Identification of a specific biomarker of *Acinetobacter baumannii* Global Clone 1 by machine learning and PCR related to metabolic fitness of ESKAPE pathogens.

Verónica Elizabeth Álvarez[1]#, María Paula Quiroga[1,2] and Daniela Centrón[1]

[1]Laboratorio de Investigaciones en Mecanismos de Resistencia a Antibióticos (LIMRA), 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.

[2]Nodo de Bioinformática. Instituto de Investigaciones en Microbiología y Parasitología Médica, Facultad de Medicina, Universidad de Buenos Aires - Consejo Nacional de Investigaciones Científicas y Técnicas (IMPaM, UBA-CONICET), Ciudad Autónoma de Buenos Aires, Argentina.

Running Head: Identification of *Acinetobacter baumannii* GC1

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

Word count for the abstract: 198

Word count for the text: 6973

ABSTRACT

Since the emergence of high-risk clones worldwide, constant investigations have been undertaken to comprehend the molecular basis that led to their prevalent dissemination in nosocomial settings over time. So far, the complex and multifactorial genetic traits of this type of epidemic clones have only allowed the identification of biomarkers with low specificity. A machine learning algorithm was able to recognize unequivocally a biomarker for the early and accurate detection of *Acinetobacter baumannii* Global Clone 1 (GC1), one of the most disseminated high-risk clones. Support Vector Machine identified the U1 sequence with 367 nucleotides length that matched a fragment of the *moaCB* gene, which encodes the molybdenum cofactor biosynthesis C and B proteins. U1 differentiates specifically between *A. baumannii* GC1 and non-GC1 strains, becoming a suitable biomarker capable of being translated into clinical settings as a molecular typing method for early diagnosis based on PCR as shown here. Since the metabolic pathways of Mo enzymes have been recognized as putative therapeutic targets for ESKAPE pathogens, our findings highlighted that machine learning can be also useful in intricate knowledge gaps of high-risk clones and implies noteworthy support to the literature to identify challenging nosocomial biomarkers for other multidrug-resistant high-risk clones.

IMPORTANCE

A. baumannii GC1 is an important high-risk clone that rapidly develops extreme drug resistance in the nosocomial niche. Furthermore, several strains were identified worldwide in environmental samples exacerbating the risk of human interactions. Early

diagnosis is mandatory to limit its dissemination and to outline appropriate antibiotic stewardship schedules. A region of 367 bp length (U1) within the *moaCB* gene not subjected to Lateral Genetic Transfer or to antibiotic pressures was successfully found by Support Vector Machine algorithm that predicts *A. baumannii* GC1 strains. PCR assays have confirmed that U1 specifically identifies *A. baumannii* GC1 strains. At the same time, research on the group of Mo enzymes proposed this metabolic pathway related to superbug´s metabolism as a potential future drug target site for ESKAPE pathogens due to its central role in bacterial fitness during infection. These findings confirmed the importance of machine learning applied to the burden of the rise of antibiotic resistance.

51 **INTRODUCTION**

52 *Acinetobacter baumannii* is an opportunistic and nosocomial Gram-negative pathogen
53 that causes a wide range of nosocomial infections. It is included in the group of ESKAPE
54 pathogens (1, 2). Nosocomial infections by this species have increased in recent years,
55 which added to its ability to acquire and spread antibiotic resistance genes to all families
56 of antibiotics, are considered a serious global threat worldwide (3, 4). The majority of *A.*
57 *baumannii* isolates that are broadly resistant to antibiotics belong to two pandemic
58 clones, known as global clones 1 (GC1) and 2 (GC2) (3, 5, 6). Recent epidemiological
59 studies of carbapenem-resistant *A. baumannii* (CRAB) isolates, revealed that *A. baumannii*
60 GC1 is the prevalent CRAB clone in several countries (7–10). In addition, *A. baumannii* GC1
61 strains have been found worldwide in environmental samples from water, soil, and
62 animals (11, http://www.acinetobacterbaumannii.no/).

63 Over time, molecular methods with different degrees of resolution have been used
64 to type *A. baumannii* strains such as amplified fragment length polymorphism analysis,
65 ribotyping, macrorestriction analysis by pulsed-field gel electrophoresis, multiplex
66 polymerase chain reactions (PCR), multilocus sequence typing (MLST) and more recently
67 whole genome sequencing (WGS) (12–18). A typing scheme based on two multiplex PCR
68 targeting three genes (*ompA*, *csuE*, and $bla_{\text{OXA-51-like}}$) has been used for the assignment of
69 *A. baumannii* isolates into two major PCR-based groups corresponding to *A. baumannii*
70 GC1 and GC2 (19). Also, since correlation has been detected between particular $bla_{\text{OXA-51-}}$
71 $_{\text{like}}$ alleles and some epidemic lineages, sequence analysis of the $bla_{\text{OXA-51-like}}$ gene has been
72 proposed as a useful typing method for *A. baumannii* isolates (19–21). In agreement to

73 this, a study conducted on sixty *A. baumannii* isolates collected worldwide demonstrated
74 that isolates belonging to *A. baumannii* GC1 encoded enzymes from the OXA-69 cluster
75 which included OXA-69, OXA-92, OXA-107, OXA-110 and OXA-112 enzymes (20). However,
76 not all the isolates encoding the OXA-69 cluster belonged to *A. baumannii* GC1 such as *A.*
77 *baumannii* A92 strain that belongs to GC2 (20). An additional typing method of *A.*
78 *baumannii* GC1 strains, consisted in the detection of 108 bp deletion in the 5'-conserved
79 segment (5'-CS) of the class 1 integron located in AbaR3 genomic island (22). Nevertheless,
80 since AbaR3 is not present in all *A. baumannii* GC1 strains, this approach serves as a
81 marker for some diverged lineages within *A. baumannii* GC1 clone (22). All the previously
82 described methods based on PCR include target genes that are subjected to Lateral
83 Genetic Transfer and/or antibiotic pressure, which represents a limitation for the
84 specificity of the technique.

85 With the increasing throughput and decreasing cost of DNA sequencing, large
86 numbers of bacterial genomes have been submitted to public databases (23, 24).
87 Genome-wide studies of DNA variation related to antibiotic resistance phenotypes have
88 garnered high public interest, especially since several multidrug-resistant strains have
89 emerged worldwide (25). New candidate biomarkers leading to the identification of
90 resistant pathogens require the study and development of fast, easily applicable, and
91 accurate tools. Furthermore, with the help of computational algorithms, such studies can
92 be conducted at a much larger scale producing more significant results (26–29). Machine
93 learning (ML) algorithms and statistics have been used increasingly to build models that
94 correlate genomic variations with phenotypes that may help to predict bacterial

95 phenotypes and genotypes (30–36). In supervised ML, each learning sample includes the
96 outcome (class label) of interest, and it is used to build a prediction model (37). The model
97 takes an outcome measurement (e.g., a bacterium having a resistant phenotype or
98 genotype) and tries to learn from the available data (e.g., information on genomic
99 mutations) to predict the outcome measure. The developed model is then applied to new
100 and unseen data. The goal of the algorithm is to train a model that accurately foresees the
101 correct outcome for any input (38).

102 Support Vector Machine (SVM) is a supervised learning algorithm formally
103 characterized by a separating hyperplane that divides binary data to solve both
104 classification and regression problems. SVM aims to correctly classify samples based on
105 examples in the training dataset (39–41). On the other hand, Set Covering Machine (SCM)
106 is a supervised learning algorithm that uses a greedy approach to produce
107 uncharacteristically sparse rule-based models from the input data (42). Both algorithms
108 have been applied to several biological knowledge gaps and proved to be accurate in
109 predicting novel antibacterial agents (43), antibiotic resistance genes (32, 33, 44–46),
110 identification of microorganisms (47–49), cancer diagnosis (50–53), among others (30, 47,
111 54–57).

112 ML can deal with large and diverse datasets to extract relevant information (58).
113 Given the wide use of ML in biology, and the absence of accurate identifiers for early
114 diagnosis of *A. baumannii* GC1 strains, our study aimed to assess whether ML could be
115 applied to process thousands of genomes to identify a suitable *A. baumannii* GC1
116 biomarker. We also wanted to analyze whether ML could be combined with other

117 techniques such as PCR and/or qPCR with high-resolution melting (HRM) assays to provide
118 a molecular typing method capable of being translated into clinical settings. For these
119 reasons, we built predictive models for typing *A. baumannii* GC1 genomes by training SVM
120 and SCM classifiers. From these classifiers, we identified a new and specific genomic
121 biomarker for the early detection of *A. baumannii* GC1 strains by PCR technique not
122 subjected to selective pressure of antibiotics nor to Lateral Genetic Transfer.

123 **RESULTS**

124 To identify new genomic biomarkers that uniquely identify strains belonging to *A.*
125 *baumannii* GC1, we applied the SVM and SCM algorithms to the Dataset 1 (500 genomes)
126 and the Dataset 2 (4799 genomes) (Table S1 and S2). Firstly, we applied both algorithms to
127 the Dataset 1 which was composed of 200 *A. baumannii* GC1 and 300 *A. baumannii* non-
128 GC1 genomes. We aimed to predict whether a particular genome in the Dataset 1
129 belonged to *A. baumannii* GC1 or not by using SVM and SCM algorithms. The Dataset 1
130 was used as input for both algorithms during the training and the testing of the models.
131 Once we obtained accurate models that predicted putative biomarkers sequences, we
132 used the second *A. baumannii* genome collection, the Dataset 2 (Table S2). The Dataset 2
133 was composed of 312 *A. baumannii* GC1 and 4487 *A. baumannii* non-GC1 genomes. By
134 using blastn searches, we analyzed whether the predicted putative biomarker sequences
135 were also found in the Dataset 2 genomes maintaining the same pattern found in the
136 Dataset 1. This analysis was done to perform an external validation of the predictions
137 made by both algorithms. The study workflow is summarized in Figure 1.

138

139 **Obtaining unique *A. baumannii* GC1 predictive sequences with SVM**

140 The Pasteur scheme for Multilocus Sequence Typing (MLST) has the potential to
141 identify isolates belonging to *A. baumannii* GC1 providing a neat demarcation of STs
142 composing GC1 and non-*A. baumannii* clonal complexes (5, 59–62). Considering this fact,
143 we annotated the ST of each genome in the Datasets 1 and 2 and categorized them as
144 "GC1" or "non-GC1" (see Materials and Methods: "MLST classification" section). Then we
145 run the DBGWAS program using the Dataset 1 genome sequences as input; we obtained a
146 total of 1,622,573 distinct unitigs that represented sequences of diverse length of the
147 Dataset 1 taking into account the genomic variation (63). The length of the unitigs
148 obtained was between 31 and 32,759 bp. We used the variant matrix built by DBGWAS to
149 create the binary matrix used as input for the SVN algorithm. Unitigs that were found in
150 less than 50 genomes from the Dataset 1 were discarded; we kept in the input binary
151 matrix only the unitigs of size 31 to 385 bp.

152 Parameter tuning and model validations were performed using a 5-fold cross-
153 validation and a grid search over a range of given values to determine the SVM kernel and
154 hyperparameters that generated the best area under the curve (AUC). Once we obtained
155 the SVM model that best fitted the Dataset 1 genomes, we extracted the first 100 unitigs
156 that contributed most to *A. baumannii* GC1 strain prediction (Tables 1 and S3) according
157 to the values of the features weight vector. By using blastn, we corroborated if the unitigs
158 predicted by SVM were specific for *A. baumannii* GC1 detection. For this purpose, we

159 searched for the unitig sequences in the genomes of the Dataset 2 to assess which were
160 the unitigs with the greatest number of matches within each genome class (*A. baumannii*
161 GC1 or non-GC1) (Tables 2 and S4). We also annotated the genome location and gene
162 product related to the 100 unitigs (Tables 1, S3, 3 and S5).

163 **Building SCM models for *A. baumannii* GC1 identification**

164 In an additional approach, we applied the SCM algorithm implemented in Kover
165 (30). We used as input the k-mers of the 500 *A. baumannii* genomes contained in the
166 Dataset 1 considering the class label for each genome (*A. baumannii* GC1 or non-GC1). We
167 ran Kover using k-mer sizes ranging from 31 to 127. Although output rules could be
168 conjunctions (logical-AND) or disjunctions (logical-OR), we obtained 49 simple rules
169 without conjunctions or disjunctions in the models of our study (Table 3 and S5). Rules
170 depicted the presence (n=10) or the absence (n=39) of k-mers in the Dataset 1 genomes.
171 Besides, we registered the number of matches of the k-mer sequences against *A.*
172 *baumannii* GC1 and non-GC1 genomes contained in the Dataset 1 (Table 4 and S6); we
173 also annotated the k-mer sequences (Tables 4 and S6).

174 We observed that the rules obtained by the SCM models selected fragments of
175 different length that matched the loci ACICU_02924 (n=23), ACICU_02095 (n=5),
176 ABAYE3455 (n=5), ACICU_01506 (n=2) and ABAYE2468 (n=2) (Table 3 and S5). This fact
177 could be caused by point mutations contained in the loci that the SCM models associated
178 with *A. baumannii* GC1 prediction would have made as previously found (30).

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

180 We were interested in obtaining the longest DNA sequences shared by all *A.*
181 *baumannii* GC1 genomes, and at the same time, without matches among *A. baumannii*
182 non-GC1 genomes. For this purpose, unitigs and k-mer sequences obtained as candidate
183 biomarkers for *A. baumannii* GC1 using the SVM and SCM algorithms were sorted in
184 descending order according to the sequence length. Then, we considered the number of
185 *A. baumannii* GC1 genomes that matched the unitig sequence or the SCM rule and sorted
186 the candidate sequences in ascending order according to the number of *A. baumannii*
187 non-GC1 genome matches.

188 First, we analyzed the SVM results. Data related to unitigs obtained are listed in
189 Tables 1, 2, S3, S4 and, Figure 2. We named unitigs from U1 to U100. We observed that
190 the unitigs named U1 to U12 were found in 100% of *A. baumannii* GC1 genomes while
191 they were not found in *A. baumannii* non-GC1 genomes from the Dataset 1. However, U1
192 was the only unitig found in 100% of *A. baumannii* GC1 genomes within the Datasets 1
193 and 2, as well as absent in *A. baumannii* non-GC1 genomes of both datasets, being a
194 specific biomarker for *A. baumannii* GC1 identification. Despite the U8 sequence was
195 found in 100% of *A. baumannii* GC1 genomes within the Dataset 1 and 2, it was found in
196 1/4487 of *A. baumannii* non-GC1 genomes from the Dataset 2. The U8 sequence matched
197 *A. baumannii* AYE genome in the locus tag ABAYE1412 between 651 and 691 coordinates
198 (Table 1 and S3). The fragment is part of a gene that encodes a putative acyl-CoA
199 dehydrogenase protein (*acdB*-like). As U8 sequence was not found exclusively in *A.*
200 *baumannii* GC1 genomes, we discarded it as a possible biomarker of *A. baumannii* GC1. U4

201 was the sequence among these 12 unitigs that had the higher number of matches
202 (43/4487) with *A. baumannii* non-GC1 genomes from the Dataset 2.

203 The SCM results are listed in Tables 3, 4, S5, S6, and Figure 3. We observed that
204 43/49 rules targeted 100% of *A. baumannii* GC1 genomes from the Dataset 1 but also
205 targeted between 1 and 4 of *A. baumannii* non-GC1 genomes from the Dataset 1. Within
206 these 43 rules, two sets of rules targeted only 1 *A. baumannii* non-GC1 genome from the
207 Dataset 1; as an example of this, 26/49 rules targeted genome with AN: GCF_000248195.1
208 (ST 69) and 1/49 rules (R49) targeted genome with AN: GCF_000453745.1 (ST 2) (Figure 3).
209 Concerning the Dataset 2, 12/49 rules targeted 100% of *A. baumannii* GC1 genomes but
210 also targeted several *A. baumannii* non-GC1 genomes from the Dataset 2. Since no rule
211 obtained can uniquely identify *A. baumannii* GC1 genomes contained in our datasets, we
212 were unable to obtain a putative biomarker from the results of the SCM models.

213 When using the Dataset 2 to validate the results obtained by the SVM and SCM
214 models from the Dataset 1, we observed that in the case of the SVM classifier there was
215 no statistically significant difference ($P > 0.05$) in the number of matches of the unitigs
216 regardless of the genome dataset (Dataset 1 or 2) except for the unitig U69 (Tables 2 and
217 S4). In the case of U69, we found a significant difference between the number of matches
218 with *A. baumannii* non-GC1 genomes from the Dataset 1 and 2 ($P = 0.01077$). Concerning
219 the SCM classifier, we observed that there was no statistically significant difference in the
220 number of *A. baumannii* GC1 and non-GC1 genomes from the Dataset 1 and 2 that
221 matched the rules ($P > 0.05$) (Tables 4 and S6). As we mentioned, the SVM model
222 predicted the U1 sequence as a specific biomarker for *A. baumannii* GC1 genomes. The U1

223 sequence had 367 nucleotides and matched *A. baumannii* AYE genome in the locus tag
224 ABAYE1552 between coordinates 558 and 924 (Table 1 and S3). This region corresponds to
225 a fragment of the *moaCB* gene which encodes a bifunctional protein that includes the
226 molybdenum cofactor biosynthesis protein C and protein B (64). In particular, the U1
227 sequence matched the region of the *moaCB* gene, which encodes for MoaB protein. The
228 molybdenum cofactor (Moco) is an essential component of a large family of enzymes
229 involved in carbon, nitrogen, and sulfur metabolism whose biosynthetic pathway is
230 evolutionarily conserved. The MoaC protein, together with the MoaA protein, is involved
231 in the first step of Moco biosynthesis (64). Interestingly, various studies have linked in-
232 host survival of prevalent pathogenic bacteria such as *Mycobacterium tuberculosis*,
233 *Escherichia coli*, and *Salmonella enterica* to the presence of functional molybdoenzymes
234 (65–68). In our study, we found that the U1 sequence was 100% conserved in *A.*
235 *baumannii* GC1 genomes (0 SNPs) while it was a variable region with 1 to 74 SNPs in the
236 total of 4987 *A. baumannii* non-GC1 genomes from the Dataset 1 and 2 rendering 94
237 allelic variants in *A. baumannii* non-GC1 genomes. According to these results, the
238 sequence of the *moaCB* gene comprised between 558 and 924 nucleotides allowed
239 accurate discrimination for *A. baumannii* GC1 and non-GC1 genomes, becoming a
240 biomarker that differentiates both groups. Moreover, we observed that the U1 sequence
241 is conserved in all the *A. baumannii* GC1 genomes, but this is not the case for *A.*
242 *baumannii* GC2. For this reason, U1 sequence could not be used as a biomarker for the
243 early detection of *A. baumannii* GC2 strains.

244 Since 95 variants within 5299 *A. baumannii* strains were identified with multiple
245 SNPs along the entire sequence of U1, we proposed that a strategy based on PCR
246 amplification, would allow us to accurately differentiate *A. baumannii* GC1 from non-GC1
247 strains. For this reason, we designed the primer´s pair BioM_GC1_ABA F 5'-
248 TATTCATAGCCTCCTGGATGC-3' and BioM_GC1_ABA R 5'-CCAGATGAAGCGGATACTTTG-3',
249 with coordinates 559 to 914 from ABAYE1552 locus tag representing 356 bp of U1
250 sequence (position 2 to 357). By a blastn search, we identified that this primer pair
251 amplified only the U1 sequence recognized in *A. baumannii* GC1 strains amongst all the
252 variants detected so far. The closest variants showed a mismatch of one nucleotide at the
253 3' end of the reverse primer in *A. baumannii* ST163, ST411 and ST976 (AN
254 GCF_015537765, GCF_000453725 and GCF_010500415, respectively).

255 Experimental analysis, with a total of 35 *A. baumannii* strains, which were 10 *A.*
256 *baumannii* GC1 and 25 *A. baumannii* non-GC1 including 4 ST2 (GC2), 15 ST79, 5 ST119 and
257 1 ST404, was performed to test the primer´s pair BioM_GC1_ABA F and BioM_GC1_ABA F
258 (Figure S1). *A. baumannii* GC1 strains that were isolated more than 25 years apart were
259 tested such as A144 (1997), A155 (1994) and HAX25Aba (2021) (69, 70). The pair of
260 primers designed in this study only amplified by PCR the *A. baumannii* GC1 strains, and as
261 expected, the subsequent DNA sequence analysis showed a 100% identity and query
262 coverage with 356 bp of U1 as confirmed by the clustal alignment (Figure S2).
263 Interestingly, *A. baumannii* Ab103 ST119 that has two mismatches in each primer, at
264 positions 15 and 21 of the 21nts of BioM_GC1_ABA F, and at positions 9 and 21 of the

265 21nts of BioM_GC1_ABA R rendered also a negative result in this PCR´s conditions (see
266 below). This result confirmed the specificity of this pair of primers.

267 The optimum PCR amplification was established in a final volume of 25 µl
268 containing 0.625 U GoTaq® DNA Polymerase (cat. # M3005, Promega, USA), 5 µl 5 X Green
269 GoTaq® buffer with $MgCl_2$ for a final concentration of 1.5mM in the 1X reaction, 0.4 pM
270 each dNTP, 1 µM each primer, and 5 µl of DNA from the boiling of 3-4 *A. baumannii*
271 colonies in 100 µl of sterile H_2O. DNA of *A. baumannii* GC1 strain A144 and water were
272 used as positive and negative control, respectively. The PCR cycling conditions were an
273 initial denaturation at 94 °C for five minutes, followed by 30 cycles of denaturation at 94
274 °C for 45 seconds, annealing at 58 °C for 45 seconds and extension at 72 °C for 30 seconds,
275 with a final extension at 72 °C for five minutes. Alternatively, to this PCR to identify *A.*
276 *baumannii* GC1, TaqMan assays could also provide a molecular typing method capable of
277 being translated into clinical settings to differentiate *A. baumannii* GC1 and other relevant
278 GC or ST.

279 **Evaluation of the performance of the SVM and SCM models**

280 To avoid overfitting, a 5-fold cross-validation was performed on the training
281 datasets used as input in both the SVM and SCM algorithms to select the best
282 hyperparameter values with the highest AUC. Performance during testing was evaluated
283 using the best hyperparameters obtained in terms of sensitivity, specificity, accuracy,
284 precision, and F1 score (Tables 5, 6 and S7). The sensitivity of the SVM model was 1 ± 0.00,
285 indicating that 100% of *A. baumannii* GC1 genomes were correctly identified within the

286 testing dataset. Also, the SVM model achieved a high specificity (1 ± 0.00) when predicting
287 *A. baumannii* non-GC1 genomes from the testing datasets. The precision value was 1 ±
288 0.00 when the model predicted that a genome was within *A. baumannii* GC1, being
289 correct 100% of the time. Accuracy was 1 ± 0.00 indicating that 100% of *A. baumannii* GC1
290 and non-GC1 were correctly predicted. No false positives and no false negatives were
291 predicted by the SVM model.

292 Regarding the SCM models, the mean values of sensitivity, specificity, precision,
293 and accuracy were 1 ± 0.01, 0.98 ± 0.01, 0.97 ± 0.01, and 0.99 ± 0.01, respectively. The
294 mean rate of false positives was 1.11% while the mean rate of false negatives was 0.097%.
295 All the rules obtained by the SCM models predicted false positives within the testing
296 dataset (Table 6).

297 The training and the testing accuracy of the SVM and SCM models were above 0.99
298 (Figure 4), indicating that the SVM and SCM models, were not overfitted. The F1 score,
299 which is the harmonic mean of precision and recall and is commonly used to compare
300 different classification algorithms, was similarly high in the SVM and SCM models (1.00
301 and 0.99 ± 0.01, respectively). We used Fisher's Exact Test to evaluate the performance of
302 the SVM and SCM predictions comparing the actual genome classes (*A. baumannii* GC1
303 and non-GC1) typed by MLST and the predicted classes obtained by the models. In both
304 cases, we obtained $P < 2.2e-16$ (Table 5, 6 and S7) indicating that the SVM and SCM models
305 could significantly classify *A. baumannii* GC1 and non-GC1 strains. Despite these results, as
306 the aim of this work was to find a biomarker that uniquely identifies *A. baumannii* GC1

307 genomes, the SVM model performed better than the SCM models since it did not predict
308 false positives or false negatives.

309 **DISCUSSION**

310 High-risk clones or also named "superbugs" are dangerous clonal complexes with
311 epidemic behavior equipped with exceptional resources both to infect the host and to
312 evolve to extreme drug resistance phenotypes over time in the nosocomial niche (1, 71–
313 77). The molecular understanding of the genetic and/or transcriptomic traits that lead to
314 these capabilities remains still unknown (78, 79). Our study showed that ML applied to the
315 study of high-risk clones can help not only in the identification of thoroughly accurate
316 biomarkers but also contribute to disentangling molecular pathways that drive to
317 epidemic lineages in the nosocomial niche which have not yet been completely
318 deciphered. Accordingly, these findings could be used as therapeutic targets to reduce the
319 dissemination of lineages with epidemic behavior. This is the case of the U1 sequence
320 identified in the present study, which corresponded to 367 bp of the *moaCB* gene that
321 encodes a bifunctional protein that includes the molybdenum cofactor biosynthesis
322 protein C and protein B (64). Mononuclear molybdoenzymes (Mo enzymes) occur in
323 organisms in all domains of life, where they mediate essential cellular functions such as
324 energy generation and detoxification reactions (80). It has been shown that in bacterial
325 pathogens, several processes such as the molybdate uptake, the cofactor biosynthesis,
326 and the activities of Mo enzymes, affect fitness in the host as well as virulence (80). In
327 addition to many studies on Mo enzymes that identified their crucial role in pathogenic
328 species such as *E. coli*, *S. enterica*, *Campylobacter jejuni*, and *Mycobacterium tuberculosis*,

329 some reports recently identified that these enzymes also contribute to the survival of
330 ESKAPE pathogens (66, 81–85). Experimental studies must be undertaken to investigate
331 the role of the *moaCB* gene in the virulence and fitness of *A. baumannii* which is also
332 included in the ESKAPE group. Based on our results and previous experimental data in
333 other pathogenic species (80), we can hypothesize that the U1 fragment of the *moaCB*
334 gene in *A. baumannii* GC1 may be related to an essential metabolic pathway that play a
335 vital role in the maintenance of epidemic clones in the hospital environment. Since it has
336 been found that most of the Mo enzymes belong to groups that are unique to
337 prokaryotes, these have been proposed as promising targets for the development of new
338 antibiotic agents (80).

339 Given the variability observed in the biomarker U1 between *A. baumannii* GC1 and
340 the 94 variants in *A. baumannii* non-GC1 strains, we propose a simple molecular biology
341 strategy of one step of PCR amplification, to accurately differentiate *A. baumannii* GC1
342 from non-GC1 strains without performing MLST, or WGS and comparative genomics. The
343 strategy proposed here, that was experimentally tested through this study, is accessible to
344 a wide range of clinical and/or research laboratories. In addition, as the methodology
345 consists of a single PCR, the detection of *A. baumannii* GC1 strains can be performed from
346 the colony or directly from the clinical sample, giving the possibility of an early and simple
347 diagnosis of this lineage.

348 Due to the increasing availability of bacterial WGS data, very active research has
349 emerged on the use of this tool for genotype-phenotype prediction of antibiotic
350 susceptibility (30, 33, 34, 44, 54, 86–88); however, there is no data available concerning

351 the identification of high-risk clones based in WGS data excluding the MLST. Previous PCR-
352 based studies used *bla*$_{OXA-51}$ as one out of the three targeted genes to discriminate strains
353 of *A. baumannii* GC1, GC2 and, GC3 (18, 19), and also the detection of deletion of 108 bp
354 in the 5'-conserved segment (5'-CS) of the class 1 integron for identification of *A.*
355 *baumannii* GC1 strains (19) have been proposed. Since *bla*$_{OXA-51}$ has been found in other
356 species (89–91) and the deletion of 108 bp as a biomarker differentiates partially two
357 lineages within *A. baumannii* GC1 (7), and considering that both can be subjected to
358 Lateral Genetic Transfer events, the analysis provided by ML in the present study supports
359 a more accurate and solid implement to evaluate the presence of *A. baumannii* GC1
360 strains. Accordingly, U1 is part of an essential gene in *A. baumannii* GC1 genome (*moaCB*).
361 This suggests that the optimization of metabolic genes from the core genome may be
362 related to the exceptional abilities of high-risk clones. Interestingly, our data suggest that
363 the accessory genome such as genomic islands or transposons involved in pathogenicity or
364 antibiotic resistance, at least in *A. baumannii* GC1 strains, would not have played a causal
365 role in the adaptation of this lineage to the hospital niche over time.

366 In our work, we applied two ML algorithms that differ substantially in
367 methodology. The SVM algorithm used the number of occurring unitigs in *A. baumannii*
368 GC1 and *A. baumannii* non-GC1 genomes to learn and predict clonal membership of the
369 strains, while the SCM model used a greedy approach to construct conjunction or
370 disjunction rules to find the most concise set of k-mers that allows for accurate *A.*
371 *baumannii* GC1 or *A. baumannii* non-GC1 genome prediction. Previously, methods that
372 combined the use of the SVM and SCM algorithms, and the representation of genomic

373 data as k-mers were used to find genomic biomarkers to identify antibiotic resistance (30)
374 or to predict antibiotic resistance from WGS data (54, 86). We ran the SCM algorithm
375 through the Kover program with k-mer lengths between 31 and 127 nucleotides to be able
376 to analyze all the possible rules obtained from these k-mers lengths. Also, we ran the SVM
377 algorithm using unitigs that usually corresponded to a longer sequence than the individual
378 equivalent k-mers. Unitigs are defined as the longest sequences that can be obtained
379 when k-mers overlap by exactly k-1 nucleotides (63, 92, 93). In k-mer-based genome
380 representations, the main downside is that the representation contains a lot of
381 redundancy, since many k-mers are always present or absent simultaneously (e.g., gene
382 deletion/ insertion). In this sense, it has been proposed to replace k-mers with unitigs (63,
383 92, 93).

384 We also proved by using statistical methods that the SVM and SCM models could
385 significantly classify *A. baumannii* GC1 isolates ($P < 0.05$). While the SVM classifier
386 predicted the U1 sequence as a specific biomarker for *A. baumannii* GC1 genomes, none
387 of the rules obtained with the SCM models was able to uniquely identify *A. baumannii* GC1
388 genomes. All the rules obtained by the SCM models matched in *A. baumannii* GC1 and *A.*
389 *baumannii* non-GC1 genomes from both the Dataset 1 and 2. Due to this result, it was not
390 possible to obtain a sequence that could be used as a specific biomarker for *A. baumannii*
391 GC1 strains from the rules obtained by the SCM models. A key step for the successful
392 implementation of ML algorithms is the preparation of the input datasets (58, 94–96). In
393 our study, we faced two issues related to the preparation of the Datasets 1 and 2. On one
394 hand, a limitation in the program DBGWAS during the preparation of the matrix used in

395 the SVM algorithm meant that only a total of 500 genomes could be included in the
396 training dataset. Since our goal was to use the same training dataset for the SVM and SCM
397 models, the 500 genomes were also used as input in the SCM algorithm. On the other
398 hand, the scarcity of *A. baumannii* GC1 genomes available in GenBank compared to *A.*
399 *baumannii* non-GC1 genomes caused the Dataset 2 to have visibly more genomes
400 representing *A. baumannii* non-GC1 class. Despite this fact, the results obtained from the
401 Dataset 1 and 2 remained consistent in 100% ($P > 0.05$) of the cases for the SVM model
402 and more than 91.83% ($P > 0.05$) of the cases for the SCM models. Perhaps the limitation
403 in the number of genomes in the training dataset is one of the reasons why the SCM
404 models did not have enough samples to learn from *A. baumannii* GC1 genomes and
405 therefore could not find a rule that uniquely identifies them. Conversely, the numbers of
406 isolates predicted to be *A. baumannii* GC1 or non-GC1 by the SVM model using the unitig
407 U1 were the same as the result obtained by the MLST typification technique. This fact
408 indicated that the SVM model obtained was excellent in the classification of *A. baumannii*
409 GC1 and non-GC1 genomes. It will be interesting to study in future works how the Dataset
410 1 splitting strategy (unitigs or k-mers), the number of genomes of each class (*A. baumannii*
411 GC1 and non-GC1) in the Dataset 1, and the total number of *A. baumannii* genomes in the
412 Dataset 1 impact on the SVM and SCM models predictions and performance. One possible
413 approach could be using as input the binary matrix obtained from the unitigs representing
414 the Dataset 1 genomes generated by the program DBGWAS in the program Kover and
415 then to analyze the SCM model results. In the same way, we could obtain k-mer profiles
416 (k-mer sizes ranging from 31 to 127) from the Dataset 1 genomes and the k-mer matrixes

417 associated with each profile using the DSK k-mer counter (97). DSK is used by Kover to
418 internally compute k-mer profiles from the input genomes (98). Then we could use the k-
419 mer matrixes as input in the SVM algorithm. As result, we would obtain new putative
420 sequence biomarkers from both approaches, and we could compare them with the ones
421 obtained in the current work.

422 In conclusion, these results suggest that the application of ML to identify
423 biomarkers for high-risk clones or superbugs can also be used at an exploratory level of
424 great precision since can provide novel understandings about bacterial adaptation to the
425 nosocomial niche. In turn, this data can contribute to the experimental work with the
426 possibility to be further translated to the clinical settings. The SVM algorithm made
427 genetic predictions based on the presence or absence of short genomic sequences present
428 in *A. baumannii* GC1 and non-GC1 genomes, that detected a biomarker, U1, not related to
429 Lateral Genetic Transfer neither to accessory genome nor to antibiotic pressures that
430 uniquely identifies the CG1 strains, which in agreement with previous experimental works
431 on the group of Mo enzymes, showed that the application of ML could be a powerful tool
432 to discover new therapeutic targets for the development of new antibiotic agents.

433 **MATERIALS AND METHODS**

434 **MLST classification.** Multilocus sequence typing (MLST) of all genomes in the data
435 collection was performed *in silico* using the mlst software developed by T. Seemann
436 (https://github.com/tseemann/mlst) and the Pasteur's MLST database and schema for *A.*
437 *baumannii* (https://pubmlst.org/organisms/acinetobacter-baumannii). *A. baumannii*

438 sequence type (ST) numbers ST 1, 19, 20, 81, 94, 328, 460, 623, 315, 717 and 1106 were
439 classified into *A. baumannii* GC1 and others ST into *A. baumannii* non-GC1 genomes
440 (Tables S1 and S2).

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

443 The Dataset 1 was composed of 200 *A. baumannii* GC1 and 300 *A. baumannii* non-
444 GC1 genomes obtained from Genbank and typified by MLST as previously described. These
445 genomes were retrieved from the GenBank assembly database filtering by *Acinetobacter*
446 *baumannii* in the search by organism option
447 (https://www.ncbi.nlm.nih.gov/assembly/organism/, last accessed in July 2021). *A.*
448 *baumannii* GC1 genomes included 18 genomes used in a previous study (Álvarez et al.,
449 2020) and 182 *A. baumannii* genomes as scaffolds and contigs (Table S1). *A. baumannii*
450 non-GC1 genomes included five genomes belonging to other high-risk epidemic clones
451 such as ACICU (AN: CP031380.1) as representative of GC2, Naval-13 (AN:
452 AMDR01000001.1) as representative of GC3, AB33405 (NZ_JPXZ00000000.1) as
453 representative of local epidemic clone CC113 and both ATCC 17978 (AN: CP018664.1), and
454 A118 (AN: AEOW01000000) as sporadic clones. We also included 205 *A. baumannii* non-
455 GC1 genomes as scaffolds and contigs (Table S1). The Dataset 2 was composed of 312 *A.*
456 *baumannii* GC1 and 4487 *A. baumannii* non-GC1 genomes (Table S2) retrieved from the
457 GenBank assembly database (last accessed in July 2021). The number of genomes in this
458 dataset was limited by *A. baumannii* GC1 and non-GC1 genomes available in the GenBank
459 database at the time of the query. We used the Dataset 2 to validate using blastn searches

460 the results obtained by the SVM model. (Table S2). STs were numbered according to the
461 Pasteur scheme for MLST (Table S1 and S2).

462 **Machine learning analyses.**

463 The SVM classifier is based on the maximization of the margin around the
464 hyperplane ($w^T x + b$) separating samples or instances of the different classes (57, 99). Each
465 instance $i = 1,...,m$ consists of an N-dimensional feature vector x_i and a class label $y_i \in \{+1,$
466 $-1\}$. The maximization of the margin corresponds to the following minimization:

(1)

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

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

467 In this soft-margin SVM equation, ξ_i is a penalty for misclassification or
468 classification within the margin. Parameter C sets the weight of this penalty. The resulting
469 weight vector $w*$ encodes the contributions of all features to the classifier (57).

470 We created a Python script using scikit-learn (100) to run the SVM algorithm. The
471 script evaluated the classifier through a 5-fold cross-validation. In detail, the data was split
472 into five consecutive folds (without shuffling) using the Python scikit-learn (sklearn) KFold
473 function (https://scikit-learn.org/stable/modules/cross_validation.html#k-fold), and five
474 models were built. Each fold was used once as a test set while the four remaining folds
475 formed the training set. During each of the five iterations, the hyperparameter tuning was
476 done using a 5-fold cross-validated grid search using the GridSearchCV function
477 implemented in the sklearn.model_selection package (https://scikit-

learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html#sklearn.model_selection.GridSearchCV) to find the best hyperparameters. We evaluated linear, polynomial (with a default degree of 3), radial basis function (RBF), and sigmoid kernels. We considered values between 0.01 and 100 for the penalty parameter of the error term (C) and values between 0.000001 and 10 for the gamma parameter. The predictions of all five iterations were compared using the AUC score (https://scikit-learn.org/stable/modules/model_evaluation.html#roc-metrics). Finally, we built up the classifier from the entire training set using the best hyperparameters (with the highest AUC) identified through cross-validation and applied the best model to the test set. The highest AUC were obtained using kernel = linear, gamma = 0.000001 and C = 0.01 (code available on GitHub at https://github.com/vealvarez/SVM_GC1).

The SCM (42), 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

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

495 Given a set *R* of candidate rules, the SCM algorithm attempts to find a model that
496 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
497 otherwise while using the smallest number of rules in *R* (101).

498 We used the program Kover which implements the SCM algorithm (30, 98). Kover
499 combines the SCM algorithm with the k-mer representation of genomes, which reveals
500 uncharacteristically sparse models that explicitly highlight the relationship between
501 genomic variations and the phenotype of interest (30). We ran the program using the
502 Dataset 1 (see Input dataset preparation below) for k-mer sizes from 31 to 127 (taking
503 only the odd numbers between them). The smallest value of k was set to 31 since
504 extensive testing has shown that this size is optimal for bacterial genome assembly and
505 has been employed for studies based on reference-free bacterial genome comparisons
506 (92, 102). The greatest k-mer size was set to 127 since it is the maximum value accepted
507 as a parameter in Kover. We chose only odd values of k to avoid the formation of
508 palindromes (103). For each k-mer size, we split the Dataset 1 into a training dataset (2/3
509 of the genomes) and a testing dataset (1/3 of the genomes). Then, we trained the 49
510 models corresponding to each k-mer using a conjunction/disjunction model type. The best
511 conjunctive and/or disjunctive model for each k-mer was selected using 5-fold cross-
512 validation to determine the optimal rule scoring function with default parameters.

513 **Input dataset preparation for ML models.** We split the 500 genomes included in the
514 Dataset 1 into unitigs by using the program DBGWAS (63). Unitigs are stretches of DNA
515 shared by the strains in a dataset. The DBGWAS method proposes connecting the overlaps
516 of k-mers in a compressed de Bruijn graph (DBG) so that k-mers are extended using the

517 adjacent sequence information in the population, forming unitigs present in the same set
518 of samples as their constituent k-mers. During the first step of the DBGWAS process, the
519 program built a variant matrix, where each variant is a pattern of the presence/absence of
520 unitigs in each genome present in the Dataset 1 (63). We wrote a Python script (code
521 available on GitHub at https://github.com/vealvarez/SVM_GC1) to format the variant
522 matrix and create a presence/absence (coded with the values 1/0 respectively) binary
523 matrix with unitigs as columns (features) and the accession numbers of the genomes as
524 rows (instances). We used the binary matrix as input for the SVM algorithm. In this matrix,
525 we discarded the data about low frequency unitigs (unitigs found in less than 50 *A.*
526 *baumannii* GC1 and non-GC1 genomes). We also integrated the MLST data corresponding
527 to the Dataset 1 genomes and created a two-column matrix used as input for the SVM
528 algorithm. The first column of the matrix contained the accession number of the genomes,
529 and the second column contained a binary variable that indicated whether each genome
530 was typed or not as *A. baumannii* GC1 (-1 = *A. baumannii* non-GC1 genome / 1 = *A.*
531 *baumannii* GC1 genome) according to MLST typing.

532 For the SCM approach, we first packaged the Dataset 1 sequences stored in FASTA
533 files into a Kover dataset using the create from-contigs command. This command also
534 received a tab-separated value (TSV) with the Dataset 1 genome classification according to
535 MLST typing (*A. baumannii* GC1 or non-GC1). The first column of the TSV file described the
536 genome accession number and the second column had the value 1 whether the genome
537 belonged to *A. baumannii* GC1 or the value 0 otherwise. We ran Kover for k-mer sizes
538 from 31 to 127. For each k-mer size, Kover constructed a reference-free input matrix

539 based on k-mer profiles generated with the DSK k-mer counting software. A k-mer
540 presence/absence binary matrix based on the Dataset 1 genomes was then created and
541 used as input for the SCM models of each k-mer size.

542 **Unitig selection using SVM for putative biomarkers analysis.** After obtaining the SVM
543 model with the best hyperparameters, the values of the features weight vector (referred
544 to as the hyperplane normal vector $w*$ in the formula (1)) were accessed through the
545 attribute sklearn.model_selection.GridSearchCV.best_estimator.coef_. The values were
546 sorted from highest to lowest and used to decide the relevance of each unitig sequence
547 (associated with each weight value) during the model prediction (104, 105). It is worth
548 mentioning that in our study, sequence unitigs were used as features in the models. The
549 positive sign of a feature weight value indicates that the feature contributes to *A.*
550 *baumannii* GC1 class prediction (represented by the value 1) and the negative sign
551 indicates that the feature contributes to *A. baumannii* non-GC1 class (represented by the
552 value -1) prediction (106). Considering the above mentioned, 100 unitig sequences with
553 the highest weight values were selected to be analyzed as putative biomarkers of *A.*
554 *baumannii* GC1 genomes.

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

561 number of *A. baumannii* non-GC1 strains predicted to be *A. baumannii* non-GC1, FP (false
562 positives) was the number of *A. baumannii* non-GC1 strains predicted to be *A. baumannii*
563 GC1, and FN (false negatives) was the number of *A. baumannii* GC1 strains predicted to be
564 *A. baumannii* non-GC1.

565 **Blastn searches.** Blastn searches (107) were done using the Dataset 1 and Dataset 2 as
566 subjects, and unitigs/k-mers that contributed most to *A. baumannii* GC1 genome
567 prediction according to the SVM/SCM models were used as queries. We identified
568 whether unitigs/k-mers matched known genes or intergenic regions and provide their
569 putative function when possible. AYE (AN: CU459141.1) and ACICU (CP000863.1) genomes
570 were used as a reference to annotate the genome location and gene product related to *A.*
571 *baumannii* GC1 and non-GC1 genomes, respectively. In case the unitig was not found in
572 the AYE genome, AB0057 (CP001182.2) genome was used instead. Also, we counted the
573 number of *A. baumannii* GC1 and non-GC1 genomes matched by each unitig. We
574 considered a cut-off E-value of E^{-10}, 100% of identity, and 100% of query cover. To analyze
575 the target of the primers designed below, Blastn searches (107) were done by using
576 primer's sequence as query and the Nucleotide collection (nr/nt) database of GenBank, or
577 the Dataset 1 and Dataset 2 as the subject. Finally, Blastn searches (107) were also done
578 by using the fragment of the U1 genomic biomarker of *A. baumannii* GC1 amplified with
579 those primers (excluding the sequence of the primers) as query, and the Dataset 1 and
580 Dataset 2 as the subject.

581 **Primer design.** The primers to amplify the U1 genomic biomarker of *A. baumannii* GC1
582 were designed by using the Oligo primer analysis software version 6.22 (108, 109).

583 **SUPPLEMENTARY MATERIAL**

584 Supplemental material is available online only.

585 **ACKNOWLEDGMENTS** The authors' work was supported by the grant PUE 2522 from
586 CONICET given to IMPaM.

587 We are grateful to the technical assistance of Dr. Gabriela Camicia and Nicolás Donis from
588 CONICET.

589 DC, VA, and PQ contributed to the conception and design of the study. VA performed the
590 script development. VA and PQ analyzed the results. DC and VA structured the work,
591 wrote, and coordinated the drafts of the manuscript and did the final edition. VA wrote
592 the results section. All authors contributed to the manuscript revision, read, and approved
593 the submitted version.

594 **REFERENCES**

595 1. Álvarez VE, Quiroga MP, Galán AV, Vilacoba E, Quiroga C, Ramírez MS, Centrón D.
596 2020. Crucial Role of the Accessory Genome in the Evolutionary Trajectory of
597 Acinetobacter baumannii Global Clone 1. Front Microbiol 11.

598 2. Douraghi, Masoumeh, Johanna, J, Aris, Parisa, Asadian, Mahla, Ghourchian,
599 Sedighe, Ghourchian, Hamidian M. 2020. Accumulation of Antibiotic Resistance
600 Genes in Carbapenem- Resistant Acinetobacter baumannii Isolates Belonging to.
601 mSphere 5:e00164-20.

602 3. Hamidian M, Nigro SJ. 2019. Emergence, molecular mechanisms and global spread

603 of carbapenem-resistant acinetobacter baumannii. Microb Genomics. Microbiology
604 Society.

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

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

611 6. Zarrilli R, Pournaras S, Giannouli M, Tsakris A. 2013. Global evolution of multidrug-
612 resistant Acinetobacter baumannii clonal lineages. Int J Antimicrob Agents 41:11–
613 19.

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

617 8. Provasi Cardoso J, Cayô R, Girardello R, Gales AC. 2016. Diversity of mechanisms
618 conferring resistance to β-lactams among OXA-23-producing Acinetobacter
619 baumannii clones. Diagn Microbiol Infect Dis 85:90–97.

620 9. Rodríguez C, Nastro M, A. Flores S, Rodriguez M, Spinozzi M, Bruni G, L. López A,
621 David V, S. Aiassa M, A. Marqués I, R. Navarro O, Paniccia L, Famiglietti A, Grupo
622 colaborativo Acinetobacter Argentina. 2018. Epidemiología molecular de aislados
623 de Acinetobacter baumannii resistentes a carbapenems en Argentina. Rev Argent

624 Microbiol https://doi.org/10.1016/j.ram.2017.12.004.

625 10. Hamed SM, Hussein AFA, Al-Agamy MH, Radwan HH, Zafer MM. 2022. Genetic
626 Configuration of Genomic Resistance Islands in Acinetobacter baumannii Clinical
627 Isolates From Egypt. Front Microbiol 13:2601.

628 11. Koong J, Johnson C, Rafei R, Hamze M, Myers GSA, Kenyon JJ, Lopatkin AJ, Hamidian
629 M. 2021. Phylogenomics of two ST1 antibiotic-susceptible non-clinical
630 Acinetobacter baumannii strains reveals multiple lineages and complex
631 evolutionary history in global clone 1. Microb Genomics 7.

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

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

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

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

643 16. Bartual SG, Seifert H, Hippler C, Luzon MAD, Wisplinghoff H, Rodríguez-Valera F.

644 2005. Development of a multilocus sequence typing scheme for characterization of
645 clinical isolates of Acinetobacter baumannii. J Clin Microbiol 43:4382–4390.

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

650 18. 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
651 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
652 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
653 and molecular characterization of Acinetobacter seifertii in Taiwan. J Antimicrob
654 Chemother https://doi.org/10.1093/jac/dkaa432.

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

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

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

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

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

671 24. Lupolova N, Lycett SJ, Gally DL. 2019. A guide to machine learning for bacterial host
672 attribution using genome sequence data. Microb Genomics 5.

673 25. Holt KE, Wertheim H, Zadoks RN, Baker S, Whitehouse CA, Dance D, Jenney A,
674 Connor TR, Hsu LY, Severin J, Brisse S, Cao H, Wilksch J, Gorrie C, Schultz MB,
675 Edwards DJ, Nguyen K Van, Nguyen TV, Dao TT, Mensink M, Minh V Le, Nhu NTK,
676 Schultsz C, Kuntaman K, Newton PN, Moore CE, Strugnell RA, Thomson NR. 2015.
677 Genomic analysis of diversity, population structure, virulence, and antimicrobial
678 resistance in Klebsiella pneumoniae , an urgent threat to public health . Proc Natl
679 Acad Sci 112:E3574–E3581.

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

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

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

692 29. Hernández M, Quijada NM, Rodríguez-Lázaro D, Eiros JM. 2020. Bioinformatics of
693 next generation sequencing in clinical microbiology diagnosis. Rev Argent Microbiol
694 52:150–161.

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

698 31. Lees JA, T. M, Tien B, Corander J. 2020. Improved Prediction of Bacterial Genotype-
699 Phenotype Associations Using Interpretable Pangenome-Spanning Regressions
700 11:1–22.

701 32. Khaledi A, Weimann A, Schniederjans M, Asgari E, Kuo T, Oliver A, Cabot G, Kola A,
702 Gastmeier P, Hogardt M, Jonas D, Mofrad MR, Bremges A, McHardy AC, Häussler S.
703 2020. Predicting antimicrobial resistance in Pseudomonas aeruginosa with machine
704 learning-enabled molecular diagnostics . EMBO Mol Med 12.

705 33. Hyun JC, Kavvas ES, Monk JM, Palsson BO. 2020. Machine learning with random
706 subspace ensembles identifies antimicrobial resistance determinants from pan-

707 genomes of three pathogens. PLoS Comput Biol 16.

708 34. Van Camp PJ, Haslam DB, Porollo A. 2020. Prediction of Antimicrobial Resistance in
709 Gram-Negative Bacteria From Whole-Genome Sequencing Data. Front Microbiol
710 11:1–13.

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

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

718 37. Saravanan R, Sujatha P. 2019. A State of Art Techniques on Machine Learning
719 Algorithms: A Perspective of Supervised Learning Approaches in Data Classification,
720 p. 945–949. *In* Proceedings of the 2nd International Conference on Intelligent
721 Computing and Control Systems, ICICCS 2018. Institute of Electrical and Electronics
722 Engineers Inc.

723 38. Macesic N, Polubriaginof F, Tatonetti NP. 2017. Machine learning: Novel
724 bioinformatics approaches for combating antimicrobial resistance. Curr Opin Infect
725 Dis 30:511–517.

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

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

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

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

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

741 44. Kouchaki S, Yang YY, Walker TM, Walker AS, Wilson DJ, Peto TEA, Crook DW, Clifton
742 DA, Hoosdally SJ, Gibertoni Cruz AL, Carter J, Grazian C, Kouchaki S, Walker TM,
743 Fowler PW, Clifton DA, Iqbal Z, Hunt M, Smith EG, Rathod P, Jarrett L, Matias D,
744 Cirillo DM, Borroni E, Battaglia S, Ghodousi A, Spitaleri A, Cabibbe A, Tahseen S,
745 Nilgiriwala K, Shah S, Rodrigues C, Kambli P, Surve U, Khot R, Niemann S, Kohl T,
746 Merker M, Hoffmann H, Molodtsov N, Plesnik S, Ismail N, Omar SV, Joseph L,
747 Marubini E, Thwaites G, Thuong TNT, Ngoc NH, Srinivasan V, Moore D, Coronel J,
748 Solano W, Gao GF, He G, Zhao Y, Ma A, Liu C, Zhu B, Laurenson I, Claxton P,

749 Wilkinson RJ, Lalvani A, Posey J, Gardy J, Werngren J, Paton N, Jou R, Wu MH, Lin
750 WH, Ferrazoli L, De Oliveira RS, Arandjelovic I, Comas I, Drobniewski F, Gao Q,
751 Sintchenko V, Supply P, Van Soolingen D. 2019. Application of machine learning
752 techniques to tuberculosis drug resistance analysis. Bioinformatics 35:2276–2282.

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

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

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

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

766 49. Uysal Ciloglu F, Saridag AM, Kilic IH, Tokmakci M, Kahraman M, Aydin O. 2020.
767 Identification of methicillin-resistant: Staphylococcus aureus bacteria using surface-
768 enhanced Raman spectroscopy and machine learning techniques. Analyst
769 145:7559–7570.

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

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

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

779 53. Xu H, Park S, Hwang TH. 2020. Computerized Classification of Prostate Cancer
780 Gleason Scores from Whole Slide Images. IEEE/ACM Trans Comput Biol Bioinforma
781 17:1871–1882.

782 54. 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.
783 Evaluation of Machine Learning Models for Predicting Antimicrobial Resistance of
784 Actinobacillus pleuropneumoniae From Whole Genome Sequences. Front Microbiol
785 11.

786 55. Li Y, Kong Y, Zhang M, Yan A, Liu Z. 2016. Using Support Vector Machine (SVM) for
787 Classification of Selectivity of H1N1 Neuraminidase Inhibitors. Mol Inform 35:116–
788 124.

789 56. Hicks AL, Wheeler N, Sánchez-Busó L, Rakeman JL, Harris SR, Grad YH. 2019.
790 Evaluation of parameters affecting performance and reliability of machine learning-

791 based antibiotic susceptibility testing from whole genome sequencing data. PLoS
792 Comput Biol 15:1–21.

793 57. Bron EE, Smits M, Niessen WJ, Klein S. 2015. Feature Selection Based on the SVM
794 Weight Vector for Classification of Dementia. IEEE J Biomed Heal Informatics
795 19:1617–1626.

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

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

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

803 61. Hamidian M, Hall RM. 2018. The AbaR antibiotic resistance islands found in
804 Acinetobacter baumannii global clone 1 – Structure, origin and evolution. Drug
805 Resist Updat 41:26–39.

806 62. Hamidian M, Hall RM. 2021. Dissemination of novel Tn7 family transposons carrying
807 genes for synthesis and uptake of fimsbactin siderophores among Acinetobacter
808 baumannii isolates. Microb Genomics 7.

809 63. Jaillard M, Lima L, Tournoud M, Mahé P, van Belkum A, Lacroix V, Jacob L. 2018. A
810 fast and agnostic method for bacterial genome-wide association studies: Bridging

811 the gap between k-mers and genetic events. PLoS Genet 14:1–28.

812 64. Wuebbens MM, Liu MTW, Rajagopalan K V., Schindelin H. 2000. Insight into
813 molybdenum cofactor deficiency provided by the crystal structure of the
814 molybdenum cofactor biosynthesis protein MoaC. Structure 8:709–718.

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

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

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

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

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

830 70. Arivett BA, Fiester SE, Ream DC, Centrón D, Ramírez MS, Tolmasky ME, Actis LA.

831 2015. Draft Genome of the Multidrug-Resistant Acinetobacter baumannii Strain
832 A155 Clinical Isolate. Genome Announc 3:e00212-15.

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

836 72. Mathers A, Peirano G, Mathers AJ, Peirano G, Pitout DD. 2015. The Role of Epidemic
837 Resistance Plasmids and International High-Risk Clones in the Spread of Multidrug-
838 Resistant ... The Role of Epidemic Resistance Plasmids and International High- Risk
839 Clones in the Spread of Multidrug-Resistant Enterobacteriaceae
840 https://doi.org/10.1128/CMR.00116-14.

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

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

849 75. Oteo J, Pérez-Vázquez M, Bautista V, Ortega A, Zamarrón P, Saez D, Lara N, Ramiro
850 R, Aracil B, Campos J. 2016. The spread of KPC-producing Enterobacteriaceae in
851 Spain: WGS analysis of the emerging high-risk clones of Klebsiella pneumoniae

852 ST11/KPC-2, ST101/KPC-2 and ST512/KPC-3. J Antimicrob Chemother 71:3392–
853 3399.

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

858 77. Heiden SE, Hübner NO, Bohnert JA, Heidecke CD, Kramer A, Balau V, Gierer W,
859 Schaefer S, Eckmanns T, Gatermann S, Eger E, Guenther S, Becker K, Schaufler K.
860 2020. A Klebsiella pneumoniae ST307 outbreak clone from Germany demonstrates
861 features of extensive drug resistance, hypermucoviscosity, and enhanced iron
862 acquisition. Genome Med 12.

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

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

871 80. Zhong Q, Kobe B, Kappler U. 2020. Molybdenum Enzymes and How They Support
872 Virulence in Pathogenic Bacteria. Front Microbiol 11:1–23.

873 81. Contreras I, Toro CS, Troncoso G, Mora GC. 1997. Salmonella typhi mutants
874 defective in anaerobic respiration are impaired in their ability to replicate within
875 epithelial cells. Microbiology 143:2665–2672.

876 82. Rosas-Magallanes V, Stadthagen-Gomez G, Rauzier J, Barreiro LB, Tailleux L, Boudou
877 F, Griffin R, Nigou J, Jackson M, Gicquel B, Neyrolles O. 2007. Signature-tagged
878 transposon mutagenesis identifies novel Mycobacterium tuberculosis genes
879 involved in the parasitism of human macrophages. Infect Immun 75:504–507.

880 83. Brodin P, Poquet Y, Levillain F, Peguillet I, Larrouy-Maumus G, Gilleron M, Ewann F,
881 Christophe T, Fenistein D, Jang J, Jang M-S, Park S-J, Rauzier J, Carralot J-P,
882 Shrimpton R, Genovesio A, Gonzalo-Asensio JA, Puzo G, Martin C, Brosch R, Stewart
883 GR, Gicquel B, Neyrolles O. 2010. High Content Phenotypic Cell-Based Visual Screen
884 Identifies Mycobacterium tuberculosis Acyltrehalose-Containing Glycolipids
885 Involved in Phagosome Remodeling. PLoS Pathog 6:e1001100.

886 84. Dutta NK, Mehra S, Didier PJ, Roy CJ, Doyle LA, Alvarez X, Ratterree M, Be NA,
887 Lamichhane G, Jain SK, Lacey MR, Lackner AA, Kaushal D. 2010. Genetic
888 Requirements for the Survival of Tubercle Bacilli in Primates. J Infect Dis 201:1743–
889 1752.

890 85. Levillain F, Poquet Y, Mallet L, Mazères S, Marceau M, Brosch R, Bange F-C, Supply
891 P, Magalon A, Neyrolles O. 2017. Horizontal acquisition of a hypoxia-responsive
892 molybdenum cofactor biosynthesis pathway contributed to Mycobacterium
893 tuberculosis pathoadaptation. PLOS Pathog 13:e1006752.

894 86. Macesic N, Bear Don't Walk OJ, Pe'er I, Tatonetti NP, Peleg AY, Uhlemann A-C.
895 2020. Predicting Phenotypic Polymyxin Resistance in Klebsiella pneumoniae
896 through Machine Learning Analysis of Genomic Data. mSystems 5.

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

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

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

906 90. Lee YT, Kuo SC, Chiang MC, Yang SP, Chen CP, Chen TL, Fung CP. 2012. Emergence
907 of carbapenem-resistant non-baumannii species of Acinetobacter harboring a
908 blaOXA-51-like gene that is intrinsic to A. baumannii. Antimicrob Agents Chemother
909 56:1124–1127.

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

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

915 93. Jaillard M, Tournoud M, Lima L, Lacroix V, Veyrieras JB, Jacob L. 2017. Representing
916 Genetic Determinants in Bacterial GWAS with Compacted De Bruijn Graphs. bioRxiv
917 1–24.

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

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

922 96. Schmidt J, Marques MRG, Botti S, Marques MAL. 2019. Recent advances and
923 applications of machine learning in solid-state materials science. npj Comput Mater
924 5.

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

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

930 99. Vapnik VN. 2000. The Nature of Statistical Learning TheoryThe Nature of Statistical
931 Learning Theory. Springer New York.

932 100. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M,
933 Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher
934 M, Perrot M, Duchesnay É. 2011. Scikit-learn: Machine learning in Python. J Mach

935 Learn Res 12:2825–2830.

936 101. Drouin A, Giguère S, Déraspe M, Marchand M, Tyers M, Loo VG, Bourgault A-M,
937 Laviolette F, Corbeil J. 2016. Aditional File of Predictive computational phenotyping
938 and biomarker discovery using reference-free genome comparisons. BMC Genomics
939 17:754.

940 102. Boisvert S, Raymond F, Godzaridis É, Laviolette F, Corbeil J. 2012. Ray Meta:
941 Scalable de novo metagenome assembly and profiling. Genome Biol 13.

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

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

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

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

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

952 108. Rychlik W, Rhoads RE. 1989. A computer program for choosing optimal
953 oligonudeotides for filter hybridization, sequencing and in vitro amplification of
954 DNA. Nucleic Acids Res 17:8543–8551.

955 109. Rychlik W, Spencer WJ, Rhoads RE. 1990. Optimization of the annealing
956 temperature for DNA amplification in vitro; Nucleic Acids Res 18:6409–6412.

957 **TABLES**

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

Unitig ID	Sequence	Length	Match Location	Gene Name	Gene Product
U1	TTATTCATAGCCTCCTGGATGCGGTAAAT CACGGCAAACATCAAAAATATGGACTAA TGCGGGTAAAATAGCGGCCATTGATTCG CTTGCTCCACCACGGCTTCCCGGTAACGT CACTACTAACGAACGGTCAATAAACCCT GCTACTCCGCGTGACATAGCAGCATATG GCGTGCGTTTTTGTCCAAATGAGCGAGC AGCTTCCATTAAGCCATCTAATTTACGTT CTAAAAGAGGCTCTAGCGTATCAACTGT AATGTCACGCTTACCAATTCCTGTCCCAC CCACGGTCATAATGCAAGCATATGATTT GCTCAACTCAAGCACTAAGTTTTTCAAAG TATCCGCTTCATCTGGCAAAATTTGG	367	ABAYE1552:558..924	*moaCB*	Bifunctional protein [Includes: molybdenum cofactor biosynthesis protein C; molybdenum cofactor biosynthesis protein B]

U2	ACCGGCAAAAATTGCAACCCCAATAAAACCGTTCACCCAACCGTTCATAAACTTTTGCATA	61	ABAYE1557:1..61	-	Putative permease (drug/metabolite transporter)
U3	ACGTAAGCCATATTGAGAGGCTGCAAGTTCGACACGAGCAACATCACCCAATCGGGTAACC	61	ABAYE1173:777..837	-	Putative Multidrug/solvent efflux pump membrane transporter
U4	AAACACTGGGAAATATACTTGAAAAATATCAACAACCCCGAACGTATGCGGGTTGCTCAGT	61	ABAYE3468:844..904	-	Conserved hypothetical protein; putative exported protein
U5	GTATATCTGGCACAAAGCAAAAGGCTTTCATGCCTTACAACATCAACAAACCCGTAAA	58	ABAYE1549:957..1014	*moaA*	Molybdopterin biosynthesis, protein A
U6	CATAACCCCTAAAGTACGATGGCTTGATACGGCATCCCACATTGTTAAGCTA	52	ABAYE1636:975..1026	*cydB*	Cytochrome d terminal oxidase polypeptide subunit II
U7	GAACTATGGCAATACCGATCAGTAAAGATAGAAGTACCTGCCAG	44	ABAYE2207:192..235	-	Putative permease
U8	TCCTTTAGAAAATTTGGTGGGACAAGAAGGCGAAGGTTATA	41	ABAYE1412:651..691	-	Putative acyl-CoA dehydrogenase protein (*acdB*-like)
U9	TCATGTCGAGCTAAGCCAGTACCAGCAAAAGCTGAAG	37	ABAYE1194:1095..1131	-	Putative two-component system sensor histidine kinase
U10	CCCCCCGAACTACATCAATATGATCATATAAGGCC	35	ABAYE1494:441..475	-	Putative outer membrane porin, receptor for Fe(III)-coprogen, Fe(III)-ferrioxamine B and Fe(III)-rhodotrulic acid uptake (FhuE)

Table 2. Analysis of putative *A. baumannii* 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 *A. baumannii* GC1 genome prediction according to the values of the features weight vector, the number of *A. baumannii* GC1 and non-GC1 genomes typified by MLST that matched the unitig within the Dataset 1 and 2 using blastn and the p-values

976 of Fisher's exact test using a significance level of 0.05. Fisher's exact test was calculated in
977 R considering the nominal variables "Dataset Source" (Dataset 1 or Dataset 2) and
978 "Matched" (Yes or No). The total number of *A. baumannii* GC1 and non-GC1 genomes that
979 matched / not matched the unitigs in each dataset was used for calculation. Total data of
980 the 100 unitigs predicted by SVM is detailed in Supplementary Table S4.

981

	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 the Dataset 1 not differs from proportion of GC1 matches in the Dataset 2)**	**P-value (proportion of non-GC1 genome matches in the Dataset 1 not differs from proportion of non-GC1 matches in the 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

982

983 **Table 3. Rules obtained by the SCM models.** The table details the rule ID, the rule output
984 from the SCM model, k-mer length and the location where the rule sequence matched in
985 *A. baumannii* AYE genome (AN:CU459141.1) or *A. baumannii* ACICU genome (AN:
986 CP000863.1), and the gene name corresponding to the genome region matched and the
987 gene product of the 10 larger k-mer sequences targeted by the SCM rules. *A. baumannii*

988 AYE and ACICU genomes were used as *A. baumannii* GC1 and non-GC1 references,
989 respectively, to locate the k-mer sequences. The prefix "REGION" was used when the
990 match occurred either in an intergenic region or in a combination of intergenic region and
991 a gene. The complete list of the 49 rules obtained by the SCM models is detailed in
992 Supplementary Table S5.

993

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 *A. baumannii* 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. Fisher's exact test was calculated in R considering the nominal variables "Dataset Source" (Dataset 1 or Dataset 2) and "Matched" (Yes or No). The total number of *A. baumannii* GC1 and non-GC1 genomes that matched / not matched the rules in each dataset was used for calculation. Total data of the 49 SCM rules is detailed in Supplementary Table S6.

	Dataset 1		Dataset 2			
Rule ID	**Number of *A. baumannii* GC1 genomes matching the rule**	**Number of *A. baumannii* non-GC1 genomes matching the rule**	**Number of *A. baumannii* GC1 genomes matching the rule**	**Number of *A. baumannii* non-GC1 genomes matching the rule**	**P-value (proportion of *A. baumannii* GC1 genome matches in the Dataset 1 not differs from proportion of GC1 matches in the Dataset 2)**	**P-value (proportion of *A. baumannii* non-GC1 genome matches in the Dataset 1 not differs from proportion of non-GC1 matches in the 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 the Dataset 1 using the best performing SVM model. Correlation between MLST typing and model prediction was calculated using Fisher's exact test in R level using a significance level of 0.05. The nominal variables "MLST Typing" and "Prediction" were considered during Fisher's exact test calculation. The variable "MLST Typing" represented the genomes typed as *A. baumannii* GC1 (positive label) or non-GC1 (negative label) by MLST technique (true class). On the other hand, the variable "Prediction" represented the genomes predicted to be *A. baumannii* GC1 or non-GC1 by the SVM model (predicted class). We used the number of True Positives (TP), False Positives (FP), False Negatives (FN) and True Negatives (TN) in 2x2 contingency table. Null hypothesis used to evaluate the correlation between MLST typing and model prediction was: "True class (MLST Typing) and Predicted class are

1019 independent, knowing the value of one variable does not help to predict the value of the

1020 other variable".

									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	$P < 2.2e^{-16}$

1021 ACC: Accuracy; SENS: Sensitivity; SPE: Specificity; PRE: Precision; F1: F1-Score; TP: True Positives, TN: True Negatives; FP: False Positives;

1022 FN: False Negatives.

1023 **Table 6. Prediction metrics on test dataset partitions from the Dataset 1 using the best**

1024 **performing SCM models.** Correlation between MLST typing and model prediction was

1025 calculated using Fisher's exact test in R using a significance level of 0.05. The nominal

1026 variables "MLST Typing" and "Prediction" were considered during Fisher's exact test

1027 calculation. The variable "MLST Typing" represented the genomes typed as *A. baumannii*

1028 GC1 (positive label) or *A. baumannii* non-GC1 (negative label) by MLST technique (true

1029 class). On the other hand, the variable "Prediction" represented the genomes predicted to

1030 be *A. baumannii* GC1 or non-GC1 by the SCM model (predicted class). We used the

1031 number of True Positives (TP), False Positives (FP), False Negatives (FN) and True

1032 Negatives (TN) in 2x2 contingency table. Null hypothesis used to evaluate the correlation

1033 between MLST typing and model prediction was: "True class (MLST Typing) and Predicted

1034 class are independent, knowing the value of one variable does not help to predict the

1035 value of the other variable". Total metrics of the 49 rules obtained by the SCM models is

1036 detailed in Supplementary Table S7.

1037

Rule ID	SENS	SPE	PRE	ACC	F1	TP	TN	FP	FN	Number of isolates typified by MLST: GC1	Number of isolates typified by MLST: 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	$P < 2.2e^{-16}$

1038 ACC: Accuracy; SENS: Sensitivity; SPE: Specificity; PRE: Precision; F1: F1-Score; TP: True Positives, TN: True Negatives; FP: False Positives;

1039 FN: False Negatives.

1040 **FIGURES LEGENDS**

1041 **Figure 1. Diagram of the workflow for *A. baumannii* GC1 biomarker discovery using**
1042 **machine learning.** The diagram indicates the steps we followed during the present work
1043 related to genomes collections used, dataset preparation, ML analysis and the design of a
1044 method for *A. baumannii* GC1 strain identification.

1045 **Figure 2. Percentage of genomes in the Dataset 1 and Dataset 2 matching unitig**
1046 **sequences U1 to U12. (A)** *A. baumannii* GC1 genomes, **(B)** *A. baumannii* non-GC1
1047 genomes.

1048 **Figure 3. Percentage of genomes in the Dataset 1 and Dataset 2 matching the SCM rules.**
1049 **(A)** *A. baumannii* GC1 genomes, **(B)** *A. baumannii* non-GC1 genomes. We observed that
1050 within the 43 rules which targeted 100% of *A. baumannii* GC1 genomes from the Dataset
1051 1, two set of rules targeted only 1 non-GC1 from the same dataset. Rules R125, R123,
1052 R121, R119, R117, R115, R113, R111, R109, R107, R105, R103, R101, R99, R97, R95, R89,
1053 R83, R81, R79, R77, R73, R69, R63, R39 and, R35 targeted genome with AN:
1054 GCF_000248195.1 (ST 69) and R49 targeted genome with AN: GCF_000453745.1 (ST 2).
1055 Regarding the Dataset 2, rules R93, R91, R87, R85, R75, R61, 211 R59, R57, R55, R53, R49
1056 and R31 targeted 100% of *A. baumannii* GC1 genomes but also targeted several *A.*
1057 *baumannii* non-GC1 genomes from the Dataset 2 (66/4487, 28/4487, 28/4487, 28/4487,
1058 213 28/4487, 26/4487, 26/4487, 14/4487, 27/4487, 26/4487, 13/4487, 42/4487 genomes,
1059 respectively).

1060 **Figure 4. Mean Accuracy of the SVM and SCM models.** Bars with blue color show the
1061 mean accuracy for the models on the training dataset. Bars with green color are the
1062 accuracy of the models on the test dataset. The models were run with the best
1063 hyperparameters selected from the 5-fold cross-validation. The error bars are standard
1064 deviations. **(A)** SVM model, **(B)** SCM models.

1065 **SUPPLEMENTAL FIGURES LEGENDS.**

1066 **Figure S1. *A. baumannii* GC1 biomarker identification by PCR.** PCR amplification of
1067 several *A. baumannii* strains is shown. Samples order at the upper wells correspond to
1068 Ladder, A144 ST1, A155 ST1, HAX19AbA ST1, HAX26Aba ST2, HAX28Aba ST2, A118 ST404,

1069 A110 ST1, A185 ST1, A325 ST119, A376 ST119, A103 ST119, ladder. Samples order at the
1070 lower wells correspond to Ladder, HAX22Aba ST1, HAX25Aba ST1, HAX27Aba ST1,
1071 Ab33405 ST79, Ab66 ST79, 186 ST79, F33943 ST79, A171 ST79, A177 ST79, A182 ST79,
1072 A384 ST79, negative control, ladder. Agarose gel 1.5%, electrophoresis in TAE buffer at
1073 80V during 45 min. 1 Kbp DNA ladder (#K0178, Inbio Highway, Argentina).

1074 **Figure S2. Alignment of the DNA sequence of U1 identified in *A. baumannii* strains**
1075 **tested by PCR.** The primers pair used in the PCR are shown, which targets coordinates 559
1076 to 914 from ABAYE1552 locus tag representing positions 2 to 357 of the 367 bp length U1
1077 sequence identified in *A. baumannii* GC1 amplified by PCR. Sequence type (ST) according
1078 to Pasteur's MLST schemes and Global Clones (GCs) are indicated in brackets. *The ST
1079 does not belong to a GC.

1080 **SUPPLEMENTAL TABLES LEGENDS.**

1081 **Table S1. *A. baumannii* genomes used as input dataset for SVM algorithm within the**
1082 **Dataset 1 genome collection.** GenBank accession number and classification of the 500
1083 genomes in ML analysis. Genomes were classified as *A. baumannii* GC1 and non-GC1
1084 according MLST Pasteur scheme.

1085 **Table S2. *A. baumannii* genomes used as input dataset for SVM algorithm within the**
1086 **Dataset 2 genome collection**. Genomes were classified as *A. baumannii* GC1 and non-GC1
1087 according MLST Pasteur scheme. ST Pasteur column with "-" value indicates that the
1088 genome does not yet have a ST assigned in the PubMLST database.

1089 **Table S3. Putative GC1 biomarkers obtained by SVM.** The table details the unitig ID,
1090 unitig sequence, unitig length, the location where the unitig sequence matched in *A.*
1091 *baumannii* AYE genome (AN:CU459141.1) or *A. baumannii* AB0057 genome (AN:
1092 CP001182.2) and the gene name corresponding to the genome region matched and the
1093 gene product. *A. baumannii* AYE strain genome was used as *A. baumannii* GC1 reference
1094 to locate the unitigs sequences. However, when the unitig was not found in the AYE
1095 genome, *A. baumannii* AB0057 strain genome was used instead. The prefix "REGION" was
1096 used when the match occurred either in an intergenic region or in a combination of
1097 intergenic region and a gene.

1098 **Table S4. Analysis of putative GC1 biomarkers obtained by SVM matches within the**
1099 **Dataset 1 and 2.** The table details the unitig ID of the first 100 unitigs that contributed
1100 most to *A. baumannii* GC1 genome prediction according to the values of the features
1101 weight vector, the number of *A. baumannii* GC1 and non-GC1 genomes typified by MLST
1102 that matched the unitig within the Dataset 1 and 2 using blastn and the p-values of
1103 Fisher's exact test using a significance level of 0.05. Fisher's exact test was calculated in R
1104 considering the nominal variables "Dataset Source" (Dataset 1 or Dataset 2) and
1105 "Matched" (Yes or No). The total number of *A. baumannii* GC1 and non-GC1 genomes that
1106 matched / not matched the unitigs in each dataset was used for calculation.

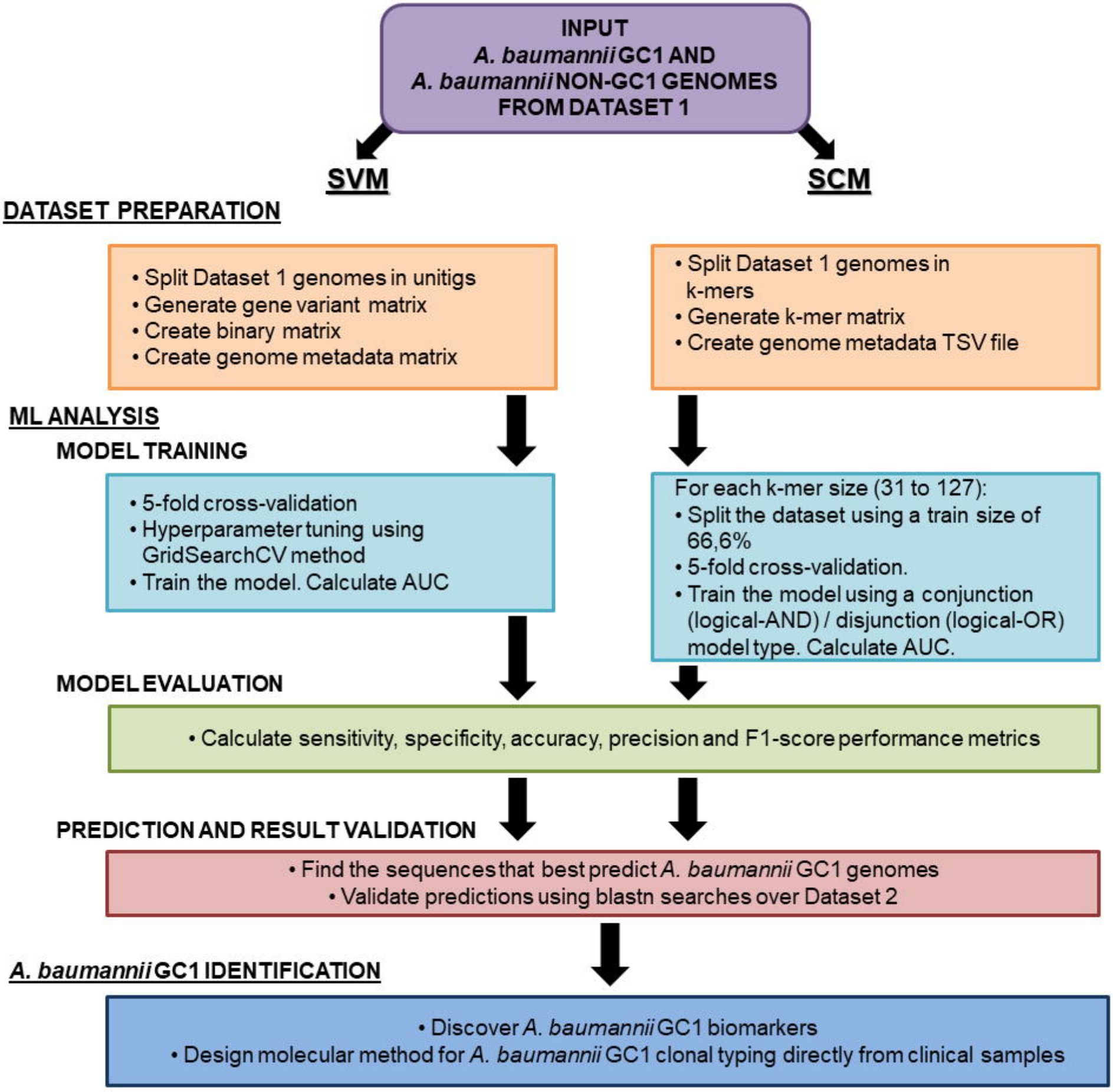
1107 **Table S5. Rules obtained by the SCM models.** The table details the rule ID, the rule output
1108 from the SCM models, k-mer length and the location where the rule sequence matched in
1109 *A. baumannii* AYE genome (AN:CU459141.1) or *A. baumannii* ACICU genome (AN:
1110 CP000863.1) and the gene name corresponding to the genome region matched and the

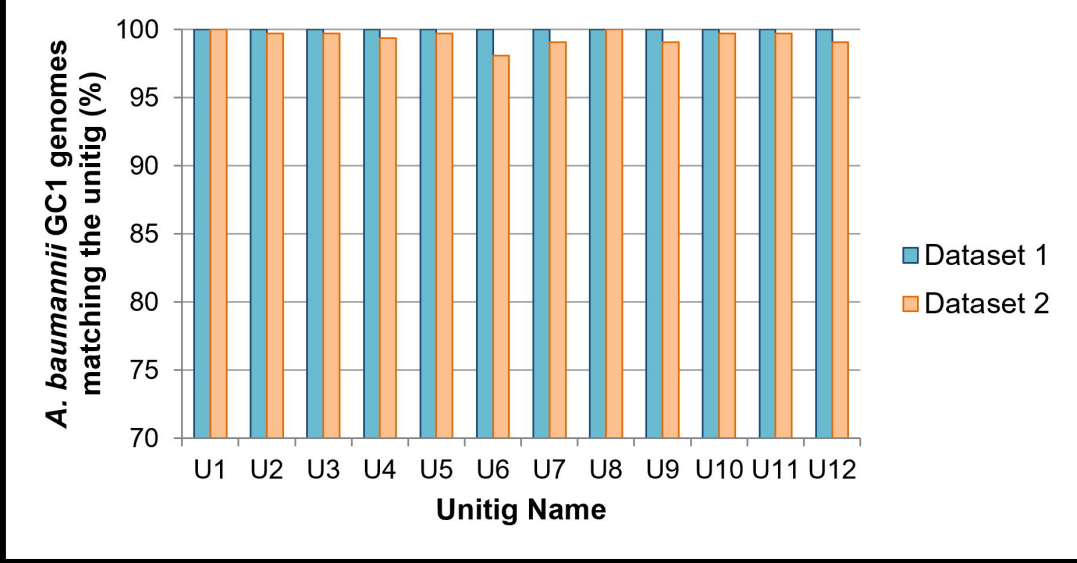
1111 gene product. *A. baumannii* AYE and ACICU genomes were used as *A. baumannii* GC1 and
1112 non-GC1 references, respectively, to locate the k-mer sequences. The prefix "REGION" was
1113 used when the match occurred either in an intergenic region or in a combination of
1114 intergenic region and a gene.

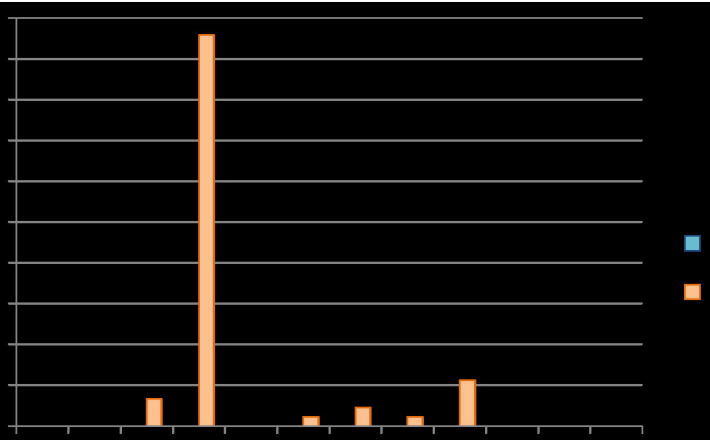
1115 **Table S6. Analysis of the SCM rules matches within the Dataset 1 and 2.** The table details
1116 the rule ID of the 100 larger k-mer sequences targeted by the SCM rules, the number of *A.*
1117 *baumannii* GC1 and non-GC1 genomes typified by MLST that matched the rule within the
1118 Datasets 1 and 2 using blastn and the p-values of Fisher's exact test using a significance
1119 level of 0.05. Fisher's exact test was calculated in R considering the nominal variables
1120 "Dataset Source" (Dataset 1 or Dataset 2) and "Matched" (Yes or No). The total number of
1121 *A. baumannii* GC1 and non-GC1 genomes that matched / not matched the rules in each
1122 dataset was used for calculation.

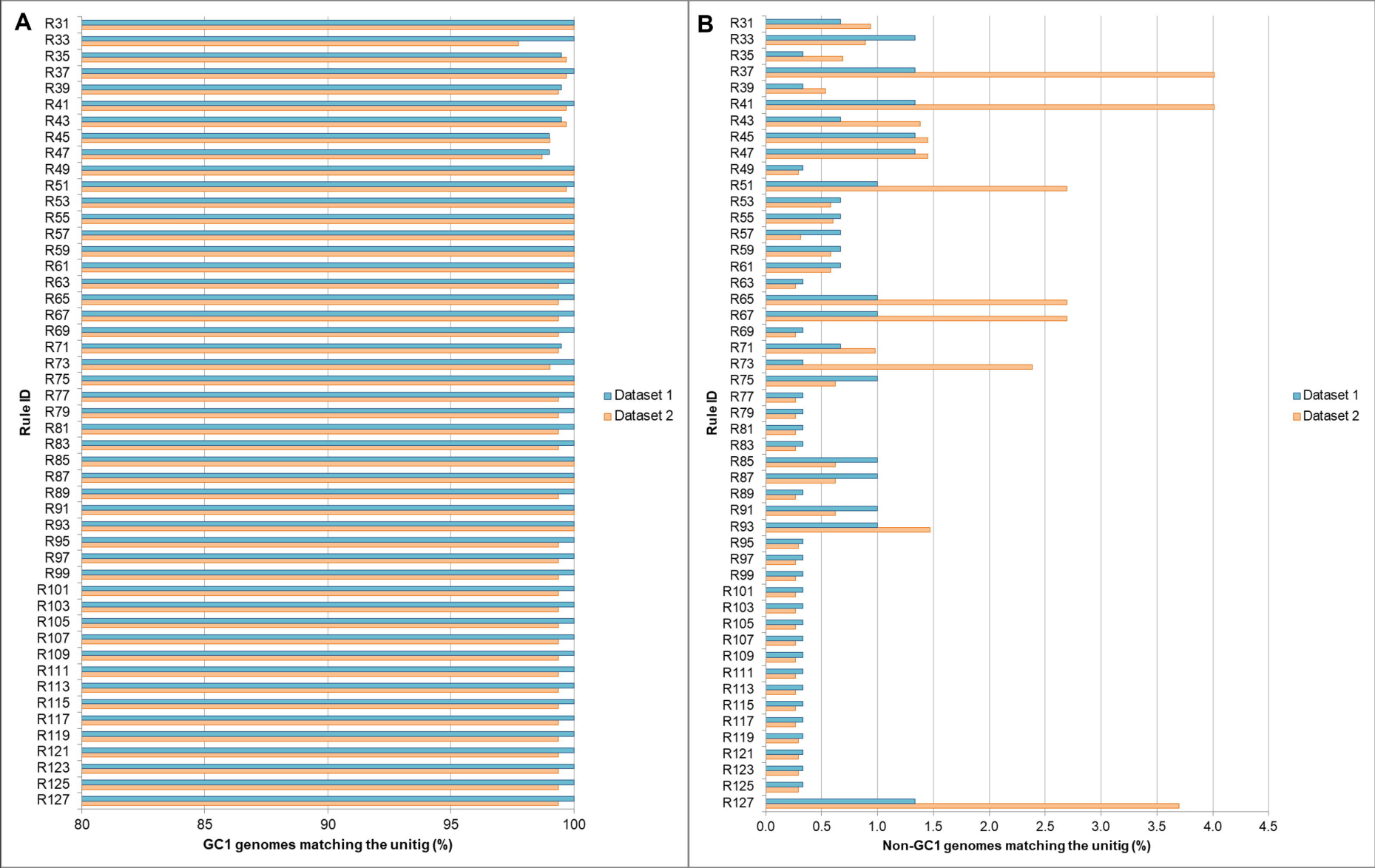
1123 **Table S7. Prediction metrics on test dataset partitions from the Dataset 1 using the best**
1124 **performing SCM models.** Correlation between MLST typing and model prediction was
1125 calculated using Fisher's exact test in R using a significance level of 0.05. The nominal
1126 variables "MLST Typing" and "Prediction" were considered during Fisher's exact test
1127 calculation. The variable "MLST Typing" represented the genomes typed as *A. baumannii*
1128 GC1 (positive label) or *A. baumannii* non-GC1 (negative label) by MLST technique (true
1129 class). On the other hand, the variable "Prediction" represented the genomes predicted to
1130 be *A. baumannii* GC1 or non-GC1 by the SCM model (predicted class). We used the
1131 number of True Positives (TP), False Positives (FP), False Negatives (FN) and True
1132 Negatives (TN) in 2x2 contingency table. Null hypothesis used to evaluate the correlation

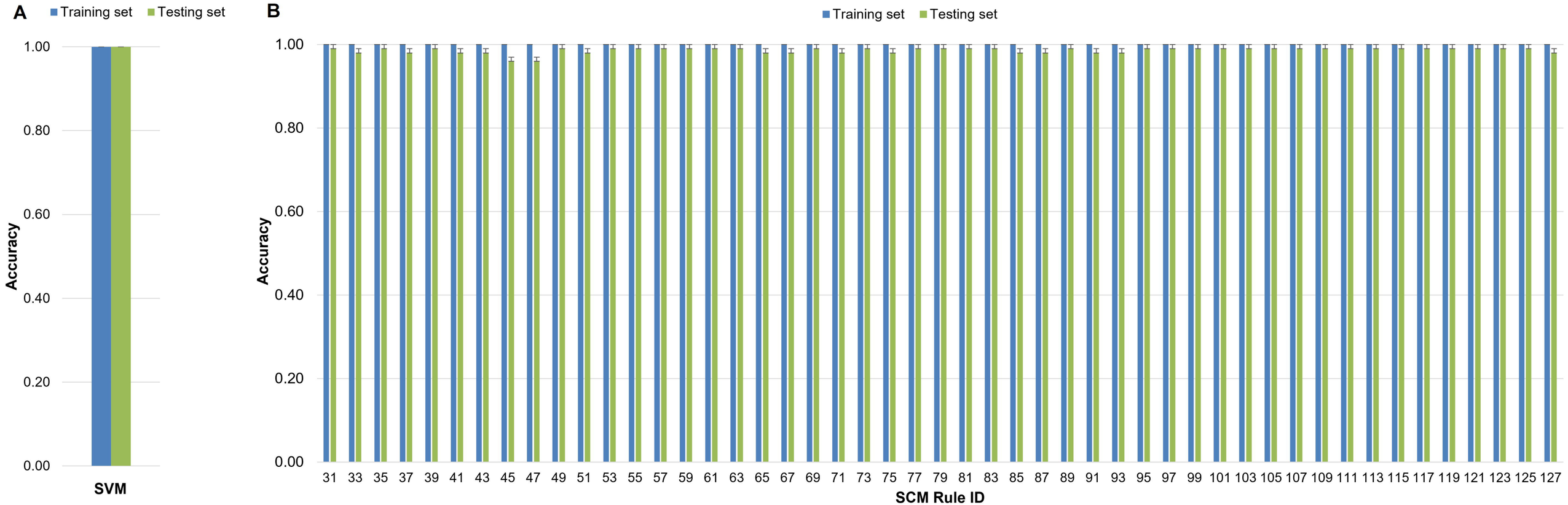
1133 between MLST typing and model prediction was: “True class (MLST Typing) and Predicted
1134 class are independent, knowing the value of one variable does not help to predict the
1135 value of the other variable”.

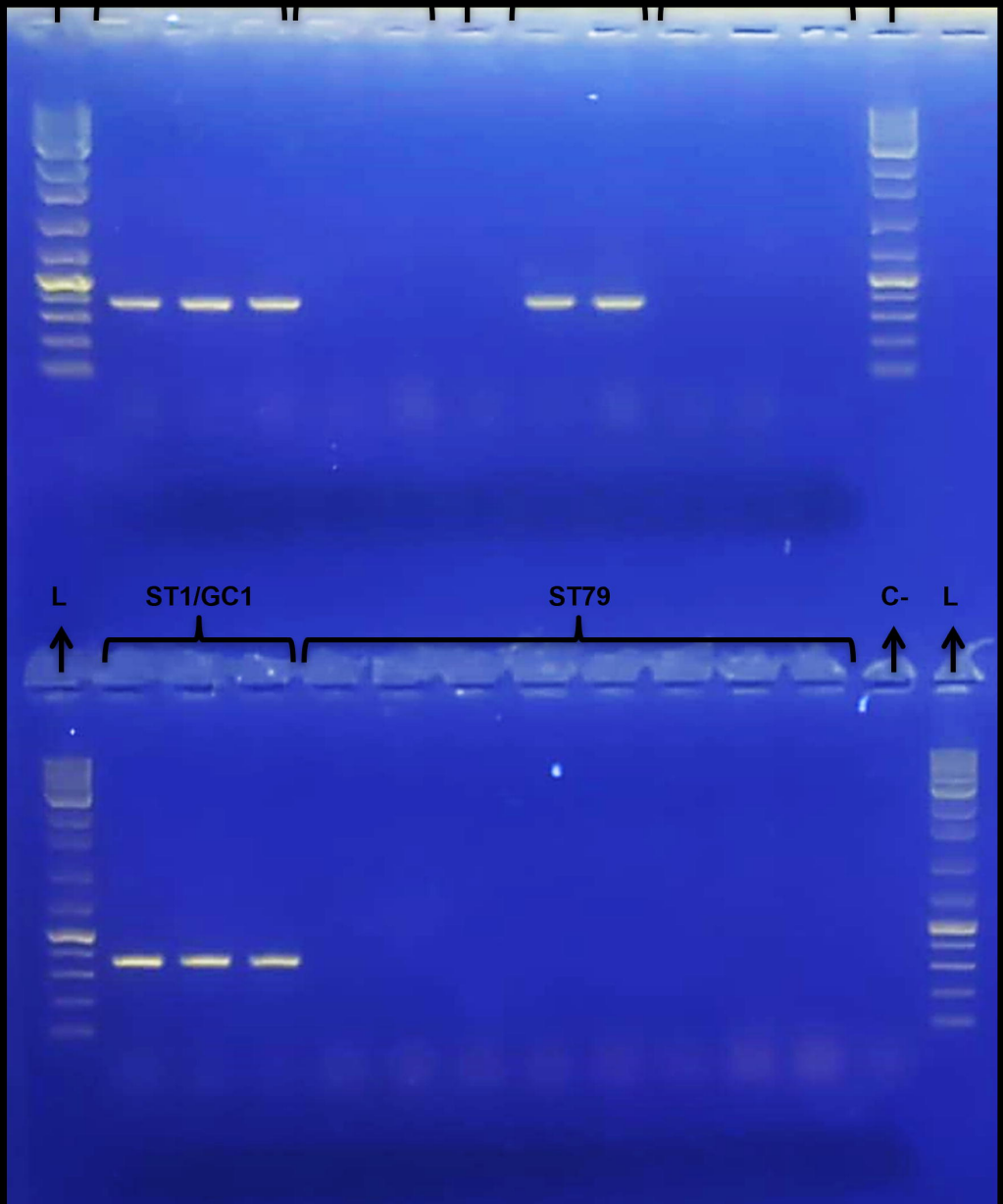

INPUT
A. baumannii GC1 AND
A. baumannii NON-GC1 GENOMES
FROM DATASET 1
SVM
SCM
DATASET PREPARATION
• Split Dataset 1 genomes in unitigs
• Generate gene variant matrix
• Create binary matrix
• Create genome metadata matrix
• Split Dataset 1 genomes in k-mers
• Generate k-mer matrix
• Create genome metadata TSV file
ML ANALYSIS
MODEL TRAINING
• 5-fold cross-validation
• Hyperparameter tuning using GridSearchCV method
• Train the model. Calculate AUC
For each k-mer size (31 to 127):
• Split the dataset using a train size of 66,6%
• 5-fold cross-validation.
• Train the model using a conjunction (logical-AND) / disjunction (logical-OR) model type. Calculate AUC.
MODEL EVALUATION
• Calculate sensitivity, specificity, accuracy, precision and F1-score performance metrics
PREDICTION AND RESULT VALIDATION
• Find the sequences that best predict A. baumannii GC1 genomes
• Validate predictions using blastn searches over Dataset 2
A. baumannii GC1 IDENTIFICATION
• Discover A. baumannii GC1 biomarkers
• Design molecular method for A. baumannii GC1 clonal typing directly from clinical samples

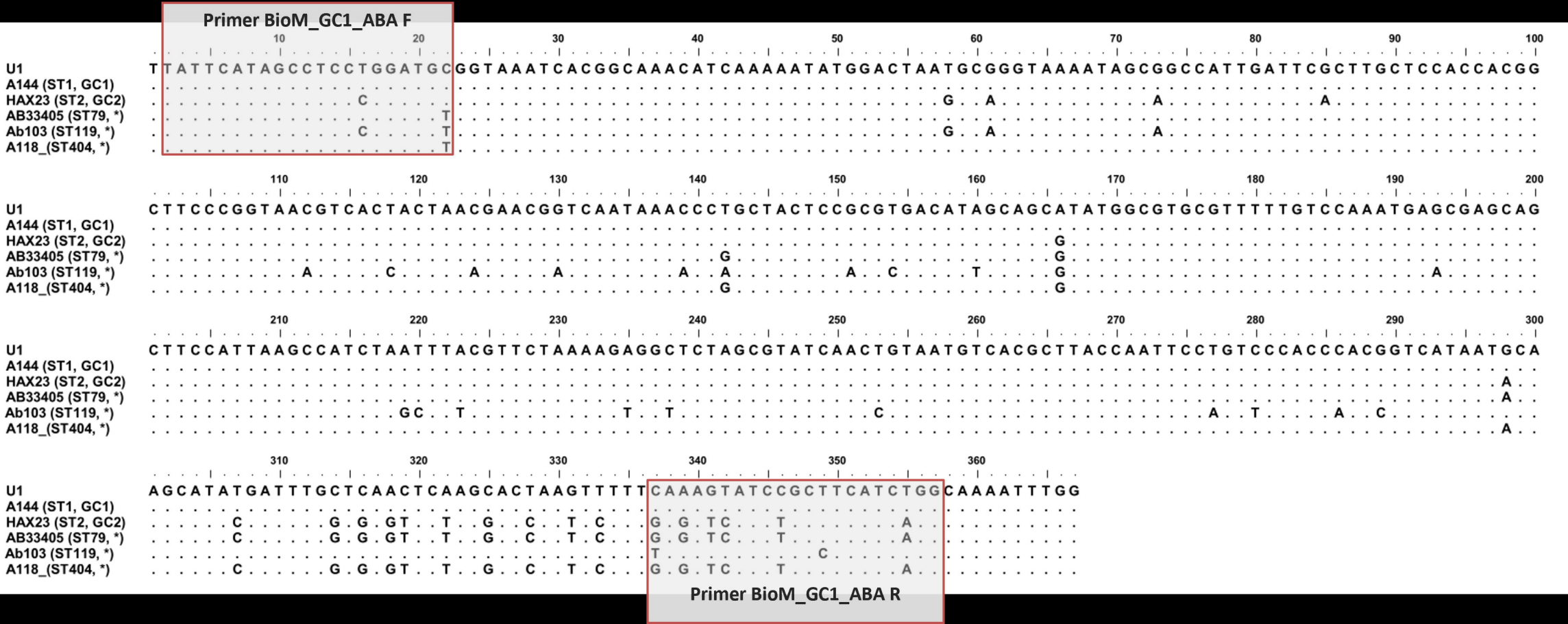

A. baumannii GC1 genomes matching the unitig (%)
100
95
90
85
80
75
70
U1
U2
U3
U4
U5
U6
U7
U8
U9
U10
U11
U12
Unitig Name
Dataset 1
Dataset 2




A
Rule ID
R31
R33
R35
R37
R39
R41
R43
R45
R47
R49
R51
R53
R55
R57
R59
R61
R63
R65
R67
R69
R71
R73
R75
R77
R79
R81
R83
R85
R87
R89
R91
R93
R95
R97
R99
R101
R103
R105
R107
R109
R111
R113
R115
R117
R119
R121
R123
R125
R127
80
85
90
95
100
GC1 genomes matching the unitig (%)
Dataset 1
Dataset 2
B
Rule ID
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
Non-GC1 genomes matching the unitig (%)
Dataset 1
Dataset 2


A
Training set
Testing set
1.00
0.80
0.60
0.40
0.20
0.00
Accuracy
SVM
B
Training set
Testing set
1.00
0.80
0.60
0.40
0.20
0.00
Accuracy
31 33 35 37 39 41 43 45 47 49 51 53 55 57 59 61 63 65 67 69 71 73 75 77 79 81 83 85 87 89 91 93 95 97 99 101 103 105 107 109 111 113 115 117 119 121 123 125 127
SCM Rule ID


L
ST1/GC1
ST79
C-
L


Primer BioM_GC1_ABA F
U1
A144 (ST1, GC1)
HAX23 (ST2, GC2)
AB33405 (ST79, *)
Ab103 (ST119, *)
A118_(ST404, *)
Primer BioM_GC1_ABA R