Identifying the genetic determinants of particular phenotypes in microbial genomes with very small training sets

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

[1]Department of Biology, University of Ottawa, Ottawa, ON K1N 6N5, Canada

[2]Department of Mathematics and Statistics, University of Ottawa, Ottawa, ON K1N 6N5, Canada

[⋆]Corresponding author: `sarisbro@uottawa.ca`

Abstract

Background Machine learning (ML) encompasses a large set of algorithms that aim at discovering complex patterns between elements within large data sets without any prior assumptions or modeling. However, some scientific disciplines still produce small data sets: in particular, empirical studies that try to link complex phenotypes such as virulence or drug resistance to individual sets of protein-coding genes (proteomes) typically have very small sample sizes. To date, it is unknown how ML performs in such cases.

Results To address this question, we evaluated the performance of adaptive boosting, a general ML classifier, on two data sets containing both the phenotype and the complete proteome of a small number of individuals. To assess the impact of proteome size, we contrasted a small genome (a virus: influenza) with a larger one (a bacterium: pseudomonas). In order to analyze large proteomes, we developed a chunking algorithm. With the influenza data, we were able to rediscover amino acid sites experimentally implicated in three different complex phenotypes (infectivity, transmissibility, and pathogenicity). However, results for the much larger pseudomonas proteome, pertaining to three types of drug resistance (Ciprofloxacin, Ceftazidime, and Gentamicin), proved unstable, depended on a number of assumptions, and were not always biologically sensible.

Conclusions Our results show that ML algorithms such as adaptive boosting can be used to successfully identify the genetic determinants of microbes with small proteomes (viruses). Our chunking algorithm improved runtimes by an order of magnitude without sacrificing accuracy. Yet we found that the size of bacterial proteomes pushed ML to its limits in the face of small number of individuals. The use of these algorithms should probably be limited to preliminary or exploratory analysis, as long as both phenotyping and sequencing are too costly to perform on more individuals.

Keywords: drug resistance; influenza virus; *Pseudomonas aeruginosa*; adaptive boosting; machine learning

Background

An overarching goal of biology is to be able to predict an individual's phenotype from its genotype, in a given environment [1]. A first step in this direction is now possible thanks to the ever decreasing costs of sequencing complete genomes [2]. In spite of this, two hurdles can still be identified. First, it is becoming more expensive to analyze the sequencing data hence produced – mostly because computational power does not scale with the increasing amount of data generated, but maybe also because current algorithms show the same lack of scalability [3]. Second, while genomics is a reality, little progress has been made towards developing phenomics [4], the corresponding large-scale phenotyping of the individuals just sequenced. This sets two constraints to predicting an individual's phenotype from its genome, or similarly, determining the genetic determinants of a phenotype: (i) this prediction needs to be done for large genomes, potentially up to 150Gbp [5] – *i.e.*, for a very large number of predictors p, and (ii) a small number of individuals – *i.e.*, for a very small sample size n. This so-called "large p, small n" problem is not new, as it 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 [6], that is, a general approach that can identify patterns in data without defining them *a priori* [7]. However, it is still unclear whether similar approaches can be used with phenotyping / sequencing data, and what kind of limitations may exist.

A large number of machine learning approaches exist [7], and are gaining popularity in biology [8, 9]. Deep learning must be ruled out where data sets suffer from the small n problem. While some statisticians claim that adaptive boosting [10] is the "best off-the-shelf classifier in the world" [11, 7], it is probable that no single classifier can be considered perfect in all situations. However, our recent success with adaptive boosting

on limited sample sizes [12] prompted us to further investigate this algorithm, in the context of genomic data. Technically, adaptive boosting 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 [10], *i.e.* a classifier with an accuracy that can come close to 99% [13].

As a supervised learning algorithm, adaptive boosting is trained on a data set for which correct assignments are known. For instance, consider a set of genomic data coming from organisms for which the phenotype such as drug resistance is known. We wish to classify (predict) 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 mutations at particular sites) are ranked by decreasing importance to fit the final model [14]. 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 [8] for phenotype prediction. A second advantage of boosting algorithms is their accuracy in the case of high-dimensional data, such as microarray [15] or single nucleotide polymorphism studies [16]. 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). However, boosting has rarely been evaluated in the context of large p, small n, and has had very few applications in genetics: the same PubMed search with "boosting SNP" as in [17] returns only 17 results in July 2017.

Here we evaluate the capacity of adaptive boosting to find the genetic determinants of particular phenotypes in two microbes with very different proteome sizes: 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). In each case, classifiers of three discrete phenotypes were fitted using genomic data. Because of the paucity of such data, where individuals are both phenotyped and sequenced, only small sample sizes are available. We take advantage of an influenza database backed by experimental validations [18], and a recent study of *P. aeruginosa* genomics [19], to train the adaptive boosting algorithm. We modify slightly the original adaptive boosting algorithm to make it amenable to analyzing genomics-sized data sets, and evaluate performance with respect to either experimental validations (influenza), or both gene annotations and cross-validations (pseudomonas). We discuss the advantages and limitations of adaptive boosting in this context.

Methods

The influenza data

Ten Influenza A strains were retrieved from the Influenza Research Database [18] using their search tool based on phenotype characteristics. The strains retrieved were of viral samples with experimental evidence supporting an increase of infectivity, transmissibility, and pathogenicity. In the case of pathogenicity, the detection of a polybasic cleavage site was used as a proxy. While the presence of a polybasic cleavage site alone does not indicate an increase of pathogenicity, it is nonetheless present in highly pathogenic strains [20].

The phenotypes were coded as binary variables, as phenotypic data were only available as a Yes / No statement. Our analysis was performed blindly, as no indication of any particular mutation was included in the strain name (Table 1). Note that our selection of strains tried to minimize class imbalance, as both infectivity and pathogenicity have a 1:1 ratio, while transmissibility data are a bit more uneven (3:7; Table 1).

The corresponding genomic data were downloaded from the Influenza Virus Resource [21]. 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. Furthermore, avian strains were not considered due to the large difference between avian and mammalian subtypes [22] which would invariably affect the results of this study.

The segments of each strain were individually aligned using MUSCLE 3.8.31 [23] 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 ensure that (i) the concatenation proceeded smoothly, and (ii) to prevent a mismatching of protein segments between the different influenza strains, and thus prevent distortion of the alignments. The concatenated alignment contained 4,475 amino acid positions.

The pseudomonas data

The proteome alignment of *P. aeruginosa* was generated by concatenating the coding regions of the 26 *P. aeruginosa* genomes previously published [19]. With a reference database of PA14 (`www.pseudomonas.com`; [24]), an alignment for each open reading frame (ORF) was created using an in-house pipeline that: (i) stored BLASTn 2.2.30 [25]

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 [26]). Following scaffold assembly, (iv) sequences were aligned using MUSCLE 3.8.31 [23]. Any alignment 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: `github.com/nylander/catfasta2phyml`). This resulted in an alignment containing 5,944 of the 5,977 ORFs in 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 also been characterized phenotypically, with respect to their antibiotic resistance to three different drugs: Ciprofloxacin, Ceftazidime, and Gentamicin [19]. All three are broad-spectrum drugs, used to treat patients infected by *P. aeruginosa*, and 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 enzymes such as DNA gyrase and topoisomerase IV, involved in the replication of DNA [27]; cephalosporins interrupt the synthesis of the peptidoglycan layer forming the bacterial cell wall [28];

aminoglycosides bind to the 30s ribosomal subunit and inhibit protein synthesis [29]. 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. As the algorithms used to identify the genetic determinants of phenotypes require categorical data, MIC distributions were discretized. Based on the shape of these distributions (Figure 3), three categories (low, medium, and high) were defined by setting two MIC thresholds (Table 2). As in the case on viral data, these choices were made to try and avoid class imbalance. Because these choices are somewhat arbitrary, four different sets of thresholds were assessed (see Table 2). These thresholds further allowed us to gauge the robustness / stability of our results to these assumptions.

Predictive modeling

Adaptive boosting (AdaBoost) [10] was used to construct a model to predict a particular phenotype from genomic information (proteome). This machine learning algorithm is implemented in the R package adabag 4.1 [30]. As an iterated process, the total number of iterations (m_{final}) was left to its default value of 100 – unless otherwise stated. All scripts were run in R 3.2.3 [31], and are available from `https://github.com/sarisbro`.

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.

One limitation of adaptive boosting is that this algorithm typically requires a large amount of memory to run, in particular for large datasets such as the *P. aeruginosa* alignment. To alleviate this large memory footprint, alignments were split in chunks of pre-specified sizes, ranging from 75 to 175 amino acids (by increments of five) for the influenza alignment, and either 500 or 5000 amino acids for the pseudomonas data. Each classifier was first run on each chunk as a first pass, hereby producing a set of positions of interest. Only those with an importance value > 1 were kept, 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). The effect of chunk size was extensively assessed in the case of the influenza alignment, which is smaller than *P. aeruginosa*'s. Both stability of the results and runtimes were recorded.

All influenza phenotypes were discrete, and hence easily amenable to the analysis with adaptive boosting. In the case of *P. aeruginosa* however, MICs are continuous variables that need to be discretized before analysis. For this, the distribution of $\log_2$ MIC of each drug was first plotted to determine three categories of 'low', 'medium', and 'high' MIC. 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 the results.

Finally, performance of the machine learning algorithm was assessed using ten-fold cross-validations. For this, the sequence alignment is divided into ten parts (sequences), nine being used for training and the remaining one for testing. That process is then repeated for all ten subsets [30]. In the context of our chunking procedure, cross-validation was performed on the second pass of the AdaBoost algorithm.

Results and Discussion

Chunking improves runtimes without affecting results

To determine how chunking affects the predictions of the adaptive boosting algorithm in the case of the influenza data, 20 chunk sizes were compared, ranging from 75 to 175 by increments of five. This was repeated for each of the three phenotypes. As expected from increased memory requirements, increasing chunk sizes also increases runtimes exponentially (Figure 1), so that smaller chunk sizes can reduce runtime by at least by a factor ten (1 $\log_{10}$ unit). An ANCOVA showed that runtimes are similarly affected across all phenotypes (similar slopes: $P = 0.8530$), but that pathogenicity runs the slowest (intercept: 3.86; $P = 1.49 \times 10^{-10}$), followed by transmission (intercept: 2.69; $P = 4.32 \times 10^{-9}$) and infectivity (intercept: 2.35; $P = 1.51 \times 10^{-7}$). This potentially indicates a large difference in the amount of information that is available for the algorithm to determine amino acids of interest.

Qualitatively, while there were differences in terms of the amino acids predicted to be important for each phenotype, these were minor differences (Figure 2; see Venn diagrams in insets, and distributions of importance values at extreme chunk sizes). For instance, in the case on infectivity (Figure 2A), the six most important sites at chunk size 75 were among the top eleven sites at intermediate (125) and largest (175) chunk sizes. Furthermore, the top site, HA 108, was the most important at all these chunk sizes, while PB2 637 and 667 were always ranked second or third. However, the other sites showed a dramatic drop in importance. The stability of the top three sites may be due to their relatively large importance values, reflecting the presence of strong information in the data. Similarly, the fourth through to fifteenth sites show both similar and smaller importance values, suggesting their unstable ranking is due to limited information content

in the data. While this pattern also held true for pathogenicity, that was not the case for transmissibility. As such, it is difficult to argue that only large changes in ranked importance should be used to evaluate the relative value of these predictions.

Lastly, the main difference that could be discerned from increasing chunk sizes was a decrease in sensitivity in the first pass of the algorithm, and maybe an increase in its final specificity. For the case of infectivity, the algorithm found a total of about 275 sites at chunk size 75, and roughly 150 sites at chunk size 175 (Figure 2A). The same pattern was found for the two other phenotypes, transmissibility and pathogenicity. This reduction amounts to a decreased sensitivity in the first pass of the algorithm. However, more sites with importance values > 1 tended to be found after the second pass. Furthermore, in the first two phenotypes, infectivity and transmissibility, only two predicted sites were supported by experimental evidence, and their ranking did not change much. Yet, for pathogenicity, the smallest and the largest chunk sizes predicted three experimentally validated sites (two were found at intermediate chunk sizes), but they tended to have higher rankings at larger chunk sizes. Statistics are difficult to compute on such a small number of sites, so that this pattern of increased specificity at larger chunk sizes should be further investigated.

Rediscovery of influenza determinants

For each phenotype, a number of the most important amino acid positions that were detected also had experimental evidence supporting our results (Figure 2). Infectivity had the smallest number of sites detected, but the largest number of experimentally validated positions. Among those that are unambiguously associated with infectivity were PB2 667 [32] and PB2 627 [33]. PB2 9 is a determinant for mitochondrial import, which can play a role in innate immune responses to viral infection [34] and hence, infectivity. Likewise,

mutations at PB2 105 are known to affect pathogenicity and RNAP II degradation [35], while also having a potential role in infectivity. Site PB2 339 is involved in a triplet of mutations that are responsible for replication [36], which could impact infectivity [37]. Position PB1 757 is known to be involved in contact between PB2 and PB1 [38], but its role in infectivity is not clear. For transmissibility, PB2 627 was detected by the analysis. Both known transmissibility mutations for PB2 627 (K and E) [39] were present in the strains tested, but PB2 627 has also been associated with changes in pathogenicity. HA 246 has been linked to receptor specificity and affinity [40]. PB2 391 [41] is also potentially meaningful. For pathogenicity, we already saw the role of PB2 105 [35] and PB2 627 [39]. The second most important site, PB1 215, is also supported by experimental evidence [42]. Site PB2 667 [32] was also shown to affect the pathogenicity of the strains. Site PB2 339, which was linked to infectivity, could also have an effect on virulence [37].

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 are false positives – but absence of experimental evidence is not evidence of absence. On the other hand, a number of key sites listed in the Influenza Research Database are also missed by our machine learning algorithm. For instance, PB2 256 is known to increase polymerase activity, and hence boost infectivity, at least in pigs [43]. Likewise, PB2 28, 274, 526, and 607 do the same, but in birds [44]. As our alignment only contains sequences isolated from mammals, it is possible that some of the sites we uncover are highly specific to these particular hosts. 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 some known sites, supported by experimental evidence, are rediscovered by adaptive boosting for all three phenotypes

considered here.

Detecting novel determinants in *P. aeruginosa*

Given the encouraging performance of adaptive boosting on the influenza data, for which experimental evidence supports some of the identified genetic determinants of three complex phenotypes (infectivity, transmissibility, and pathogenicity) with a small number of strains (ten), we ran our algorithm on an alignment of previously sequenced bacteria for which we had access to MIC values for three antibiotics (Ciprofloxacin, Ceftazidime, and Gentamicin) [19]. For each phenotype (MIC value; Figure 3), we ran the 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 stability and detection. As with the influenza alignment, increasing chunk size tended to lower sensitivity in the first pass of the algorithm, as fewer positions of interest were found (Additional file 1, 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 (Additional file 1, main panels). The Additional file 2 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 (Additional file 2). Some of these predictions are sensible: DNA gyrase subunit B (gyrB) is known to be involved in drug resistance [45, 27], hypothetical protein PA14_40040 (see Ceftazidime phenotype runs) is known for its involvement in antibiotic biosynthesis processes [46], and tonB2 is inferred to be an iron transporter which potentially affects bacterial drug resistance [47]. 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 [48]. What is more unexpected though is to identify genes that are not supposed to be involved in the mode of action of a specific drug: for instance, Ciprofloxacin, as a fluoroquinolone, is known to disrupt gyrB [27], but it is also the only drug in our results for which gyrB was not identified (Table 3). Likewise, sbrR encodes a factor involved in swimming ability [49], 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 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 (26) leading to unstable training.

To better quantify general performance of the machine learning algorithm 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, during cross-validation, class imbalance can be quite high for resistance to the three drugs, leading to no predicted strain. This can be seen *e.g.* at medium MIC values for Ceftazidime and Gentamicin:

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

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 machine learning algorithm did not have a good performance, but 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 do not explain neither the instability of the pseudomonas results, nor their lack of complete biological sensibility.

A possibility is that adaptive boosting is actually overfitting the data: this happens when a particular classifier accommodates all the most minute singularities of a data set. While adaptive boosting is generally considered to be robust to overfitting [50, 13], this can occur when too many iterations are performed [14] – which prompted some authors to consider m_{final} as the main tuning parameter of adaptive boosting [51]. To assess whether overfitting was responsible for the poor pseudomonas results, we reran the 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 5 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 is 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 is trpI (PA14_00460, Table 3, underlined), a transcriptional activator implicated in Tryptophan biosynthesis. Mutations in this gene lead to reduced (albeit modest) swimming motility [52], implicated in ciprofloxacin resistance [53]. 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 previously analysis, which suggests that, for a given discretization scheme of the MIC values (phenotypes), these gene lists are fairly robust to the number of iteration (m_{final}), and that overfitting is probably not an issue in this application.

Conclusions

In order to find the genetic determinants of particular phenotypes, we implemented and tested a machine learning approach based on adaptive boosting (the AdaBoost algorithm), which has recently proved to be quite successful when analyzing small data sets [12]. The machine learning approach applied here can be characterized as being unbiased, in that it gauges the importance of every single position in a proteome without any *a priori* assumptions – just as RNA-seq studies are often characterized [54, 55, 56]. However, because each proteome contains a very large number of positions, this algorithm 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 approaches typically focus on sequence windows centered on the traits of interest, thereby reducing the feature space [57, 58]. Here, we did not have such prior knowledge to pre-define features, so we took a more agnostic approach by using the entire alignment of polymorphic positions. This however prompted us to develop the chunking algorithm,

where the initial alignment is chopped up into smaller parts or chunks. Adaptive boosting 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.

We then showed with a small genomic influenza alignment (ten viruses $\times$ 4,392 polymorphic positions) that adaptive boosting correctly identified some positions supported by experimental evidence for three complex phenotypes [59], but that the top results also included some potential false positives, and missed known sites. Pushing the algorithm further on the pseudomonas alignment (26 individuals *times* 511,780 polymorphic positions) also allowed us to identify positions in genes known, or predicted to be, involved in the phenotypes assayed (such as gyrB; [60]), but (i) most results proved quite unstable and biologically unreasonable, (ii) cross-validation suggested fairly high error rates – even if they were much lower than making random assignments, and (iii) that overfitting could be ruled out. Thus, it is recommended that small training sets only be used in preliminary or exploratory analyses, where obtaining a sufficiently sized sample size is otherwise too difficult or costly. These preliminary and/or exploratory analyses will allow for a quicker throughput of project ideas at a potentially reduced time and monetary cost, but might blossom when both phenotyping and genome sequencing costs are low enough to perform these analyses on many more individuals.

Declarations

Acknowledgements

We thank Jeremy Dettman and Rees Kassen for sharing with us their MIC and sequencing data, as well as the Center for Advanced Computing and Compute Ontario for providing us access to their servers.

Funding

This work was supported by the University of Ottawa's Undergraduate Research Opportunity Program (GSL, MH) and assistantships (JD), and the Natural Sciences Research Council of Canada (SAB).

Availability of data and materials

The data and source code used in this study are available from `github.com/sarisbro`.

Authors' contributions

SAB conceived the research; GSL, MH, JD and SAB wrote parts of the R programming code; JD participated in method design and data handling; GSL and MH performed the data analyses. All authors wrote parts and edited the complete manuscript, before reading and approving the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable. We have no human or animal data involved.

References

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

2. Goodwin, S., McPherson, J.D., McCombie, W.R.: Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet **17**(6), 333–51 (2016). doi:10.1038/nrg.2016.49

3. Muir, P., Li, S., Lou, S., Wang, D., Spakowicz, D.J., Salichos, L., Zhang, J., Weinstock, G.M., Isaacs, F., Rozowsky, J., *et al.*: The real cost of sequencing: scaling computation to keep pace with data generation. Genome biology **17**(1), 53 (2016)

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

5. Hidalgo, O., Pellicer, J., Christenhusz, M., Schneider, H., Leitch, A.R., Leitch, I.J.: Is there an upper limit to genome size? Trends in Plant Science (2017)

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

7. Hastie, T., Tibshirani, R., Friedman, J.: The elements of statistical learning. In: Overview of Supervised Learning, pp. 9–41. Springer, ??? (2009)

8. Goldstein, B.A., Navar, A.M., Carter, R.E.: Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. Eur Heart J (2016). doi:10.1093/eurheartj/ehw302

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

10. Freund, Y., Schapire, R.E., *et al.*: Experiments with a new boosting algorithm. In: ICML, vol. 96, pp. 148–156 (1996)

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

12. Shoji, A., Aris-Brosou, S., Culina, A., Fayet, A., Kirk, H., Padget, O., Juarez-Martinez, I., Boyle, D., Nakata, T., Perrins, C.M., Guilford, T.: Breeding phenology and winter activity predict subsequent breeding success in a trans-global migratory seabird. Biol Lett **11**(10) (2015). doi:10.1098/rsbl.2015.0671

13. Zhou, Z.-H.: Ensemble Methods: Foundations and Algorithms. CRC press, ??? (2012)

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

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

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

17. Goldstein, B.A., Polley, E.C., Briggs, F.B.S.: Random forests for genetic association studies. Stat Appl Genet Mol Biol **10**(1), 32 (2011). doi:10.2202/1544-6115.1691

18. Northrop Grumman Health IT, V.T. J. Craig Venter Institute: Influenza Research Database (2017). `https://www.fludb.org/brc/home.spg?decorator=influenza`

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

20. Stech, O., Veits, J., Weber, S., Deckers, D., Schröer, D., Vahlenkamp, T.W., Breithaupt, A., Teifke, J., Mettenleiter, T.C., Stech, J.: 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 **83**(11), 5864–5868 (2009)

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

22. Subbarao, K., Klimov, A., Katz, J., Regnery, H., Lim, W., Hall, H., Perdue, M., Swayne, D., Bender, C., Huang, J., *et al.*: Characterization of an avian in-

fluenza A (H5N1) virus isolated from a child with a fatal respiratory illness. Science **279**(5349), 393–396 (1998)

23. Edgar, R.C.: Muscle: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res **32**(5), 1792–7 (2004). doi:10.1093/nar/gkh340

24. 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**(D1), 646–53 (2016). doi:10.1093/nar/gkv1227

25. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J.: Basic local alignment search tool. J Mol Biol **215**(3), 403–10 (1990). doi:10.1016/S0022-2836(05)80360-2

26. Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T., Carey, V.J.: Software for computing and annotating genomic ranges. PLoS Comput Biol **9**(8), 1003118 (2013). doi:10.1371/journal.pcbi.1003118

27. Dalhoff, A.: Global fluoroquinolone resistance epidemiology and implictions for clinical use. Interdiscip Perspect Infect Dis **2012**, 976273 (2012). doi:10.1155/2012/976273

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

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

30. Alfaro, E., Gámez, M., García, N.: adabag: An R package for classification with boosting and bagging. Journal of Statistical Software **54**(2), 1–35 (2013)

31. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2015). R Foundation for Statistical Computing. https://www.R-project.org/

32. Sharshov, K., Romanovskaya, A., Uzhachenko, R., Durymanov, A., Zaykovskaya, A., Kurskaya, O., Ilinykh, P., Silko, N., Kulak, M., Alekseev, A., *et al.*: Genetic and biological characterization of avian influenza H5N1 viruses isolated from wild birds and poultry in Western Siberia. Archives of virology **155**(7), 1145–1150 (2010)

33. Bogs, J., Kalthoff, D., Veits, J., Pavlova, S., Schwemmle, M., Mänz, B., Mettenleiter, T.C., Stech, J.: 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 **85**(20), 10691–10698 (2011)

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

35. Llompart, C.M., 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 **88**(6), 3455–63 (2014). doi:10.1128/JVI.02263-13

36. Yamaji, R., Yamada, S., Le, M.Q., Li, C., Chen, H., Qurnianingsih, E., Nidom, C.A., Ito, M., Sakai-Tagawa, Y., Kawaoka, Y.: Identification of PB2 mutations responsible for the efficient replication of H5N1 influenza viruses in human lung epithelial cells. J Virol **89**(7), 3947–56 (2015). doi:10.1128/JVI.03328-14

37. Li, J., Ishaq, M., Prudence, M., Xi, X., Hu, T., Liu, Q., Guo, D.: 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 **144**(1), 123–129 (2009)

38. Sugiyama, K., Obayashi, E., Kawaguchi, A., Suzuki, Y., Tame, J.R.H., Nagata, K., Park, S.-Y.: Structural insight into the essential PB1-PB2 subunit contact of the influenza virus RNA polymerase. EMBO J **28**(12), 1803–11 (2009). doi:10.1038/emboj.2009.138

39. Herfst, S., Schrauwen, E.J., Linster, M., Chutinimitkul, S., de Wit, E., Munster, V.J., Sorrell, E.M., Bestebroer, T.M., Burke, D.F., Smith, D.J., *et al.*: Airborne transmission of influenza A/H5N1 virus between ferrets. science **336**(6088), 1534–1541 (2012)

40. Hensley, S.E., Das, S.R., Bailey, A.L., Schmidt, L.M., Hickman, H.D., Jayaraman, A., Viswanathan, K., Raman, R., Sasisekharan, R., Bennink, J.R., Yewdell, J.W.: Hemagglutinin receptor binding avidity drives influenza A virus antigenic drift. Science **326**(5953), 734–6 (2009). doi:10.1126/science.1178258

41. Ellebedy, A.H.: Impact of Adjuvants on the Antibody Responses to Pre-pandemic H5N1 Influenza Vaccines. The University of Tennessee Health Science Center, ??? (2011)

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

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

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

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

46. Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A.F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S.-Y., Lopez, R., Hunter, S.: Interproscan 5: genome-scale protein function classification. Bioinformatics **30**(9), 1236–40 (2014). doi:10.1093/bioinformatics/btu031

47. de Carvalho, C.C.C.R., Fernandes, P.: Siderophores as "Trojan Horses": tackling multidrug resistance? Front Microbiol **5**, 290 (2014). doi:10.3389/fmicb.2014.00290

48. Kos, V.N., McLaughlin, R.E., Gardner, H.A.: 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. http://aac.asm.org/content/early/2016/04/05/AAC.03113-15.full.pdf+html

49. McGuffie, B.A., Vallet-Gely, I., Dove, S.L.: factor and anti- factor that control swarming motility and biofilm formation in pseudomonas aeruginosa. J Bacteriol **198**(5), 755–65 (2015). doi:10.1128/JB.00784-15

50. Schapire, R.E., Freund, Y.: Boosting: Foundations and Algorithms. MIT press, ??? (2012)

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

52. Yeung, A.T.Y., Torfs, E.C.W., Jamshidi, F., Bains, M., Wiegand, I., Hancock, R.E.W., Overhage, J.: Swarming of *pseudomonas aeruginosa* is controlled by a broad spectrum of transcriptional regulators, including metr. J Bacteriol **191**(18), 5592–602 (2009). doi:10.1128/JB.00157-09

53. Amini, S., Hottes, A.K., Smith, L.E., Tavazoie, S.: Fitness landscape of antibiotic tolerance in pseudomonas aeruginosa biofilms. PLoS Pathog **7**(10), 1002298 (2011). doi:10.1371/journal.ppat.1002298

54. Ameur, A., Wetterbom, A., Feuk, L., Gyllensten, U.: Global and unbiased detection of splice junctions from RNA-seq data. Genome Biol **11**(3), 34 (2010). doi:10.1186/gb-2010-11-3-r34

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

56. Torres-Oliva, M., Almudi, I., McGregor, A.P., 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 **17**, 392 (2016). doi:10.1186/s12864-016-2646-x

57. Leung, M.K.K., Xiong, H.Y., Lee, L.J., Frey, B.J.: Deep learning of the tissue-regulated splicing code. Bioinformatics **30**(12), 121–9 (2014). doi:10.1093/bioinformatics/btu277

58. Xiong, H.Y., Alipanahi, B., Lee, L.J., Bretschneider, H., Merico, D., Yuen, R.K.C., Hua, Y., Gueroussov, S., Najafabadi, H.S., Hughes, T.R., Morris, Q., Barash, Y., Krainer, A.R., Jojic, N., Scherer, S.W., Blencowe, B.J., Frey, B.J.: RNA splicing. the human splicing code reveals new insights into the genetic determinants of disease. Science **347**(6218), 1254806 (2015). doi:10.1126/science.1254806

59. Stockwell, D.R., Peterson, A.T.: Effects of sample size on accuracy of species distribution models. Ecological modelling **148**(1), 1–13 (2002)

60. Arais, L.R., Barbosa, A.V., Carvalho, C.A., Cerqueira, A.M.F.: Antimicrobial resistance, integron carriage, and gyrA and gyrB mutations in *Pseudomonas aeruginosa* isolated from dogs with otitis externa and pyoderma in Brazil. Vet Dermatol **27**(2), 113–731 (2016). doi:10.1111/vde.12290

Tables

Table 1. 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/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

Table 2. The different sets of MIC thresholds employed to assess the robustness of the classification results in the case of *P. aeruginosa*. Shown are the thresholds used on a $\log_2$ MIC scale: two thresholds separated by a slash (/) imply the use of three discrete categories (low, medium, and high MIC); for instance, setting 1 for Ciprofloxacin means that MIC is low when ≤ -1, high when MIC > 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*

Figures

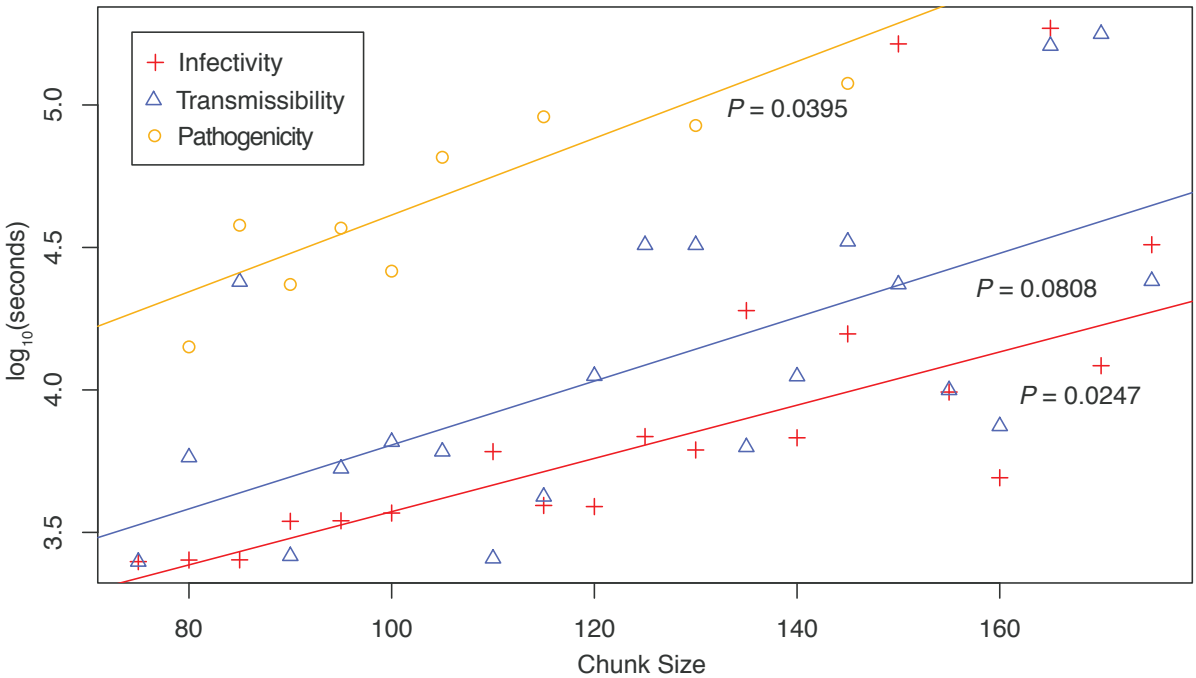


Figure 1. Impact of chunk size on the runtime of the machine learning algorithm for the influenza data. Runtime is shown for infectivity (red), transmissibility (blue), and pathogenicity (orange). Solid lines represent the robust regressions; their P-values are also shown.

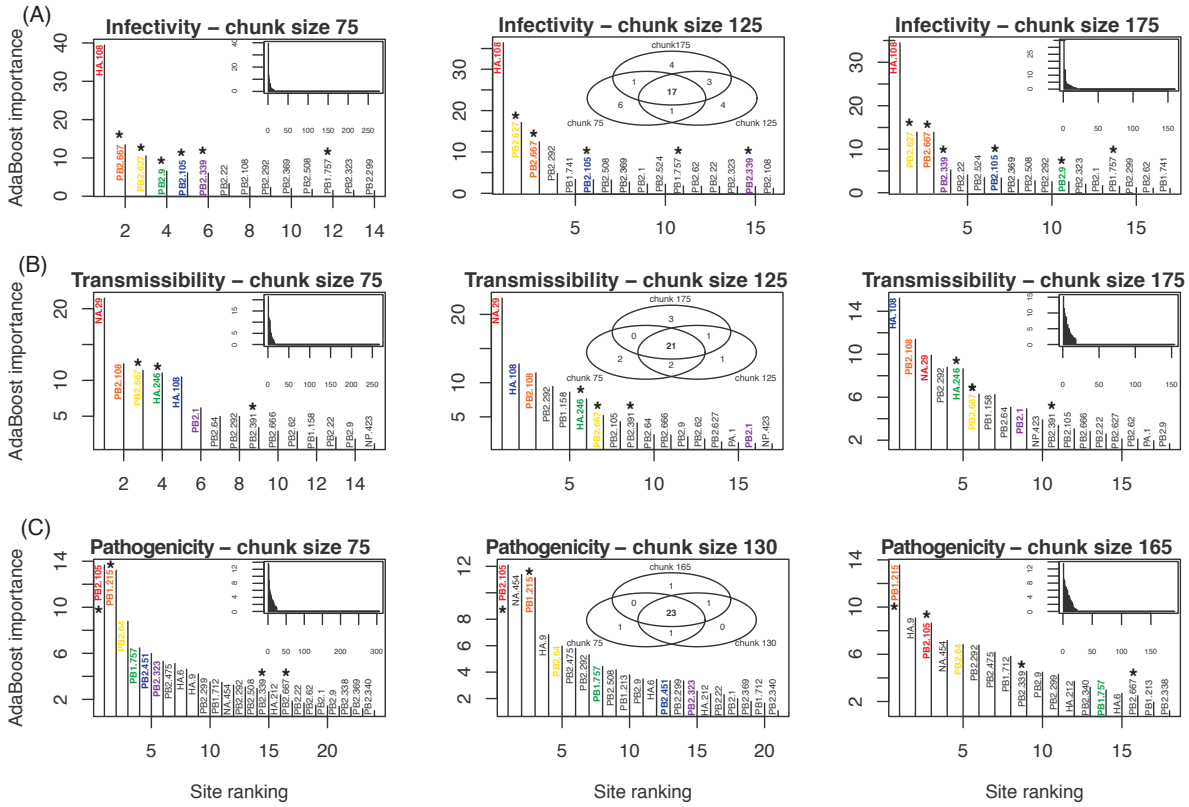


Figure 2. Effect of chunk size on the most important genetic determinants of influenza phenotypes. The genes and sites identified are shown for: (A) for infectivity, (B) for transmissibility, and (C) for pathogenicity. Only the smallest (75 amino acids), intermediate (125), and largest (175) chunk sizes are shown. Only the most important sites (importance > 1) are shown. The identifiers (gene name dot amino acid position) of top six sites found in the smallest chunk size are colored and bolded from warmer to cooler shades; these patterns are shown for the two other chunk sizes. Sites supported by experimental evidence are indicated by an asterisk. Insets show the whole distribution of importance values for all the positions of interest found in each analysis (at the extremes of the chunk size range), and the overlap of sites found at the three chunk sizes (Venn diagrams, middle column).

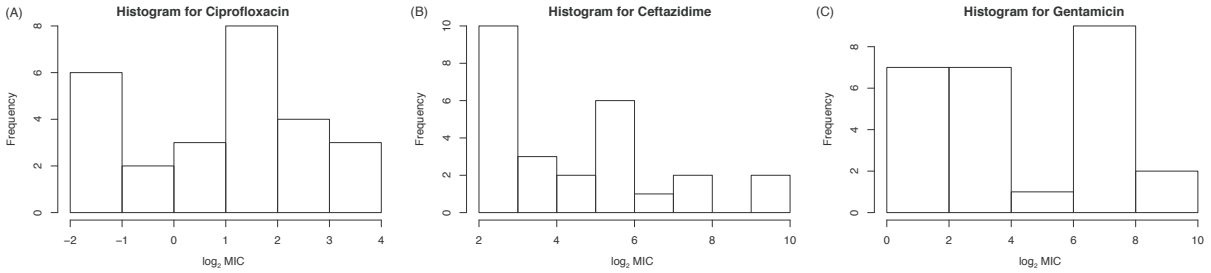


Figure 3. Distributions of $\log_2$ MICs for *P. aeruginosa* across the 26 strains from [19]. (A) for Ciprofloxacin. (B) for Ceftazidime. (C) for Gentamicin. These empirical distributions were used to determine MIC thresholds (Table 2). Note that the scales on the y-axis vary slightly.

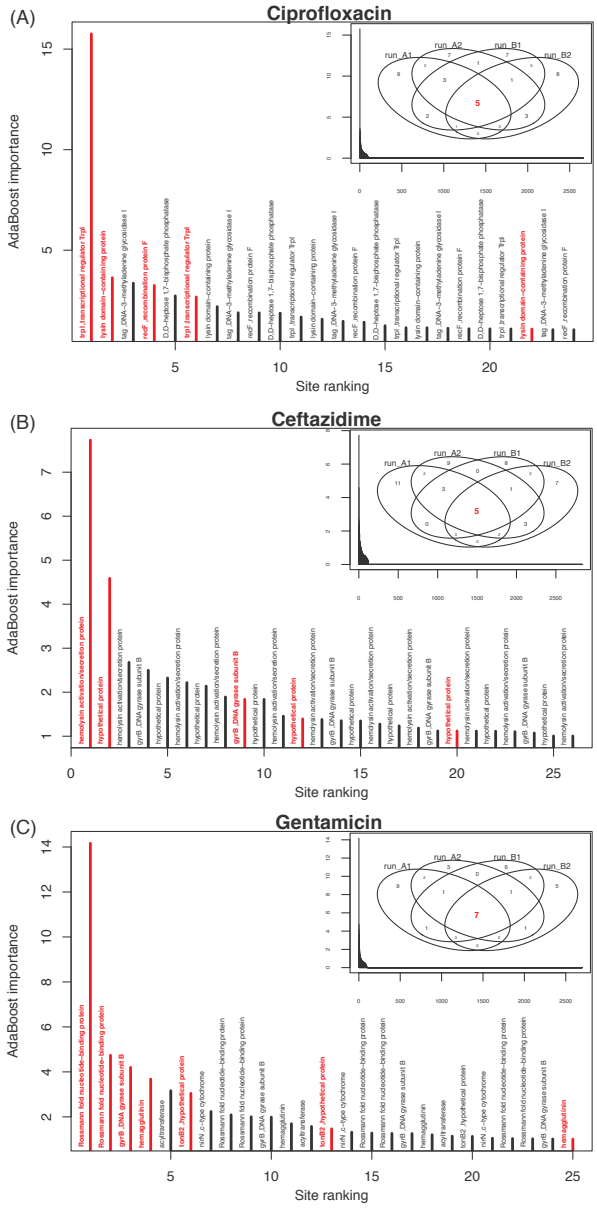


Figure 4. Importance values for the genetic determinants of drug resistance in *P. aeruginosa*. The genes identified are shown for: (A) for Ciprofloxacin, (B) for Ceftazidime, and (C) for Gentamicin. In each panel, the main figure shows the sorted importance values for the most important sites (importance > 1) as found in run A1 (replicate 1 under chunk size of 5000). The complete distribution of importance values for all amino acid positions of interest is shown as an inset, which also shows the Venn diagram (intersect) for the top 25 sites found in run A2 (second replicate under chunk size of 5000), and also B1 and B2 (two replicates under chunk size of 1000). The number of sites commonly identified by all four runs is shown in red; the genes in which these sites occur are labeled in red in the main figure. See Additional file 1 for details.

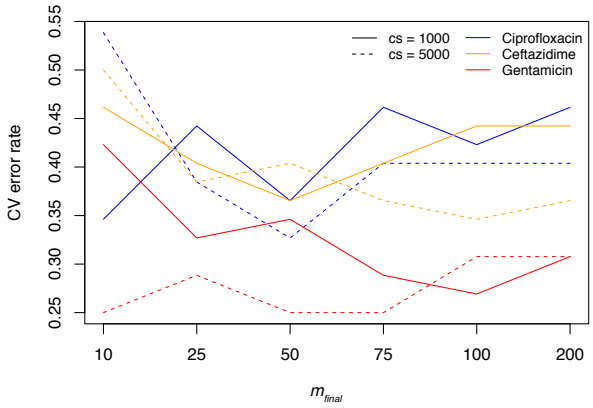
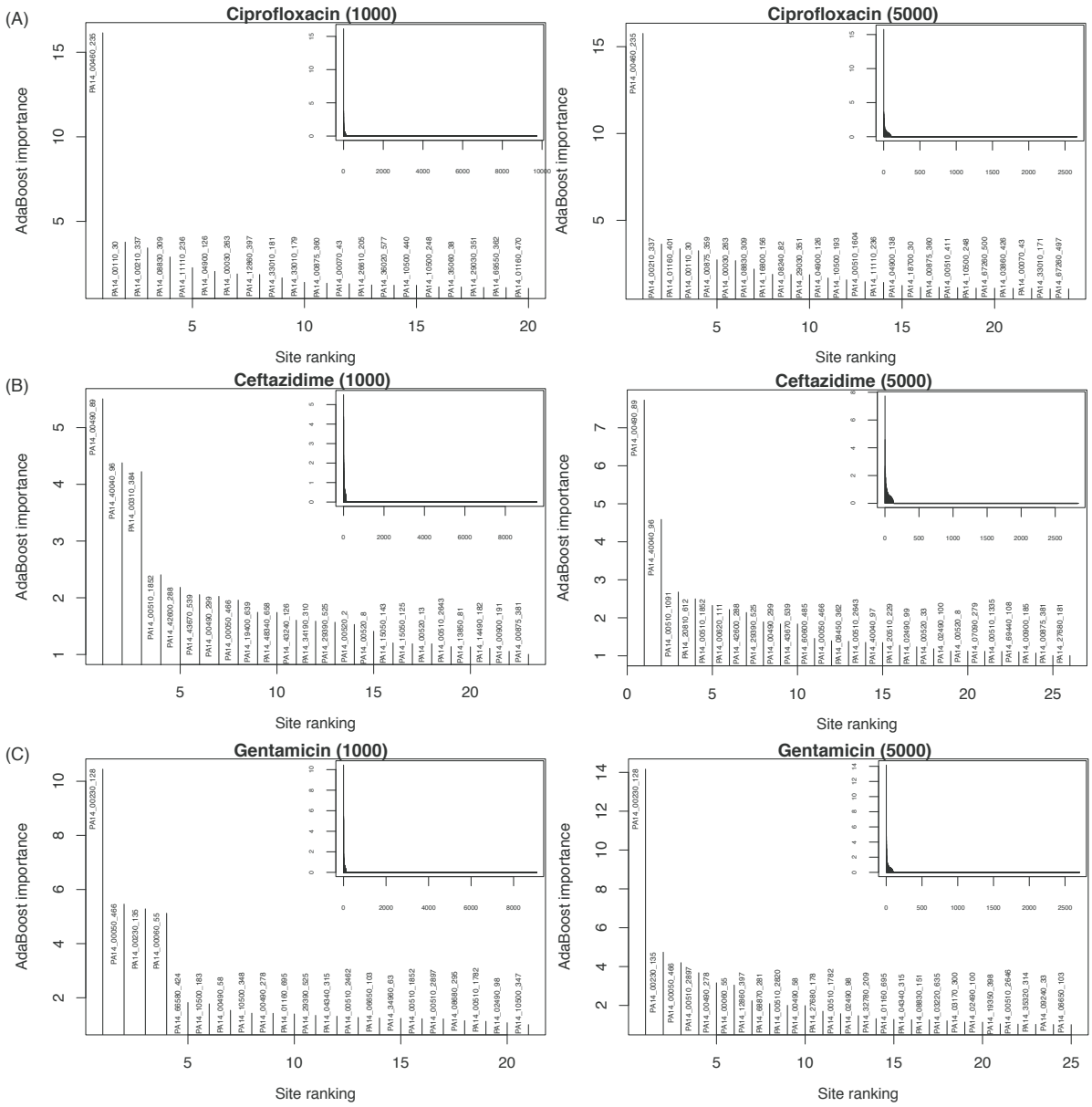
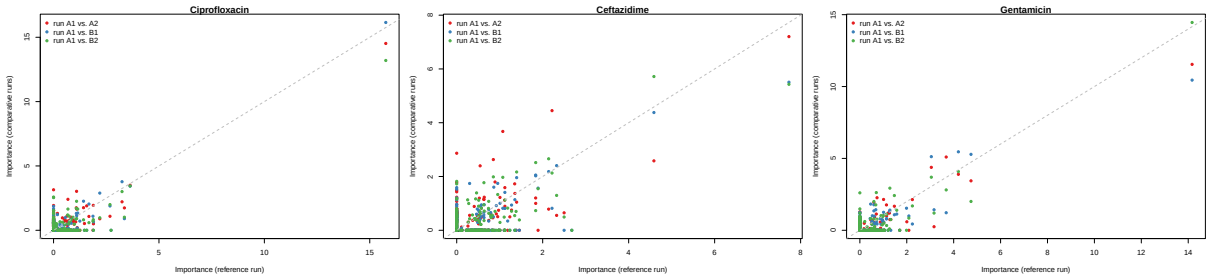


Figure 5. Cross validation errors rates as a function of total number of iteration for the *P. aeruginosa* data. In each case, the mean ten-fold cross-validation error rates were computed over two replicates, and are shown for Ciprofloxacin (blue), Ceftazidime (orange), and Gentamicin (red), with a chunk size of 1000 (solid lines) or 5000 (broken lines).

Additional files



Additional file 1. Importance values for the genetic determinants of drug resistance in *P. aeruginosa*. The genes and sites identified are shown for: (A) for Ciprofloxacin, (B) for Ceftazidime, and (C) for Gentamicin. In each panel, the main figure shows the sorted importance values for the most important sites (importance > 1). Results for chunk size of 1000 are shown on the left, and chunk size 5000 (as in Figure 4) on the right. The complete distribution of importance values for all the amino acid positions of interest is shown as insets.



Additional file 2. Stability of importance values for the genetic determinants of drug resistance in *P. aeruginosa*. The sites identified are shown for: (A) for Ciprofloxacin, (B) for Ceftazidime, and (C) for Gentamicin. In each panel, the comparison of importance values is between run A1 (replicate 1 under chunk size of 5000, the reference run) and comparative runs A2 (red), B1 (replicate 1 under chunk size of 1000; blue), and B2 (green).