Machine learning applied to molecular identification of *Acinetobacter baumannii* Global Clone 1

Verónica Elizabeth Álvarez[1]#, María Paula Quiroga[1] and Daniela Centrón[1]

[1]Laboratorio de Investigaciones en Mecanismos de Resistencia a Antibióticos, Instituto de Investigaciones en Microbiología y Parasitología Médica, Facultad de Medicina, Universidad de Buenos Aires - Consejo Nacional de Investigaciones Científicas y Tecnológicas (IMPaM, UBA-CONICET), Buenos Aires, Argentina.

Running Head: Identification of *Acinetobacter baumannii* GC1

#Address correspondence to Verónica E. Álvarez, valvarez@gmail.com

Word count for the abstract: 220

Word count for the text: 4292

ABSTRACT

Since the emergence of clones with epidemic behavior, or "high-risk clones", uninterrupted investigations have been undertaken to understand the genetic and biological basis that led their widespread in nosocomial settings at a global scale. However, its complex genetic and ecological nature only allowed the identification of partial biomarkers with low specificity. The implication of our study is that we successfully applied machine learning algorithms to recognize genetic biomarkers for the identification of *Acinetobacter baumannii* Global Clone 1, one of the most widespread high-risk clones. Support Vector Machine model predicted U1 sequence with 367 nucleotides length as a specific biomarker for *A. baumannii* GC1 strains. U1 sequence matched AYE genome in ABAYE1552 locus between 558 and 924 coordinates that corresponds to a fragment of the *moaCB* gene which encodes a bifunctional protein that includes the molybdenum cofactor biosynthesis protein C and protein B. This region of the gene *moaCB* allows to differentiate between the *A. baumannii* GC1 and non-GC1 strains, becoming a suitable biomarker capable of being translated into clinical settings as a molecular typing method such as PCR. Our findings highlight that machine learning can be fruitfully useful even in complex gaps of knowledge of epidemic clones and signify a noteworthy contribution to the literature that could be used to identify challenging nosocomial biomarkers for other multidrug resistant pandemic clones.

IMPORTANCE

High-risk clones play a major role in the spread of resistance. The factors driving the emergence and the dissemination capacity of these clones are not well understood. The

dissemination of *A. baumannii* strains belonging to high-risk GC1 is a global concern. Molecular methods have been used to type GC1 clones such as a particular deletion in *intI1* that identify among islands in *comM* with a Tn*6019* backbone. However, since some GC1 strains do not harbor the genomic island AbaR, and *intI1* is very common in clinical samples, a more specific biomarker is necessary. The significance of our research is that we successfully applied machine learning methods to predict *A. baumannii* GC1. Our findings highlight that machine learning can be successfully applied not only to find high-risk clones putative biomarkers but also in disentangling genetic traits that could be used as therapeutic targets to reduce dissemination of lineages with epidemic behavior.

INTRODUCTION

Acinetobacter baumannii is a Gram-negative opportunistic and nosocomial pathogen that causes a broad range of nosocomial infections mainly implicated in ventilator-associated pneumonia, catheter-related bloodstream, urinary tract, and wound infections (Álvarez et al., 2020; Douraghi et al., 2020). Infections by this species have increased in recent years, which, added to its ability to acquire and spread antimicrobial resistance genes to all families of antibiotics, it is considered a serious global threat worldwide. (Hamidian and Nigro, 2019; Yakkala et al., 2019). The majority of *A. baumannii* isolates that are broadly resistant to antibiotics belong to two pandemic disseminated clones, known as global clones 1 (GC1) and 2 (GC2) (2, 5, 6). Recent epidemiological studies of carbapenem resistant *A. baumannii* (CRAB) isolates from Argentina, revealed that GC1 is the main widespread CRAB clone in our country (7).

Over time molecular methods with different degree of resolution have been used to type *A. baumannii* strains such as amplified fragment length polymorphism analysis, ribotyping, macrorestriction analysis by pulsed-field gel electrophoresis, multiplex polymerase chain reactions (PCRs) and multilocus sequence typing (8–14). A typing scheme based on two multiplex polymerase chain reactions (PCRs) targeting three genes (*ompA*, *csuE*, and $bla_{\text{OXA-51-like}}$) has been found a convenient method for rapid assignment of *A. baumannii* isolates into two major PCR-based groups (Gs) corresponding to GC1 (group G1) and GC2 (group G2) (15). Since evident correlation has been detected between particular $bla_{\text{OXA-51-like}}$ enzymes and particular epidemic lineages, sequence analysis of the $bla_{\text{OXA-51-like}}$ gene has also been proposed as a useful typing method for *A. baumannii* isolates (15–17). In agreement, a study conducted on sixty *A. baumannii* isolates collected

worldwide demonstrated that isolates belonging to GC1 encoded enzymes from the OXA-69 cluster which included OXA-69, OXA-92, OXA-107, OXA-110 and OXA-112 enzymes (16). However, not all the isolates encoding the OXA-69 cluster belonged to GC1 such as A92 that belongs to GC2 (16). Concerning also the molecular identification of GC1 strains, another PCR approach consisted in the detection of a deletion of 108 bp in the 5′-conserved segment (5′-CS) of the class 1 integron located in AbaR3 genomic island (18); since AbaR3 is not present in all GC1 strains, or some isolates with an intact 5′-CS, or even with a deletion were both present among GC1 strains, this approach serves only as a marker for some diverged lineages within the GC1 clone (18).

With the increasing throughput and decreasing cost of DNA sequencing, whole genome sequencing (WGS) may be an alternative for rapidly identification of pathogens of all types of phyla based on a small sample of DNA (19). Genome-wide studies of DNA variation related to antibiotic resistance phenotypes are one of the most concerning and have garnered high public interest, especially since several multidrug-resistant strains have emerged worldwide (20). New candidate biomarkers leading to the identification of resistant pathogens requires the study and development of fast, easily applicable and accurate tools. Furthermore, with the help of computational algorithms, such studies can be conducted at a much larger scale and produce more significant results (21–24). Machine learning (ML) algorithms and statistics have been used increasingly to build models that correlate genomic variations with phenotypes and may help to predict bacterial phenotypes and genotypes (25–31). In supervised ML, each learning sample includes the outcome (class label) of interest, and it is used to build a prediction model (32). The model takes an outcome measurement (e.g., a bacterium having a resistant phenotype or genotype) and

tries to learn from the available data (e.g., information on genomic mutations) to predict the outcome measure. The developed model is then applied to new, unseen data. The goal of the algorithm is to learn a model that accurately foresee the correct outcome for any input.

Support Vector Machine (SVM) is a supervised learning algorithm formally characterized by a separating hyperplane that divides binary data to solve both classification and regression problems. SVM aims at correctly classifying the samples based on examples in the training dataset (33–35). On the other hand, the Set Covering Machine (SCM) is a supervised learning algorithm that uses a greedy approach to produce uncharacteristically sparse rule-based models from the input data (36). Both algorithms have been applied to several biological problems and proved to be accurate in predicting novel antibacterials (37), antimicrobial resistance (27, 28, 38–40), identification and classification of microorganisms (41–43), cancer diagnosis (44–47) among others (25, 41, 48–51).

In this study we built predictive models of *A. baumannii* GC1 genome typing by training SVM and SCM machine learning classifiers. From these classifiers, we inferred a new genomic biomarker for GC1 strains not subjected to selective pressure such as antimicrobial determinants. We hypothesized that ML models could be used to elucidate genetic determinants that could be useful for identifying high-risk clones such as *A. baumannii* GC1.

RESULTS

The purpose of our study was to try to identify new genomic biomarkers that uniquely identify strains within the *A. baumannii* GC1. Having a set of *A. baumannii*

genomes, we aimed to predict whether a genome belonged to GC1 clone or not, by using SVM and SCM models and to obtain which biomarker sequences contribute most to the prediction. We used two *A. baumannii* previously sequenced genome collections named Dataset 1 (Table S1) and Dataset 2 (Table S2). In a first step, Dataset 1 was used as input for the SVM and SCM algorithms during training and testing the models. Once we obtained accurate models that predicted putative biomarkers sequences, we used Dataset 2 to perform an external validation of the model predictions. Using blastn searches, we analyzed whether the putative biomarker sequences were also found in Dataset 2 genomes maintaining the same pattern found in Dataset 1. The study workflow is summarized in Figure 1.

Obtaining GC1 predictive sequences with SVM model

According to the Pasteur scheme for MLST, GC1 can be interrelated with clonal complex (CC) 1 (52, 53). CC1 was initially composed of five STs (ST1, ST7, ST8, ST19 and ST20). The neat demarcation of these five STs from other non-CC1 STs endorsed the potential of this scheme to identify isolates belonging to CC1, and subsequently to GC1 (52, 53). Considering this fact, we annotated the ST of each genome in Dataset 1 and categorized them as “GC1” or “non-GC1”. Then we run DBGWAS program (54) using Dataset 1 genome sequences as input. DBGWAS uses an extended k-mer-based Genome-Wide Association Study (GWAS) method producing interpretable genetic variants associated with distinct phenotypes. Relying on compacted De Bruijn Graphs (cDBG), the method built a single cDBG from all the genomes included in Dataset 1. The graph nodes, which are called unitigs, represented sequences of variable length of Dataset 1 taking into account the genomic variation (54). A total of 1,622,573 distinct unitigs were obtained from

the 500 input genomes. The length of the unitigs obtained was between 31 and 32,759 bp. We used the variant matrix built by DBGWAS to create the binary matrix used as input for the SVN algorithm. Unitigs found in less than 50 genomes were discarded keeping in the input binary matrix only the unitigs of size 31 to 385 bp. In the binary input matrix, the rows represented Dataset 1 genomes, and the columns depicted presence “1” or absence “0” of each unitig in a given genome. The columns and the rows are referred to as “features” and “instances”, respectively, in ML literature. We also used a two-column binary matrix as input for the SVM algorithm. This matrix indicated for each genome whether it belonged to the CG1 (value = 1) or not (value = -1). We optimized the hyperparameters of SVM models on the training set and assessed the best values using a cross-validated grid search. Once obtained the SVM model that best fitted Dataset 1 genomes, we extracted the first 100 unitigs that contributed most to the GC1 genome prediction (Tables 1 and S3) according to the values of the features weight vector. We used blastn to search the unitig sequences in Dataset 1 and Dataset 2 genomes and account the number of matches of the unitigs in each genome class (GC1 or non-GC1) to find the most representative unitigs within the GC1 genomes (Tables 2 and S4). AYE genome (AN: CU459141.1) was used as reference to annotate the genome location and gene product related to the 100 unitigs. In case the unitig was not found in AYE genome, AB0057 (CP001182.2) genome was used instead.

Building SCM models for GC1 identification

In this approach, we applied the SCM algorithm implemented in Kover (25) to GC1 genome identification. Kover combines SCM algorithm with the k-mer representation of genomes, which reveals uncharacteristically sparse models that explicitly highlight the relationship between genomic variations and the phenotype of interest (25). We used as

input the k-mers of the 500 *A. baumannii* genomes contained in Dataset 1 and the class label for each genome (GC1 or non-GC1). We ran Kover using k-mer sizes ranged from 31 to 127. Although output rules could be conjunctions (logical-AND) or disjunctions (logical-OR), we obtained 49 simple rules without conjunctions or disjunctions in the models of our study (Table 3 and S5). Rules depicted the presence (n=10) or the absence (n=39) of k-mers in Dataset 1 genomes. Blastn was used to search the k-mer sequences targeted by the rules in Dataset 1 and Dataset 2 genomes and account the number of matches in each genome class (GC1 or non-GC1) (Table 4 and S6). We annotated k-mer sequences by mapping them against AYE genome as GC1 reference or ACICU genome (CP000863.1) as *A. baumannii* non-GC1 reference (Tables 4 and S6).

It is worth to highlight the sequence overlap in the rules obtained by the SCM models that resulted in the selection of fragments of different length matching the loci ACICU_02924 (n=23), ACICU_02095 (n=5), ABAYE3455 (n=5), ACICU_01506 (n=2) and ABAYE2468 (n=2) (Table 3 and S5). The overlapping is caused by a point mutation contained in the loci that SCM models associated with *A. baumannii* GC1 prediction (25).

Selection of candidate biomarkers for rapid detection of *A. baumannii* GC1

We were interested in obtaining the longest DNA sequences with the highest number of matches among *A. baumannii* GC1 genomes and at the same time the least number of matches among *A. baumannii* non-GC1 genomes. For this purpose, unitigs and k-mer sequences obtained as candidate GC1 predictors in SVM and SCM models were sorted in descending order according to sequence length and the number of GC1 genomes matching the DNA sequence or rule, and ascending order according to the non-GC1 genomes matches. First, we analyzed SVM results. We named unitigs U1 through U100

(Tables 1 and S3). We observed that unitigs named U1 to U12 were found in 100% of GC1 genomes within Dataset 1 and were not found in non-GC1 genomes from Dataset 1. However, U1 was the only unitig found in 100% of the GC1 genomes within Dataset 1 and Dataset 2, and it was not found in any non-GC1 genome of both datasets being a specific biomarker for GC1 genomes (Tables 2 and S4). Despite the sequence U8 was also found in 100% of the GC1 genomes within Dataset 1 and Dataset 2, it was found in 1/4487 non-GC1 genomes from Dataset2. On the other hand, unitigs named U2, U3, U5, U10 and U11 were found in 311/312 GC1 genomes from Dataset 2. The unitig U3 was found in 3/4487 non-GC1 genomes from Dataset 2 while U2, U5, U10 and U11 were not found in non-GC1 genomes from this dataset. Regarding U4 and U6, they were found in 310 and 306 GC1 genomes, respectively, from Dataset 2 while U7, U9 and U12 were found only in 309 GC1 genomes from Dataset 2. U4 was the sequence between these 12 unitigs been analyzed that had the higher number of matches (43/4487) among the non-GC1 genomes from Dataset 2 (Tables 2 and S4).

Regarding SCM results, rules R127, R125, R123, R121, R119, R117, R115, R113, R111, R109, R107, R105, R103, R101, R99, R97, R95, R93, R91, R89, R87, R85, R83, R81, R79, R77, R75, R73, R69, R67, R65, R63, R61, R59, R57, R55, R53, R51, R49, R41, R37, R33 and R31 targeted 100% of *A. baumannii* GC1 genomes from Dataset 1 but also targeted between 1 and 4 *A. baumannii* non-GC1 genomes from Dataset 1 (Table 4 and S6). We observed that within these 43 rules, two set of rules targeted only 1 non-GC1 from Dataset 1. Rules R125, R123, R121, R119, R117, R115, R113, R111, R109, R107, R105, R103, R101, R99, R97, R95, R89, R83, R81, R79, R77, R73, R69, R63 targeted genome with AN: GCF_000248195.1 (ST 69) and R49 targeted genome with AN:

GCF_000453745.1 (ST 2). With respect to Dataset 2, rules R93, R91, R87, R85, R75, R61, R59, R57, R55, R53, R49 and R31 targeted 100% of GC1 but also targeted several *A. baumannii* non-GC1 genomes from Dataset 2 (66/4487, 28/4487, 28/4487, 28/4487, 28/4487, 26/4487, 26/4487, 14/4487, 27/4487, 26/4487, 13/4487, 42/4487 genomes, respectively). Since no rule obtained can uniquely identify *A. baumannii* GC1 genomes contained in our datasets, we were unable to obtain a putative biomarker from the results of SCM models.

When using Dataset 2 to validate the results obtained by SVM and SCM models from Dataset 1, we observed that in the case of SVM the proportion of GC1 and non-GC1 genomes matched by the unitigs remained at 100% of cases, respectively ($p > 0.05$) (Tables 2 and S4). With regard to SCM, the rules suggested by the models maintained the proportions of GC1 and non-GC1 genomes from Dataset 2 matched in 97.97% and 91.83% of the cases, respectively ($p > 0.05$) (Tables 4 and S6). As we mentioned, SVM model predicted U1 sequence as a specific biomarker for *A. baumannii* GC1 genomes. U1 sequence had 367 nucleotides and matched AYE genome in ABAYE1552 locus between 558 and 924 coordinates (Table 1 and S3). This region corresponds to a fragment of the *moaCB* gene which encodes a bifunctional protein that includes the molybdenum cofactor biosynthesis protein C and protein B (55). The molybdenum cofactor (Moco) is an essential component of a large family of enzymes involved in important transformations in carbon, nitrogen and sulfur metabolism. The Moco biosynthetic pathway is evolutionarily conserved and found in archaea, eubacteria and eukaryotes. The MoaC protein, together with the MoaA protein, is involved in the first step of Moco biosynthesis (55). On the other hand, essays done in *Escherichia coli* using precise in-frame deletions of the individual

genes of the *moa*ABCDE operon, indicated that MoaB is not required for Moco biosynthesis (56). Various studies have linked in-host survival of prevalent pathogenic bacteria such as *Mycobacterium tuberculosis*, *E. coli*, and *Salmonella enterica* to the presence of functional molybdoenzymes (57–60). Briefly, according to these results, the region of the gene *moaCB* comprised between nucleotides 558 and 924 differs between the GC1 and non-GC1 genomes, becoming a biomarker that differentiates both groups.

Performance evaluation of SVM and SCM models

To avoid overfitting, 5-folds cross-validation was performed on the training datasets used as input in SVM and SCM models to select the best hyperparameter values. The training datasets were subsets of the Dataset 1 chosen randomly by the algorithms. The best hyperparameter values selected from the five-folds cross-validation were chosen to evaluate the performance of the models. Testing performance was evaluated in terms of sensitivity, specificity, accuracy, precision and F1 score (Tables 5, 6 and S7). The sensitivity of SVM model was 1 ± 0.00, indicating that 100% of the GC1 genomes were correctly identified within the testing dataset. Also, the model performed well predicting non-GC1 genomes included in the testing datasets with a specificity value of 1 ± 0.00. The precision value was 1 ± 0.00, in other words, when the model predicted that a genome was within the GC1, it was correct 100% of the time. Accuracy was 1 ± 0.00 indicating that 100% of GC1 and non-GC1 were correctly predicted. No false positives and no false negatives were predicted by SVM model.

Regarding SCM models, the mean values of sensitivity, specificity, precision, and accuracy were 1 ± 0.01, 0.98 ± 0.01, 0.97 ± 0.01, 0.99 ± 0.01, respectively. The mean rate

of false positives was 1.11% and false negatives was 0.097%. All the rules obtained by SCM models predicted false positives within the testing dataset (Table 6).

The training accuracy and the testing accuracy of SVM and SCM models were above 0.99 (Figure 2), indicating that both the SVM and SCM models were not overfitted. The F1 score, which is the harmonic mean of precision and recall and is commonly used to compare different classification algorithms, is similarly high in SVM and SCM models (1.00 and 0.99 ± 0.01, respectively). Correlations between MLST typing and the model predictions of testing datasets (Table 5, 6 and S7) indicated that both SVM and SCM models could significantly classify *A. baumannii* GC1 and non-GC1 isolates ($p < 0.01$). Despite these results, as the aim of this work was to find a biomarker that uniquely identifies the GC1 genomes, SVM model performed better than SCM models since it did not predict false positives or false negatives.

DISCUSSION

High-risk clones are dangerous clonal complex with epidemic behavior equipped with exceptional resources both to infect the host and to evolve to extreme antimicrobial resistance phenotypes over time in the nosocomial niche (1, 61–67). The molecular understanding of the genetic and/or transcriptomic traits that lead to these capabilities of bacterial species that cause hospital-acquired infections remains still unknown (68, 69). ML applied to the study of high-risk clones can help not only in the identification of potential biomarkers but also in disentangling genetic traits that could be used as therapeutic targets to reduce dissemination of lineages with epidemic behavior. With increasing availability of bacterial whole-genome sequencing (WGS) data, there has been active investigation into using this tool for genotype-phenotype prediction of antimicrobial susceptibility (25, 28,

29, 38, 48, 70–72); however, there is not data available concerning the identification of high-risk clones based in WGS data excluding the MLST. Since it is not clear which genetic properties convert a clone in a pandemic one, the use of ML can be an interesting tool that can help in disentangling intrinsic genetic properties of these pandemic clones such as GC1. In a previous PCR-based study (15) they used $bla_{\text{OXA-51}}$ as one out of the three targeted genes to discriminate strains of *A. baumannii* GC1, GC2 and GC3. Since this gene has been found in other species (73–75) due to events of the Lateral Genetic Transfer, the analysis provided by ML in the present study supports a more accurate and solid implement to evaluate the presence of GC1.

In our work, we applied two ML algorithms that differ substantially in methodology. The SVM algorithm used the number of occurring unitigs in *A. baumannii* GC1 and *A. baumannii* non-GC1 genomes to learn and predict clonal membership of the strains, while the SCM model used a greedy approach to construct conjunction or disjunction rules to find the most concise set of k-mers that allows for accurate *A. baumannii* GC1 or *A. baumannii* non-GC1 prediction. Previously, methods that combined the use of SVM and SCM algorithms, and the representation of genomic data as k-mers were used to find genomic biomarkers to identify antibiotic resistance (25) or to predict antibiotic resistance from WGS data (48, 70).. We ran SCM algorithm through the Kover program with k-mer lengths between 31 and 127 nucleotides to be able to analyze all the possible rules obtained from these k-mers lengths. Also, we ran SVM algorithm using unitigs that usually corresponds to a longer sequence than the individual equivalent k-mers. Unitigs are defined as the longest sequence(s) that can be obtained when k-mers overlap by exactly k-1 nucleotides (54, 76, 77). In k-mer-based genome representations, the main

downside is that the representation contains a lot of redundancy, due to the fact the many k-mers are always present or absent simultaneously (e.g., gene deletion/ insertion). In this sense, it has been proposed to replace k-mers by unitigs (54, 76, 77).

The results of correlation between MLST genome typing and the model predictions of GC1 genomes from Dataset 1 indicated that both SVM and SCM models could significantly classify *A. baumannii* GC1 isolates ($p < 0.01$). Despite this, none of the rules obtained with the SCM model was able to uniquely identify the GC1 genomes since all the rules matched in *A. baumannii* GC1 and *A. baumannii* non-GC1 genomes from both Dataset 1 and Dataset 2. A key step for the successful implementation of ML algorithms is the preparation of the input datasets (78–81). In our study we faced two issues related to preparation of Datasets 1 and 2. On the one hand, a limitation in the program DBGWAS during the preparation of the matrix used in SVM algorithm meant that only a total of 500 genomes could be included in the training SVM dataset. Since our goal was to use the same training dataset for SVM and SCM models, the 500 genomes were also used as input in SCM algorithm. On the other hand, the scarcity of *A. baumannii* GC1 genomes available in GenBank compared to *A. baumannii* non-GC1 genomes caused Dataset 2 to have visibly more genomes representing *A. baumannii* non-GC1 class. Despite this fact, the results obtained from Dataset 1 and Dataset 2 remained consistent in 100% ($p > 0.05$) of the cases for SVM and more than 91.83% ($p > 0.05$) of the cases for SCM. Perhaps the limitation in the number of genomes in the training dataset is one of the reasons why the SCM model did not have enough sample to learn from the *A. baumannii* GC1 genomes, and therefore could not find a rule that uniquely identifies them. Conversely, the numbers of isolates predicted to be *A. baumannii* GC1 or non-GC1 by SVM model using unitig U1 were exactly the same

as the result obtained by MLST typification technique. This indicated that the SVM model obtained was excellent in classifying *A. baumannii* GC1 and non-GC1 genomes. It would be interesting to study in future works the impact produced in SVM and SCM model predictions the way of splitting Dataset 1 genomes (unitigs or k-mers), the number of genomes of each class (GC1 and non-GC1) present in Dataset 1, and the total number of *A. baumannii* genomes in Dataset 1. We could analyze SCM results using as input in the program Kover the binary matrix obtained from the unitigs representing Dataset 1 genomes generated by the program DBGWAS during the present work. In the same way, we could obtain k-mer profiles (k-mer sizes ranging from 31 to 127) from Dataset 1 genomes and k-mer matrixes associated to each profile using the DSK k-mer counter (82). DSK is used by Kover to internally compute k-mer profiles from the input genomes (83). Then we could use the k-mer matrixes as input in SVM algorithm. We could compare the new putative *A. baumannii* GC1 biomarkers obtained by SCM and SVM models and the performance of the new models with the results of the current work.

In conclusion, we obtained a SVM model that made genotypic predictions based on the presence or absence of short genomic sequences present in *A. baumannii* GC1 and non-GC1 genomes, that identified a biomarker that uniquely identifies CG1 strains. Given the length of the biomarker found, it is well-suited for translation to the clinical settings to be used in typified methods such as PCR.

MATERIALS AND METHODS

Data collection. To perform an accurate ML analysis in order to identify an *A. baumannii* GC1 biomarker discovery, we defined two datasets to do our studies.

Dataset 1 was composed by 200 *A. baumannii* GC1 and 300 *A. baumannii* non-GC1 genomes typified by multilocus sequence typing (MLST) used in a previous study (Álvarez et al., 2020). GC1 genomes included 18 genomes as follows: (i) A144 and A155 draft genomes from Argentina (84, 85); (ii) 14 GC1 complete genomes downloaded from GenBank which corresponded to strains AYE (AN: CU459141.1), D36 (AN: CP012952.1), A1 (AN: CP010781.1), AB307-0294 (AN: CP001172.2), AB5075-UW (AN: CP008706.1), AB0057 (AN: CP001182.1), USA15 (AN: NZ_CP020595.1), A85 (AN: NZ_CP021782.1), A388 (AN: NZ_CP024418.1), AR_0083 (AN: NZ_CP027528.1), DA33382 (AN: NZ_CP030106.1), 9102 (AN: NZ_CP023029.1), 11W359501 (AN: CP041035.1), NCTC13421 (AN: NZ_LS483472.1); and (iii) two highly quality genomes available for GC1, strains NIPH 527 and NIPH 290, isolated in 1984 and 1994, respectively (APQW00000000.1 and APRD00000000.1). We also included 182 GC1 genomes as scaffolds and contigs (Table S1). GC1 genomes belonged to sequence type (ST) 1 (n=152), ST 19 (n=8), ST 20 (n=8), ST 81 (n=14) and ST 623 (n=18). Non-GC1 genomes included five genomes belonging to other high-risk epidemic clones such as ACICU (AN: CP031380.1) as representative of GC2, Naval-13 (AN: AMDR01000001.1) as representative of GC3, AB33405 (NZ_JPXZ00000000.1) as representative of local epidemic clone CC113 and both ATCC 17978 (AN: CP018664.1), and A118 (AN: AEOW01000000) as sporadic clones. We also included 205 non-GC1 genomes as scaffolds and contigs. The non-GC1 included strains belonging to ST 2 (n=169), ST 3 (n=8), ST 6 (n=1), ST 10 (n=3), ST 16 (n=1), ST 23 (n=2), ST 25 (n=6), ST 32 (n=3), ST 33 (n=1), ST 34 (n=1), ST 38 (n=1), ST 39 (n=1), ST 49 (n=3), ST 52 (n=5), ST 54 (n=1), ST 69 (n=1), ST 78 (n=2), ST 79 (n=6), ST 85 (n=3), ST 113 (n=3), ST 136 (n=2), ST 155 (n=1), ST 156 (n=1), ST 158 (n=1), ST 164 (n=1), ST 187 (n=1), ST 203 (n=1), ST 215 (n=2), ST 218

(n=1), ST 221 (n=1), ST 241 (n=3), ST 250 (n=1), ST 255 (n=2), ST 267 (n=1), ST 388 (n=1), ST 403 (n=1), ST 404 (n=1), ST 405 (n=1), ST 406 (n=9), ST 411 (n=1), ST 412 (n=2), ST 413 (n=1), ST 414 (n=1), ST 415 (n=3), ST 416 (n=2), ST 417 (n=7), ST 422 (n=2), ST 425 (n=1), ST 427 (n=1), ST 428 (n=1), ST 429 (n=1), ST 431 (n=1), ST 437 (n=2), ST 438 (n=1), ST 490 (n=1), ST 499 (n=1), ST 500 (n=1), ST 510 (n=1), ST 513 (n=1), ST 515 (n=1), ST 516 (n=1), ST 524 (n=1), ST 575 (n=1), ST 639 (n=1), ST 665 (n=1), ST 738 (n=1), ST 821 (n=2), ST 938 (n=1), ST 1035 (n=1), ST 1140 (n=1), ST 1156 (n=1), ST 1157 (n=1) and ST 1168 (n=1). STs were numbered according to the Pasteur scheme for MLST (Table Genomes).

Dataset 2 was composed by 312 *A. baumannii* GC1 and 4487 *A. baumannii* non-GC1 genomes (Table_Genomes2). These genomes were retrieved from the GenBank assembly database filtering by *Acinetobacter baumannii* in the search by organism option (https://www.ncbi.nlm.nih.gov/assembly/organism/, last accessed in July 2021). The number of genomes in this dataset was limited by the *A. baumannii* GC1 and non-GC1 genomes available in the GenBank assembly database. The GC1 genomes included ST 1 (n=215), ST 19 (n=28), ST 20 (n=14), ST 81 (n=21), ST 94 (n=16), ST 315 (n=3), ST 328 (n=9), ST 460 (n=1), ST 717 (n=3) and ST 1106 (n=2). The non-GC1 genomes included ST 2 (n=3007), ST 3 (n=14), ST 6 (n=2), ST 9 (n=1), ST 10 (n=68), ST 11 (n=2), ST 12 (n=1), ST 15 (n=38), ST 16 (n=24), ST 23 (n=13), ST 25 (n=141), ST 32 (n=18), ST 33 (n=2), ST 34 (n=1), ST 35 (n=1), ST 36 (n=3), ST 37 (n=1), ST 39 (n=1), ST 40 (n=15), ST 45 (n=4), ST 46 (n=24), ST 47 (n=1), ST 48 (n=6), ST 49 (n=16), ST 52 (n=32), ST 54 (n=3), ST 56 (n=2), ST 57 (n=2), ST 63 (n=1), ST 71 (n=1), ST 77 (n=4), ST 78 (n=74), ST 79 (n=116), ST 80 (n=1), ST 85 (n=22), ST 98 (n=1), ST 103 (n=10), ST 105 (n=1), ST 106 (n=6), ST

107 (n=9), ST 108 (n=8), ST 109 (n=2), ST 111 (n=2), ST 113 (n=8), ST 119 (n=1), ST 126 (n=8), ST 129 (n=6), ST 132 (n=3), ST 136 (n=4), ST 138 (n=4), ST 142 (n=4), ST 145 (n=2), ST 149 (n=6), ST 150 (n=8), ST 151 (n=1), ST 152 (n=4), ST 154 (n=2), ST 155 (n=1), ST 156 (n=26), ST 158 (n=15), ST 162 (n=1), ST 163 (n=1), ST 164 (n=28), ST 172 (n=3), ST 187 (n=22), ST 193 (n=6), ST 195 (n=2), ST 203 (n=8), ST 213 (n=3), ST 215 (n=20), ST 216 (n=2), ST 218 (n=1), ST 221 (n=8), ST 229 (n=2), ST 237 (n=1), ST 238 (n=2), ST 239 (n=5), ST 240 (n=3), ST 241 (n=4), ST 250 (n=18), ST 255 (n=1), ST 265 (n=1), ST 267 (n=4), ST 271 (n=1), ST 285 (n=2), ST 286 (n=1), ST 294 (n=2), ST 307 (n=2), ST 309 (n=1), ST 317 (n=3), ST 318 (n=1), ST 329 (n=1), ST 336 (n=1), ST 338 (n=2), ST 341 (n=1), ST 342 (n=1), ST 345 (n=2), ST 348 (n=3), ST 370 (n=1), ST 372 (n=2), ST 374 (n=13), ST 375 (n=1), ST 388 (n=5), ST 400 (n=8), ST 402 (n=1), ST 403 (n=1), ST 406 (n=25), ST 411 (n=1), ST 412 (n=5), ST 415 (n=10), ST 416 (n=13), ST 417 (n=9), ST 419 (n=1), ST 422 (n=3), ST 428 (n=2), ST 429 (n=1), ST 437 (n=15), ST 438 (n=1), ST 464 (n=3), ST 492 (n=3), ST 494 (n=2), ST 497 (n=1), ST 498 (n=4), ST 499 (n=33), ST 500 (n=11), ST 501 (n=1), ST 503 (n=1), ST 505 (n=2), ST 506 (n=1), ST 508 (n=1), ST 509 (n=2), ST 512 (n=2), ST 513 (n=1), ST 514 (n=16), ST 515 (n=1), ST 521 (n=1), ST 522 (n=1), ST 523 (n=1), ST 525 (n=1), ST 526 (n=1), ST 527 (n=1), ST 528 (n=1), ST 529 (n=1), ST 532 (n=1), ST 533 (n=1), ST 535 (n=1), ST 536 (n=1), ST 537 (n=2), ST 563 (n=1), ST 570 (n=7), ST 571 (n=3), ST 575 (n=5), ST 578 (n=2), ST 600 (n=1), ST 604 (n=4), ST 621 (n=1), ST 622 (n=14), ST 624 (n=1), ST 625 (n=2), ST 626 (n=1), ST 632 (n=4), ST 636 (n=42), ST 638 (n=2), ST 641 (n=1), ST 650 (n=1), ST 651 (n=1), ST 652 (n=3), ST 653 (n=2), ST 654 (n=1), ST 656 (n=1), ST 657 (n=1), ST 658 (n=1), ST 659 (n=1), ST 660 (n=1), ST 661 (n=1), ST 662 (n=1), ST 664 (n=7), ST 665 (n=1), ST 668 (n=1), ST 669 (n=2), ST 670 (n=1), ST 671 (n=1), ST 673 (n=1), ST

674 (n=1), ST 690 (n=1), ST 721 (n=1), ST 730 (n=5), ST 742 (n=1), ST 745 (n=4), ST 763 (n=1), ST 764 (n=1), ST 810 (n=1), ST 821 (n=3), ST 823 (n=1), ST 825 (n=1), ST 848 (n=2), ST 866 (n=1), ST 875 (n=1), ST 876 (n=1), ST 878 (n=1), ST 879 (n=1), ST 880 (n=1), ST 911 (n=3), ST 920 (n=1), ST 922 (n=2), ST 923 (n=1), ST 927 (n=1), ST 939 (n=1), ST 941 (n=2), ST 943 (n=1), ST 963 (n=1), ST 972 (n=1), ST 973 (n=2), ST 976 (n=1), ST 979 (n=1), ST 991 (n=18), ST 1018 (n=1), ST 1022 (n=2), ST 1048 (n=2), ST 1057 (n=1), ST 1077 (n=1), ST 1079 (n=1), ST 1083 (n=2), ST 1088 (n=16), ST 1093 (n=2), ST 1095 (n=1), ST 1108 (n=1), ST 1118 (n=1), ST 1126 (n=1), ST 1130 (n=1), ST 1137 (n=1), ST 1142 (n=1), ST 1143 (n=1), ST 1153 (n=1), ST 1158 (n=1), ST 1165 (n=1), ST 1179 (n=1), ST 1184 (n=1), ST 1190 (n=1), ST 1191 (n=1), ST 1195 (n=1), ST 1197 (n=1), ST 1230 (n=3), ST 1232 (n=1), ST 1233 (n=1), ST 1252 (n=1), ST 1258 (n=1), ST 1278 (n=1), ST 1302 (n=1), ST 1303 (n=1), ST 1337 (n=1), ST 1356 (n=1), ST 1426 (n=1), ST 1430 (n=3), ST 1431 (n=1), ST 1432 (n=3), ST 1433 (n=1), ST 1436 (n=1), ST 1437 (n=1), ST 1438 (n=1), ST 1447 (n=1), ST 1448 (n=2), ST 1472 (n=1), ST 1479 (n=1), ST 1487 (n=1), ST 1512 (n=2), ST 1535 (n=1), ST 1542 (n=1), ST 1543 (n=1), ST 1544 (n=1), ST 1545 (n=1), ST 1546 (n=1), ST 1547 (n=1), ST 1548 (n=1), ST 1549 (n=1), ST 1550 (n=3), ST 1554 (n=2), ST 1555 (n=4), ST 1573 (n=1), ST 1580 (n=9), ST 1596 (n=1) and new ST not assigned yet in the Pasteur MLST scheme (n=5). We used Dataset 2 to validate the results obtained by SVM model. (Table S2).

Machine learning analyses.

The SVM classifier is based on maximization of the margin around the hyperplane ($w^T x + b$) separating samples or instances of the different classes (51, 86). Each instance i =

(1)

1,...,m consists of an N-dimensional feature vector x_i and a class label $y_i \in \{+1, -1\}$. The maximization of the margin corresponds to the following minimization:

$$w^*, b^*, \xi^* = arg \min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} \xi_i$$

$$s.t.\ \mathrm{y}_i(w^T x_i + b) \geq 1 - \xi_i; \quad \xi_i \geq 0;\ i = 1, \ldots, m$$

In this soft-margin SVM equation, ξ_i is a penalty for misclassification or classification within the margin. Parameter C sets the weight of this penalty. The resulting weight vector $w*$ encodes the contributions of all features to the classifier (51).

We created a Python script using scikit-learn (87) to run the SVM algorithm. The script performed the hyperparameter tuning using a 5-fold cross-validated grid search (GridSearchCV method implemented in sklearn.model_selection) to find the best hyperparameters. Dataset 1 was split into 5 randomly folds. Each fold was then used once as a testing set while the 4 remaining folds formed the training set. The training and testing process was repeated five times (code available on GitHub at https://github.com/vealvarez/SVM_GC1).

The SCM (36), is a learning algorithm that produces models that are conjunctions (logical-AND) or disjunctions (logical-OR) of boolean-valued rules $r : \mathbb{R}^d \rightarrow \{0,1\}$. Let us use $h(x)$ to denote the output of model h on genome x. When h consists of a conjunction (i.e., a logical-AND) of a set R^* of rules r^*, we have

$$h(x) = \bigwedge_{r^* \in R^*} r^*(x)$$

Whereas, for a disjunction (i.e., a logical-OR) of rules, we have

$$h(x) = \bigvee_{r^* \in R^*} r^*(x)$$

Given a set R of candidate rules, the SCM attempts to find a model that minimizes the empirical error rate $R_S \stackrel{\text{def}}{=} \frac{1}{m}\sum_{i=1}^{m} I\left[h(x_{i)} \neq y_i\right]$, where I[True] = 1 and 0 otherwise, while using the smallest number of rule in R (88).

We used the program Kover which implements SCM (25, 83). We ran the program using Dataset 1 (see Input dataset preparation below) for k-mer sizes from 31 to 127 (taking only the odd numbers between them). The smallest value of k was set to 31 since extensive testing has shown that this size is optimal for bacterial genome assembly and has been employed for studies based on reference-free bacterial genome comparisons (76, 89). The greatest k-mer size was set to 127 since it is the maximum value accepted as a parameter in Kover. We chose only odd values of k to avoid the formation of palindromes (90). For each k-mer size, we split Dataset 1 into a train dataset (2/3 of the genomes) and a testing dataset (1/3 of the genomes). Then, we trained the 49 models corresponding to each k-mer using a conjunction / disjunction model type. The best conjunctive and/or disjunctive model for each k-mer was selected using 5-fold cross-validation to determine the optimal rule scoring function with default parameters.

Input dataset preparation for ML models. We split the 500 genomes included in Dataset 1 into unitigs by using the program DBGWAS (54). Unitigs are stretches of DNA shared by the strains in a dataset. The DBGWAS method proposes connecting the overlaps of k-mers in a compressed de Bruijn graph (DBG) so that k-mers are extended using adjacent sequence information in the population, forming unitigs present in exactly the same set of

samples as their constituent k-mers. During the first step of DBGWAS process, the program built a variant matrix, where each variant is a pattern of presence/absence of unitigs in each genome present in Dataset 1 (54). We wrote a Python script to format the variant matrix and create a presence/absence (coded with the values 1/0 respectively) binary matrix with unitigs as columns (features) and the accession numbers of the genomes as rows (instances). We used the binary matrix as input of the SVM algorithm. In this matrix, we discarded the data about low frequency unitigs (unitigs found in less than 50 GC1 and non-GC1 genomes). We also integrated the MLST data corresponding to Dataset 1 genomes and created a two-column matrix used as input for SVM. The first column of the matrix contained the accession number of the genomes, and the second column contained a binary variable that indicated whether each genome was typed or not as GC1 (-1 = non-GC1 genome / 1 = GC1 genome) according to MLST.

For the SCM approach, we first packaged Dataset 1 sequences stored in FASTA files into a Kover dataset using the create from-contigs command. This command also received a tab-separated value (TSV) with Dataset 1 genome classification according MLST typing (GC1 or non-GC1). The first column of the TSV file described the genome accession number and the second column had the value 1 whether the genome belonged to GC1 or the value 0 otherwise. We ran Kover for k-mer sizes from 31 to 127. For each k-mer size, Kover constructed a reference-free input matrix based on k-mer profiles generated with the DSK k-mer counting software. A k-mer presence / absence binary matrix based on Dataset 1 genomes was then created and used as input for the SCM models of each k-mer size.

Unitig selection using SVM for putative biomarkers analysis. After obtaining the linear SVM model with the best hyperparameters, the values of the features weight vector (referred as the hyperplane normal vector $w*$ in the formula (1)) were accessed through the attribute sklearn.model_selection.GridSearchCV.best_estimator.coef_. The values were sorted from highest to lowest and used to decide the relevance of each unitig sequence (associated with each weight value) during the model prediction (91, 92). It is worth mentioning that in our study, sequence unitigs were used as features in the models. The positive sign of a feature weight value indicates that the feature contributes to GC1 class prediction (represented by the value 1) and negative sign indicates that the feature contributes to non-GC1 class (represented by the value -1) prediction (93). Considering the above mentioned, 100 unitig sequences with the highest weight values were selected to be analyzed as putative biomarkers of *A. baumannii* GC1 genomes.

Machine learning performance metrics. The performance of the SVM and SCM models was evaluated in terms of sensitivity, specificity, accuracy, precision and F1-score. They were defined as: sensitivity = TP/(TP + FN), specificity = TN/(TN + FP), accuracy = (TP + TN)/(TP + FP + TN + FN), precision = TP/ (TP + FP) and F1 score = 2 * (precision * sensitivity)/(precision + sensitivity). Where TP (true positives) was the number of GC1 strains predicted to be GC1, TN (true negatives) was the number of non-GC1 strains predicted to be non-GC1, FP (false positives) was the number of non-GC1 strains predicted to be GC1, and FN (false negatives) was the number of GC1 strains predicted to be non-GC1.

MLST genome typing. MLST data was obtained in silico using the program mlst (https://github.com/tseemann/mlst#mlst) which scans genome files against traditional PubMLST database for *A. baumannii* species using the Pasteur MLST scheme.

Blastn searches. Blastn searches (94) were done using Dataset 1 and Dataset 2 as subject and unitigs/k-mers that contributed most to *A. baumannii* GC1 genome prediction according to SVM/SCM models were used as query. We identified whether unitigs/k-mers matched known genes or intergenic regions and provide their putative function when possible. Also, we counted the number of GC1 and non-GC1 genomes matched by each unitig. We considered a cut-off E-value of E^{-10}, 100% of identity and 100% of query cover.

SUPPLEMENTARY MATERIAL

Supplemental material is available online only

ACKNOWLEGMENTS

The authors' work was supported by the grant PUE 2522 from the CONICET given to IMPAM.

DC, VA and PQ contributed to the conception and design of the study. VA performed the scripts development. VA and PQ analyzed the results. DC and VA structured the work, wrote, and coordinated the drafts of the manuscript and did the final edition. VA wrote the result's section. All authors contributed to the manuscript revision, read, and approved the submitted version.

REFERENCES

1. Álvarez VE, Quiroga MP, Galán AV, Vilacoba E, Quiroga C, Ramírez MS, Centrón D. 2020. Crucial Role of the Accessory Genome in the Evolutionary Trajectory of Acinetobacter baumannii Global Clone 1. Front Microbiol 11.

2. Hamidian M, Nigro SJ. 2019. Emergence, molecular mechanisms and global spread of carbapenem-resistant acinetobacter baumannii. Microb Genomics. Microbiology Society.

3. Douraghi M, Kenyon JJ, Aris P, Asadian M, Ghourchian S, Hamidian M. 2020. Accumulation of Antibiotic Resistance Genes in Carbapenem-Resistant Acinetobacter baumannii Isolates Belonging to Lineage 2, Global Clone 1, from Outbreaks in 2012–2013 at a Tehran Burns Hospital . mSphere 5.

4. Yakkala H, Samantarrai D, Gribskov M, Siddavattam D. 2019. Comparative genome analysis reveals niche-specific genome expansion in Acinetobacter baumannii strains. PLoS One 14:e0218204.

5. Holt K, Kenyon JJ, Hamidian M, Schultz MB, Pickard DJ, Dougan G, Hall R. 2016. Five decades of genome evolution in the globally distributed, extensively antibiotic-resistant Acinetobacter baumannii global clone 1. Microb Genomics 2:e000052.

6. Zarrilli R, Pournaras S, Giannouli M, Tsakris A. 2013. Global evolution of multidrug-resistant Acinetobacter baumannii clonal lineages. Int J Antimicrob Agents 41:11–19.

7. Rodríguez C, Nastro M, A. Flores S, Rodriguez M, Spinozzi M, Bruni G, L. López A, David V, S. Aiassa M, A. Marqués I, R. Navarro O, Paniccia L, Famiglietti A, Grupo colaborativo Acinetobacter Argentina. 2018. Epidemiología molecular de

aislados de Acinetobacter baumannii resistentes a carbapenems en Argentina. Rev Argent Microbiol https://doi.org/10.1016/j.ram.2017.12.004.

8. Dijkshoorn L, Nemec A, Seifert H. 2007. An increasing threat in hospitals: Multidrug-resistant Acinetobacter baumannii. Nat Rev Microbiol 5:939–951.

9. Bouvet PJM, Grimont PAD. 1986. Taxonomy of the genus Acinetobacter with the recognition of Acinetobacter baumannii sp. nov., Acinetobacter haemolyticus sp. nov., Acinetobacter johnsonii sp. nov., and Acinetobacter junii sp. nov. and emended descriptions of Acinetobacter calcoaceticus and Acinetobacter lwoffii. Int J Syst Bacteriol 36:228–240.

10. Gerner-Smidt P, Tjernberg I, Ursing J. 1991. Reliability of phenotypic tests for identification of Acinetobacter species. J Clin Microbiol 29:277–282.

11. Bouvet PJM, Grimont PAD. 1987. Identification and biotyping of clinical isolates of Acinetobacter. Ann Inst Pasteur Microbiol 138:569–578.

12. Bartual SG, Seifert H, Hippler C, Luzon MAD, Wisplinghoff H, Rodríguez-Valera F. 2005. Development of a multilocus sequence typing scheme for characterization of clinical isolates of Acinetobacter baumannii. J Clin Microbiol 43:4382–4390.

13. Gozalan A, Unaldı O, Guldemir D, Aydogan S, Kuzucu C, Cakirlar FK, Açıkgoz ZC, Durmaz R. 2020. Molecular Characterization of Carbapenem-Resistant *Acinetobacter baumannii* blood culture isolates from three hospitals in Turkey. Jpn J Infect Dis https://doi.org/10.7883/yoken.jjid.2020.478.

14. Li L-H, Yang Y-S, Sun J-R, Huang T-W, Huang W-C, Chen F-J, Wang Y-C, Kuo T-H, Kuo S-C, Chen T-L, Lee Y-T, Chang Y-Y, Yang Y-S, Liu Y-M, Kuo S-C, Liu C-

P, Chen T-L, Lee Y-T, Chang Y-Y, Yang Y-S, Liu Y-M, Kuo S-C, Liu C-P, Chen T-L, Lee Y-T. 2020. Clinical and molecular characterization of Acinetobacter seifertii in Taiwan. J Antimicrob Chemother https://doi.org/10.1093/jac/dkaa432.

15. Turton JF, Gabriel SN, Valderrey C, Kaufmann ME, Pitt TL. 2007. Use of sequence-based typing and multiplex PCR to identify clonal lineages of outbreak strains of Acinetobacter baumannii. Clin Microbiol Infect 13:807–815.

16. Evans BA, Hamouda A, Towner KJ, Amyes SGB. 2008. OXA-51-like β-lactamases and their association with particular epidemic lineages of Acinetobacter baumannii. Clin Microbiol Infect 14:268–275.

17. Hamouda A, Evans BA, Towner KJ, Amyes SGB. 2010. Characterization of epidemiologically unrelated Acinetobacter baumannii isolates from four continents by use of multilocus sequence typing, pulsed-field gel electrophoresis, and sequence-based typing of blaOXA-51-like genes. J Clin Microbiol 48:2476–2483.

18. Hamidian M, Wynn M, Holt KE, Pickard D, Dougan G, Hall RM. 2014. Identification of a marker for two lineages within the GC1 clone of Acinetobacter baumannii. J Antimicrob Chemother 69:557–558.

19. Michalowitz A, Yang J, Castaneda P, Litrenta J. 2020. Existing and emerging methods of diagnosis and monitoring of pediatric musculoskeletal infection. Injury. Elsevier Ltd.

20. Holt KE, Wertheim H, Zadoks RN, Baker S, Whitehouse CA, Dance D, Jenney A, Connor TR, Hsu LY, Severin J, Brisse S, Cao H, Wilksch J, Gorrie C, Schultz MB, Edwards DJ, Nguyen K Van, Nguyen TV, Dao TT, Mensink M, Minh V Le, Nhu

NTK, Schultsz C, Kuntaman K, Newton PN, Moore CE, Strugnell RA, Thomson NR. 2015. Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in Klebsiella pneumoniae , an urgent threat to public health . Proc Natl Acad Sci 112:E3574–E3581.

21. Ezewudo M, Borens A, Chiner-Oms Á, Miotto P, Chindelevitch L, Starks AM, Hanna D, Liwski R, Zignol M, Gilpin C, Niemann S, Kohl TA, Warren RM, Crook D, Gagneux S, Hoffner S, Rodrigues C, Comas I, Engelthaler DM, Alland D, Rigouts L, Lange C, Dheda K, Hasan R, McNerney R, Cirillo DM, Schito M, Rodwell TC, Posey J. 2018. Integrating standardized whole genome sequence analysis with a global Mycobacterium tuberculosis antibiotic resistance knowledgebase. Sci Rep 8.

22. Cario CL, Chen E, Leong L, Emami NC, Lopez K, Tenggara I, Simko JP, Friedlander TW, Li PS, Paris PL, Carroll PR, Witte JS. 2020. A machine learning approach to optimizing cell-free DNA sequencing panels: With an application to prostate cancer. BMC Cancer 20.

23. Ransom EM, Potter RF, Dantas G, Burnham CAD. 2020. Genomic Prediction of Antimicrobial Resistance: Ready or Not, Here It Comes! Clin Chem 66:1278–1289.

24. Hernández M, Quijada NM, Rodríguez-Lázaro D, Eiros JM. 2020. Bioinformatics of next generation sequencing in clinical microbiology diagnosis. Rev Argent Microbiol 52:150–161.

25. Drouin A, Giguère S, Déraspe M, Marchand M, Tyers M, Loo VG, Bourgault A-M, Laviolette F, Corbeil J. 2016. Predictive computational phenotyping and biomarker

discovery using reference-free genome comparisons. BMC Genomics 17:754.

26. Lees JA, T. M, Tien B, Corander J. 2020. Improved Prediction of Bacterial Genotype-Phenotype Associations Using Interpretable Pangenome-Spanning Regressions 11:1–22.

27. Khaledi A, Weimann A, Schniederjans M, Asgari E, Kuo T, Oliver A, Cabot G, Kola A, Gastmeier P, Hogardt M, Jonas D, Mofrad MR, Bremges A, McHardy AC, Häussler S. 2020. Predicting antimicrobial resistance in Pseudomonas aeruginosa with machine learning-enabled molecular diagnostics . EMBO Mol Med 12.

28. Hyun JC, Kavvas ES, Monk JM, Palsson BO. 2020. Machine learning with random subspace ensembles identifies antimicrobial resistance determinants from pan-genomes of three pathogens. PLoS Comput Biol 16.

29. Van Camp PJ, Haslam DB, Porollo A. 2020. Prediction of Antimicrobial Resistance in Gram-Negative Bacteria From Whole-Genome Sequencing Data. Front Microbiol 11:1–13.

30. Aytan-Aktug D, Clausen PTLC, Bortolaia V, Aarestrup FM, Lund O. 2020. Prediction of Acquired Antimicrobial Resistance for Multiple Bacterial Species Using Neural Networks. mSystems 5.

31. Lakin SM, Kuhnle A, Alipanahi B, Noyes NR, Dean C, Muggli M, Raymond R, Abdo Z, Prosperi M, Belk KE, Morley PS, Boucher C. 2019. Hierarchical Hidden Markov models enable accurate and diverse detection of antimicrobial resistance sequences. Commun Biol 2.

32. Saravanan R, Sujatha P. 2019. A State of Art Techniques on Machine Learning

Algorithms: A Perspective of Supervised Learning Approaches in Data Classification, p. 945–949. *In* Proceedings of the 2nd International Conference on Intelligent Computing and Control Systems, ICICCS 2018. Institute of Electrical and Electronics Engineers Inc.

33. González L, Angulo C, Velasco F, Català A. 2005. Unified dual for bi-class SVM approaches. Pattern Recognit 38:1772–1774.

34. Scholkopf B, Smola A. 2001. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. Cambridge, MA MIT Press https://doi.org/10.1109/tnn.2005.848998.

35. Boser BE, Guyon IM, Vapnik VN. 1992. Training algorithm for optimal margin classifiers, p. 144–152. *In* Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory. Publ by ACM, New York, New York, USA.

36. Marchand M, Shawe-Taylor J. 2000. The Set Covering Machine. CrossRef List Deleted DOIs 1:723–746.

37. Ivanenkov YA, Zhavoronkov A, Yamidanov RS, Osterman IA, Sergiev P V., Aladinskiy VA, Aladinskaya A V., Terentiev VA, Veselov MS, Ayginin AA, Kartsev VG, Skvortsov DA, Chemeris A V., Baimiev AK, Sofronova AA, Malyshev AS, Filkov GI, Bezrukov DS, Zagribelnyy BA, Putin EO, Puchinina MM, Dontsova OA. 2019. Identification of Novel Antibacterials Using Machine Learning Techniques. Front Pharmacol 10:1–15.

38. Kouchaki S, Yang YY, Walker TM, Walker AS, Wilson DJ, Peto TEA, Crook DW, Clifton DA, Hoosdally SJ, Gibertoni Cruz AL, Carter J, Grazian C, Kouchaki S,

Walker TM, Fowler PW, Clifton DA, Iqbal Z, Hunt M, Smith EG, Rathod P, Jarrett L, Matias D, Cirillo DM, Borroni E, Battaglia S, Ghodousi A, Spitaleri A, Cabibbe A, Tahseen S, Nilgiriwala K, Shah S, Rodrigues C, Kambli P, Surve U, Khot R, Niemann S, Kohl T, Merker M, Hoffmann H, Molodtsov N, Plesnik S, Ismail N, Omar SV, Joseph L, Marubini E, Thwaites G, Thuong TNT, Ngoc NH, Srinivasan V, Moore D, Coronel J, Solano W, Gao GF, He G, Zhao Y, Ma A, Liu C, Zhu B, Laurenson I, Claxton P, Wilkinson RJ, Lalvani A, Posey J, Gardy J, Werngren J, Paton N, Jou R, Wu MH, Lin WH, Ferrazoli L, De Oliveira RS, Arandjelovic I, Comas I, Drobniewski F, Gao Q, Sintchenko V, Supply P, Van Soolingen D. 2019. Application of machine learning techniques to tuberculosis drug resistance analysis. Bioinformatics 35:2276–2282.

39. Her HL, Wu YW. 2018. A pan-genome-based machine learning approach for predicting antimicrobial resistance activities of the Escherichia coli strains. Bioinformatics 34:i89–i95.

40. Yang Y, Niehaus KE, Walker TM, Iqbal Z, Walker AS, Wilson DJ, Peto TEA, Crook DW, Smith EG, Zhu T, Clifton DA. 2018. Machine learning for classifying tuberculosis drug-resistance from DNA sequencing data. Bioinformatics 34:1666–1671.

41. Zvezdanova ME, Arroyo MJ, Méndez G, Guinea J, Mancera L, Muñoz P, Rodríguez-Sánchez B, Escribano P. 2020. Implementation of maldi-tof mass spectrometry and peak analysis: Application to the discrimination of cryptococcus neoformans species complex and their interspecies hybrids. J Fungi 6:1–12.

42. Lorenz B, Ali N, Bocklitz T, Rösch P, Popp J. 2020. Discrimination between pathogenic and non-pathogenic E. coli strains by means of Raman microspectroscopy. Anal Bioanal Chem 412:8241–8247.

43. Uysal Ciloglu F, Saridag AM, Kilic IH, Tokmakci M, Kahraman M, Aydin O. 2020. Identification of methicillin-resistant: Staphylococcus aureus bacteria using surface-enhanced Raman spectroscopy and machine learning techniques. Analyst 145:7559–7570.

44. Guyon I, Weston J, Barnhill S, Labs T, Bank R. 2013. Gene Selection for Cancer Classification using Support Vector Machines. Biotechnol Bioeng 110:fmvi-fmvi.

45. Henneges C, Bullinger D, Fux R, Friese N, Seeger H, Neubauer H, Laufer S, Gleiter CH, Schwab M, Zell A, Kammerer B. 2009. Prediction of breast cancer by profiling of urinary RNA metabolites using Support Vector Machine-based feature selection. BMC Cancer 9:1–11.

46. Aswathy MA, Jagannath M. 2021. An SVM approach towards breast cancer classification from H&E-stained histopathology images based on integrated features. Med Biol Eng Comput https://doi.org/10.1007/s11517-021-02403-0.

47. Xu H, Park S, Hwang TH. 2020. Computerized Classification of Prostate Cancer Gleason Scores from Whole Slide Images. IEEE/ACM Trans Comput Biol Bioinforma 17:1871–1882.

48. Liu Z, Deng D, Lu H, Sun J, Lv L, Li S, Peng G, Ma X, Li J, Li Z, Rong T, Wang G. 2020. Evaluation of Machine Learning Models for Predicting Antimicrobial Resistance of Actinobacillus pleuropneumoniae From Whole Genome Sequences.

Front Microbiol 11.

49. Li Y, Kong Y, Zhang M, Yan A, Liu Z. 2016. Using Support Vector Machine (SVM) for Classification of Selectivity of H1N1 Neuraminidase Inhibitors. Mol Inform 35:116–124.

50. Hicks AL, Wheeler N, Sánchez-Busó L, Rakeman JL, Harris SR, Grad YH. 2019. Evaluation of parameters affecting performance and reliability of machine learning-based antibiotic susceptibility testing from whole genome sequencing data. PLoS Comput Biol 15:1–21.

51. Bron EE, Smits M, Niessen WJ, Klein S. 2015. Feature Selection Based on the SVM Weight Vector for Classification of Dementia. IEEE J Biomed Heal Informatics 19:1617–1626.

52. Diancourt L, Passet V, Nemec A, Dijkshoorn L, Brisse S. 2010. The population structure of Acinetobacter baumannii: expanding multiresistant clones from an ancestral susceptible genetic pool. PLoS One 5:e10034.

53. Karah N, Wai SN, Uhlin BE. 2021. CRISPR-based subtyping to track the evolutionary history of a global clone of Acinetobacter baumannii. Infect Genet Evol 90:104774.

54. Jaillard M, Lima L, Tournoud M, Mahé P, van Belkum A, Lacroix V, Jacob L. 2018. A fast and agnostic method for bacterial genome-wide association studies: Bridging the gap between k-mers and genetic events. PLoS Genet 14:1–28.

55. Wuebbens MM, Liu MTW, Rajagopalan K V., Schindelin H. 2000. Insight into molybdenum cofactor deficiency provided by the crystal structure of the

molybdenum cofactor biosynthesis protein MoaC. Structure 8:709–718.

56. Kozmin SG, Schaaper RM. 2013. Genetic characterization of moaB mutants of Escherichia coli. Res Microbiol 164:689–694.

57. Winter SE, Thiennimitr P, Winter MG, Butler BP, Huseby DL, Crawford RW, Russell JM, Bevins CL, Adams LG, Tsolis RM, Roth JR, Bäumler AJ. 2010. Gut inflammation provides a respiratory electron acceptor for Salmonella. Nature 467:426–429.

58. Winter SE, Winter MG, Xavier MN, Thiennimitr P, Poon V, Keestra AM, Laughlin RC, Gomez G, Wu J, Lawhon SD, Popova IE, Parikh SJ, Adams LG, Tsolis RM, Stewart VJ, Bäumler AJ. 2013. Host-derived nitrate boosts growth of E. coli in the inflamed gut. Science (80-) 339:708–711.

59. Denkel LA, Rhen M, Bange FC. 2013. Biotin sulfoxide reductase contributes to oxidative stress tolerance and virulence in Salmonella enterica serovar Typhimurium. Microbiol (United Kingdom) 159:1447–1458.

60. Williams M, Mizrahi V, Kana BD. 2014. Molybdenum cofactor: A key component of Mycobacterium tuberculosis pathogenesis? Crit Rev Microbiol. Taylor & Francis.

61. Navon-Venezia S, Kondratyeva K, Carattoli A. 2017. Klebsiella pneumoniae: A major worldwide source and shuttle for antibiotic resistance. FEMS Microbiol Rev. Oxford University Press.

62. Mathers A, Peirano G, Mathers AJ, Peirano G, Pitout DD. 2015. The Role of Epidemic Resistance Plasmids and International High-Risk Clones in the Spread of Multidrug-Resistant ... The Role of Epidemic Resistance Plasmids and International

High- Risk Clones in the Spread of Multidrug-Resistant Enterobacteriaceae https://doi.org/10.1128/CMR.00116-14.

63. Abdouchakour F, Aujoulat F, Licznar-Fajardo P, Marchandin H, Toubiana M, Parer S, Lotthé A, Jumas-Bilak E. 2018. Intraclonal variations of resistance and phenotype in Pseudomonas aeruginosa epidemic high-risk clone ST308: A key to success within a hospital? Int J Med Microbiol 308:279–289.

64. Villa L, Feudi C, Fortini D, Brisse S, Passet V, Bonura C, Endimiani A, Mammina C, Ocampo AM, Jimenez JN, Doumith M, Woodford N, Hopkins K, Carattoli A. 2017. Diversity, virulence, and antimicrobial resistance of the KPCproducing klebsiella pneumoniae ST307 clone. Microb Genomics 3.

65. Oteo J, Pérez-Vázquez M, Bautista V, Ortega A, Zamarrón P, Saez D, Lara N, Ramiro R, Aracil B, Campos J. 2016. The spread of KPC-producing Enterobacteriaceae in Spain: WGS analysis of the emerging high-risk clones of Klebsiella pneumoniae ST11/KPC-2, ST101/KPC-2 and ST512/KPC-3. J Antimicrob Chemother 71:3392–3399.

66. Cao H, Xia T, Li Y, Xu Z, Bougouffa S, Lo YK, Bajic VB, Luo H, Woo PCY, Yana A. 2019. Uncoupled Quorum Sensing Modulates the Interplay of Virulence and Resistance in a Multidrug-Resistant Clinical Pseudomonas aeruginosa Isolate Belonging to the MLST550 Clonal Complex. Antimicrob Agents Chemother 63.

67. Heiden SE, Hübner NO, Bohnert JA, Heidecke CD, Kramer A, Balau V, Gierer W, Schaefer S, Eckmanns T, Gatermann S, Eger E, Guenther S, Becker K, Schaufler K. 2020. A Klebsiella pneumoniae ST307 outbreak clone from Germany demonstrates

features of extensive drug resistance, hypermucoviscosity, and enhanced iron acquisition. Genome Med 12.

68. Gato E, Vázquez-Ucha JC, Rumbo-Feal S, Álvarez-Fraga L, Vallejo JA, Martínez-Guitián M, Beceiro A, Vivas JR, Sola Campoy PJ, Pérez-Vázquez M, Iglesias JO, Rodiño-Janeiro BK, Romero A, Poza M, Bou G, Pérez A. 2020. Kpi, a chaperone-usher pili system associated with the worldwide-disseminated high-risk clone Klebsiella pneumoniae ST-15. Proc Natl Acad Sci U S A 117:17249–17259.

69. de Lagarde M, Vanier G, Arsenault J, Fairbrother JM. 2021. High risk clone: A proposal of criteria adapted to the one health context with application to enterotoxigenic escherichia coli in the pig population. Antibiotics 10:1–19.

70. Macesic N, Bear Don’t Walk OJ, Pe’er I, Tatonetti NP, Peleg AY, Uhlemann A-C. 2020. Predicting Phenotypic Polymyxin Resistance in Klebsiella pneumoniae through Machine Learning Analysis of Genomic Data. mSystems 5.

71. Deelder W, Christakoudi S, Phelan J, Benavente ED, Campino S, McNerney R, Palla L, Clark TG. 2019. Machine learning predicts accurately mycobacterium tuberculosis drug resistance from whole genome sequencing data. Front Genet 10.

72. Mahé P, Tournoud M. 2018. Predicting bacterial resistance from whole-genome sequences using k-mers and stability selection. BMC Bioinformatics 19:383.

73. Leski TA, Bangura U, Jimmy DH, Ansumana R, Lizewski SE, Li RW, Stenger DA, Taitt CR, Vora GJ. 2013. Identification of blaOXA-51-like, blaOXA-58, bla DIM-1, and blaVIM carbapenemase genes in hospital Enterobacteriaceae isolates from Sierra Leone. J Clin Microbiol 51:2435–2438.

74. Lee YT, Kuo SC, Chiang MC, Yang SP, Chen CP, Chen TL, Fung CP. 2012. Emergence of carbapenem-resistant non-baumannii species of Acinetobacter harboring a blaOXA-51-like gene that is intrinsic to A. baumannii. Antimicrob Agents Chemother 56:1124–1127.

75. Lee YT, Turton JF, Chen TL, Wu RCC, Chang WC, Fung CP, Chen CP, Cho WL, Huang LY, Siu LK. 2009. First identification of blaOXA-51-like in non-baumannii acinetobacter spp. J Chemother 21:514–520.

76. Mahé P, Tournoud M. 2018. Predicting bacterial resistance from whole-genome sequences using k-mers and stability selection. BMC Bioinformatics 19:383.

77. Jaillard M, Tournoud M, Lima L, Lacroix V, Veyrieras JB, Jacob L. 2017. Representing Genetic Determinants in Bacterial GWAS with Compacted De Bruijn Graphs. bioRxiv 1–24.

78. Ajiboye AR, Abdullah-Arshah R, Qin H, Isah-Kebbe H. 2015. Evaluating the Effect of Dataset Size on Predictive Model Using. Int J Softw Eng Comput Sci 1:75–84.

79. Kolisetty VV, Rajput DS. 2020. A review on the significance of machine learning for data analysis in big data. Jordanian J Comput Inf Technol 6:41–57.

80. Zhang Y, Ling C. 2018. A strategy to apply machine learning to small datasets in materials science. npj Comput Mater 4:28–33.

81. Schmidt J, Marques MRG, Botti S, Marques MAL. 2019. Recent advances and applications of machine learning in solid-state materials science. npj Comput Mater 5.

82. Rizk G, Lavenier D, Chikhi R. 2013. DSK: k-mer counting with very low memory usage. Bioinformatics 29:652–653.

83. Drouin A, Letarte G, Raymond F, Marchand M, Corbeil J, Laviolette F. 2019. Interpretable genotype-to-phenotype classifiers with performance guarantees. Sci Rep 9:1–13.

84. Vilacoba E, Déraspe M, Traglia GM, Roy PH, Ramírez S. 2014. Draft Genome Sequence of an International Clonal Lineage 1 Acinetobacter baumannii Strain from Argentina 2:13–14.

85. Arivett BA, Fiester SE, Ream DC, Centrón D, Ramírez MS, Tolmasky ME, Actis LA. 2015. Draft Genome of the Multidrug-Resistant Acinetobacter baumannii Strain A155 Clinical Isolate. Genome Announc 3:e00212-15.

86. Vapnik VN. 2000. The Nature of Statistical Learning TheoryThe Nature of Statistical Learning Theory. Springer New York.

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

88. Drouin A, Giguère S, Déraspe M, Marchand M, Tyers M, Loo VG, Bourgault A-M, Laviolette F, Corbeil J. 2016. Aditional File of Predictive computational phenotyping and biomarker discovery using reference-free genome comparisons. BMC Genomics 17:754.

89. Boisvert S, Raymond F, Godzaridis É, Laviolette F, Corbeil J. 2012. Ray Meta:

Scalable de novo metagenome assembly and profiling. Genome Biol 13.

90. Miller JR, Koren S, Sutton G. 2010. Assembly algorithms for next-generation sequencing data. Genomics. NIH Public Access.

91. Guyon I, Weston J, Barnhill S, Vapnik V. 2002. Gene selection for cancer classification using support vector machines. Mach Learn 46:389–422.

92. Chang Y, Lin C. 2008. Feature Ranking Using Linear SVM. Featur Rank Using Linear SVM 53–64.

93. Zisserman A. 2013. Lecture 2 : The SVM classifier. C19 Mach Learn Lect Hilar 2015, Oxford Univ 40.

94. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215:403–410.

TABLES

Table 1. Putative GC1 biomarkers obtained by SVM. The table details the unitig ID, unitig sequence, unitig length, the location where the unitig sequence matched in the AYE genome (AN:CU459141.1) or the AB0057 genome (AN: CP001182.2) and the gene name corresponding to the genome region matched and the gene product of the first 10 unitigs that contributed most to the GC1 genome prediction according to the values of the features weight vector. AYE strain genome was used as GC1 reference to locate the unitigs sequences. However, when the unitig was not found in the AYE genome, AB0057 strain genome was used instead. The prefix "REGION" was used when the match occurred either in an intergenic region or in a combination of intergenic region and a gene. The complete list of the 100 unitigs predicted by SVM is detailed in Supplementary Table S3.

Unitig ID	Sequence	Length	Match Location	Gene Name	Gene Product

U1	TTATTCATAGCCTCCTGGATGCG GTAAATCACGGCAAACATCAAA AATATGGACTAATGCGGGTAAAA TAGCGGCCATTGATTCGCTTGCT CCACCACGGCTTCCCGGTAACGT CACTACTAACGAACGGTCAATAA ACCCTGCTACTCCGCGTGACATA GCAGCATATGGCGTGCGTTTTTG TCCAAATGAGCGAGCAGCTTCCA TTAAGCCATCTAATTTACGTTCTA AAAGAGGCTCTAGCGTATCAACT GTAATGTCACGCTTACCAATTCC TGTCCCACCCACGGTCATAATGC AAGCATATGATTTGCTCAACTCA AGCACTAAGTTTTTCAAAGTATC CGCTTCATCTGGCAAAATTTGG	367	ABAYE1552: 558..924	*moaCB*	Bifunctional protein [Includes: molybdenum cofactor biosynthesis protein C; molybdenum cofactor biosynthesis protein B]
U2	ACCGGCAAAAATTGCAACCCCAA TAAAACCGTTCACCCAACCGTTC ATAAACTTTTGCATA	61	ABAYE1557: 1..61	-	Putative permease (drug/metabolite transporter)
U3	ACGTAAGCCATATTGAGAGGCTG CAAGTTCGACACGAGCAACATCA CCCAATCGGGTAACC	61	ABAYE1173: 777..837	-	Putative Multidrug/solvent efflux pump membrane transporter
U4	AAACACTGGGAAATATACTTGAA AAATATCAACAACCCCGAACGTA TGCGGGTTGCTCAGT	61	ABAYE3468: 844..904	-	Conserved hypothetical protein; putative exported protein
U5	GTATATCTGGCACAAAGCAAAAG GCTTTCATGCCTTACAACATCAA CAAACCCGTAAA	58	ABAYE1549: 957..1014	*moaA*	Molybdopterin biosynthesis, protein A
U6	CATAACCCCTAAAGTACGATGGC TTGATACGGCATCCCACATTGTT AAGCTA	52	ABAYE1636: 975..1026	*cydB*	Cytochrome d terminal oxidase polypeptide subunit II
U7	GAACTATGGCAATACCGATCAGT AAAGATAGAAGTACCTGCCAG	44	ABAYE2207: 192..235	-	Putative permease
U8	TCCTTTAGAAAATTTGGTGGGAC AAGAAGGCGAAGGTTATA	41	ABAYE1412: 651..691	-	Putative acyl-CoA dehydrogenase protein (acdB-like)
U9	TCATGTCGAGCTAAGCCAGTACC AGCAAAAGCTGAAG	37	ABAYE1194: 1095..1131	-	Putative two-component system sensor histidine kinase
U10	CCCCCCGAACTACATCAATATGA TCATATAAGGCC	35	ABAYE1494: 441..475	-	Putative outer membrane porin, receptor for Fe(III)-coprogen, Fe(III)-ferrioxamine B and Fe(III)-rhodotrulic acid uptake (FhuE)

889

Table 2. Analysis of putative GC1 biomarkers obtained by SVM matches within the Dataset 1 and 2. The table details the unitig ID of the first 10 unitigs that contributed most to the GC1 genome prediction according to the values of the features weight vector, the number of GC1 and non-GC1 genomes typified by MLST that matched the unitig within the Dataset 1 and 2 using blastn and the p-values of Fisher's exact test using a significance level of 0.05. Correlation between non-GC1 genome matches in Datasets 1 and 2 were calculated using Fisher's exact test in R. Total data of the 100 unitigs predicted by SVM is detailed in Supplementary Table S4.

	Dataset 1		**Dataset 2**			
Unitig ID	**Number of GC1 genomes matching the unitig**	**Number of non-GC1 genomes matching the unitig**	**Number of GC1 genomes matching the unitig**	**Number of non-GC1 genomes matching the unitig**	**P-value (proportion of GC1 genome matches in Dataset 1 not differs from proportion of GC1 matches in Dataset 2)**	**P-value (proportion of non-GC1 genome matches in Dataset 1 not differs from proportion of non-GC1 matches in Dataset 2)**
U1	200/200	0/300	312/312	0/4487	p = 1	p = 1
U2	200/200	0/300	311/312	0/4487	p = 1	p = 1
U3	200/200	0/300	311/312	3/4487	p = 1	p = 1
U4	200/200	0/300	310/312	43/4487	p = 0.523	p = 0.1106
U5	200/200	0/300	311/312	0/4487	p = 1	p = 1
U6	200/200	0/300	306/312	1/4487	p = 0.0862	p = 1
U7	200/200	0/300	309/312	2/4487	p = 0.2845	p = 1
U8	200/200	0/300	312/312	1/4487	p = 1	p = 1
U9	200/200	0/300	309/312	5/4487	p = 0.2845	p = 1
U10	200/200	0/300	311/312	0/4487	p = 1	p = 1

Table 3. Rules obtained by SCM. The table details the rule ID, the rule output from the

SCM model, k-mer length and the location where the rule sequence matched in the AYE genome (AN:CU459141.1) or ACICU genome (AN: CP000863.1), and the gene name corresponding to the genome region matched and the gene product of the 10 larger k-mer sequences targeted by the SCM rules. AYE and ACICU genomes were used as GC1 and non-GC1 references, respectively, to locate the k-mer sequences. The prefix "REGION" was used when the match occurred either in an intergenic region or in a combination of intergenic region and a gene. The complete list of the 49 rules obtained by SCM is detailed in Supplementary Table S5.

Rule ID	Rule	Length	Match Location	Gene Name	Gene Product
R127	Presence(AAAAAAGCATGTTTGAAACATGCTTTTTTATTTTATGGCGTTAAACCAACAGGATTGCGATACCAGCTCTGAATTAGCAAAGCCGCGGCAAAACTATCGGCTGACAACTTCTTGGCACGGCCTTGTT)	127	AYE(CU459141.1): REGION:3509175..3509301	-	Putative Holliday junction resolvase
R125	Absence(AATGATTAACAGTACAGGGAAACTAGCAATGAGAAGTTGCATCAAAATGCCTTGACGTTGTGGTGCTGTGCCCTCTACAACAACATTTTGCTTGTTTAAGCTTGGCATAAGTTCAGTGTCTTCAA)	125	ACICU_02924: 125..249	-	ATP-dependent Zn protease
R123	Absence(AATGATTAACAGTACAGGGAAACTAGCAATGAGAAGTTGCATCAAAATGCCTTGACGTTGTGGTGCTGTGCCCTCTACAACAACATTTTGCTTGTTTAAGCTTGGCATAAGTTCAGTGTCTTC)	123	ACICU_02924: 127..249	-	ATP-dependent Zn protease
R121	Absence(AATGATTAACAGTACAGGGAAACTAGCAATGAGAAGTTGCATCAAAATGCCTTGACGTTGTGGTGCTGTGCCCTCTACAACAACATTTTGCTTGTTTAAGCTTGGCATAAGTTCAGTGTCT)	121	ACICU_02924: 129..249	-	ATP-dependent Zn protease
R119	Absence(AATGATTAACAGTACAGGGAAACTAGCAATGAGAAGTTGCATCAAAATGCCTTGACGTTGTGGTGCTGTGCCCTCTACAACAACATTTTGCTTGTTTAAGCTTGGCATAAGTTCAGTGT)	119	ACICU_02924: 131..249	-	ATP-dependent Zn protease

R117	Absence(AACAGTACAGGGAAACTAGCAATGAGAAGTTGCATCAAAATGCCTTGACGTTGTGGTGCTGTGCCCTCTACAACAACATTTTGCTTGTTTAAGCTTGGCATAAGTTCAGTGTCTTCA)	117	ACICU_02924: 126..242	-	ATP-dependent Zn protease
R115	Absence(AACAGTACAGGGAAACTAGCAATGAGAAGTTGCATCAAAATGCCTTGACGTTGTGGTGCTGTGCCCTCTACAACAACATTTTGCTTGTTTAAGCTTGGCATAAGTTCAGTGTCTT)	115	ACICU_02924: 128..242	-	ATP-dependent Zn protease
R113	Absence(AACAGTACAGGGAAACTAGCAATGAGAAGTTGCATCAAAATGCCTTGACGTTGTGGTGCTGTGCCCTCTACAACAACATTTTGCTTGTTTAAGCTTGGCATAAGTTCAGTGTC)	113	ACICU_02924: 130..242	-	ATP-dependent Zn protease
R111	Absence(AAGACACTGAACTTATGCCAAGCTTAAACAAGCAAAATGTTGTTGTAGAGGGCACAGCACCACAACGTCAAGGCATTTTGATGCAACTTCTCATTGCTAGTTTCCCTGTAC)	111	ACICU_02924: 128..238	-	ATP-dependent Zn protease
R109	Absence(AAGACACTGAACTTATGCCAAGCTTAAACAAGCAAAATGTTGTTGTAGAGGGCACAGCACCACAACGTCAAGGCATTTTGATGCAACTTCTCATTGCTAGTTTCCCTGT)	109	ACICU_02924: 128..236	-	ATP-dependent Zn protease

Table 4. Analysis of the SCM rules matches within the Dataset 1 and 2. The table details the rule ID of the 10 larger k-mer sequences targeted by the SCM rules, the number of GC1 and non-GC1 genomes typified by MLST that matched the rule within the Datasets 1 and 2 using blastn and the p-values of Fisher's exact test using a significance level of 0.05. Correlation between non-GC1 genome matches in Datasets 1 and 2 were calculated using Fisher’s exact test in R. Total data of the 49 SCM rules is detailed in Supplementary Table S6.

	Dataset 1		Dataset 2			
Rule ID	**Number of GC1 genomes matching the rule**	**Number of non-GC1 genomes matching the rule**	**Number of GC1 genomes matching the rule**	**Number of non-GC1 genomes matching the rule**	**P-value (proportion of GC1 genome matches in Dataset 1 not differs from proportion of GC1 matches in Dataset 2)**	**P-value (proportion of non-GC1 genome matches in Dataset 1 not differs from proportion of non-GC1 matches in Dataset 2)**
R127	200/200	4/300	310/312	166/4487	p = 0.523	p = 0.03405
R125	200/200	1/300	310/312	13/4487	p = 0.523	p = 0.5964
R123	200/200	1/300	310/312	13/4487	p = 0.523	p = 0.5964
R121	200/200	1/300	310/312	13/4487	p = 0.523	p = 0.5964
R119	200/200	1/300	310/312	13/4487	p = 0.523	p = 0.5964
R117	200/200	1/300	310/312	12/4487	p = 0.523	p = 0.5693
R115	200/200	1/300	310/312	12/4487	p = 0.523	p = 0.5693
R113	200/200	1/300	310/312	12/4487	p = 0.523	p = 0.5693
R111	200/200	1/300	310/312	12/4487	p = 0.523	p = 0.5693
R109	200/200	1/300	310/312	12/4487	p = 0.523	p = 0.5693

Table 5. Prediction metrics on test dataset partitions from Dataset 1 using the best performing SVM model. Correlation between MLST typing and model prediction were calculated using Fisher's exact test in R.

									Number of isolates typified by MLST		
ACC	**SENS**	**SPE**	**PRE**	**F1**	**TP**	**TN**	**FP**	**FN**	**GC1**	**non-GC1**	**Correlation of MLST typing to the model prediction**
1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	52	48	0	0	52	48	$p < 2.2e^{-16}$

ACC: Accuracy; SENS: Sensitivity; SPE: Specificity; PRE: Precision; F1: F1-Score; TP: True Positives, TN: True Negatives; FP: False Positives; FN: False Negatives.

Table 6. Prediction metrics on test dataset partitions from Dataset 1 using the best performing SCM models. Correlation between MLST typing and model prediction were calculated using Fisher's exact test in R. Total metrics of the 49 rules obtained by SCM is

detailed in Supplementary Table S7.

										Number of isolates typified by MLST		
Rule ID	**SENS**	**SPE**	**PRE**	**ACC**	**F1**	**TP**	**TN**	**FP**	**FN**	**GC1**	**non-GC1**	**Correlation of MLST typing to the model prediction**
R127	1.00	0.96	0.95	0.98	0.97	72	91	4	0	72	95	$p < 2.2e^{-16}$
R125	1.00	0.99	0.99	0.99	0.99	72	94	1	0	72	95	$p < 2.2e^{-16}$
R123	1.00	0.99	0.99	0.99	0.99	72	94	1	0	72	95	$p < 2.2e^{-16}$
R121	1.00	0.99	0.99	0.99	0.99	72	94	1	0	72	95	$p < 2.2e^{-16}$
R119	1.00	0.99	0.99	0.99	0.99	72	94	1	0	72	95	$p < 2.2e^{-16}$
R117	1.00	0.99	0.99	0.99	0.99	72	94	1	0	72	95	$p < 2.2e^{-16}$
R115	1.00	0.99	0.99	0.99	0.99	72	94	1	0	72	95	$p < 2.2e^{-16}$
R113	1.00	0.99	0.99	0.99	0.99	72	94	1	0	72	95	$p < 2.2e^{-16}$
R111	1.00	0.99	0.99	0.99	0.99	72	94	1	0	72	95	$p < 2.2e^{-16}$
R109	1.00	0.99	0.99	0.99	0.99	72	94	1	0	72	95	$p < 2.2e^{-16}$

ACC: Accuracy; SENS: Sensitivity; SPE: Specificity; PRE: Precision; F1: F1-Score; TP: True Positives, TN: True Negatives; FP: False Positives; FN: False Negatives.

FIGURE LEGENDS

Figure 1. Diagram of the workflow for GC1 biomarker discovery using machine learning. The diagram indicates the steps we followed during the present work related to genomes collections used, dataset preparation, ML analysis and the design of a method for *A. baumannii* GC1 strain identification.

Figure 2. Mean Accuracy of SVM and SCM models. Bars with blue color show the mean accuracy for the models on the training dataset. Bars with green color are the accuracy of the models on the test dataset. The models were run with the best hyperparameters values selected from the five-folds cross-validation. The error bars are standard deviations. **(A)** SVM model, **(B)** SCM model.

SUPPLEMENTAL TABLES LEGENDS.

Table S1. A. baumannii genomes used as input dataset for SVM algorithm within the Group 1 genome collection. GenBank accession number and classification of the 500 genomes in ML analysis. Genomes were classified as GC1 and non-GC1 according MLST Pasteur scheme.

Table S2. A. baumannii genomes used as input dataset for SVM algorithm within the Group 2 genome collection. Genomes were classified as GC1 and non-GC1 according MLST Pasteur scheme. ST Pasteur column with "-" value indicates that the genome does not yet have a ST assigned in the PubMLST database.

Table S3. Putative GC1 biomarkers obtained by SVM. The table details the unitig ID, unitig sequence, unitig length, the location where the unitig sequence matched in the AYE genome (AN:CU459141.1) or the AB0057 genome (AN: CP001182.2) and the gene name corresponding to the genome region matched and the gene product. AYE strain genome was used as GC1 reference to locate the unitigs sequences. However, when the unitig was not found in the AYE genome, AB0057 strain genome was used instead. The prefix "REGION" was used when the match occurred either in an intergenic region or in a combination of intergenic region and a gene.

Table S4. Analysis of putative GC1 biomarkers obtained by SVM matches within the Dataset 1 and 2. The table details the unitig ID, the number of GC1 and non-GC1 genomes typified by MLST that matched the unitig within the datasets 1 and 2 using blastn and the p-values of Fisher's exact test using a significance level of 0.05. Correlation between non-GC1 genome matches in Datasets 1 and 2 were calculated using Fisher’s exact test in R.

Table S5. Rules obtained by SCM. The table details the rule ID, the rule output from the SCM model, k-mer length and the location where the rule sequence matched in the AYE genome (AN:CU459141.1) or ACICU genome (AN: CP000863.1) and the gene name corresponding to the genome region matched and the gene product. AYE and ACICU genomes were used as GC1 and non-GC1 references, respectively, to locate the k-mer sequences. The prefix "REGION" was used when the match occurred either in an intergenic region or in a combination of intergenic region and a gene.

Table S6. Analysis of the SCM rules matches within the Dataset 1 and 2. The table details the rule ID, the number of GC1 and non-GC1 genomes typified by MLS that matched the rule within the datasets 1 and 2 using blastn and the p-values of Fisher's exact test using a significance level of 0.05. Correlation between non-GC1 genome matches in Datasets 1 and 2 were calculated using Fisher’s exact test in R.

Table S7. Prediction metrics on test dataset partitions from Dataset 1 using the best performing SCM models. Correlation between MLST typing and model prediction were calculated using Fisher’s exact test in R.

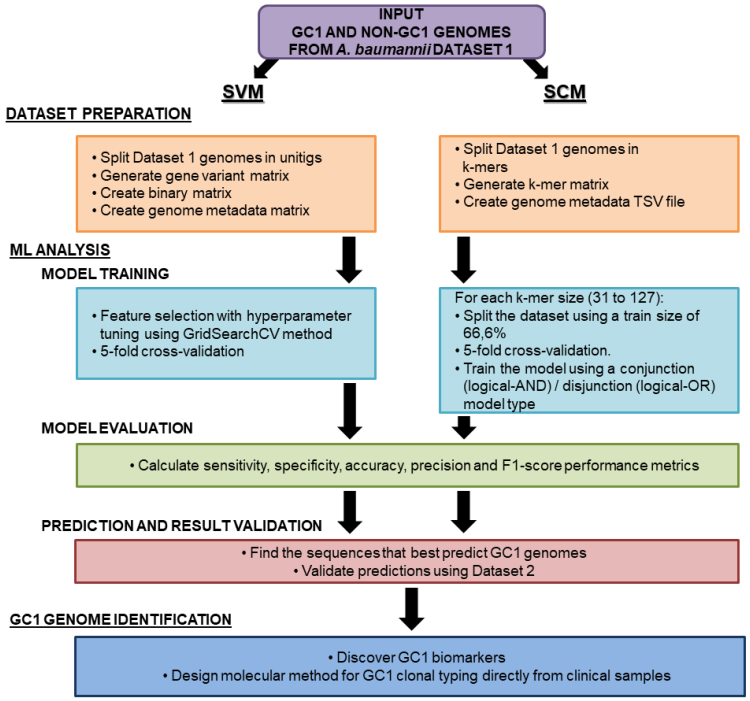

Figure 1. Diagram of the workflow for GC1 biomarker discovery using machine learning. The diagram indicates the steps we followed during the present work related to genomes collections used, dataset preparation, ML analysis and the design of a method for *A. baumannii* GC1 strain identification.

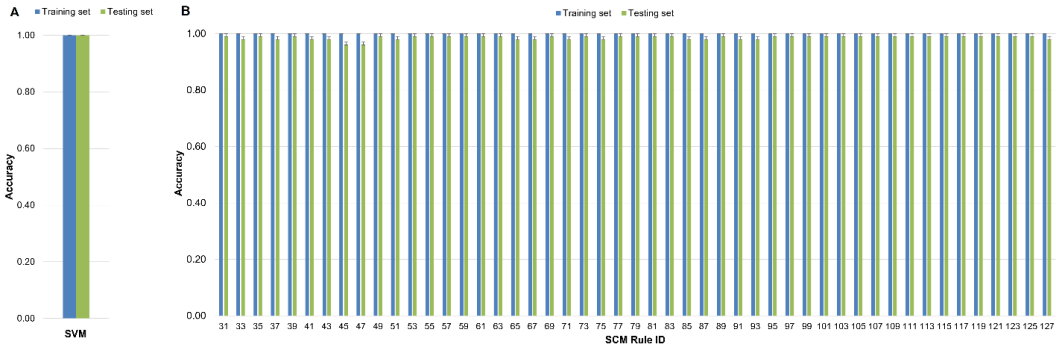

Figure 2. Mean Accuracy of SVM and SCM models. Bars with blue color show the mean accuracy for the models on the training dataset. Bars with green color are the accuracy of the models on the test dataset. The models were run with the best

hyperparameters values selected from the five-folds cross-validation. The error bars are standard deviations. **(A)** SVM model, **(B)** SCM model.