching goal in biology is to predict an individual's phenotype from its genotype, in a given environment [1], or from a given genetic makeup [2]. An obvious first step in this direction is to find the genetic determinants of each phenotype of interest – which is what genome-wide association studies (GWAS's) have endeavored to achieve over the past ten years [3]. At their foundation, GWAS's rely on the analysis of millions of variants in the genome, without any prior knowledge of their involvement with a particular phenotype, over a sample of unrelated individuals: as such, GWAS's are often qualified of performing an "unbiased scan of the genome" [4]. While GWAS's have some limitations that may be shared by alternative approaches (they assume that common diseases are caused by common variant [4], which may result in failing to explain most of the phenotypic variance [5]), their main issue may lie in their reliance on single nucleotide polymorphism (SNP) data, that only offer a partial snapshot of the genome, and that may not even contain the causative agents of the phenotype under study. The immediate alternative would be whole-genome sequencing (WGS) [6], predicted to replace SNP genotyping [3]. However, changing from focal SNP's to whole genome information will radically increase the number of tests to be performed, and might lead to statistical complications in controlling false discovery, as acknowledged at the very outset of GWAS [7]. Recently, it was suggested that machine learning could be used to predict susceptibility to some cancers in humans [8]. Following up on these steps, we here hypothesize that machine learning is able to go beyond predicting genomic inheritance within a cohort of individuals, by predicting the genetic determinants of particular phenotypes using WGS data.

A large number of machine learning approaches exist [9], and are gaining popularity in biology [10, 11]. For instance, in the work just cited above, the authors resorted to neural networks [8]. While some statisticians claim that adaptive boosting [12] is the "best off-the-shelf classifier in the world" [13, 9], it is probable that no single classifier can be considered perfect in all situations – a situation known as the *no free lunch theorem* [14]. Technically, adaptive boosting (AB) relies on an iterated process where linear decisions are fitted in the space of genomic features. At each iteration, individual observations that were misclassified in the previous iteration are emphasized, so that the algorithm learns from past errors. While each iteration typically leads to a weak classifier that is just a bit better than chance, the final classifier takes advantage of combining these weak classifiers to improve (*boost*) their performance and construct a strong one [12], *i.e.* a classifier with an accuracy that can come close to 99% [15]. While

our recent success with AB [16] prompted us to further investigate this algorithm in the context of WGS data, we also aimed at comparing its performance with more popular algorithms, such as random forests [17] (RF). This algorithm is essentially based on decision trees [18], combined with a bootstrap procedure (bagging) aimed at increasing the stability of the predictions, and a random subsetting of predictors to decorrelate the bagged trees, that are then averaged to produce the final classifier.

As any supervised learning algorithm, AB and RF are trained on a labeled data set, *i.e.* one for which correct assignments are known. For instance, consider some WGS data coming from an organism for which drug resistance (the phenotype) is known. We wish to predict (classify) drug resistant individuals, given their genomic features (*i.e.*, mutations at particular sites). This trained classifier can then be used to predict drug resistance in new individuals, based solely on their genomic information. During the learning process, features (again, these are the mutations at particular sites) are ranked by decreasing importance to fit the final model [19]. The use of importance values as a method of feature selection is intrinsic to the boosting algorithm [20]. The algorithm weighs the influence of each feature based on their relative importance during the creation of the model [20], and thus uses only the most important sites. The end result is that we have both a model for predicting phenotype, and a means for identifying a ranked list of the most important features [10] for phenotype prediction. A second advantage of boosting algorithms is their accuracy in the case of high-dimensional data, such as microarray [21] or SNP studies [22]. In these contexts, the data are typically large matrices that contain measurements (light intensities in the case of microarray studies) for thousands of genes (features), assayed in hundreds of individuals (samples). Due to the potential for many features to have only a small effect [23], one would predict that adaptive boosting has a better predictive ability than random forests. However, studies that aim at predicting the genetic determinants of particular phenotypes based on boosting typically have small sample sizes (but see [3] for what the future may hold). Indeed, as technology allows us to increase the number of features on which to train these algorithms (number of SNP ascertained / genome size in WGS), accurately phenotyping the sequenced individuals is still challenging [24]. This sets our goal of using machine learning to predict the genetic determinants of phenotypes in the particular light of what is known as the "large p, small n" problem, where p stands for the number of features (parameters; here, the genome of an organism), and n is the sample size, that is, the number of individuals that have been both phenotyped and sequenced. This context emerged at the turn of the century with the development of microarray technology – and was already addressed at that time by resorting to machine learning [25].

However, it is still unclear whether similar approaches can be used with phenotyping / WGS data, and what kind of limitations may exist.

In order to better understand these limitations, we evaluate the ability of two classifiers, AB and RF, to predict the genetic determinants of particular phenotypes in two microbes with very different genome sizes (large p): the Influenza A virus, that has about a dozen protein-coding genes, and the *Pseudomonas aeruginosa* bacterium, that has $\sim 6,000$ genes. In both cases, we retrieved complete genome sequences of phenotyped individuals. These phenotypes pertained either to their capacity to infect a human host (influenza), or their resistance to particular antibiotics (pseudomonas). Although these two biological systems are very different, both are expected to have a highly complex genetic basis: the molecular determinants of influenza virulence and pathogenesis can span its entire genome [26, 27], and bacterial drug resistance can involve many distinct mechanisms [28]. Here we aim at pushing the limits of these algorithms, and assessing their ability to find the determinants of such complex phenotypes. In both the viral and the bacterial cases, classifiers of three discrete phenotypes were fitted to genomic data. Due to the paucity of data where individuals are both phenotyped and sequenced, only small sample sizes are available (small n). We take advantage of an influenza database backed by experimental validations [29], and a recent study of *P. aeruginosa* genomics [30], to train both the adaptive boosting and the random forest algorithms. We modify both original algorithms to make them amenable to analyzing WGS-sized data sets, and alter the RF algorithm to further stabilize its predictions by introducing a Repeated Random Forest (RRF) algorithm. We then evaluate the performance of these algorithms with respect to either experimental validations (influenza), or both gene annotations and cross-validations (pseudomonas). We discuss the advantages and limitations of our modified algorithms in this context of whole genomes (large p) and relatively small sample sizes (n).

Results and Discussion

Determination of the RRF thresholds on the influenza data

The influenza data represent our gold standard here, as we can run our machine learning algorithms on them, and compare the model predictions with experimental evidence (Table 1). Because both algorithms (AB and RRF) essentially return a list of sites by decreasing importance, we can use the influenza data to

determine optimal thresholds for predicting genuine genetic determinants. Two thresholds were developed for the RRF algorithm: the percentile and the consensus thresholds (AB only has the former). This was done by maximizing the number of true positives, *i.e.* sites that (i) have high mean Gini index [our "percentile threshold"] and (ii) sites that are found in at least a certain number of RF runs of the RRF [our "consensus threshold"] – given of course that these sites are backed by experimental evidence (Table 1). By varying both thresholds simultaneously, we determined that using the 90th percentile (top 10% important sites) and consensus threshold around 50% maximized the number of true positives while minimizing the false positives (Figures S1-S10). Increasing the percentile threshold to 95% dramatically increased false negatives, while decreasing it to the 85th percentile only led to more false positives. Likewise, we found that increasing the consensus threshold only decreased the proportion of false positives, without affecting sensitivity. We henceforth used these two RRF thresholds (50% consensus, 90th percentile).

Chunking improves runtimes, minimally affecting site ranking

As these algorithms can have large computational requirements (AB in particular), we implemented a chunking algorithm. For this, each data set was subdivided into smaller parts (*chunks*), on which each ML algorithm was run in a first pass to determine a set of positions of interests. A second pass ran the same ML algorithm on the entire set of positions of interests to determine the final important sites over the entire data set. To determine how chunking affects both the runtimes and the predictions of the machine learning algorithms on the influenza data, 20 chunk sizes were compared, ranging from 75 (or 80 in the case of RRF) to 175 by increments of five. Being much faster than AB, the RRF algorithm was further tested beyond the initial 20 chunks, up until the sequence alignment was analyzed in full. This was repeated for each of the three influenza phenotypes. As expected from its associated increased memory requirements, increasing chunk sizes also increases runtimes exponentially for both AB (Figure 1A) and RRF (Figure 1B). Smaller chunk sizes can reduce runtime by about an order of magnitude (1 $\log_{10}$ unit) for both AB and RF. More specifically, an ANCOVA showed that runtimes are similarly affected across all phenotypes in AB ($P = 0.855$). However, while the slopes of these regressions are similar, the analyses for the pathogenicity phenotype ran the slowest (intercept: 3.85; $P = 1.12 \times 10^{-5}$), followed by transmissibility (intercept: 2.68; $P = 9.26 \times 10^{-7}$) and infectivity (intercept: 2.35; $P = 4.61 \times 10^{-6}$). Similar results were observed with RRF, where again all slopes were similar ($P = 0.973$), pathogenicity ran

the slowest (intercept: 1.175, $P < 2 \times 10^{-16}$), followed by infectivity (intercept: 1.174, $P < 2 \times 10^{-16}$) and transmissibility (intercept: 1.168, $P < 2 \times 10^{-16}$). As the feature space is the same (the same amino acid alignment), it must be the distribution of the phenotypes (*labels*) that impacted runtimes. However, base runtimes (*i.e.*, their intercepts) do not increase due to class imbalance, as the most imbalanced phenotype (transmissibility; Table 5) has an intermediate intercept (Figure 1A). Finally, note that RRF was actually so fast at small chunk sizes that there was a significant overhead associated with our parallelization of the algorithm, as linear models fitted to only chunk sizes less than 200 were highly significant ($P < 2 \times 10^{-16}$), with negative slopes (Figure 1B).

Not only were runtimes significantly and similarly reduced by chunking across phenotypes and algorithms, but differences in terms of which amino acids were predicted to be the most important were also similar (Figures 2-3; see Venn diagrams in insets, and distributions of importance values for the largest chunk size). In the case of AB, the six most important sites determining infectivity at the smallest chunk size (75) were among the top fifteen sites at intermediate (125) and largest (175) chunk sizes (Figure 2A). Furthermore, the top site, HA 108, was the most important at all of these chunk sizes, while PB2 627 and 667 were always ranked second or third. The RRF results showed a similar pattern in terms of which sites were the most important (Figure 3). In the case of infectivity, the four top sites (HA 108, HA 536, NA 84 and NA 340) were consistently the most important. After the fourth site, a large drop in importance was observed, suggesting that limited information was available at those sites. But while this drop was also observed for transmissibility at extreme chunk sizes, it was not observed anywhere else, suggesting that the first large drop in ranked importance should not be used to evaluate the relative merit of these predictions. As expected from its repeated nature, the RRF predictions were more stable than those under AB across chunk sizes (compare Venn insets in Figures 3 and 2, respectively).

One tendency that emerged when increasing chunk sizes was a decrease in sensitivity for the first pass of the algorithm, and perhaps an increase in its final specificity for both AB and RRF. In the case of infectivity for instance, the algorithms found a total of about 275 sites at chunk size 75, and roughly 150 sites at chunk size 175 (Figure 2, insets with complete distributions) for AB, and 130 sites with RRF. (Figure 3, likewise). The same pattern was observed for the two other phenotypes, transmissibility and pathogenicity. Although beyond the scope of this work, it is possible that an optimal chunk size exists in terms of area under the Receiver Operating Characteristic curve (balance between sensitivity and specificity); however, larger labeled data sets would be required for testing this possibility.

Rediscovery of influenza determinants

So far the results showed that the predictions made by AB were less stable across chunk sizes than those from RRF, but they did not tell us anything about their accuracy. While the results in Figures 2-3 suggested that more true positive sites were found by AB, with an average across chunk sizes of 5.7, 2, and 2 for each phenotype, than by RRF, with similar averages at 3.7, 1.3, and 4.3, Table 1 suggests a more nuanced view, where RRF found 44% more sites backed by experimental evidence (from nine to thirteen) than AB. Note finally that AB seemed to be better at finding sites in PB2 (the longer gene), while RRF's rediscoveries were more evenly spread. Unfortunately, with such small numbers, the significance of these results is difficult to gauge.

Focusing on each phenotype, infectivity had the largest number of sites detected, but also the largest number of experimentally validated positions (Table 1). Among those that are unambiguously associated with infectivity were PB2 667 [31] and PB2 627 [32]. PB2 9 is a determinant for mitochondrial import, which can play a role in innate immune responses to viral infection [33] and hence, infectivity. Site PB2 339 is involved in a triplet of mutations that are responsible for replication [34], which could impact infectivity [35]. Position PB1 757 is known to be involved in contact between PB2 and PB1 [36], but its role in infectivity is not clear. Specific to RRF, HA 212 is involved during the HA-mediated membrane fusion of Influenza A [37], and mutations in HA 163 cause a decrease in steric interference during the propagation of the virus [38]. Along with having the largest number of experimentally validated sites being detected, infectivity was also the phenotype with the most stable predictions among the three phenotypes. For transmissibility, HA 246 has been linked to receptor specificity and affinity [39], and PB2 391 [40] is also potentially meaningful. Likewise, mutations at PB2 105 are known to affect pathogenicity and RNAP II degradation [41], while also having a potential role in infectivity. PB1 215, is also supported by experimental evidence [42]. Site PB2 667 [31] was also shown to affect the pathogenicity of the strains. RRF also found PB1 375, which is one of the unique mutations in the 1918 influenza virus [43], PB1 215, which leads to increased the pathogenicity [42], while NS1 212 is implicated in the activation of the PI3K/Akt pathway [44].

However, a number of sites detected by our approach, sometimes with consistent high importance values (*e.g.*, HA 108 for infectivity; NA 29 for transmissibility), are, to our knowledge, not supported by any experimental evidence. As a result, it is possible that these predictions are false positives – but absence of experimental evidence is not evidence of absence. We note that with this small data set,

cross-validation could not be performed to gauge the validity of these results: by splitting the samples ten times (equivalent to a leave-one-out resampling strategy), class imbalance increased dramatically, and the probability of drawing monomorphic alignment chunks (from which the algorithm cannot recover) also increased. On the other hand, a number of key sites listed in the Influenza Research Database are also missed by our machine learning algorithms (Table 1). Furthermore, additional sites, which were not detected here, are known genetic determinants, but in other species. For instance, PB2 256 is known to increase polymerase activity, and hence boost infectivity, at least in pigs [45]. Likewise, PB2 28, 274, 526, and 607 do the same, but in birds [46]. As our alignment essentially contains sequences isolated from humans, it is possible that some of the sites we uncover are highly specific to this particular host. However, among these last five positions in PB2, only 526 was found to be polymorphic: small n also means that sampling rare mutations is almost impossible. Our results are therefore promising in that most of the known influenza sites (16 out of 20, or 80%), *i.e.* those supported by experimental evidence, were rediscovered by our algorithms.

Unimpressive performance of AB on *P. aeruginosa*

Given the encouraging performance of our predictive modeling on the influenza data, for which experimental evidence supported some of the identified genetic determinants of three complex phenotypes (infectivity, transmissibility, and pathogenicity) with a small number of strains ($n = 10$), we analyzed an alignment of previously sequenced bacteria ($n = 26$), for which we had access to MIC values for three antibiotics (Ciprofloxacin, Ceftazidime, and Gentamicin) [30]. We first employed AB, as this algorithm performed well in a recent small sample size application [16]. For each phenotype (MIC value; Figure 4, top row), we ran the AB algorithm twice with a chunk size of 5000 ("runs A"), and twice with a chunk size of 1000 ("runs B"). This allowed us to test how chuck size affects which sites are detected, and how stable this detection is across different chunk sizes. As with the influenza alignment, increasing chunk size tended to lower sensitivity in the first pass of the algorithm, as fewer positions of interest had high importance (Figure S11, insets). On the other hand, sensitivity seemed to be restored, in larger chunk sizes, after the second pass, as more sites with importance values > 1 were found for all phenotypes (Figure S11, main panels). Figure S12 shows that only the most important sites were identified with very similar importance values across the four runs (and hence the two chunk sizes). When only the top 25

sites were compared across these four runs, only five to seven sites were shared (Figure S13). Some of these predictions are sensible: DNA gyrase subunit B (gyrB) is known to be involved in drug resistance [47, 48], hypothetical protein PA14_40040 (see Ceftazidime phenotype runs) is known for its involvement in antibiotic biosynthesis processes [49], and tonB2 is inferred to be an iron transporter which potentially affects bacterial drug resistance [50]. Detecting genes that were previously unknown to be involved in drug resistance is not uncommon, as a recent study of antibiotic resistant *P. aeruginosa* found that 12% of the assayed strains carried novel genetic determinants [51]. What is more unexpected though is that some of these genes are not supposed to be involved in the mode of action of these specific drug: for instance, Ciprofloxacin, as a fluoroquinolone, is known to disrupt gyrB [48], but it is also the only drug in our results for which gyrB was not identified (Table 3). Likewise, sbrR is a factor involved in swimming ability [52], and would hence had been expected to be involved in the resistance to cephalosporins such as Ceftazidime, not the aminoglycoside Gentamicin (Table 3).

All these results were obtained with AB under one means of discretizing the MIC distributions (setting 1; Table 2). By using two other discretization schemes of the MIC distributions, it is striking that there was almost no consistency among the results (Table 3). The only genes identified under all three settings were PA14_40040, for Ceftazidime, and *tonB2* for Gentamicin. Even well-known factors such as gyrB were not identified under all three settings. Potential nonexclusive reasons for this lack of consistency include class imbalance and a very small number of strains ($n = 26$) leading to unstable training.

To better quantify general performance of AB in the case of *P. aeruginosa*, we finally performed a ten-fold cross-validation (CV) analysis on the 26 strains. The confusion matrix, which depicts the predicted number of strains in each MIC category (low / medium / high) in rows, and observed numbers in columns, showed that under setting 1, class imbalance can be quite high for resistance to the three drugs, systematically leading to one of the three MIC categories with absolutely no prediction. This can be seen *e.g.* at medium MIC values for Ceftazidime and Gentamicin (middle row):

$$Ciprofloxacin: \begin{bmatrix} 0 & 0 & 0 \\ 0 & 3 & 3 \\ 2 & 6 & 12 \end{bmatrix} Ceftazidime: \begin{bmatrix} 9 & 1 & 1 \\ 0 & 0 & 0 \\ 1 & 4 & 9 \end{bmatrix} Gentamicin: \begin{bmatrix} 9 & 1 & 2 \\ 0 & 0 & 0 \\ 5 & 1 & 10 \end{bmatrix}$$

The resulting error rates for predicting the correct MIC categories were 42.31% 30.77%, and 26.92%, respectively. As there are three discrete MIC categories, with little class imbalance, the error rate for

a random classification should be close to 67%. The AB algorithm did not have a good performance, even if it still did better than chance alone at predicting the MIC category (low / high) of an unknown bacterial strain from its proteome only. Yet, these performances explained neither the instability of the pseudomonas results, nor their lack of complete biological sensibility.

One possibility is that AB was actually overfitting the data: this happens when a particular classifier accommodates all the most minute singularities of a data set. While AB is generally considered to be robust to overfitting [53, 15], this can occur when too many iterations are performed [19] – which prompted some authors to consider m_{final} as the AB's main tuning parameter [54]. To assess whether overfitting was responsible for the poor pseudomonas results, we reran the AB analyses under setting 1 (Table 2) with different numbers of iterations ($m_{final} \in \{10, 25, 50, 100, 200\}$), both in the first and second passes of the chunking algorithm. To further assess the potential interaction with chunking, these additional analyses were run for both chunk sizes (1000 and 5000). Figure S14 shows that most analyses show a convex (concave up) CV error rate, which entails the existence of an optimal m_{final}, that allowed optimal errors rates to be as low as 25%. Different chunk sizes had different optimal m_{final}: for Ciprofloxacin, *e.g.*, the minimum CV error rate was at $m_{final} = 50$ at chunk size of 5000, but at the edge of the m_{final} interval tested for chunk size 1000. Importantly, under both chunk sizes, the only unambiguously identified gene was *trpI* (PA14_00460, Table 3, underlined), a transcriptional activator implicated in Tryptophan biosynthesis. Mutations in this gene lead to reduced (albeit modest) swimming motility [55], implicated in ciprofloxacin resistance [56]. Table 3 shows that for the two other drugs, under optimal m_{final} values, the genes identified with both chunk sizes were already among the top ten genes identified by the previous analysis, which suggests that, for a given discretization scheme of the MIC values (phenotypes), these gene lists were fairly robust to the number of iterations (m_{final}), and that overfitting was probably not an issue in this application.

RRF outperforms AB on *P. aeruginosa*

AB showed some obvious shortcomings when analyzing the pseudomonas data, some of which could be linked to our discretization of the MIC curves, in addition to large memory footprints and runtimes. RRF allowed us to address all these issues: as this algorithm ran almost four orders of magnitude faster than AB (Figure 1), it was possible to perform a more thorough search of the discretization space by varying

systematically the position of the two thresholds (θ_1, θ_2) (Figure 4A-C). The pattern of Out-of-bag error, with triangles of low errors along the diagonal (Figure 4D-F), showed that these data should be analyzed with a single MIC threshold: dissimilar (θ_1, θ_2) thresholds led to high errors; similar (θ_1, θ_2) thresholds led to low errors. This pattern was not unexpected given the coarseness of the MIC distributions (Figure 4A-C). While both Ceftazidime and Gentamicin had clear threshold minimizing the Out-of-bag error, multiple choices existed for Ciprofloxacin. To identify genetic determinants at high drug levels (the mutations conferring the highest levels of resistance), a high threshold was chosen (Figure 4A,D).

Figure 4G-I shows the most important sites identified by RRF. To determine which of these sites are potential candidates for drug resistance, we used the empirical rules derived from the analysis of influenza data, as those analyses were informed by experimental evidence. As above, we expected that most of the true positive sites (i) are found in at least 50% of the repeated parts of RRF, and (ii) are in the top 10% of the global distribution of importance (Gini) values. For influenza, these two rules allowed us to capture all but one of the experimentally-validated sites, while minimizing the number of false positives (see Figure S4). Under these two rules of thumb for site discovery, we found lists of sites (Figure 4) that are completely different from those found with AB (Table 3). However, these lists of sites were very stable across a wide range of chunk sizes (Figure S15), and their content made sense in light of what is known about these three drugs, which are commonly employed to treat *P. aeruginosa* infections of cystic fibrosis patients (Table 4). Finally, Out-of-bag errors rates were all $< 10\%$ (Figure 4), when CV error rates under AB were at least 25%. These lists of candidate genetic determinants for drug resistance are still long, with many hypothetical proteins, and hence remain problematic for experimental validation, but large and deep mutational screens have already started to revolutionize the field [57].

Conclusions

In order to find the genetic determinants of particular phenotypes in the context of a large feature space (large p) and small sample sizes (small n), we implemented and tested two machine learning algorithms based on Adaptive Boosting, which has recently proved to be quite successful when analyzing small data sets [16], and Random Forests. Our use of these machine learning algorithms can be characterized as being unbiased, in that it gauges the importance of every single position in a proteome without any *a priori* assumptions – in the same way as are often characterized GWAS [4] or RNA-seq studies [58, 59, 60].

However, because each proteome contains a very large number of positions, these algorithms could not be run 'as is' on the entire alignment, even after removing invariant positions. In regulatory genomics, where the objective is to uncover splice junctions, such machine learning algorithms typically focus on sequence windows centered on the traits of interest, thereby reducing the feature space [61, 62]. Here, we did not consider such prior knowledge (when it existed) to pre-define features, and instead took a more agnostic approach, focusing on the entire alignment of polymorphic positions. The resulting computational burden prompted us to develop the chunking algorithm, especially in the case of AB, where the initial alignment is chopped up into smaller parts or chunks. Each algorithm, AB and RRF, is run in a first pass on each chunk, to determine a set of positions of interest, which are then collated for a second pass to rank these positions by their importance in predicting a particular phenotype. Here we showed that chunking improves runtimes, without qualitatively affecting performance, both in terms of which positions are identified, and their importance values – and may even increase sensitivity of model predictions. Although our RRF algorithm may be reminiscent of the iterated RF algorithm (iRF) [63], it is less sophisticated but yet, definitely improves the stability of model predictions.

We then showed by analyzing a small genomic influenza alignment ($n = 10$ viruses for $p = 4,392$ polymorphic positions) that adaptive boosting correctly identified some positions supported by experimental evidence for three complex phenotypes [64], but that the top results also included some potential false positives, and missed some known sites. Pushing the algorithms further on the pseudomonas alignment ($n = 26$ individuals for $p = 511,780$ polymorphic positions) also allowed us to identify positions in genes known, or predicted to be, involved in the phenotypes assayed, with more sensible results obtained with RRF than with AB. This latter algorithm indeed exhibited (i) unstable and biologically unreasonable results, (ii) with cross-validation suggesting fairly high error rates – even if they were much lower than making random assignments, (iii) while overfitting could be ruled out. Even as RRF returns reasonable predictions, there is still a lot of uncertainty as to where the limits of these algorithms are: our results are, given the small sample sizes, surprisingly good when $p/n \approx 500$ (influenza), and still quite impressive when $p/n \approx 25,000$. There is a factor of 50 between these two ratios, so that more real-life data sets should be analyzed. For now, it can only be recommended that data sets with $p/n \geq 500$ be used only in preliminary or exploratory analyses. These preliminary and/or exploratory analyses may allow for a quicker discovery process of the genetic determinants of complex phenotypes. Far from being restricted to finding resistance genes and mutations, our approach is amenable to predicting any kind of

phenotype/genotype relationships, including disease-causing mutations. However, such a use of machine learning to determine phenotype/genotype relationships might only blossom when both phenotyping and genome sequencing costs (time and money) are low enough to perform these analyses on many more individuals.

Materials and Methods

The influenza data

Ten Influenza A strains were retrieved from the Influenza Research Database [29] using their search tool based on phenotype characteristics. The retrieved strains included all viral samples with experimental evidence supporting an increase of infectivity, transmissibility, and pathogenicity, the three main phenotypes in this database (retrieved Sep 2016). The detection of a polybasic cleavage site was used as a proxy for pathogenicity; while the presence of such a site alone does not indicate an increase of pathogenicity, it is nonetheless present in highly pathogenic strains [65]. The phenotypes were encoded as binary variables, since phenotypic data were only available as a Yes / No statement. Our analyses were performed blindly, as no indication of any particular mutation was included in the strain name (Table 5). Note that our selection of strains tried to minimize class imbalance, so that both infectivity and pathogenicity have a 1:1 ratio, while transmissibility data are a bit more uneven (3:7; Table 5).

The corresponding genomic data were downloaded from the Influenza Virus Resource [66]. We focused on amino acid data, assuming that phenotypic differences are caused by nonsynonymous mutations. Only the ten most common proteins found in all influenza strains (PB2, PB1, PA, HA, NP, NA, M1, M2, NS1, and NS2) were retrieved. This was done to ensure that the results obtained could be applied to the widest selection of strains possible. The selected viral strains were essentially from human hosts. This was done to prevent any potentially confounding factors from arising due to the different cellular targets [67], or due to the large sequence difference between avian and mammalian subtypes [68]. Strains containing only one segment (*e.g.*, *PA/Fort Monmouth/1/47-MA(H1N1)*) were discarded.

The segments of each strain were individually aligned with MUSCLE 3.8.31 [69] to ensure accuracy, and were then concatenated into a single alignment. Any missing segment in any strain was replaced with a row of gap characters to (i) ensure proper concatenation, and (ii) prevent a mismatching of protein

segments between the different influenza strains, and thus prevent distortion of the alignments. After the segment concatenation, invariant sites were removed from the alignment. This was done to reduce the computational time required for adaptive boosting, as sites without mutations do not contain any information relevant to the analysis.

The pseudomonas data

The proteome alignment of *P. aeruginosa* was generated by concatenating the coding regions of the 26 *P. aeruginosa* genomes previously published [30]. With a reference database of PA14 (`www.pseudomonas.com`; [70]), an alignment for each open reading frame (ORF) was created using an in-house pipeline that: (i) stored BLASTn 2.2.30 [71] results for each ORF of each non-PA14 genome, (ii) discarded any results with identity $< 90\%$, (iii) assembled alignments for each ORF ensuring a genome's sequence was used only once. This was achieved by first building a scaffold from genomic sequence with only one BLASTn result, then extending incomplete scaffolds, *i.e.*, those that did not cover the full range of their respective PA14 reference sequence. Extensions were done using BLASTn results that did not correlate with higher percent identity to another incomplete scaffold nor overlapped the current scaffold by more than 30 nucleotides. Scoring of genomic ranges and overlaps was performed using Bioconductor's function GRanges [72]). Following scaffold assembly, (iv) sequences were aligned using MUSCLE 3.8.31 [69]. Any aligned ORF with $< 50\%$ of strains having non-gap characters in at least 90% of reference sites (established via PA14 sequence) were discarded. Lastly, the remaining aligned ORFs were concatenated with the perl script `catfasta2phyml.pl` (by Johan Nylander: `https://github.com/nylander/catfasta2phyml/commit/5035eb`). This resulted in an alignment containing 5,944 of the 5,977 ORFs in the PA14 reference genome, and a total of 1,974,843 amino acid positions. Gene annotations were obtained from the file `UCBPP-PA14.csv` available at `www.pseudomonas.com/downloads/pseudomonas/pgd_r_16_2/Pseudomonas/complete/gtf-complete.tar.gz`.

Each of these 26 strains had previously been characterized phenotypically, with respect to their antibiotic resistance to three different drugs: Ciprofloxacin, Ceftazidime, and Gentamicin [30]. All three are broad-spectrum drugs, used to treat patients infected by *P. aeruginosa*, and all three belong to different families of antibiotics (Ciprofloxacin is a fluoroquinolone, Ceftazidime is a cephalosporin, and Gentamicin is an aminoglycoside). As such, each drug has a different mode of action: fluoroquinolones inhibit en-

zymes such as DNA gyrase and topoisomerase IV, involved in the replication of DNA [48]; cephalosporins interrupt the synthesis of the peptidoglycan layer forming the bacterial cell wall [73]; aminoglycosides bind to the 30s ribosomal subunit and inhibit protein synthesis [74]. Hence, different genes can be expected to be involved in the resistance to these antibiotics. Resistance had been quantified by means of minimum inhibitory concentration (MIC) assays, where a growth medium containing antibiotics is serially diluted, in two-fold steps, before an equal volume of overnight bacterial culture be inoculated into each dilution. After at least 16 hours of growth in these conditions, the MIC is defined as the minimum antibiotic concentration that does not permit growth.

Predictive modeling

Two machine learning algorithms were used to construct a model to predict each phenotype from the genomic information (proteome) in both data sets, influenza and pseudomonas. The first was AB [12], as implemented in the R package adabag [75], while the second was the RF [17] algorithm from the R package randomForest 4.6-14 [76]. All scripts were run in R 3.5 [77], and are available from `https://github.com/sarisbro`, alongside the data used.

The features included in both machine learning algorithms were the same: the amino acids of each proteome alignment. To keep track of site identity, each alignment was stored as a matrix, where column names contained the name of each ORF and the amino acid position within each ORF. As only polymorphic positions in the alignments are potentially informative, invariant sites were first discarded. This left 4,392 polymorphic positions in the influenza alignment, and 511,780 in the *P. aeruginosa* alignment.

While the AB algorithm used was unaltered, with the total number of iterations left to its default value (m_{final} = number of sites in the alignment), we slightly modified the RF algorithm. Indeed, due to the stochastic nature of this algorithm, RF can lead to different rankings of the most important features across different runs of the same model on the same data. To alleviate this issue, we ran each random forest model ten times, and combined the results of this RRF algorithm in the following manner. We rank features (amino acid sites) according to their importance (mean Gini index) across all ten replicates. A similar modification of AB could have been attempted, but was not pursued here due to large memory footprint and runtimes of this algorithm.

Indeed, one limitation of most machine learning algorithms is that they can require a large amount

of memory to run, especially in the case of data sets with large numbers of features, such as with the *P. aeruginosa* alignment. To alleviate this issue, alignments were split into sequential chunks of pre-specified sizes, ranging from 75 to 175 amino acids (by increments of five) for the influenza data, and chunk sizes being either 1000 or 5000 (AB), or ranging from 80 to 4000 (RRF) amino acids for the pseudomonas data. Note that these splits are largely random as segments of the influenza genome were "randomly" concatenated (by convention, segments are ordered by decreasing length, just like chromosomes in Eukaryotes), and protein-coding genes in pseudomonas were "randomly" concatenated (based on their order in the PA14 strain). All the computing in this step can be parallelized, so that the analysis of each chunk was forked over eight threads with the R foreach package 1.4.4 [78]. The RRF algorithm was first run on each chunk as a first pass, hereby producing a set of positions of interest. Only those with a Gini index > 0.075 were kept (first pass feature selection), collated, and each classifier was run a second time on these, to produce the final set of most important sites (*i.e.*, predictors of each phenotype with a Gini index > 0.075 – second pass feature selection).

The performance of the machine learning algorithms was assessed using ten-fold cross-validations (AB) and Out-of-bag errors (RRF). For cross-validations, the alignment was divided into ten sets of sequences (samples), nine sets being used for training and the remaining one for testing. That process was then repeated for all ten subsets [75]. In the context of our chunking procedure, cross-validation was performed on the second pass of the AB algorithm. Out-of-bag errors were computed on the predictions based on the bootstrapped trees that were not included during training.

As the evolution of antimicrobial resistance in *P. aeruginosa* can include many understudied sites [79], only the influenza data have sufficient experimental validations to which we can compare our predictions and thus determine their accuracy. In order to learn from this data set how to best balance true positives (sites that are known to be experimentally validated and are detected) and false positives (sites that are detected but are not experimentally validated) from the distribution of their importance values, we defined two thresholds. First, the expectation is that the most important sites will mostly have true positives, so the first threshold used the influenza data to determine what top percentile of ranked importance values maximizes the number of true positive. We refer to this threshold as the "percentile threshold." Then, to minimize the number of false positives among these top sites based on the repeated nature of the RRF, we implemented a second threshold, the "consensus threshold," which works as follows: all replicates of the RRF were run independently, and only sites that were predicted at a certain percentile threshold (*i.e.*,

a consensus) of all the runs were logged. Among these, only those that were found, say in 90% of the runs (9 runs out of 10), were considered as "genetic determinants." Intuitively, the higher this consensus threshold, the lower the number of false positives. We then varied both threshold on the influenza data to (i) maximize the number of true positives (percentile threshold) while (ii) minimizing the number of false positive (consensus threshold). We employed these thresholds to identify genetic determinants in the pseudomonas data.

Finally for the pseudomonas data, one additional step was required. As the algorithms used to identify the genetic determinants of phenotypes require categorical data, MIC distributions were discretized. To do so, and in the case of AB first, the distribution of $\log_2$ MIC of each drug was first plotted, and appeared to be trimodal; hence, it seemed natural to design a classification with three categories: 'low', 'medium', and 'high' MIC. The boundaries between these categories were determined to minimize class imbalance, and hence guarantee that each discrete category had similar numbers of samples. Because of the relative subjectivity in determining these categories, three different sets of MIC thresholds were employed to assess the robustness of the results. All analyses were run four times to further assess robustness, and stability of which sites were identified. However, when doing so, the number of classification categories goes from two (influenza) to three (pseudomonas), without any statistical justification. To address this point, a more thorough search was performed using RRF – as RRF is much faster than AB. For this, the range of MIC values for each drug was discretized into bins of width 0.125 (on a $\log_2$ scale of MIC values), and two thresholds (θ_1, θ_2) were defined, hereby dividing the distribution of MIC values into three domains. An initial classifier was then run for all combination of thresholds with $\theta_1 < \theta_2$, and classification errors (RRF: Out-of-bag errors) were logged, and used to define MIC thresholds to perform the final analyses. A small chunk size (100) and a regular RF algorithm was used to speed up these computations (see Results).

Acknowledgments

We thank Jeremy Dettman and Rees Kassen for sharing with us both their MIC and sequencing data, as well as the Center for Advanced Computing and Compute Ontario for providing us access to their servers. We are also grateful to Berthin Biyong, Graham Colby, Jeremy Dettman, Rees Kassen, and Matti Ruuskanen for discussions. This work was supported by the University of Ottawa's Undergraduate

Research Opportunity Program (GSL, MH) and teaching assistantships (JD), and the Natural Sciences Research Council of Canada (SAB).

References

1. Stranger BE, Stahl EA, Raj T. Progress and promise of genome-wide association studies for human complex trait genetics. Genetics. 2011;187(2):367–383.

2. Lippert C, Sabatini R, Maher MC, Kang EY, Lee S, Arikan O, et al. Identification of individuals by trait prediction using whole-genome sequencing data. Proc Natl Acad Sci U S A. 2017;114(38):10166–10171. doi:10.1073/pnas.1711125114.

3. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 Years of GWAS Discovery: Biology, Function, and Translation. Am J Hum Genet. 2017;101(1):5–22. doi:10.1016/j.ajhg.2017.06.005.

4. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. Am J Hum Genet. 2012;90(1):7–24. doi:10.1016/j.ajhg.2011.11.029.

5. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461(7265):747–53. doi:10.1038/nature08494.

6. Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat Rev Genet. 2010;11(6):415–25. doi:10.1038/nrg2779.

7. Risch N, Merikangas K. The future of genetic studies of complex human diseases. Science. 1996;273(5281):1516–7.

8. Kim BJ, Kim SH. Prediction of inherited genomic susceptibility to 20 common cancer types by a supervised machine-learning method. Proc Natl Acad Sci U S A. 2018;115(6):1322–1327. doi:10.1073/pnas.1717960115.

9. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. In: Overview of supervised learning. Springer; 2009. p. 9–41.

10. Goldstein BA, Navar AM, Carter RE. Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. Eur Heart J. 2016;doi:10.1093/eurheartj/ehw302.

11. Angermueller C, Pärnamaa T, Parts L, Stegle O. Deep learning for computational biology. Mol Syst Biol. 2016;12(7):878.

12. Freund Y, Schapire RE, et al. Experiments with a new boosting algorithm. In: ICML. vol. 96; 1996. p. 148–156.

13. Breiman L, et al. Arcing classifier (with discussion and a rejoinder by the author). The annals of statistics. 1998;26(3):801–849.

14. Wolpert DH. The lack of a priori distinctions between learning algorithms. Neural computation. 1996;8(7):1341–1390.

15. Zhou ZH. Ensemble methods: foundations and algorithms. CRC press; 2012.

16. Shoji A, Aris-Brosou S, Culina A, Fayet A, Kirk H, Padget O, et al. Breeding phenology and winter activity predict subsequent breeding success in a trans-global migratory seabird. Biol Lett. 2015;11(10). doi:10.1098/rsbl.2015.0671.

17. Liaw A, Wiener M. Classification and Regression by randomForest. R News. 2002;2(3):18–22.

18. Aris-Brosou S, Kim J, Li L, Liu H. Predicting the reasons of customer complaints: a first step toward anticipating quality issues of *in vitro* diagnostics assays with machine learning. JMIR Med Inform. 2018;6(2):e34. doi:10.2196/medinform.9960.

19. Collaboration A, et al. The evolution of boosting algorithms-from machine learning to statistical modelling. Methods Inf Med. 2014;53(6):419–427.

20. Elith J, Leathwick JR, Hastie T. A working guide to boosted regression trees. Journal of Animal Ecology. 2008;77(4):802–813.

21. Li H, Luan Y. Boosting proportional hazards models using smoothing splines, with applications to high-dimensional microarray data. Bioinformatics. 2005;21(10):2403–9. doi:10.1093/bioinformatics/bti324.

22. Binder H, Benner A, Bullinger L, Schumacher M. Tailoring sparse multivariable regression techniques for prognostic single-nucleotide polymorphism signatures. Statistics in medicine. 2013;32(10):1778–1791.

23. Frankel N, Erezyilmaz DF, McGregor AP, Wang S, Payre F, Stern DL. Morphological evolution caused by many subtle-effect substitutions in a transcriptional enhancer. Nature. 2011;474. doi:10.1038/nature10200.

24. Houle D, Govindaraju DR, Omholt S. Phenomics: the next challenge. Nat Rev Genet. 2010;11(12):855–66. doi:10.1038/nrg2897.

25. Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods for the classification of tumors using gene expression data. Journal of the American statistical association. 2002;97(457):77–87.

26. Kamal RP, Katz JM, York IA. Molecular determinants of influenza virus pathogenesis in mice. In: Influenza Pathogenesis and Control-Volume I. Springer; 2014. p. 243–274.

27. Schrauwen EJ, de Graaf M, Herfst S, Rimmelzwaan GF, Osterhaus AD, Fouchier RA. Determinants of virulence of influenza A virus. European journal of clinical microbiology & infectious diseases. 2014;33(4):479–490.

28. Munita JM, Arias CA. Mechanisms of antibiotic resistance. Microbiol Spectr. 2016;4(2). doi:10.1128/microbiolspec.VMBF-0016-2015.

29. Northrop Grumman Health IT VT J Craig Venter Institute. Influenza Research Database; 2017. Available from: https://www.fludb.org/brc/home.spg?decorator=influenza.

30. Dettman JR, Rodrigue N, Aaron SD, Kassen R. Evolutionary genomics of epidemic and nonepidemic strains of *Pseudomonas aeruginosa*. Proc Natl Acad Sci U S A. 2013;110(52):21065–70. doi:10.1073/pnas.1307862110.

31. Sharshov K, Romanovskaya A, Uzhachenko R, Durymanov A, Zaykovskaya A, Kurskaya O, et al. Genetic and biological characterization of avian influenza H5N1 viruses isolated from wild birds and poultry in Western Siberia. Archives of virology. 2010;155(7):1145–1150.

32. Bogs J, Kalthoff D, Veits J, Pavlova S, Schwemmle M, Mänz B, et al. Reversion of PB2-627E to-627K during replication of an H5N1 Clade 2.2 virus in mammalian hosts depends on the origin of the nucleoprotein. Journal of virology. 2011;85(20):10691–10698.

33. Long JCD, Fodor E. The PB2 subunit of the Influenza A virus RNA polymerase is imported into the mitochondrial matrix. J Virol. 2016;90(19):8729–38. doi:10.1128/JVI.01384-16.

34. Yamaji R, Yamada S, Le MQ, Li C, Chen H, Qurnianingsih E, et al. Identification of PB2 mutations responsible for the efficient replication of H5N1 influenza viruses in human lung epithelial cells. J Virol. 2015;89(7):3947–56. doi:10.1128/JVI.03328-14.

35. Li J, Ishaq M, Prudence M, Xi X, Hu T, Liu Q, et al. Single mutation at the amino acid position 627 of PB2 that leads to increased virulence of an H5N1 avian influenza virus during adaptation in mice can be compensated by multiple mutations at other sites of PB2. Virus research. 2009;144(1):123–129.

36. Sugiyama K, Obayashi E, Kawaguchi A, Suzuki Y, Tame JRH, Nagata K, et al. Structural insight into the essential PB1-PB2 subunit contact of the influenza virus RNA polymerase. EMBO J. 2009;28(12):1803–11. doi:10.1038/emboj.2009.138.

37. Godley L, Pfeifer J, Steinhauer D, Ely B, Shaw G, Kaufmann R, et al. Introduction of intersubunit disulfide bonds in the membrane-distal region of the influenza hemagglutinin abolishes membrane fusion activity. Cell. 1992;68(4):635–645.

38. Gambaryan A, Robertson J, Matrosovich M. Effects of egg-adaptation on the receptor-binding properties of human influenza A and B viruses. Virology. 1999;258(2):232–239.

39. Hensley SE, Das SR, Bailey AL, Schmidt LM, Hickman HD, Jayaraman A, et al. Hemagglutinin receptor binding avidity drives influenza A virus antigenic drift. Science. 2009;326(5953):734–6. doi:10.1126/science.1178258.

40. Ellebedy AH. Impact of adjuvants on the antibody responses to pre-pandemic H5N1 influenza vaccines. The University of Tennessee Health Science Center; 2011.

41. Llompart CM, Nieto A, Rodriguez-Frandsen A. Specific residues of PB2 and PA influenza virus polymerase subunits confer the ability for RNA polymerase II degradation and virus pathogenicity in mice. J Virol. 2014;88(6):3455–63. doi:10.1128/JVI.02263-13.

42. Hulse-Post D, Franks J, Boyd K, Salomon R, Hoffmann E, Yen H, et al. Molecular changes in the polymerase genes (PA and PB1) associated with high pathogenicity of H5N1 influenza virus in mallard ducks. Journal of virology. 2007;81(16):8515–8524.

43. Taubenberger JK, Reid AH, Lourens RM, Wang R, Jin G, Fanning TG. Characterization of the 1918 influenza virus polymerase genes. Nature. 2005;437(7060):889–893.

44. Shin YK, Liu Q, Tikoo SK, Babiuk LA, Zhou Y. Influenza A virus NS1 protein activates the phosphatidylinositol 3-kinase (PI3K)/Akt pathway by direct interaction with the p85 subunit of PI3K. Journal of General Virology. 2007;88(1):13–18.

45. Manzoor R, Sakoda Y, Nomura N, Tsuda Y, Ozaki H, Okamatsu M, et al. PB2 protein of a highly pathogenic avian influenza virus strain A/chicken/Yamaguchi/7/2004 (H5N1) determines its replication potential in pigs. J Virol. 2009;83(4):1572–8. doi:10.1128/JVI.01879-08.

46. Leung BW, Chen H, Brownlee GG. Correlation between polymerase activity and pathogenicity in two duck H5N1 influenza viruses suggests that the polymerase contributes to pathogenicity. Virology. 2010;401(1):96–106. doi:10.1016/j.virol.2010.01.036.

47. Collin F, Karkare S, Maxwell A. Exploiting bacterial DNA gyrase as a drug target: current state and perspectives. Appl Microbiol Biotechnol. 2011;92(3):479–97. doi:10.1007/s00253-011-3557-z.

48. Dalhoff A. Global fluoroquinolone resistance epidemiology and implictions for clinical use. Interdiscip Perspect Infect Dis. 2012;2012:976273. doi:10.1155/2012/976273.

49. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014;30(9):1236–40. doi:10.1093/bioinformatics/btu031.

50. de Carvalho CCCR, Fernandes P. Siderophores as “Trojan Horses”: tackling multidrug resistance? Front Microbiol. 2014;5:290. doi:10.3389/fmicb.2014.00290.

51. Kos VN, McLaughlin RE, Gardner HA. The elucidation of mechanisms of ceftazidime resistance among clinical isolates of *Pseudomonas aeruginosa* using genomic data. Antimicrobial Agents and Chemotherapy. 2016;doi:10.1128/AAC.03113-15.

52. McGuffie BA, Vallet-Gely I, Dove SL. factor and anti- factor that control swarming motility and biofilm formation in *Pseudomonas aeruginosa*. J Bacteriol. 2015;198(5):755–65. doi:10.1128/JB.00784-15.

53. Schapire RE, Freund Y. Boosting: Foundations and algorithms. MIT press; 2012.

54. Mayr A, Hofner B, Schmid M, et al. The importance of knowing when to stop. Methods of Information in Medicine. 2012;51(2):178–186.

55. Yeung ATY, Torfs ECW, Jamshidi F, Bains M, Wiegand I, Hancock REW, et al. Swarming of *Pseudomonas aeruginosa* is controlled by a broad spectrum of transcriptional regulators, including MetR. J Bacteriol. 2009;191(18):5592–602. doi:10.1128/JB.00157-09.

56. Amini S, Hottes AK, Smith LE, Tavazoie S. Fitness landscape of antibiotic tolerance in *Pseudomonas aeruginosa* biofilms. PLoS Pathog. 2011;7(10):e1002298. doi:10.1371/journal.ppat.1002298.

57. Lee JM, Huddleston J, Doud MB, Hooper KA, Wu NC, Bedford T, et al. Deep mutational scanning of hemagglutinin helps predict evolutionary fates of human H3N2 influenza variants. Proc Natl Acad Sci U S A. 2018;doi:10.1073/pnas.1806133115.

58. Ameur A, Wetterbom A, Feuk L, Gyllensten U. Global and unbiased detection of splice junctions from RNA-seq data. Genome Biol. 2010;11(3):R34. doi:10.1186/gb-2010-11-3-r34.

59. Costa V, Aprile M, Esposito R, Ciccodicola A. RNA-Seq and human complex diseases: recent accomplishments and future perspectives. Eur J Hum Genet. 2013;21(2):134–42. doi:10.1038/ejhg.2012.129.

60. Torres-Oliva M, Almudi I, McGregor AP, Posnien N. A robust (re-)annotation approach to generate unbiased mapping references for RNA-seq-based analyses of differential expression across closely related species. BMC Genomics. 2016;17:392. doi:10.1186/s12864-016-2646-x.

61. Leung MKK, Xiong HY, Lee LJ, Frey BJ. Deep learning of the tissue-regulated splicing code. Bioinformatics. 2014;30(12):i121–9. doi:10.1093/bioinformatics/btu277.

62. Xiong HY, Alipanahi B, Lee LJ, Bretschneider H, Merico D, Yuen RKC, et al. RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. Science. 2015;347(6218):1254806. doi:10.1126/science.1254806.

63. Basu S, Kumbier K, Brown JB, Yu B. Iterative random forests to discover predictive and stable high-order interactions. Proc Natl Acad Sci U S A. 2018;115(8):1943–1948. doi:10.1073/pnas.1711236115.

64. Stockwell DR, Peterson AT. Effects of sample size on accuracy of species distribution models. Ecological modelling. 2002;148(1):1–13.

65. Stech O, Veits J, Weber S, Deckers D, Schröer D, Vahlenkamp TW, et al. Acquisition of a polybasic hemagglutinin cleavage site by a low-pathogenic avian influenza virus is not sufficient for immediate transformation into a highly pathogenic strain. Journal of virology. 2009;83(11):5864–5868.

66. Bao Y, Bolotov P, Dernovoy D, Kiryutin B, Zaslavsky L, Tatusova T, et al. The influenza virus resource at the National Center for Biotechnology Information. J Virol. 2008;82(2):596–601. doi:10.1128/JVI.02005-07.

67. van Riel D, Munster VJ, de Wit E, Rimmelzwaan GF, Fouchier RA, Osterhaus AD, et al. Human and avian influenza viruses target different cells in the lower respiratory tract of humans and other mammals. The American journal of pathology. 2007;171(4):1215–1223.

68. Subbarao K, Klimov A, Katz J, Regnery H, Lim W, Hall H, et al. Characterization of an avian influenza A (H5N1) virus isolated from a child with a fatal respiratory illness. Science. 1998;279(5349):393–396.

69. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32(5):1792–7. doi:10.1093/nar/gkh340.

70. Winsor GL, Griffiths EJ, Lo R, Dhillon BK, Shay JA, Brinkman FSL. Enhanced annotations and features for comparing thousands of Pseudomonas genomes in the Pseudomonas genome database. Nucleic Acids Res. 2016;44(D1):D646–53. doi:10.1093/nar/gkv1227.

71. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–10. doi:10.1016/S0022-2836(05)80360-2.

72. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, et al. Software for computing and annotating genomic ranges. PLoS Comput Biol. 2013;9(8):e1003118. doi:10.1371/journal.pcbi.1003118.

73. Yotsuji A, Mitsuyama J, Hori R, Yasuda T, Saikawa I, Inoue M, et al. Mechanism of action of cephalosporins and resistance caused by decreased affinity for penicillin-binding proteins in *Bacteroides fragilis*. Antimicrob Agents Chemother. 1988;32(12):1848–53.

74. Kotra LP, Haddad J, Mobashery S. Aminoglycosides: perspectives on mechanisms of action and resistance and strategies to counter resistance. Antimicrob Agents Chemother. 2000;44(12):3249–56.

75. Alfaro E, Gámez M, García N. adabag: An R Package for Classification with Boosting and Bagging. Journal of Statistical Software. 2013;54(2):1–35.

76. Liaw A, Wiener M. Classification and Regression by randomForest. R News. 2002;2(3):18–22.

77. R Core Team. R: A Language and Environment for Statistical Computing; 2015. Available from: https://www.R-project.org/.

78. Microsoft, Weston S. foreach: Provides Foreach Looping Construct for R; 2017. Available from: https://CRAN.R-project.org/package=foreach.

79. Melnyk AH, McCloskey N, Hinz AJ, Dettman J, Kassen R. Evolution of cost-free resistance under fluctuating drug selection in *Pseudomonas aeruginosa*. mSphere. 2017;2(4). doi:10.1128/mSphere.00158-17.

80. Takeda M, Leser GP, Russell CJ, Lamb RA. Influenza virus hemagglutinin concentrates in lipid raft microdomains for efficient viral fusion. Proceedings of the National Academy of Sciences. 2003;100(25):14610–14617.

81. Poole E, Elton D, Medcalf L, Digard P. Functional domains of the influenza A virus PB2 protein: identification of NP-and PB1-binding sites. Virology. 2004;321(1):120–133.

82. Li K, Guan Y, Wang J, Smith G, Xu K, Duan L, et al. Genesis of a highly pathogenic and potentially pandemic H5N1 influenza virus in eastern Asia. Nature. 2004;430(6996):209–213.

83. Gianfrani C, Oseroff C, Sidney J, Chesnut RW, Sette A. Human memory CTL response specific for influenza A virus is broad and multispecific. Human immunology. 2000;61(5):438–452.

84. Min JY, Li S, Sen GC, Krug RM. A site on the influenza A virus NS1 protein mediates both inhibition of PKR activation and temporal regulation of viral RNA synthesis. Virology. 2007;363(1):236–243.

85. Hale BG, Randall RE, Ortín J, Jackson D. The multifunctional NS1 protein of influenza A viruses. Journal of General Virology. 2008;89(10):2359–2376.

86. Heeb S, Fletcher MP, Chhabra SR, Diggle SP, Williams P, Cámara M. Quinolones: from antibiotics to autoinducers. FEMS Microbiol Rev. 2011;35(2):247–74. doi:10.1111/j.1574-6976.2010.00247.x.

87. Nguyen AT, O'Neill MJ, Watts AM, Robson CL, Lamont IL, Wilks A, et al. Adaptation of iron homeostasis pathways by a *Pseudomonas aeruginosa* pyoverdine mutant in the cystic fibrosis lung. J Bacteriol. 2014;196(12):2265–76. doi:10.1128/JB.01491-14.

88. Delcour AH. Outer membrane permeability and antibiotic resistance. Biochim Biophys Acta. 2009;1794(5):808–16. doi:10.1016/j.bbapap.2008.11.005.

89. Wozniak M, Tiuryn J, Wong L. An approach to identifying drug resistance associated mutations in bacterial strains. BMC Genomics. 2012;13 Suppl 7:S23. doi:10.1186/1471-2164-13-S7-S23.

90. Seyedmohammad S, Fuentealba NA, Marriott RAJ, Goetze TA, Edwardson JM, Barrera NP, et al. Structural model of FeoB, the iron transporter from *Pseudomonas aeruginosa*, predicts a cysteine lined, GTP-gated pore. Biosci Rep. 2016;36(2). doi:10.1042/BSR20160046.

91. Rouch DA, Cram DS, DiBerardino D, Littlejohn TG, Skurray RA. Efflux-mediated antiseptic resistance gene qacA from *Staphylococcus aureus*: common ancestry with tetracycline- and sugar-transport proteins. Mol Microbiol. 1990;4(12):2051–62.

92. Galle M, Carpentier I, Beyaert R. Structure and function of the Type III secretion system of Pseudomonas aeruginosa. Curr Protein Pept Sci. 2012;13(8):831–42.

93. Alvarez-Ortega C, Wiegand I, Olivares J, Hancock REW, Martínez JL. Genetic determinants involved in the susceptibility of *Pseudomonas aeruginosa* to beta-lactam antibiotics. Antimicrob Agents Chemother. 2010;54(10):4159–67. doi:10.1128/AAC.00257-10.

94. Kubacka A, Diez MS, Rojo D, Bargiela R, Ciordia S, Zapico I, et al. Understanding the antimicrobial mechanism of TiO-based nanocomposite films in a pathogenic bacterium. Sci Rep. 2014;4:4134. doi:10.1038/srep04134.

95. Hall CW, Zhang L, Mah TF. PA3225 Is a Transcriptional Repressor of Antibiotic Resistance Mechanisms in *Pseudomonas aeruginosa*. Antimicrob Agents Chemother. 2017;61(8). doi:10.1128/AAC.02114-16.

96. Srinivasan VB, Mondal A, Venkataramaiah M, Chauhan NK, Rajamohan G. Role of oxyRKP, a novel LysR-family transcriptional regulator, in antimicrobial resistance and virulence in *Klebsiella pneumoniae*. Microbiology. 2013;159(Pt 7):1301–14. doi:10.1099/mic.0.065052-0.

97. Laehnemann D, Peña-Miller R, Rosenstiel P, Beardmore R, Jansen G, Schulenburg H. Genomics of rapid adaptation to antibiotics: convergent evolution and scalable sequence amplification. Genome Biol Evol. 2014;6(6):1287–301. doi:10.1093/gbe/evu106.

Tables

Table 1. Performance of the algorithms on the analysis of the influenza data. This table lists the genes and amino acid positions known to be involved in the three phenotypes studied here, and which one of these were rediscovered by our algorithms. For AB, chunk sizes of 75, 125, and 175 were used to calculate the importance values of each site for adaptive boosting. An importance threshold of 1 was used to determine whether a site was a potential genetic determinant. For RRF, chunk sizes of 80, 125, and 175 were used with a threshold of the 90th percentile and a 60% consensus. Data on experimental validations are from the Influenza Research Database [29]. Genes are ordered by segment size.

Site	AB	RRF	Phenotype	Reference
PB2 9	✓	✓	Infectivity	[33]
PB2 105	✓	✓	Pathogenicity/Infectivity	[41]
PB2 339	✓	✓	Infectivity	[35]
PB2 391	✓		Transmissibility	[40]
PB2 627	✓	✓	Infectivity	[32]
PB2 667	✓		Infectivity	[31]
PB1 215	✓	✓	Pathogenicity	[42]
PB1 375		✓	Pathogenicity	[43]
PB1 757	✓		Infectivity	[36]
HA 163		✓	Pathogenicity/Infectivity	[38]
HA 212		✓	Pathogenicity/Infectivity	[37]
HA 246	✓	✓	Transmissibility	[39]
HA 536		✓	Infectivity	[80]
NP 400			Pathogenicity	[81]
NA 49		✓	Transmissibility	[82]
NA 75			Transmissibility	[83]
M2 31		✓	Pathogenicity/Infectivity	[82]
NS1 127			Pathogenicity	[84]
NS1 195			Transmissibility/Infectivity	[85]
NS1 212		✓	Pathogenicity/Infectivity	[44]

Table 2. The different sets of MIC thresholds employed to assess the robustness of the classification results with AB in the case of *P. aeruginosa*. Shown are the thresholds θ_1 / θ_2 used on a $\log_2$ MIC scale: for instance, setting 1 for Ciprofloxacin means that MIC is low when ≤ -1 ($\theta_1 = -1$), high when MIC > 1.5 ($\theta_2 = 1.5$), and medium in-between.

Drug	Setting 1	Setting 2	Setting 3
Ciprofloxacin	-1/1.5	0/2	-1/2
Ceftazidime	4/6	6/8	4/8
Gentamicin	4/6	6/8	4/8

Table 3. Gene lists of the most important candidates for drug resistance in *P. aeruginosa*. Shown are the genes identified in all four runs under the four settings defined in Table 2. For each drug, the genes identified in all three settings are highlighted (boldface), as well as those found in two out of the three settings (italics). Gene names that are underlined (setting 1 only) are those identified during the cross-validation experiment, under both chunk sizes.

Drug	Setting 1	Setting 2	Setting 3
Ciprofloxacin	*trpI, transcriptional regulator TrpI* *lysin domain-containing protein* *tag, DNA-3-methylad. glycosidase I* recF, recombination protein F D,D-heptose 1,7-bisphos. phosphatase	hypothetical protein HIS/PHE ammonia-lyase LysR family transcriptional regulator	*trpI, transcriptional regulator TrpI* *tag, DNA-3-methylad. glycosidase I* *lysin domain-containing protein* glutamine synthetase hypothetical protein
Ceftazidime	*hemolysin activ./secret. prot* **hypothetical protein (PA_40040)** *gyrB, DNA gyrase subunit B*	**hypothetical protein (PA_40040)** sensor/response regulator hybrid	*hemolysin activ./secret. prot* **hypothetical protein (PA_40040)** *gyrB, DNA gyrase subunit B* hemagglutinin recQ, ATP-depend. DNA helicase
Gentamicin	Rossmann fold nucleotide-bind. prot. gyrB, DNA gyrase subunit B **hemagglutinin** acyltransferase **tonB2, hypothetical protein** nirN, c-type cytochrome	*sbrR, SbrR* **tonB2, hypothetical protein** **hemagglutinin** tufA, elongation factor Tu *hemolysin activ./secret. prot*	**tonB2, hypothetical protein** **hemagglutinin** *sbrR, SbrR* *hemolysin activ./secret. prot*

Table 4. Characterization of RRF drug resistance candidates in the pseudomonas data from the literature.

Drug	Gene	Evidence	Reference
Ciprofloxacin	pchE	siderophore family, extracellular iron-acquisition system, upregulation associated with exposure to natural quinolones	[86]
		iron is required for virulence and is deficient in the human lung environment of these clinical strains	[87]
	cupB3	part of an outer membrane porin family, mutation that reduce membrane permeability are linked to Gram-negative bacterial mechanisms for antibiotic resistance	[88]
	permease	reduces accumulation of drug inside cell by decreased cell wall permeability or by pumping drug out	[89]
	ABC transporter	see permease	[89]
	SH3	controls numerous protein-protein interactions, some implicated in virulence of pathogenic bacteria	[90]
	alkB	DNA repair system (fluoroquinolones prevent proper winding and unwinding of DNA during replication), also affects outer membrane lipids - and thus permeability - it may affect antibiotic resistance	[88]
	sbrR	anti-sigma factor, identified as necessary during the chronic infection of respiratory tracts	[52]
	mnmC	Part of tRNA modification, and thus protein synthesis, no obvious antibiotic or lung environment connection	
Ceftazidime	MFS	membrane translocases, include many multidrug resistant proteins of Gram-positive bacteria	[91]
	pscC	type III secretion outer membrane protein, probable general resistance candidate	[92, 88]
	algW	mutations are known to confer susceptibility to the beta-lactam family of antibiotics	[93]
	glyA2	produces anti-oxidant coenzymes, involved in cell response to TiO2-based nanocomposite antimicrobials	[94]
	lysR	associated with minimum inhibitory concentration of antibiotics and oxidative stress chemicals	[95, 96]
Gentamicin	rnt	this ribonuclease would have a logical role in degrading 30S bound by the antibiotic	
	algW	correlated to Ceftazidime resistance?	
	quinone OR	antibiotic resistance in response to antibiotics that inhibit protein synthesis - including binding of the 50S ribosomal subunit	[97]

Table 5. List of the Influenza A strains and their associated phenotypes, as used in the training of the machine learning algorithms. For pathogenicity, polybasic cleavage was used as a proxy.

Strain name	Infectivity	Transmissibility	Pathogenicity
A/HongKong/156/97	No	Yes	Yes
A/HongKong/213/2003	No	Yes	Yes
A/Indonesia/5/2005	No	No	No
A/Indonesia/7/2005	No	No	No
A/PuertoRico/8/34	Yes	No	No
A/Swine/Indiana/1726/1988	Yes	Yes	No
A/Turkey/15/2006	No	No	No
A/VietNam/1203/2004	Yes	No	Yes
A/VietNam/3046/2004	Yes	No	Yes
A/VietNam/3062/2004	Yes	No	Yes

Figures

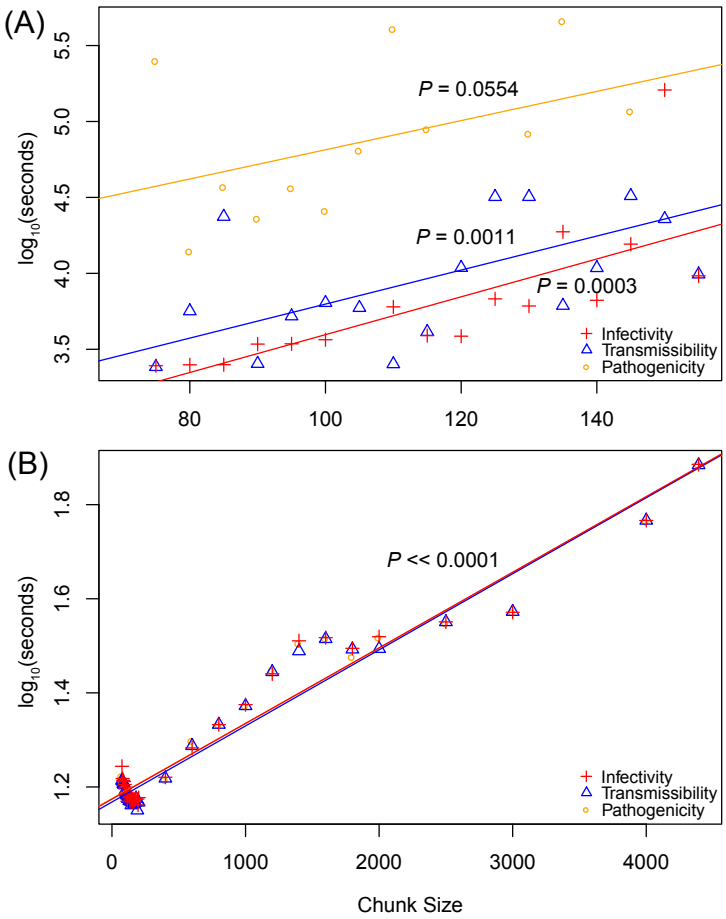

Figure 1. Impact of chunk size on the runtime of the machine learning algorithms for the influenza data. Runtimes for infectivity (red), transmissibility (blue), and pathogenicity (orange) are shown for AB (A) and RRF (B). While each data point is based on a single run, run-to-run variability is taken into account by performing linear regressions (solid lines); their P-values are also shown.

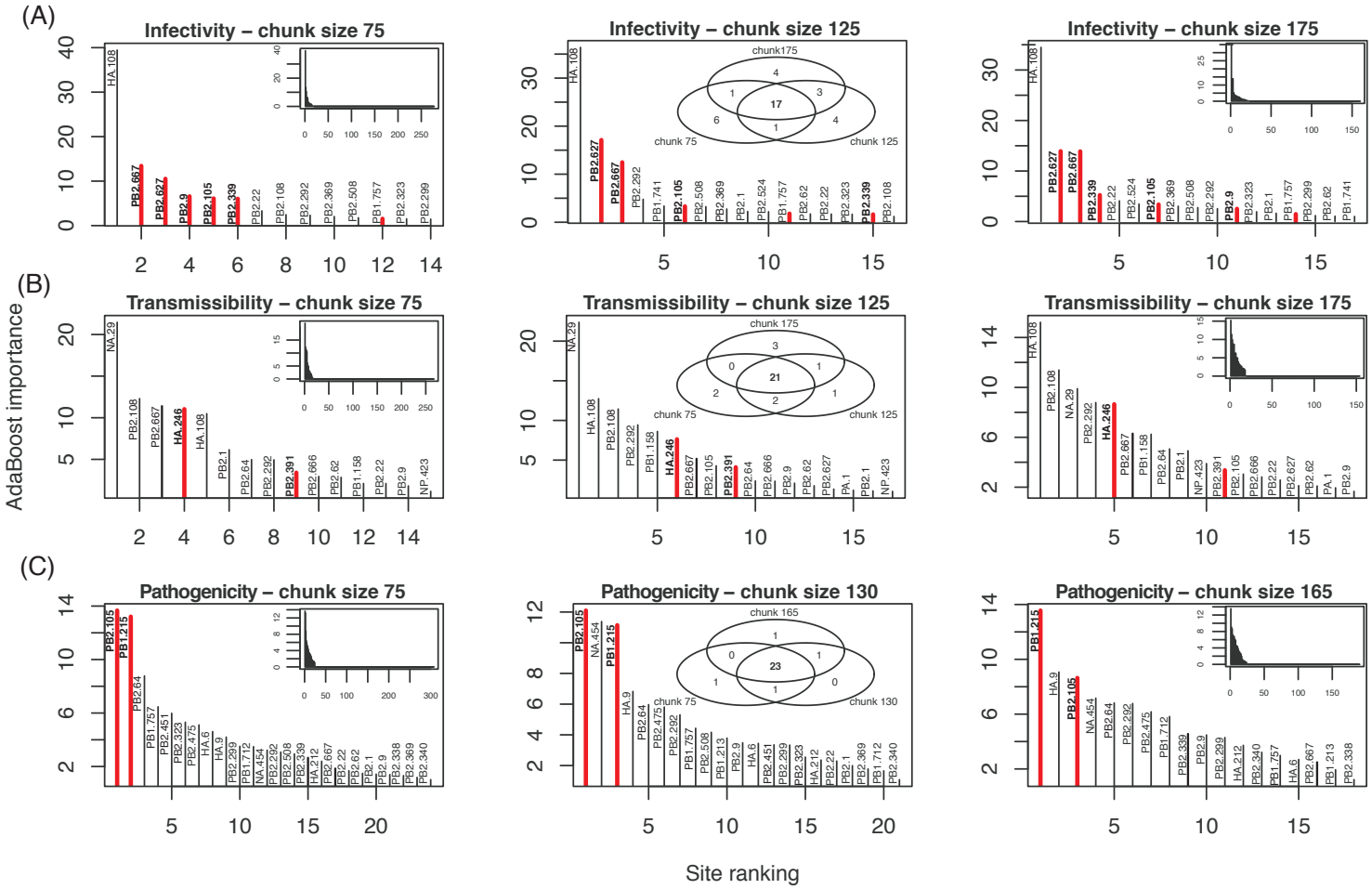

Figure 2. Effect of chunk size on the distribution of importance of sites for the AB algorithm. The genes and sites identified as genetic determinants of influenza phenotypes are shown for: (A) infectivity, (B) transmissibility, and (C) pathogenicity. Only results for the smallest (75 amino acids), intermediate (125), and largest (175) chunk sizes are shown. Only the most important sites (importance > 1) are shown in each panel, with sites backed by with experimental evidence highlighted in red. Insets show the whole distribution of importance values (left and right columns), and the Venn diagrams of the most important sites at all three chunk size (middle column).

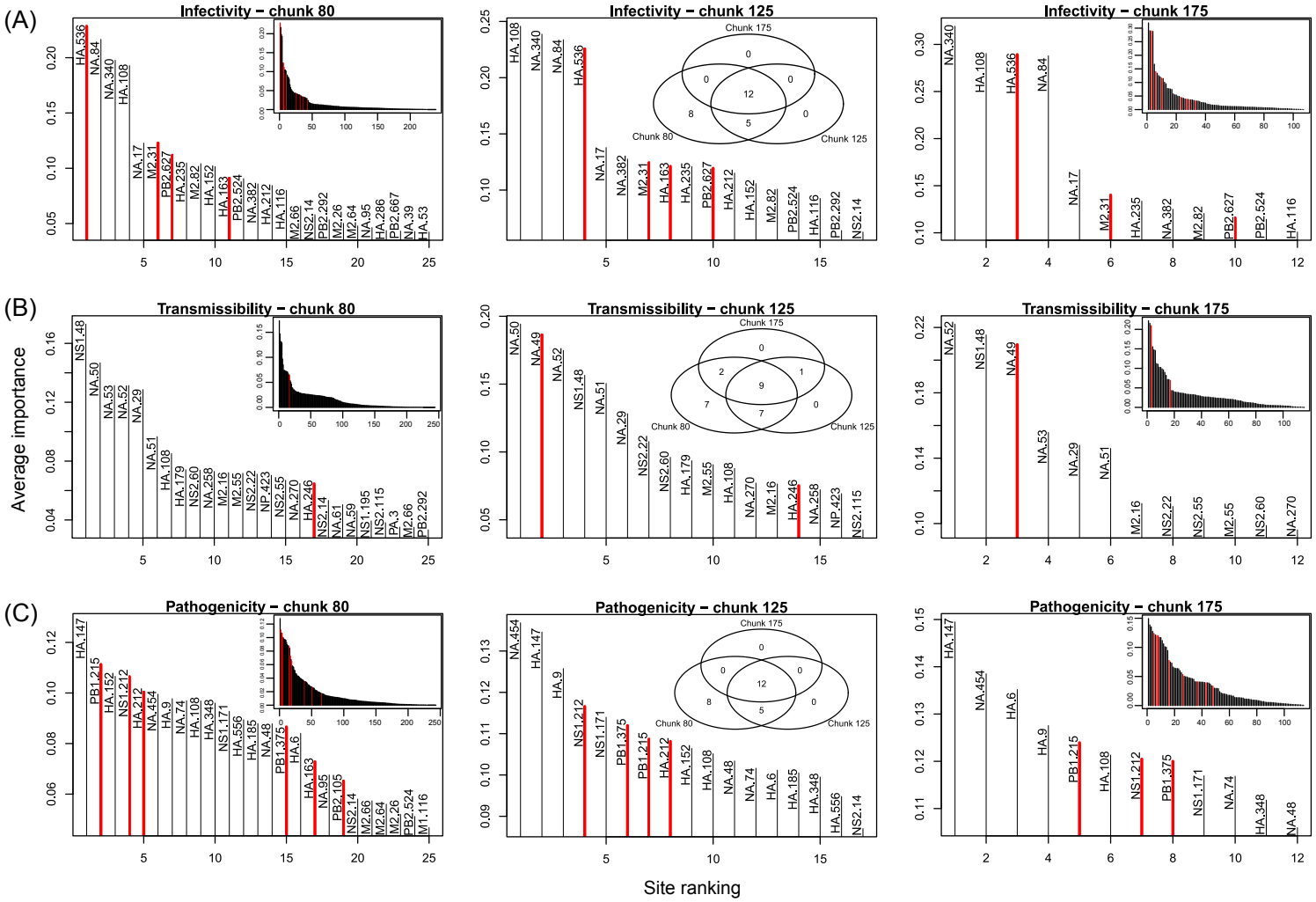

Figure 3. Effect of chunk size on the distribution of importance of sites for the RRF algorithm. The genes and sites identified as genetic determinants of influenza phenotypes are shown for: (A) infectivity, (B) transmissibility, and (C) pathogenicity. Only results for the smallest (80 amino acids), intermediate (125), and largest (175) chunk sizes are shown. Only the most important sites (Gini index in top 90th percentile of its distribution over all the sites) are shown in each panel, with sites backed by with experimental evidence highlighted in red. Insets show the whole distribution of importance values (left and right columns), and the Venn diagrams of the most important sites at all three chunk size (middle column).

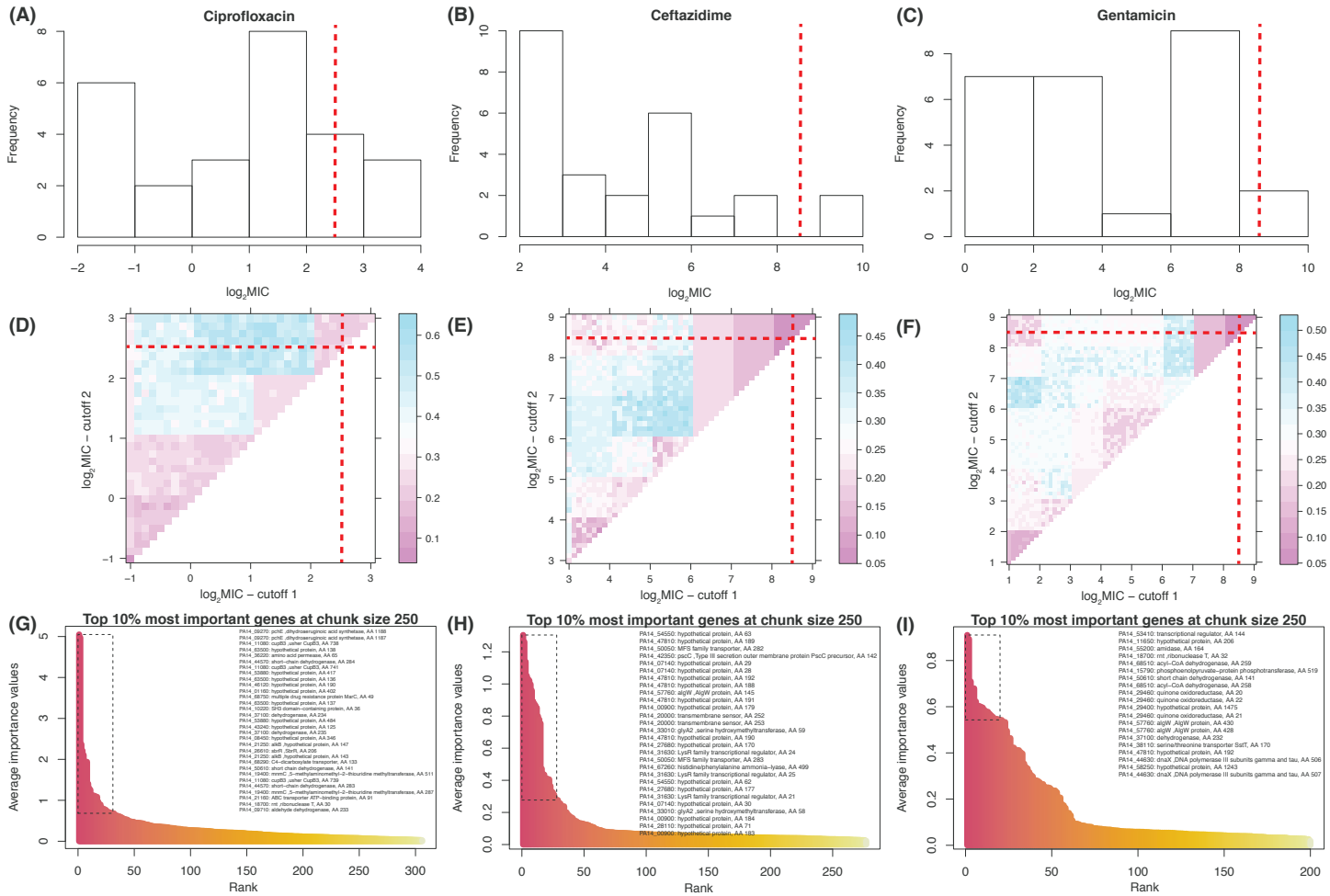

Figure 4. Analysis of the *P. aeruginosa* data across the 26 strains from [30]. The distributions of MIC values (on a $\log_2$ scale) are shown for (A) Ciprofloxacin, (B) Ceftazidime, and (C) Gentamicin. These empirical distributions were used to determine MIC thresholds for the AB analyses (Table 2). Note that the scales on the y-axis vary slightly. With RRF, a throughout search of the discretization was performed to select thresholds θ_1 / θ_2 that would minimize the Out-of-bag error for (D) Ciprofloxacin, (E) Ceftazidime, and (F) Gentamicin (color scale to the right of each panel). The θ_1 / θ_2 combinations (red dotted lines, also reported in first row) were determined visually. The top 10% most important sites (as per their Gini index) are highlighted (box with broken lines) among sites selected at the end of the first tier of the chunking algorithm for (G) Ciprofloxacin, (H) Ceftazidime, and (I) Gentamicin. These top sites are listed in the top right part of each distribution.