A first perturbome of *Pseudomonas aeruginosa*: Identification of core genes related to multiple perturbations by a machine learning approach

Jose Arturo Molina Mora (***corresponding author***)

Research Center in Tropical Diseases (CIET), University of Costa Rica, Costa Rica

Email: jose.molinamora@ucr.ac.cr

Pablo Montero-Manso

Department of Mathematics, University of A Coruña, Spain

Email: pmonteromanso@udc.es

Raquel García Batán

Research Center in Tropical Diseases (CIET), University of Costa Rica, Costa Rica

Email: raquel.garcia@ucr.ac.cr

Rebeca Campos Sánchez

Research Center in Cellular and Molecular Biology, (CIBCM), University of Costa Rica, Costa Rica

Email: rebeca.campos@ucr.ac.cr

Jose Vilar Fernández

Department of Mathematics, University of A Coruña, Spain

Email: josevilarf@udc.es

Fernando García Santamaría

Research Center in Tropical Diseases (CIET), University of Costa Rica, Costa Rica

Email: fernando.garcia@ucr.ac.cr

Abstract

Tolerance to stress conditions is vital for organismal survival, including bacteria under specific environmental conditions, antibiotics, and other perturbations. Some studies have described common modulation and shared genes during stress response to different types of disturbances (termed as perturbome), leading to the idea of central control at the molecular level. We implemented a robust machine learning approach to identify and describe genes associated with multiple perturbations or perturbome in a *Pseudomonas aeruginosa* PAO1 model.

Using microarray datasets from the Gene Expression Omnibus (GEO), we evaluated six approaches to rank and select genes: using two methodologies, data single partition (SP method) or multiple partitions (MP method) for training and testing datasets, we evaluated three classification algorithms (SVM Support Vector Machine, KNN K-Nearest neighbor and RF Random Forest). Gene expression patterns and topological features at the systems level were included to describe the perturbome elements.

We were able to select and describe 46 core response genes associated with multiple perturbations in *P. aeruginosa* PAO1 and it can be considered a first report of the *P. aeruginosa* perturbome. Molecular annotations, patterns in expression levels, and topological features in molecular networks revealed biological functions of biosynthesis, binding, and metabolism, many of them related to DNA damage repair and aerobic respiration in the context of tolerance to stress. We also discuss different issues related to implemented and assessed algorithms, including data partitioning, classification approaches, and metrics. Altogether, this work offers a different and robust framework to select genes using a machine learning approach.

Keywords: Perturbations, *Pseudomonas aeruginosa*, machine learning, gene selection, perturbome.

48 **1. Introduction**

49 Cell stress can be defined as a wide range of molecular changes that cells undergo in response to
50 environmental, physical, chemical, or biological stressors; sensing and responding to stress is critical for
51 survival [1]. These biological functions and metabolic activities are executed through complex physical and
52 regulatory interactions of genes that resemble networks [2]. Additionally, tolerance to stress conditions (i.e.
53 stressors, perturbations, or disturbances) is vital for organismal survival, including bacteria under diverse
54 environmental conditions, including antibiotics [1].

55 Several studies have revealed diverse molecular levels that can explain the general response to
56 disturbances in many organisms. However, detailed mechanisms related to responses to perturbations remain
57 poorly understood [3]. Available reports usually focus on specific stressors and relatively few studies have
58 focused on common, central, and universal determinants affected by multiple perturbations [3]. This concept
59 has been recently referred to as the perturbome [4,5]. For example, in eukaryotic organisms, including plants
60 and human cancer models, some studies have described diverse stress-response genes as common
61 modulators for different types of disturbances, suggesting a central control mechanism [2,4–6]. In prokaryotic
62 models, similar findings have been reported for *Escherichia coli* [3,7].

63 Additionally, a comprehensive study of gene-interactions allows the identification of functional
64 relationships among genes [8], their products, and the underlying biological phenomena that are critical to
65 understand phenotypes under different biological conditions [9,10]. In the context of cell stress, the response
66 to different environmental or experimental stimuli can be recognized by distinct gene expression patterns.
67 This can be inferred from transcriptomic profiling data and functional associations using high throughput
68 molecular technologies such as microarrays or RNASeq [2]. However, a challenge with these technologies is
69 the large amount of high complexity data generated. Specialized bioinformatics analyses are required to select
70 relevant information and to reduce noise that distinguishes the molecular determinants for particular
71 biological conditions. Thus, a primary objective of transcriptomic profiling is to find an optimal subset of genes
72 that could be used to characterize and classify unknown samples [11]. This gene selection is not obvious and
73 complex due to the thousands of genes to select from [12].

To study the central response determinants to different perturbations in a living organism, we used the model *Pseudomonas aeruginosa* PAO1 (the reference strain of Pseudomonals). *P. aeruginosa* is a Gram-negative gamma-proteobacterium with a noteworthy metabolic versatility and adaptability enabling the colonization of diverse niches infecting plants, animals, and humans alike [8,13]. This bacterium is an opportunistic pathogen that causes a variety of infections in immunocompromised hosts, and it is associated with complications in cystic fibrosis and is a major cause of death among patients with this disease [14]. For Pseudomonals, the molecular mechanisms associated with most biological processes remain unclear causing limited action to modulate responses, including the susceptibility to stressors, environmental factors, and experimental conditions.

In *P. aeruginosa*, the response to disturbances has been reported to be influenced by <10% of coding genes [15,16], including complex molecular networks with pleiotropic patterns [17], i.e., a single perturbation induces the modulation of hundreds of genes [15,16]. These networks can be modeled using system biology approaches to recognized emergent patterns at the molecular level, as previously reported [18–21]. Large-scale molecular networks are regularly built using databases and can be used not only to understand biological functions, regulatory mechanisms, pathways, or gene groups associated with the molecular response [22] but also to eventually manipulate it [23].

In the response to stress conditions, gene expression profiling using transcriptomic analysis and topological metrics can be used to provide a global view of transcriptional regulation [24]. The identification of hub genes and pathways associated with the network can be used to gain insights into the biological meaning of the complex interactions [24,25]. This is a key point to translate molecular patterns into the phenotypes, in particular for the response to perturbations [26].

Several studies have used machine learning algorithms at the transcriptomic level to recognize gene expression patterns [2,27–29]. However, for many common biological contexts, the applicability and utility of these machine learning approaches have not been fully explored and utilized [30], for example in the exposure to multiple stressors and the molecular response. To our knowledge, only a few studies have used feature selection methods on biological data to describe the effects of multiple perturbations in complex biological systems [4,6] and so far none in *P. aeruginosa*. A related work in *P. aeruginosa* used a machine learning

101 approach to identify sets of genes that correlate with multiple culture media [31]. Due to differences in the
102 experimental design, this study and our work are not comparable.

103 The use of microarray and other high throughput technologies data involve challenges for machine
104 learning approaches. These include the curse of dimensionality [32,33], normalization of raw values to
105 compare samples [34,35], data partitions for training and testing models [36,37], and evaluation of
106 performance [34,38]. Since comparison between the machine learning algorithms is completely variable
107 [11,30,33,39,40], Support Vector Machine (SVM) [41], K-Nearest neighbor (KNN) [40], and Random Forest (RF)
108 [42] have been successfully used with microarray gene expression data allowing the recognition of emerging
109 patterns [39].

110 Here, we hypothesize that perturbations on living cells will trigger global reprogramming of multiple
111 molecular determinants that can be sensed at the transcriptional level. The initial central response after acute
112 stress will then expand producing the global molecular rearrangement (Figure 1-A). Then, pleiotropic and
113 specific effects on gene networks will be reflected as changes in gene expression profiles and the complexity
114 of molecular regulation at other levels. Therefore, by using a machine learning approach, common molecular
115 features (for all stressors) could be identified as a central or core determinant.

116 Thus, the aims for this study were (i) to implement a machine learning approach to select genes from
117 gene expression data, and (ii) to identify and describe a subset of genes that can be associated with multiple
118 perturbations or the core perturbome in *P. aeruginosa*.

119

120 **2. Materials and Methods**

121

122 The overall methodology of the study is presented in Figure 1-B. In brief, after a data selection,
123 normalization and gene selection to define the perturbome were run. To achieve this, we considered two
124 different procedures to split the data (training and testing datasets for the machine learning approaches: SVM,
125 KNN, and RF): a single partition (SP method) and another with multiple partitions (MP method). Relations
126 between genes were represented using both large scale and small-world networks, and a final comparison
127 between conditions, an analysis of differential expression, and gene annotation were performed.

128 *2.1 Datasets*

129 To compare gene expression profiles of strain *P. aeruginosa* PAO1 when exposed to multiple
130 perturbations, the GEO database (https://www.ncbi.nlm.nih.gov/geo/) was used for a systematic selection of
131 datasets (Figure 2-A-B). Initial evaluation by the organism and mRNA profiles by GPL84 platform (Affymetrix
132 *Pseudomonas aeruginosa* PAO1 Array, with all 5549 protein-coding sequences) identified 156 series of
133 datasets with 1310 samples (Date of Access: January 25th, 2018). In a second step, data were selected
134 according to experimental design if they included perturbations, leaving only 47 series. Finally, to make
135 datasets comparable by experimental conditions, evaluation and selection were done for series with similar
136 culture conditions (Luria Bertani LB medium and exponential phase when measuring mRNA profile) and if a
137 control condition was available (without any perturbation or treatment). The final dataset was composed of
138 10 series with 71 samples (Series GSE2430, GSE3090, GSE4152, GSE5443, GSE7402, GSE10605, GSE12738,
139 GSE13252, GSE14253 and GSE36753).

140 Some series included temporal measurements (see Experiment 3 of the Figure 2-B) which we considered
141 as separate perturbations, resulting in replicates of 10 controls and 14 perturbations (Figure 3-A):
142 azithromycin with 2 series (AZM-a and AZM-b) [43,44], Hydrogen peroxide (H_2O_2) [45], copper (Cu) [46],
143 sodium hypoclorite (NaClO) [47], ortho-phenylphenol (OPP) at 20 and 60 minutes [48], colistin (COL) [49],
144 chlorhexidine diacetate (CDA) at 10 and 60 minutes [50], E-4-bromo-5-bromomethylene-3-methylfuran-2-5H-
145 one (BF8) [51] and ciprofloxacin CIP at 0, 30 and 120 minutes [19]. Details of each study (experimental design
146 and main results) are provided in Table S1.

147

148 *2.2 Pre-processing and comparison of global transcriptomic profiles*

149 To compare all the 71 samples, a first analysis was the pre-processing step using Bioconductor 3.8
150 (https://www.bioconductor.org/) in the R software (Version 3.5) with classical functions for microarrays.
151 Robust MultiArray Average algorithm (RMA) was performed in the Affy package to correct background,
152 normalization, and summarization.

153 Subsequently, a clustering algorithm of Principal Component Analysis (PCA) was implemented to
154 compare global transcriptomic profiles between perturbations and controls. PCA was done with default
155 parameters of the Caret Package (caret.r-forge.r-project.org/) in R software.
156 In order to robustly select genes that could separate experimental conditions (controls and
157 perturbations) and to identify the core response of *P. aeruginosa*, two approaches of feature selection
158 protocols were implemented, as detailed below (Figure 2-B-C).
159
160 *2.3 Gene ranking and selection by Single Partition SP method*
161 To identify genes which expression values were commonly related to multiple perturbations, a first
162 approach was implemented considering a particular partition of the dataset for training and testing sets
163 (Figure 2-B-C). Single partition was established using the last replicate of each experiment, in both control and
164 perturbation. Because there were 14 perturbations and 10 controls (71 samples including replicates or
165 individual samples), a total of 24 samples were included in the testing dataset and the remaining 47 samples
166 were included for the training dataset (66% for training and 34% for the testing set).
167 Using this partition, the ranking of genes was done by a machine learning approach using three
168 classification algorithms: SVM, KNN, and RF. A homemade script in R included these functions of the Caret
169 package. For all three algorithms, default parameters were used for training and 10-fold cross-validation was
170 included. After this, the variable importance metric was calculated for all genes using the *varImp* function,
171 associating a specific value for each gene. In the case of SVM and KNN, the same importance is calculated
172 because the function is model-free for these cases (as detailed in Caret Package), resulting in the same list of
173 genes but metrics are specifically calculated for each algorithm. For RF, the accuracy and model parameters
174 were used to calculate the metric.
175 As similarly reported in [11], to evaluate the number of genes that should be selected in the top group
176 (the first K ranked genes, as candidates for the core response by each algorithm), multiple classification
177 models were systematically run, starting only with the highest-score ranked gene. In brief, after the training
178 with one gene, model performance was evaluated by calculating metrics using the testing dataset. Then,
179 training was run again when the next ranked gene was added, and new metrics were calculated. This process

was repeated up to complete all the ranked genes. Metrics included accuracy (correct classification percentage), kappa value (inter-rater classification agreement), sensitivity, specificity, precision, recall, prevalence, and F1 score (harmonic average of the precision and recall).

Selection of the K value of top genes was based on the following criteria: (i) the stability of the metrics (priority for accuracy, kappa, and F1) when the increment of ranked genes was done, as suggested in [11], (ii) consensus value suitable for all the three algorithms (including a 10% of tolerance), and (iii) minimum number of genes as possible. After the selection of the K value, ROC (Receiver-operating characteristic) curve and AUC (Area under the curve) value were calculated for each algorithm. Finally, the selection of top K genes between algorithms was compared by metrics and- the list of genes.

2.4 Gene ranking and selection by Multiple Partitions (MP) method

In order to identify genes related to multiple perturbations independently of a single/specific partition, a second method using multiple random partitions was implemented (Figure 2-B-C). To address this, a random data selection for training and testing sets was done using the *createDataPartition* function. Partition was set to 80% (57 samples) for the training set and the remaining for the testing set (14 samples) with experimental conditions equally distributed. Then, protocols with SVM, KNN, and RF algorithms (same conditions as previously described in SP method, with 10-fold cross-validation) resumed the analysis with a final ranking of genes using the *varImp* function. Using only the top K of ranked genes (K value determined using the criteria described in the SP method), a new set of training/testing sets were used for evaluating the performance of the models and each metric was stored with the list of the K ranked genes. This full process was automatically repeated 100 times using the *replicate* function, starting with a new random partition and finishing with the list of the K genes and the metrics associated with that partition. Finally, for each algorithm, full data of all the runs were analyzed for determining the frequency of the appearance of genes (*table* function) and calculating the average and dispersion of metrics across all the 100 runs. The definitive list of the K more frequent genes was established for each algorithm after this comparison by frequency.

207 *2.5 Identification of core response genes*

208 After selection of top K genes (candidates of the perturbome) in each algorithm by SP and MP methods,
209 a comparison of genes was done using UpSets plots (https://github.com/hms-dbmi/UpSetR) in R software.
210 We called the "support" of a gene the number of algorithms (out of six approaches) in which a gene was
211 selected as a candidate. Genes with a support ≥4 (i.e., it was identified by at least four algorithms) were
212 considered as part of the final list of genes of the core perturbome. This consideration guaranteed that a gene
213 was identified by the two methods and at least by two different classification algorithms.

214 Gene relationships were represented using molecular networks using a large scale model (using a top-
215 down systems biology approach). The network was built using protein-protein interaction (PPI) graphs in
216 PseudomonasNet database (www.inetbio.org/pseudomonasnet/). Visualization of the network and
217 measurement of the topological features were performed using Cytoscape software [52].

218

219 *2.6 Description and comparison of core response genes*

220 In order to describe and compare the genes associated with the core response of *P. aeruginosa,* by
221 experimental conditions, four analyzes were established. First, PCA was evaluated again but now only
222 considering genes of the core response. Based on distribution in the case of PCA, representation of centroids
223 was done using the Kmeans algorithm.

224 Second, using the PseudomonasNet database, a small-world network was built and then exported into
225 Cytoscape software with the genes of the core response. The information of topological features (including
226 connectivity) and expression levels of kmedoids were incorporated into different versions of the network. To
227 identify central or hub genes in the network in Cytoscape, the cytoHubba app was used [53] with the top 5
228 elements according to the metrics of degree, bottleneck, and betweenness. All identified genes in any out of
229 the three approaches were assigned as hub genes.

230 Third, a classic analysis of differentially expressed genes (DEGs, $p<0.05$) was implemented in R with
231 Limma package (https://www.rdocumentation.org/packages/limma/versions/3.28.14) using empirical Bayes
232 moderated t-statistic (eBayes) with Benjamini and Hochberg method for *p*-value correction [54]. This was
233 done to compare our results with a classical approach for gene expression.

234 Finally, to give biological interpretation to the selected genes, to determine levels of expression reported
235 in databases, and to study metabolic pathways involved under each perturbation, an exhaustive annotation
236 was made using the databases PseudomonasDB (gene ontology), GEO database, KEGG annotation, and
237 particular literature. This information was integrated with the results obtained by all the analysis and the DEGs
238 to fully describe the genes that make up the core response or perturbome of *P. aeruginosa* PAO1.
239
240 *Availability of data and material*
241 The datasets generated and/or analyzed during the current study are available in:
242 • Public raw data used in this study: GEO database (https://www.ncbi.nlm.nih.gov/geo/), data Series
243 GSE2430, GSE3090, GSE4152, GSE5443, GSE7402, GSE10605, GSE12738, GSE13252, GSE14253 and
244 GSE36753.
245 • Normalized data and R Scripts: https://github.com/josemolina6/CoreResponsePae
246
247 **3. Results**
248 *3.1 Perturbome genes of* P. aeruginosa *can be identified by a machine learning approach using SP and MP*
249 *methods*
250 A total of 71 samples of 10 controls and 14 perturbations were considered for the study, with comparable
251 expression levels (Figure 3-A). Global transcriptomic profiles (all 5549 genes) were compared by a PCA,
252 revealing a mixed pattern (no separation) between perturbations and controls (Figure 3-B). Two samples (BF8
253 and Control 8) resulted in "extreme" global profiles. Two methods using machine learning (SP and MP) were
254 implemented in order to robustly rank and select genes associated with multiple perturbations in *P.*
255 *aeruginosa* (Figure 2). In each method, three classification algorithms were evaluated: SVM, KNN, and RF.
256 The first method (SP) implemented a gene ranking based on variable importance using a single/specific
257 data partition. Metrics associated with RF are shown in Figure 4, and in supplementary Figure S1 for SVM and
258 KNN. After the ranking was established, multiple classification models were run with the ranked genes (Figures
259 4-A and supplementary Figure S1-A-C). For each classification model, the stability of the three metrics was
260 evaluated to select the suitable K value of genes that could be applied to all algorithms at the same time. For

261 SVM and RF, stable values of metrics are given with at least the first 51 ranked genes, meanwhile it is 45 genes
262 for KNN. Considering a 10% of tolerance with the highest of these values, K=56 was selected as the number
263 of top genes that were included as preliminary candidates of the core response according to each algorithm.
264 With this value, metrics of each algorithm were compared (Table S2). For example, accuracy was 0.79, 0.71,
265 and 0.75 for SVM, KNN, and RF respectively in the SP method. SVM obtained a better performance according
266 to kappa, sensitivity, recall, and F1 scores, but higher values of specificity and precision resulted for RF. Also,
267 the ROC curve and AUC value were calculated (Figures 4-B and supplementary Figure S1-B-D). The best
268 performance was obtained for RF with AUC = 0.92, then 0.82 for SVM, and finally 0.76 for KNN. Since the
269 ranking by *importance* for SVM and KNN is the same (model-free strategy, see details in
270 https://topepo.github.io/caret/variable-importance.html), they shared the same list of top 56 genes.
271 Comparison between implementations showed that 21 genes were identified by the three algorithms at the
272 same time, 35 exclusively by RF, and the same number for SVM/KNN. In total 91 genes were identified by any
273 of the algorithms. The list of genes and the importance value of the top 56 ranked genes for each approach is
274 presented in Figures 4-C for RF and Figure S1-E for SVM/KNN.

275 In a second approach, the MP method was implemented using multiple random partitions. Same SVM,
276 KNN, and RF algorithms were evaluated by running 100 iterations with different partitions and top-56 more
277 frequent genes for each method were selected. Details of ranking and frequency are shown in Figure 5-A
278 (SVM) and supplementary Figure S2 (KNN and RF). Metrics for all 100 iterations are presented in Table S2,
279 Figure 5-B, and supplementary Figure S2. Accuracies for all the models were 0.66, 0.69, and 0.70 for SVM,
280 KNN, and RF respectively. Specific values for kappa, precision, recall, and F1 score were found for each
281 algorithm. When a comparison of the list of genes was done, 29 genes were identified at the same time by all
282 implementations, and 23 by both SVM and KNN.

283 To identify preliminary core response genes in *P. aeruginosa*, the lists of top 56 genes selected by each
284 algorithm and method were jointly represented using UpSet plots (Figure 6-A). A total of 118 different genes
285 were identified and 15 genes were simultaneously identified by all the algorithms. The KEGG annotation of all
286 the genes revealed functions associated with energy and DNA metabolisms, biosynthesis, ribosomal activity,
287 and other pathways (Figure 6-B). See below for more details of the annotation.

288 The distribution of all 118 genes on a large-scale molecular network is presented in Figure 7-A. Results
289 show that selected genes are associated with different subgroups of highly connected genes.
290 The topological metrics of the network showed a non-random arrangement with 90 edges (more than
291 expected by chance, enrichment p-value: < 1.0e-16), 3.9 average number of neighbors, a clustering coefficient
292 of 0.51, and others. Besides, 10 hub genes were identified among the core perturbome, which are
293 representative elements of the network. More details in Figure 7-A and Table 1.
294 The final version of the core perturbome was established by selecting genes with a support ≥4 (i.e found
295 by at least four algorithms). In total 46 genes were finally associated with core response. Notwithstanding six
296 individual classifiers were implemented (each one with specific datasets), the consensus strategy to select the
297 final genes (by the support) had no effects on the resolution of the classes (control or perturbation), as shown
298 with the PCA with the transcriptomic profile using the 46 genes.
299 In the representation as a small-world network, relationships between the 46 genes revealed different
300 topological patterns of connectivity between molecules, being *nuoC* and *nuoF* the ones with higher
301 connectivity (connection degree). In addition, only six genes had no connections between them (Figure 7-B).
302
303 *3.2 Comparisons of core response genes show the separation of global transcriptomic profiles according to*
304 *experimental conditions and biological functions related to tolerance to stress*
305 After the selection of the core response genes, comparisons between controls and perturbations were
306 done to characterize these genes by global profiles. First, the PCA was implemented to compare the global
307 profiles using the 46 genes of the perturbome, in which the distribution shows a well-defined separation
308 between controls and perturbations (Figure 7-C). A Kmedoid algorithm (k = 2) was able to identify two clusters
309 enriched by samples of each condition: one consisting entirely of perturbation samples (11 samples, blue
310 color) and the other mostly by control samples (10 controls and 3 disturbances, red color). In the case of the
311 blue cluster, the kmedoid was sample CIP-120min, meanwhile control 6 was selected for another cluster.
312 Supplementary analysis of gene expression was included by comparing levels for all core response genes
313 (Figure S3-A) and comparing expression levels of the kmedoids on the small-world network (Figure S3-B-C).

314 On the other hand, gene annotation revealed that biological processes related to most of the genes are
315 metabolism, molecule binding, and biosynthesis (Figure 8-A). At the level of pathways, genes of the core
316 perturbome have participation in processes of DNA damage response, DNA repairing, and response to
317 general/specific stimuli (Figure 6-B and Table 1). Other studies show distinct expression patterns for each
318 gene, depending on the disturbance as shown in Table 1. Finally, in order to compare the results of the
319 machine learning strategy with another approach, differential expression analysis was run. A total of 101 DEGs
320 were identified, of which 33 were shared by the core response (Figure 8-B). This means that 72% of core
321 response genes were also identified by another single and independent method. Annotation of DEGs for
322 biological processes (Figure 8-C) showed a similar profile to our machine learning approach.
323
324 **4. Discussion**
325 Living organisms face external and internal conditions that compromise cellular functions at molecular,
326 metabolic, and structural levels, disrupting their homeostasis [55,56]. The cell stress response is crucial for
327 organismal survival and complex networks are usually involved in the molecular mechanism related to
328 tolerance [1]. However, few studies have identified central and possible universal regulation of the response
329 to multiple disturbances, a concept termed as perturbome [4,5]. The common molecular response was
330 previously reported as a network of a common set of genes and pathways that can be generically associated
331 with multiple perturbations in plants [7], pathogenic bacteria [3], or cell lines models [4], and others.
332 In our approach using *P. aeruginosa,* but also applicable to other organisms, we hypothesized the
333 existence of a set of core genes regulating the response to stressors in a generic sense of different pathways.
334 This bacterium is a ubiquitous organism that causes infections in immunocompromised hosts [14]. It has a
335 high proportion (about 5%) of its genome dedicated to regulatory mechanisms, probably explaining its
336 adaptability to such a broad range of growth conditions [38]. Since the strain *P. aeruginosa* PAO1 is a clinical
337 isolate with a profile of multiresistance to many antibiotics [57], characterization of molecular mechanisms
338 involved in the tolerance to stressors in this strain could eventually help to modulate sensitivity and overcome
339 resistance. Exhaustive integration of -omics data and network analysis are required to clarify the molecular
340 mechanisms related to stress conditions and eventually use them for modulating cell response.

341 For example, we have a particular interest in *P. aeruginosa* due to we have been studying a high-risk
342 strain (called *P. aeruginosa* AG1) which was isolated in 2010 from a Costa Rican Hospital. We reported this
343 isolated in 2015 [58] and it is a Priority 1 organism [59] because it is resistant to carbapenems, a last-resort
344 antibiotic. We assembled the genome using Illumina and Nanopore sequencing [13], revealing the genomic
345 features including two integrons, which were studied in detail in another work [60]. Proteomic profiles after
346 exposure to antibiotics have been also studied [61]. Finally, we studied the transcriptome in response to
347 ciprofloxacin using RNA-Seq and systems biology [16]. We expect to integrate our results of the perturbome
348 (in the reference strain) with the Costa Rican strain (AG1) to identify the "expected general response" to
349 multiple perturbations (as a Pseudomonal), and more importantly, to identify the "exclusive response" to
350 antibiotics, to perform new assays. In the response to ciprofloxacin, we already demonstrated that energy
351 metabolism, RNA degradation, and stress response are induced, as found in other perturbations (that here
352 we call "the perturbome"), but also other more exclusive responses such as phage induction (very exclusive
353 and not part of the perturbome) are modulated with ciprofloxacin.
354 On the other hand, although nowadays RNA-Seq is one of the most common high-throughput molecular
355 strategies to analyze transcriptomic data, the selection of microarray data to analyze the perturbome of *P.*
356 *aeruginosa* was based on the availability of data in databases. First, unlike other microorganisms (for example
357 intra-cellular bacteria), *P. aeruginosa* has a very variable genome size between 6 and 7.5 Mb, which is
358 explained in part due to the great repertory of the accessory genome [13]. Because of this, it is difficult to
359 combine molecular data of multiple strains. Second, the PAO1 strain is the reference strain (model for the
360 group during more than 20 years), and no other strain has the completeness of gene annotation as found for
361 this model. Finally, transcriptomic data of PAO1 is mainly found with microarray strategies, which is
362 significantly more abundant in databases than RNA-Seq.
363
364 *4.1 Insights of algorithms to identify core response genes or perturbome*
365 In our study, initial global transcriptomic profiles showed mixed patterns between samples according to
366 their experimental condition. Because of its complexity, raw microarray data showed noise and redundant
367 information that can explain the poor resolution of the initial PCA to identify classes [33]. Thus, a first step

368 was the normalization, which was particularly critical for multi-origin samples as in our case. Comparative
369 analysis needs to consider variation that depends not only on the experimental design but also parameters
370 including laboratory/culture conditions and mRNA extraction methods [62]. Subsequently, a feature (gene)
371 selection analysis was implemented for not only identifying genes associated with multiple perturbations in
372 *P. aeruginosa,* but also to improve the performance of predictive/descriptive models capable of separate
373 control and perturbation categories. In three experiments, two or three time points were considered as
374 perturbations (Figure 3-A and the example of Experiment 3 in Figure 2-B). Although an overrepresentation of
375 that perturbation cannot be discarded, we assume they are independent assays due to the global profile is
376 known to be different over time.

377 The potential patterns of the core response were investigated using a robust machine learning approach
378 by implementing six protocols, using SP and MP methods, and SVM, KNN, and RF algorithms in each case. This
379 was a critical step because gene selection from microarray data is complex *per se*. Feature reduction remains
380 a challenging task in transcriptomic studies because thousands of genes to select from, and it introduces an
381 additional layer of complexity in the modeling task [12,63]. To avoid bias and overfitting, implementations of
382 diverse strategies of data partitioning such as bootstrapping, random partitions, and cross-validation are
383 recommended. In general, these methodologies can robustly minorate the influence from noise, outliers,
384 absence of ground truth sets, and reduce variance [2,37,64]. In the specific case of implementations for the
385 study of molecular networks, partitions are suggested not only because to counteracts the curse of
386 dimensionality but also ground truth of candidate genes or regulatory networks are rarely available to allow
387 for an unbiased assessment of these methods [37].

388 The single partition SP method consisted of particular and invariable data for training (with internal cross-
389 validation) and another to test, and it is probably the most common approach used in machine learning. In
390 the case of the multiple partitions MP method, 100 random partitions of the dataset were run. MP method
391 had a dual consideration when splitting data (multiple partitions and the internal cross-validation). This
392 method can be considered as an *ensemble based on different data partitioning*, as it had been previously
393 proposed [36]. Datasets were divided using multiple random partitioning procedures and then genes were
394 ranked. After all runs, a final feature subset is determined by calculating the frequency of features in all the

395 runs [36]. An equivalent approach was implemented by Pai and collaborators to classify gene expression data
396 in a cancer model [65].

397 However, one possible problem with the MP approach is that cross-validation results may depend on
398 the similarity of testing and training sets. For example, all the samples of the experiment can be assigned as
399 part of the same training set and none in the testing set (or vice versa, for example, the red/blue/dark-green
400 replicates of the last arrangement of the MP method in Figure 2-B). A classification prediction method is only
401 expected to learn how to predict on unseen samples that are drawn from the same distribution as training
402 samples [37,64], and the MP method could violate this assumption. SP method guarantees this because it was
403 built using the last replicate for each perturbation and control (Figure 2-B, SP method).

404 In contrast, because of the presence of several biological replicates of the same perturbation in both
405 datasets of the SP method (one in the testing and the remaining in the training set), it can be considered that
406 the conformation of the testing data set is not completely independent, i.e. every sample in the training
407 dataset has a biological "copy" (replicate) itself in the training set. This scenario can introduce a bias in the
408 performance of the classifiers and robust analyses, such as multiple partitions and/or cross-validation, are
409 required to avoid a false impression of high predictive ability of the model and hence of the reliability of the
410 inferred influences on the gene [37].

411 Thus, both SP and MP methods have drawbacks but they can be compensated using the best features of
412 each strategy to robustly select features (genes).

413 On the other hand, many algorithms for dimension reduction have been proposed [6,32,33] but no
414 standard machine learning algorithm can be selected due to multiple evaluation results on completely variable
415 metrics associated with the performance [39]. Many studies have shown that SVM, KNN, and RF generally
416 outperform other traditional supervised classifiers [30,39,66]. In our case, a variable pattern was found for
417 different metrics in all evaluated implementations. For example, based on accuracy SVM for SP method and
418 RF for MP method resulted in higher scores, which agree with other studies using machine learning and
419 biological data [39,65].

420 In our subsequent analysis, we were interested in identifying a consensus list of candidate genes for the
421 core response in *P. aeruginosa*, resulting in 46 genes of the perturbome. When global profiles were compared

422 using these genes (PCA), control and perturbation classes were clearly separated (Figure 7-C). This gene
423 number seems to be a modest number of elements (less than 1% of all the available datasets with 5549 genes)
424 but it agrees with other studies, including machine learning methods [11,30] or other approaches
425 [11,35,67,68]. Besides, 72% of core response genes were also identified as DEGs with a similar annotation
426 profile. Differences can be explained by significant fluctuations in the differential expression results as
427 previously reported, mainly because it is not a consensus strategy (only based on *p*-value) and it does not
428 incorporate the estimates of the test performance (true positive/negative rates and other metrics) on the
429 results [2].
430
431 *4.2 Biological insights of the core response genes in* P. aeruginosa*: the perturbome*
432 In the study of complex systems, the lack of specific biological information at the single gene level and
433 the interactions make the extraction of the relevant information a complicated task in this field. This situation
434 remains as an unsolved issue of molecular high-throughput technologies, in which molecular data is generated
435 at a higher rate than our ability to find patterns or extract relevant information [69]. Thus, large-scale/top-
436 down system biology-based approaches have been used as a strategy to exploit and extract relevant
437 information of ample amount of -omics data [21], letting to predict interactions between genes [23].
438 With the representation of gene-to-gene interactions, system biology allows studying not only specific
439 genes or gene clusters but also the emergent properties found in the molecular networks [17]. Thus,
440 pleiotropic patterns can be recognized towards the response to perturbations, in which large-scale networks
441 can be built [69]. In the case of the core perturbome, the representation of genes showed a complex network
442 with topological properties that defines a non-random arrangement, including hub genes and pathways
443 associated with the response to stress and stimulus, metabolism, and ribosome activity. The network can be
444 inferred as a coordinated central response to the perturbations, independently of a particular stressor.
445 Genes of the core response or perturbome can be related to a central regulation network, and as the
446 convergent point of signal transduction, transcriptional regulation, and stress-related pathways, as it has been
447 suggested [2,4,5,56]. Annotation of the 46 genes shows that most of them are functionally related to
448 biosynthesis, molecule binding, and metabolism (including an important number of hits for lipids), including

449 additional functions for the regulation of DNA damage repair, response to stimuli, and aerobic respiration.
450 Interestingly, hub genes represent these processes with high connectivity in the small-world network. For
451 example, the main functions associated with *fabA* are lipid metabolism, fatty acid, and lipid biosynthesis,
452 meanwhile for *nuoC* are aerobic respiration, electron/proton transport, and DNA damage. Finally, *nuoF* is
453 associated with aerobic respiration, electron/proton transport, and Krebs cycle.

454 In this sense, cells are equipped with systems and mechanisms to recover from the environmental stress
455 and stimuli to maintain all necessary physiological functions [70]. Other common stimuli such as low ATP, slow
456 growth, and ROS production can also occur before cells express stress-specific factors, but mediating a
457 common effect. For example, response to stress includes modulation of energy production and aerobic
458 electron transfer chain components. As it has been reported in *E. coli*, aerobic electron transport chain
459 components are down-regulated in response to growth arrest [71]. This corresponds with the global profiles
460 of expression of 46 core response genes. Also, regulation of lipid metabolism is relevant for survival in the
461 wide range of environmental conditions where bacteria thrive [72], even for biofilm-living forms [3] as *P.*
462 *aeruginosa*. Core response genes *fabG*, *lpxA*, and PA5174 could be implied in this process.

463 In the case of DNA damage repair (including the case of *cycB* and *gltP* genes), responses mediated by
464 SOS and *rpoS* help to maintain genome integrity, colonization, and virulence [19,73]. These responses are
465 activated under multiple disturbances and modulating a low energy production and shutdown of the
466 metabolism, promotes the formation of antibiotic resistance and biofilms [3]. Other related pathways for
467 some specific genes included regulation of the transcription during stress by RNA-binding proteins to
468 reprogram or shut down translation and to rescue the ribosome stalled by a variety of mechanisms induced
469 [74]. Three core response genes (*rpmH*, *tsf,* and PA2735) were annotated with these functions.

470 Jointly, the relatively few diversity of metabolic functions and pathways makes sense in order to ensure
471 redundancy and robustness in the response to stimuli. Similar results, regarding enriched pathways, have
472 been obtained in other studies with eukaryotic models, including disturbed human cell lines [4,5,56],
473 *Arabidopsis thaliana* under physical and genotoxic stresses [2], or a genome-wide association study of a
474 generic response to stress conditions [6]. In the case of prokaryotic organisms, two studies have used
475 *Escherichia coli* as a model to identify differentially expressed genes after exposure to stress conditions [3]

476 and to create networks associated with the response to fluctuating environments [7]. Differences with other
477 organisms and disturbances can suggest that response cell stress can be organismal specific, although
478 heterogeneity has also been suggested as a reasonable explanation because of differences in the response in
479 a homogeneous cell population [56,75].

480 As in our case, the response to stimuli and stressors is orchestrated by a pleiotropic modulation [20]
481 which can be associated with central regulation. Alternative mechanisms such as cross stress protection (the
482 ability of a stress condition to protect against other stressors) [7], the role of sigma factors, and specific two-
483 component systems [3] can contribute to explaining this phenomenon. The molecular response can lead to
484 regulate multiple biological activities including metabolism, replication, transcription and translation, changes
485 in membrane composition, motility, modification gene expression, expression of virulence factors, multi-drug
486 resistant phenotypes and biofilm formation, and others [56].

487 For further works, we are running experiments to validate our findings here, in which *P. aeruginosa* is
488 exposed to different perturbations and gene expression is quantified by RT-qPCR. This will also allow us to
489 build small-world networks with the representation of the core perturbome as a directed graph. Using a
490 bottom-up system biology approach using differential equations, sensitivity analysis, and other machine
491 learning algorithms, we expect to identify and manipulate the points of control the network, as we have
492 previously done in other models [76].

493 Taking all together, results of our study suggest that identification of core response genes associated
494 with multiple perturbations or perturbome in *P. aeruginosa* can define a central network available to
495 modulate a basic response that includes biological functions such as biosynthesis, binding, and metabolism,
496 many of them related to DNA damage repair and aerobic respiration. To our knowledge, this study can be
497 considered a first report of the *P. aeruginosa* perturbome.

498 Further analyses are required to explore the potential use of perturbome network to modulate
499 (positively or negatively) the response to disturbances, to model molecular circuits, to identify possible
500 biomarker genes of stressed states, and to experimentally validate our findings. Besides, this approach can be
501 used to model the perturbome in other *P. aeruginosa* strains, as we hope to run soon with a genome we
502 recently described [13], and other organisms.

503 **5. Conclusions**

504 A robust machine learning approach was implemented to identify and describe core response genes to
505 multiple perturbations in *P. aeruginosa*. Using public microarray data, two independent partition strategies
506 (single and multiple with SP and MP methods respectively), and three classification algorithms, we were able
507 to identify 46 perturbome elements. Both network analysis and functional annotations of these genes showed
508 coordinated modulation of biological processes in response to multiple perturbations, including metabolism,
509 biosynthesis, and molecule binding, associated with DNA damage repairing, and aerobic respiration, all
510 probably related to tolerance to stressors, growth arrest, and molecular regulation. We also discussed
511 different issues related to implemented and assess algorithms of normalization analysis, data partitioning,
512 classification approaches, and metrics.

513

514 **Abbreviations**

515 AUC: Area under the curve

516 AZM: Azithromycin

517 B8F: E-4-bromo-5-bromomethylene-3-methylfuran-2-5H-one

518 CDA: Chlorhexidine diacetate

519 CIP: Ciprofloxacin

520 COL: Colistin

521 Cu: Copper

522 DEGs: Differentially expressed genes

523 GEO: Gene Expression Omnibus

524 H2O2: Hydrogen peroxide

525 HC: Hierarchical Clustering

526 KNN: K-Nearest Neighbor

527 LB: Luria Bertani

528 mRNA: Messenger RNA

529 NaClO: Sodium hypochlorite

530 OPP: Ortho-phenylphenol

531 PCA: Principal Component Analysis

532 PPI: Protein-protein interaction

533 RF: Random Forest

534 ROC: Receiver-operating characteristic

535 SVM: Support Vector Machine

536

537 **Declarations**

538 *Ethics approval and consent to participate*

539 Not Applicable

540

541 *Consent for Publication*

542 Not Applicable

543

544 *Availability of data and material*

545 The datasets generated and/or analyzed during the current study are available in:

546 Public raw data used in this study: GEO database (https://www.ncbi.nlm.nih.gov/geo/), data Series GSE2430,

547 GSE3090, GSE4152, GSE5443, GSE7402, GSE10605, GSE12738, GSE13252, GSE14253 and GSE36753.

548 Normalized data and R Scripts: https://github.com/josemolina6/CoreResponsePae

549

550 *Competing interests*

551 No competing interests to declare.

552

553 *Funding*

554 This work was funded by projects B8114 Definición de la red transcripcional y de las alteraciones genómicas

555 inducidas por la ciprofloxacina en *Pseudomonas aeruginosa* AG1 and B8152 proNGS 1.0: Implementación y

evaluación de protocolos de análisis de datos de tecnologías NGS y afines para el estudio de sistemas biológicos complejos, Universidad de Costa Rica (period 2017-2020), Costa Rica.

Authors' contributions

JMM and FGS participated in the conception, design of the study, and data acquisition. JMM, PMM, RGB, and JVF participated in data analysis and interpretation. JMM, RCS, JVF, and FGS participated in the interpretation of the data analysis. JMM drafted the manuscript and all authors were involved in its revision. All authors read and approved the final manuscript.

Acknowledgments

We thank students Iosif Forero Trelles, Daniel Solano Alvarado, and Daniela Aguilar Orozco for their collaborations in different activities of the project.

References

1. DeLong, E.F. *Prokaryotes : prokaryotic physiology and biochemistry*; Springer, 2012; ISBN 9783642301407.
2. Ma, C.; Xin, M.; Feldmann, K.A.; Wang, X. Machine Learning-Based Differential Network Analysis: A Study of Stress-Responsive Transcriptomes in Arabidopsis. *Plant Cell* **2014**, *26*, 520–537, doi:10.1105/tpc.113.121913.
3. Nagar, S.D.; Aggarwal, B.; Joon, S.; Bhatnagar, R.; Bhatnagar, S. A Network Biology Approach to Decipher Stress Response in Bacteria Using *Escherichia coli* As a Model. *Omi. A J. Integr. Biol.* **2016**, *20*, 310–324, doi:10.1089/omi.2016.0028.
4. Caldera, M.; Müller, F.; Kaltenbrunner, I.; Licciardello, M.P.; Lardeau, C.H.; Kubicek, S.; Menche, J. Mapping the perturbome network of cellular perturbations. *Nat. Commun.* **2019**, *10*, doi:10.1038/s41467-019-13058-9.
5. Sadeh, S.; Clopath, C. Theory of Neuronal Perturbome: Linking Connectivity to Coding via Perturbations. *bioRxiv* **2020**, 2020.02.20.954222, doi:10.1101/2020.02.20.954222.
6. Bermingham, M.L.; Pong-Wong, R.; Spiliopoulou, A.; Hayward, C.; Rudan, I.; Campbell, H.; Wright, A.F.; Wilson, J.F.; Agakov, F.; Navarro, P.; et al. Application of high-dimensional feature selection: Evaluation for genomic prediction in man. *Sci. Rep.* **2015**, *5*, 1–12, doi:10.1038/srep10312.
7. Dragosits, M.; Mozhayskiy, V.; Quinones-Soto, S.; Park, J.; Tagkopoulos, I. Evolutionary potential, cross-stress

behavior and the genetic basis of acquired stress resistance in Escherichia coli. *Mol. Syst. Biol.* **2014**, *9*, 643–643, doi:10.1038/msb.2012.76.

8. Nogales, J.; Guðmundsson, S.; Duque, E.; Ramos, J.L.; Palsson, B.Ø. Expanding the computable reactome in Pseudomonas putida reveals metabolic cycles providing robustness. *bioRxiv* **2017**, 139121, doi:10.1101/139121.

9. Kc, K.; Li, R.; Cui, F.; Haake, A.R. GNE: A deep learning framework for gene network inference by aggregating biological information. *Bioinformatics* **2018**, 1–9.

10. Szklarczyk, D.; Gable, A.L.; Lyon, D.; Junge, A.; Wyder, S.; Huerta-Cepas, J.; Simonovic, M.; Doncheva, N.T.; Morris, J.H.; Bork, P.; et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **2019**, *47*, D607–D613, doi:10.1093/nar/gky1131.

11. Li, Y.; Wang, N.; Perkins, E.J.; Zhang, C.; Gong, P. Identification and optimization of classifier genes from multi-class earthworm microarray dataset. *PLoS One* **2010**, *5*, 1–9, doi:10.1371/journal.pone.0013715.

12. Saeys, Y.; Inza, I.; Larranaga, P. A review of feature selection techniques in bioinformatics. *Bioinformatics* **2007**, *23*, 2507–2517, doi:10.1093/bioinformatics/btm344.

13. Molina-Mora, J.-A.; Campos-Sánchez, R.; Rodríguez, C.; Shi, L.; García, F. High quality 3C de novo assembly and annotation of a multidrug resistant ST-111 Pseudomonas aeruginosa genome: Benchmark of hybrid and non-hybrid assemblers. *Sci. Rep.* **2020**, *10*, 1392, doi:10.1038/s41598-020-58319-6.

14. Kumar, S.S.; Penesyan, A.; Elbourne, L.D.H.; Gillings, M.R.; Paulsen, I.T. Catabolism of Nucleic Acids by a Cystic Fibrosis Pseudomonas aeruginosa Isolate: An adaptive pathway to cystic fibrosis sputum environment. *Front. Microbiol.* **2019**, *10*, 1–14, doi:10.3389/fmicb.2019.01199.

15. Wu, W.; Jin, S. PtrB of Pseudomonas aeruginosa suppresses the type III secretion system under the stress of DNA damage. *J. Bacteriol.* **2005**, *187*, 6058–6068, doi:10.1128/JB.187.17.6058-6068.2005.

16. Molina-Mora, J.A.; Chinchilla, D.; Chavarría, M.; Ulloa, A.; Campos-Sanchez, R.; Mora-Rodríguez, R.A.; Shi, L.; García, F. Transcriptomic determinants of the response of ST-111 Pseudomonas aeruginosa AG1 to ciprofloxacin identified by a top-down systems biology approach. *Sci. Rep.* **2020**, *10*, 1–23, doi:10.1038/s41598-020-70581-2.

17. Schulz, S.; Eckweiler, D.; Bielecka, A.; Nicolai, T.; Franke, R.; Dötsch, A.; Hornischer, K.; Bruchmann, S.; Düvel, J.; Häussler, S. Elucidation of Sigma Factor-Associated Networks in Pseudomonas aeruginosa Reveals a Modular Architecture with Limited and Function-Specific Crosstalk. *PLoS Pathog.* **2015**, *11*, 1–21, doi:10.1371/journal.ppat.1004744.

18. Brazas, M.D.; Brazas, M.D.; Hancock, R.E.W.; Hancock, R.E.W. Ciprofloxacin Induction of a Susceptibility

614 Determinant in Pseudomonas aeruginosa. *Antimicrob. Agents Chemother.* **2005**, *49*, 3222–3227,
615 doi:10.1128/AAC.49.8.3222.
616 19. Cirz, R.T.; O'Neill, B.M.; Hammond, J.A.; Head, S.R.; Romesberg, F.E. Defining the Pseudomonas aeruginosa SOS
617 response and its role in the global response to the antibiotic ciprofloxacin. *J. Bacteriol.* **2006**, *188*, 7101–7110,
618 doi:10.1128/JB.00807-06.
619 20. Molina-Mora, J.A.; Campos-Sanchez, R.; Garcia, F. Gene Expression Dynamics Induced by Ciprofloxacin and Loss
620 of Lexa Function in Pseudomonas aeruginosa PAO1 Using Data Mining and Network Analysis. In Proceedings of
621 the 2018 IEEE International Work Conference on Bioinspired Intelligence (IWOBI); IEEE, 2018; pp. 1–7.
622 21. Anupama, R.; Sajitha Lulu, S.; Mukherjee, A.; Babu, S. Cross-regulatory network in Pseudomonas aeruginosa
623 biofilm genes and TiO 2 anatase induced molecular perturbations in key proteins unraveled by a systems
624 biology approach. *Gene* **2018**, *647*, 289–296, doi:10.1016/j.gene.2018.01.042.
625 22. Fang, G.; Passalacqua, K.D.; Hocking, J.; Llopis, P.M.; Gerstein, M.; Bergman, N.H.; Jacobs-Wagner, C.
626 Transcriptomic and phylogenetic analysis of a bacterial cell cycle reveals strong associations between gene co-
627 expression and evolution. *BMC Genomics* **2013**, *14*, doi:10.1186/1471-2164-14-450.
628 23. Linde, J.; Schulze, S.; Henkel, S.G.; Guthke, R. Data- and knowledge-based modeling of gene regulatory
629 networks: An update. *EXCLI J.* **2015**, *14*, 346–378, doi:10.17179/excli2015-168.
630 24. Langfelder, P.; Horvath, S. WGCNA: An R package for weighted correlation network analysis. *BMC*
631 *Bioinformatics* **2008**, *9*, doi:10.1186/1471-2105-9-559.
632 25. Dai, H.; Zhou, J.; Zhu, B. Gene co-expression network analysis identifies the hub genes associated with immune
633 functions for nocturnal hemodialysis in patients with end-stage renal disease. *Med. (United States)* **2018**, *97*, 1–
634 8, doi:10.1097/MD.0000000000012018.
635 26. Lovelace, A.H.; Smith, A.; Kvitko, B.H. Pattern-triggered immunity alters the transcriptional regulation of
636 virulence-associated genes and induces the sulfur starvation response in pseudomonas syringae pv. tomato
637 DC3000. *Mol. Plant-Microbe Interact.* **2018**, *31*, 750–765, doi:10.1094/MPMI-01-18-0008-R.
638 27. Zhao, W.; Chen, J.J.; Perkins, R.; Wang, Y.; Liu, Z.; Hong, H.; Tong, W.; Zou, W.; Metzker, M.; Didelot, X.; et al. A
639 novel procedure on next generation sequencing data analysis using text mining algorithm. *BMC Bioinformatics*
640 **2016**, *17*, 213, doi:10.1186/s12859-016-1075-9.
641 28. Cornforth, D.M.; Dees, J.L.; Ibberson, C.B.; Huse, H.K.; Mathiesen, I.H.; Kirketerp-Møller, K.; Wolcott, R.D.;
642 Rumbaugh, K.P.; Bjarnsholt, T.; Whiteley, M. Pseudomonas aeruginosa transcriptome during human infection.
643 *Proc. Natl. Acad. Sci. U. S. A.* **2018**, *115*, doi:10.1073/pnas.1717525115.

644 29. Glaab, E.; Bacardit, J.; Garibaldi, J.M.; Krasnogor, N. Using rule-based machine learning for candidate disease
645 gene prioritization and sample classification of cancer gene expression data. *PLoS One* **2012**, *7*,
646 doi:10.1371/journal.pone.0039932.
647 30. Leung, R.K.K.; Wang, Y.; Ma, R.C.W.; Luk, A.O.Y.; Lam, V.; Ng, M.; So, W.Y.; Tsui, S.K.W.; Chan, J.C.N. Using a
648 multi-staged strategy based on machine learning and mathematical modeling to predict genotype-phenotype
649 risk patterns in diabetic kidney disease: A prospective case-control cohort analysis. *BMC Nephrol.* **2013**, *14*, 1,
650 doi:10.1186/1471-2369-14-162.
651 31. Tan, J.; Doing, G.; Lewis, K.A.; Price, C.E.; Chen, K.M.; Cady, K.C.; Perchuk, B.; Laub, M.T.; Hogan, D.A.; Greene,
652 C.S. Unsupervised Extraction of Stable Expression Signatures from Public Compendia with an Ensemble of
653 Neural Networks. *Cell Syst.* **2017**, *5*, 63-71.e6, doi:10.1016/j.cels.2017.06.003.
654 32. Kar, S.C. Comparing Prediction Accuracy for Supervised Techniques in Gene Expression Data. *Math. Theory*
655 *Model.* **2014**, *4*, 108–116.
656 33. Raza, K.; Hasan, A. A Comprehensive Evaluation of Machine Learning Techniques for Cancer Class Prediction
657 Based on Microarray Data. *Int. J. Bioinform. Res. Appl.* **2015**, *11*, 397–416, doi:10.1504/IJBRA.2015.071940.
658 34. Savli, H.; Karadenizli, A.; Kolayli, F.; Gundes, S.; Ozbek, U.; Vahaboglu, H. Expression stability of six housekeeping
659 genes: a proposal for resistance gene quantification studies of Pseudomonas aeruginosa by real-time
660 quantitative RT-PCR. *J. Med. Microbiol.* **2003**, *52*, 403–408, doi:10.1099/jmm.0.05132-0.
661 35. Casares, F.M. A Simple Method for Optimization of Reference Gene Identification and Normalization in DNA
662 Microarray Analysis. *Med. Sci. Monit. Basic Res.* **2016**, *22*, 45–52, doi:10.12659/MSMBR.897644.
663 36. Yang, P.; Zhou, B.B.; Yang, J.Y.-H.; Zomaya, A.Y. Stability of feature selection algorithms and ensemble feature
664 selection methods in bioinformatics. *Biol. Knowl. Discov. Handb. Preprocessing, mining, postprocessing Biol.*
665 *data* **2013**, 333–352, doi:10.1002/9781118617151.ch14.
666 37. Tabe-Bordbar, S.; Emad, A.; Zhao, S.D.; Sinha, S. A closer look at cross-validation for assessing the accuracy of
667 gene regulatory networks and models. *Sci. Rep.* **2018**, *8*, 1–11, doi:10.1038/s41598-018-24937-4.
668 38. Alqarni, B.; Colley, B.; Klebensberger, J.; McDougald, D.; Rice, S.A. Expression stability of 13 housekeeping genes
669 during carbon starvation of Pseudomonas aeruginosa. *J. Microbiol. Methods* **2016**, *127*, 182–187,
670 doi:10.1016/j.mimet.2016.06.008.
671 39. Noi, P.T.; Kappas, M. Comparison of random forest, k-nearest neighbor, and support vector machine classifiers
672 for land cover classification using sentinel-2 imagery. *Sensors (Switzerland)* **2018**, *18*, doi:10.3390/s18010018.
673 40. Li, L.; Weinberg, C.R.; Darden, T.A.; Pedersen, L.G. Gene selection for sample classification based on gene

674 expression data: study of sensitivity to choice of parameters of the GA/KNN method. *Bioinformatics* **2001**, *17*,
675 1131–42.
676 41. Vapnik, V.N.; Vladimir *Estimation of dependences based on empirical data*; Springer-Verlag, 1982; ISBN
677 0387907335.
678 42. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32, doi:10.1023/A:1010933404324.
679 43. Yusuf Nalca; Lothar Jänsch; Florian Bredenbruch; Robert Geffers, J.B.; Susanne Häussler Quorum-sensing
680 antagonistic activities of azithromycin in Pseudomonas aeruginosa PAO1: a global approach. *Antimicrob Agents*
681 *Chemother* **2008**, *50*, 1680–1688, doi:10.1128/AAC.50.5.1680.
682 44. Kai, T.; Tateda, K.; Kimura, S.; Ishii, Y.; Ito, H.; Yoshida, H.; Kimura, T.; Yamaguchi, K. A low concentration of
683 azithromycin inhibits the mRNA expression of N-acyl homoserine lactone synthesis enzymes, upstream of lasI or
684 rhlI, in Pseudomonas aeruginosa. *Pulm. Pharmacol. Ther.* **2009**, *22*, 483–486, doi:10.1016/j.pupt.2009.04.004.
685 45. Chang, W.; Small, D.A.; Toghrol, F.; Bentley, W.E. Microarray analysis of Pseudomonas aeruginosa reveals
686 induction of pyocin genes in response to hydrogen peroxide. *BMC Genomics* **2005**, *6*, 1–14, doi:10.1186/1471-
687 2164-6-115.
688 46. Teitzel, G.M.; Geddie, A.; De Long, S.K.; Kirisits, M.J.; Whiteley, M.; Parsek, M.R. Survival and Growth in the
689 Presence of Elevated Copper: Transcriptional Profiling of Copper-Stressed Pseudomonas aeruginosa. *J.*
690 *Bacteriol.* **2006**, *188*, 7242–7256, doi:10.1128/JB.00837-06.
691 47. Small, D.A.; Chang, W.; Toghrol, F.; Bentley, W.E. Comparative global transcription analysis of sodium
692 hypochlorite, peracetic acid, and hydrogen peroxide on Pseudomonas aeruginosa. *Appl. Microbiol. Biotechnol.*
693 **2007**, *76*, 1093–1105, doi:10.1007/s00253-007-1072-z.
694 48. Nde, C.W.; Jang, H.-J.; Toghrol, F.; Bentley, W.E. Toxicogenomic response of Pseudomonas aeruginosa to ortho-
695 phenylphenol. *BMC Genomics* **2008**, *9*, 473, doi:10.1186/1471-2164-9-473.
696 49. Cummins, J.; Reen, F.J.; Baysse, C.; Mooij, M.J.; O'Gara, F. Subinhibitory concentrations of the cationic
697 antimicrobial peptide colistin induce the pseudomonas quinolone signal in Pseudomonas aeruginosa.
698 *Microbiology* **2009**, *155*, 2826–2837, doi:10.1099/mic.0.025643-0.
699 50. Nde, C.W.; Jang, H.J.; Toghrol, F.; Bentley, W.E. Global transcriptomic response of Pseudomonas aeruginosa to
700 chlorhexidine diacetate. *Environ. Sci. Technol.* **2009**, *43*, 8406–8415, doi:10.1021/es9015475.
701 51. Pan, J.; Bahar, A.A.; Syed, H.; Ren, D. Reverting Antibiotic Tolerance of Pseudomonas aeruginosa PAO1 Persister
702 Cells by (Z)-4-bromo-5-(bromomethylene)-3-methylfuran-2(5H)-one. *PLoS One* **2012**, *7*,
703 doi:10.1371/journal.pone.0045778.

704 52. Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N.S.; Wang, J.T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T.
705 Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.*
706 **2003**, *13*, 2498–2504, doi:10.1101/gr.1239303.
707 53. Chin, C.-H.; Chen, S.-H.; Wu, H.-H.; Ho, C.-W.; Ko, M.-T.; Lin, C.-Y. cytoHubba: identifying hub objects and sub-
708 networks from complex interactome. *BMC Syst. Biol.* **2014**, *8 Suppl 4*, S11, doi:10.1186/1752-0509-8-S4-S11.
709 54. Smyth, G.K. Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray
710 Experiments. *Stat. Appl. Genet. Mol. Biol.* **2004**, *3*, 1–25, doi:10.2202/1544-6115.1027.
711 55. Richter, K.; Haslbeck, M.; Buchner, J. The Heat Shock Response: Life on the Verge of Death. *Mol. Cell* **2010**, *40*,
712 253–266, doi:10.1016/j.molcel.2010.10.006.
713 56. Vihervaara, A.; Duarte, F.M.; Lis, J.T. Molecular mechanisms driving transcriptional stress responses. *Nat. Rev.*
714 *Genet.* **2018**, *19*, 385–397, doi:10.1038/s41576-018-0001-6.
715 57. Holloway, B.W. Genetic Recombination in Pseudomonas aeruginosa. *Microbiology* **1955**, *13*, 572–581,
716 doi:10.1099/00221287-13-3-572.
717 58. Toval, F.; Guzmán-Marte, A.; Madriz, V.; Somogyi, T.; Rodríguez, C.; García, F. Predominance of carbapenem-
718 resistant Pseudomonas aeruginosa isolates carrying blaIMP and blaVIM metallo-β-lactamases in a major
719 hospital in Costa Rica. *J. Med. Microbiol.* **2015**, *64*, 37–43, doi:10.1099/jmm.0.081802-0.
720 59. World Health Organization *Guidelines for the prevention and control of carbapenem-resistant*
721 *Enterobacteriaceae, Acinetobacter baumannii and Pseudomonas aeruginosa in health care facilities*; Geneva,
722 2017;
723 60. Molina-Mora, J.-A.; Garcia-Batan, R.; Garcia, F. Genomic context of the two integrons of ST-111 Pseudomonas
724 aeruginosa AG1: A VIM-2-carrying old-acquaintance and a novel IMP-18-carrying integron. *Res. Sq.* **2020**,
725 doi:10.21203/RS.3.RS-41474/V1.
726 61. Molina-Mora, J.A.; Chinchilla-Montero, D.; Castro-Peña, C.; García, F. Two-dimensional gel electrophoresis (2D-
727 GE) image analysis based on CellProfiler. *Medicine (Baltimore).* **2020**, *99*, e23373,
728 doi:10.1097/MD.0000000000023373.
729 62. Bustin, S.A.; Benes, V.; Nolan, T.; Pfaffl, M.W. Quantitative real-time RT-PCR – a perspective. *J. Mol. Endocrinol.*
730 **2005**, *34*, 597–601, doi:10.1677/jme.1.01755.
731 63. Piao, J.; Sun, J.; Yang, Y.; Jin, T.; Chen, L.; Lin, Z. Target gene screening and evaluation of prognostic values in
732 non-small cell lung cancers by bioinformatics analysis. *Gene* **2018**, *647*, 306–311,
733 doi:10.1016/j.gene.2018.01.003.

734 64. Touw, W.G.; Bayjanov, J.R.; Overmars, L.; Backus, L.; Boekhorst, J.; Wels, M.; van Hijum, S.A.F.T. Data mining in
735 the Life Sciences with Random Forest: a walk in the park or lost in the jungle? *Brief. Bioinform.* **2013**, *14*, 315–
736 326, doi:10.1093/bib/bbs034.
737 65. Pai, S.; Hui, S.; Isserlin, R.; Shah, M.A.; Kaka, H.; Bader, G.D. netDx: interpretable patient classification using
738 integrated patient similarity networks. *Mol. Syst. Biol.* **2019**, *15*, e8497, doi:10.15252/msb.20188497.
739 66. Park, H.; Shimamura, T.; Imoto, S.; Miyano, S. Adaptive NetworkProfiler for Identifying Cancer Characteristic-
740 Specific Gene Regulatory Networks. *J. Comput. Biol.* **2017**, *25*, cmb.2017.0120, doi:10.1089/cmb.2017.0120.
741 67. Falciani, F.; Diab, A.; Sabine, V.; Williams, T.; Ortega, F.; George, S.; Chipman, J. Hepatic transcriptomic profiles
742 of European flounder (Platichthys flesus) from field sites and computational approaches to predict site from
743 stress gene responses following exposure to model toxicants. *Aquat. Toxicol.* **2008**, *90*, 92–101,
744 doi:10.1016/j.aquatox.2008.07.020.
745 68. Nota, B.; Verweij, R.A.; Molenaar, D.; Ylstra, B.; van Straalen, N.M.; Roelofs, D. Gene Expression Analysis Reveals
746 a Gene Set Discriminatory to Different Metals in Soil. *Toxicol. Sci.* **2010**, *115*, 34–40, doi:10.1093/toxsci/kfq043.
747 69. Liu, W.; Li, L.; Long, X.; You, W.; Zhong, Y.; Wang, M.; Tao, H.; Lin, S.; He, H. Construction and Analysis of Gene
748 Co-Expression Networks in Escherichia coli. *Cells* **2018**, *7*, 19, doi:10.3390/cells7030019.
749 70. Krämer, R. Bacterial stimulus perception and signal transduction: Response to osmotic stress. *Chem. Rec.* **2010**,
750 *10*, 217–229, doi:10.1002/tcr.201000005.
751 71. Schurig-Briccio, L.A.; Farías, R.N.; Rodríguez-Montelongo, L.; Rintoul, M.R.; Rapisarda, V.A. Protection against
752 oxidative stress in Escherichia coli stationary phase by a phosphate concentration-dependent genes expression.
753 *Arch. Biochem. Biophys.* **2009**, *483*, 106–110, doi:10.1016/j.abb.2008.12.009.
754 72. Parsons, J.B.; Rock, C.O. Bacterial lipids: metabolism and membrane homeostasis. *Prog. Lipid Res.* **2013**, *52*,
755 249–76, doi:10.1016/j.plipres.2013.02.002.
756 73. Storvik, K.A.M.; Foster, P.L. RpoS, the stress response sigma factor, plays a dual role in the regulation of
757 Escherichia coli's error-prone DNA polymerase IV. *J. Bacteriol.* **2010**, *192*, 3639–44, doi:10.1128/JB.00358-10.
758 74. Starosta, A.L.; Lassak, J.; Jung, K.; Wilson, D.N. The bacterial translation stress response. *FEMS Microbiol. Rev.*
759 **2014**, *38*, 1172–1201, doi:10.1111/1574-6976.12083.
760 75. Adamson, B.; Norman, T.M.; Jost, M.; Cho, M.Y.; Nuñez, J.K.; Chen, Y.; Villalta, J.E.; Gilbert, L.A.; Horlbeck, M.A.;
761 Hein, M.Y.; et al. A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the
762 Unfolded Protein Response. *Cell* **2016**, *167*, 1867-1882.e21, doi:10.1016/j.cell.2016.11.048.
763 76. Molina-Mora, J.A.; Kop-Montero, M.; Quirós-Fernández, I.; Quiros, S.; Crespo-Mariño, J.L.; Mora-Rodríguez, R.A.

764 A hybrid mathematical modeling approach of the metabolic fate of a fluorescent sphingolipid analogue to
765 predict cancer chemosensitivity. *Comput. Biol. Med.* **2018**, *97*, 8–20, doi:10.1016/j.compbiomed.2018.04.008.
766

767 **LEGENDS OF FIGURES AND LEGENDS**

768

769 **Figure 1. General pipeline for identifying core response genes in *P. aeruginosa* by a machine learning**
770 **approach.** (A) Schematic representation of hypothesis for identifying core response determinants when
771 bacteria are exposed to multiple perturbations. (B) The workflow of the machine learning approach using
772 microarray data and model fitting by SP and MP methods for identifying and describing core response genes.

773 **Figure 2. Data selection and partitioning.** (A) Microarray-data selection in GEO-NCBI database according to
774 experimental design. (B) Samples, including replicates in each experiment -perturbation and control-, were
775 separated into the training or the testing data sets using the SP (the last replicate of each condition was moved
776 to the testing dataset) or MP (random separation with multiple arrangements) methods. (C) Representation
777 of classification procedures for the SP and MP methods, including internal 10-fold cross-validation for the
778 training data set.

779 **Figure 3. Comparison of samples by global profiles with transcriptomic data of 5549 genes of *P. aeruginosa*.**
780 (A) Dispersion of intensities of samples, showing similar distribution between samples after normalization. (B)
781 Global profiles of samples were compared by PCA, using all the 5549 genes in the microarray.

782 **Figure 4**. **Evaluation of SP method for gene ranking by importance, the case for RF algorithm.** (A) Accuracy,
783 F1, and kappa values after iterations of classification with the first top 200 genes (adding genes 1-to-1). (B)
784 ROC plot using selected top 56 genes for evaluation of the performance of the algorithm. (C) Ranking and
785 importance value of top 56 genes. Similar results are shown for SVM and KNN algorithms in supplementary
786 Figure S1.

787 **Figure 5**. **Evaluation of MP method for gene ranking by frequency, case for SVM algorithm.** (A) Ranking of
788 top 56 genes by frequency after iterations of 100 data partitions and classification model fitting. (B) Dispersion
789 of metrics across 100 iterations. Similar results are shown for KNN and RF algorithms in supplementary Figure
790 S2.

791 **Figure 6. Comparison of 118 candidate genes of the core perturbome.** (A) The top 56 genes of each
792 algorithm were selected and compared using upset plots, showing the intersection of the lists
793 among classifiers. This comparison resulted in 118 distinct candidate genes. A gene was selected in
794 the final list if it was found at least by four different algorithms, i.e. the support was ≥ 4. (B) Gene
795 annotation using KEGG shows the enriched pathways found for the core perturbome genes.

796 **Figure 7. Network analysis and comparison of global transcriptomic profiles using the 46 core perturbome**
797 **genes.** (A) Distribution and support (number of algorithms) of preliminary 118 selected genes on a basal large-
798 scale network of functional associations in *P. aeruginosa.* Topological metrics of the network and identity of
799 hub genes (circles with light-blue border) are included. (B) Small-world network showing relationships
800 between the 46 core response genes and connectivity metric for each gene. (C) The global profile of samples
801 by PCA, showing the separation of conditions. For reference, centroids were plotted as triangles in each
802 cluster. More details are shown in Supplementary Figure S3.

803 **Figure 8. Annotation of core response genes and comparison with a differential expression analysis.** (A)
804 General annotation profile of identified genes by core response genes showing associated biological
805 processes. (B) Comparison of identified genes by our machine learning approach and DEGs lists. (C) General
806 annotation of DEGs showing a similar profile to our approach.

807 **Table 1.** Annotation of genes of the core perturbome *Pseudomonas aeruginosa*

808

809 **Supplementary Figures and Tables**

810 **Figure S1**. **Evaluation of SP method for gene ranking by importance, the case for SVM and KNN algorithms.**
811 (A-C) Accuracy, F1, and kappa values after iterations of classification with the first top 200 genes (adding genes
812 1-to-1). (B-D) ROC plot using selected top 56 genes for evaluation of the performance of the algorithm. (E)
813 Ranking and importance value of top 56 genes (same importance for both algorithms).

814 **Figure S2**. **Evaluation of MP method for gene ranking by frequency, the case for KNN and RF algorithms.** (A)
815 KNN evaluation of ranking of top 56 genes by frequency after iterations of 100 data partitions and
816 classification model fitting and dispersion of metrics across 100 iterations. (B) Same evaluation, the case for
817 RF algorithm.

818 **Figure S3. Systems-level description and gene expression comparison by core response genes.** (A) Gene
819 expression and clustering of the 46 core response determinants of *P. aeruginosa*. (B-C) Comparison of gene
820 expression (logR) of kmedoids for each cluster using small-world networks.

821 **Table S1.** Studies of datasets included in this work to identify the core perturbome of
822 *Pseudomonas aeruginosa*

823 **Table S2.** Metrics associated with the performance of each method and algorithm

824

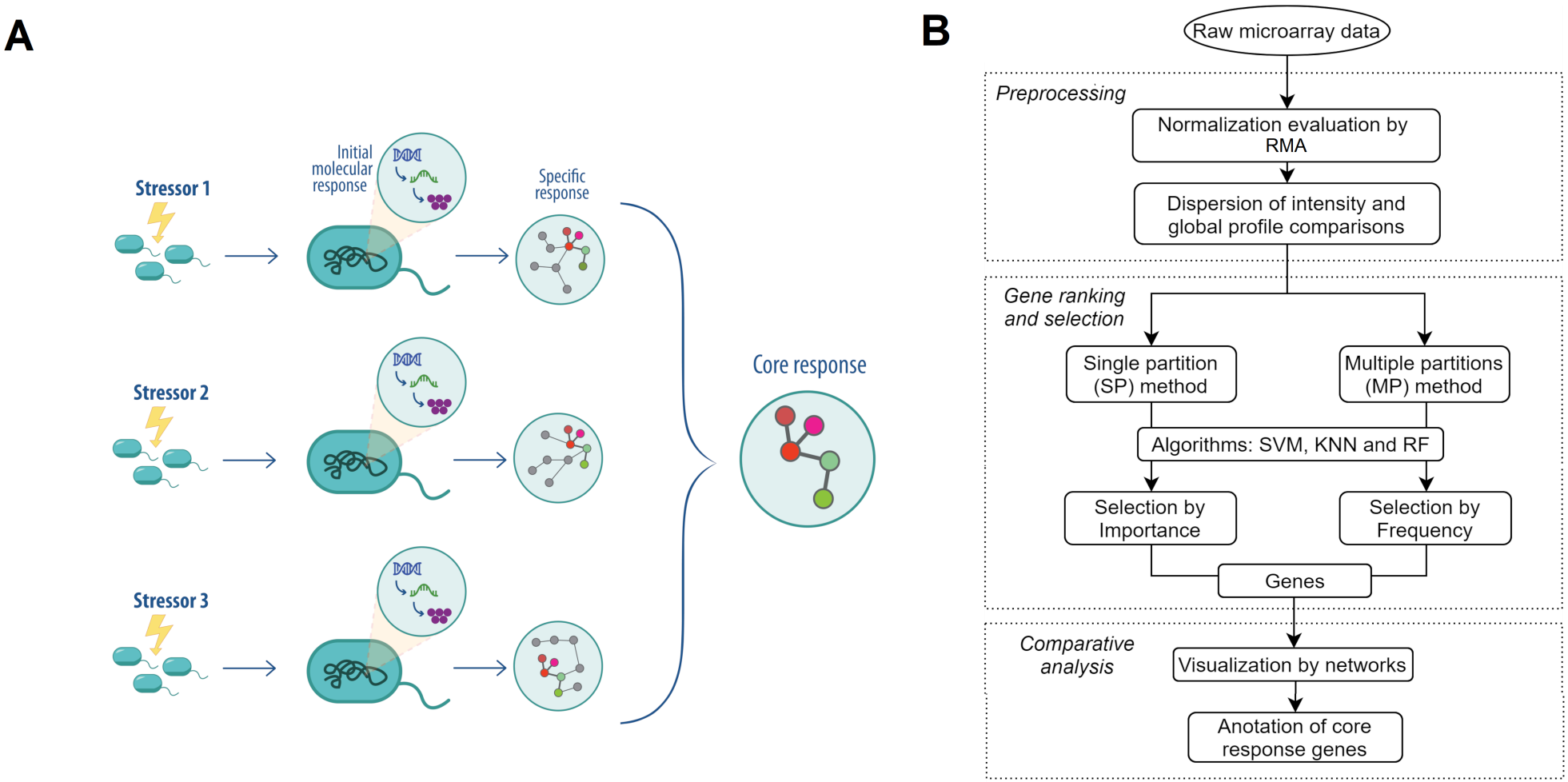
A
Stressor 1
Initial molecular response
Specific response
Stressor 2
Stressor 3
Core response
B
Raw microarray data
Preprocessing
Normalization evaluation by RMA
Dispersion of intensity and global profile comparisons
Gene ranking and selection
Single partition (SP) method
Multiple partitions (MP) method
Algorithms: SVM, KNN and RF
Selection by Importance
Selection by Frequency
Genes
Comparative analysis
Visualization by networks
Anotation of core response genes

Figure 1

A
GEO-NCBI
First selection by organism and technology:
P. aeruginosa PAO1 and microarray Accession GPL84
156 Series with 1310 samples
Second selection by experimental design:
Exposition to perturbation
47 Series with 347 samples
Final selection by experimental conditions:
Control conditions available and culture conditions
including LB medium
10 Series with 71 samples
B
Samples:
Perturbation 1:
Antibiotic A
+ Controls
Experiment 1
R1 R2 R3
Perturbation 2:
Chemical compound B
+ Controls
Experiment 2
R1 R2 R3 R4
Perturbation 3:
Antibiotic C, time X
Perturbation 4:
Antibiotic C, time Y
Experiment 3
R1 R2 R3
R1 R2 R3
+ Controls
+ Controls
SP Method: last replicate is moved to the testing dataset
Training dataset
Testing dataset
R1 R2 R1 R2 R3 R1 R2 R1 R2
R3 R4 R3 R3
MP Method: all samples are randomly separated into the
training or testing data sets
C
Data partition methods
Training
Testing
Single partition (SP) method
Dataset
Multiple partitions (MP) method
Training
Testing
Model fitting and evaluation
Training
Training
fold
Validation
fold
10-fold cross validation
Testing
Fitted
model
Evaluation by
metrics and
gene ranking

Figure 2

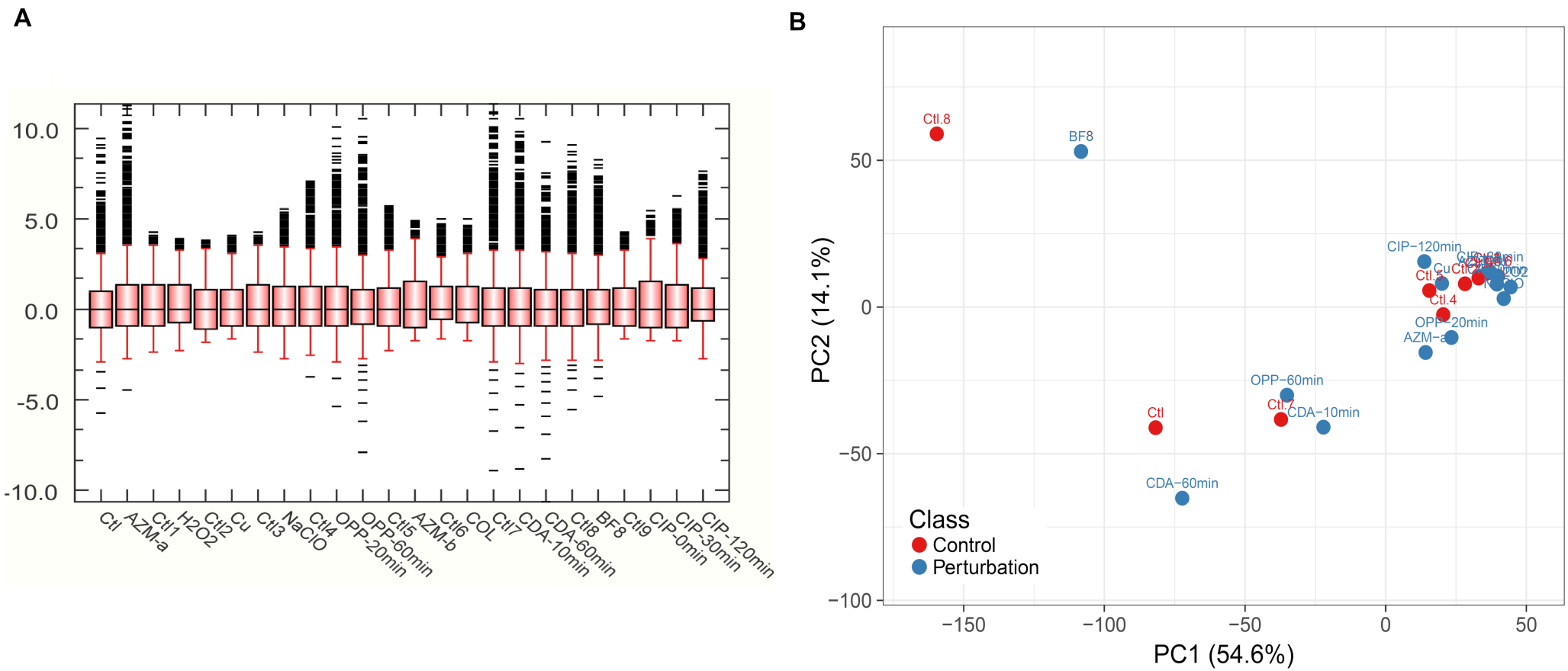
A
10.0
5.0
0.0
-5.0
-10.0
Ctl
AZM-a
Ctl1
H2O2
Ctl2
Cu
Ctl3
NaClO
Ctl4
OPP-20min
OPP-60min
Ctl5
AZM-b
Ctl6
COL
Ctl7
CDA-10min
CDA-60min
Ctl8
BF8
Ctl9
CIP-0min
CIP-30min
CIP-120min
B
PC2 (14.1%)
PC1 (54.6%)
50
0
−50
−100
−150
−100
−50
0
50
Ctl.8
BF8
CIP−120min
Ctl.5
Ctl.4
OPP−20min
AZM−a
OPP−60min
Ctl.7
CDA−10min
Ctl
CDA−60min
Class
Control
Perturbation

Figure 3

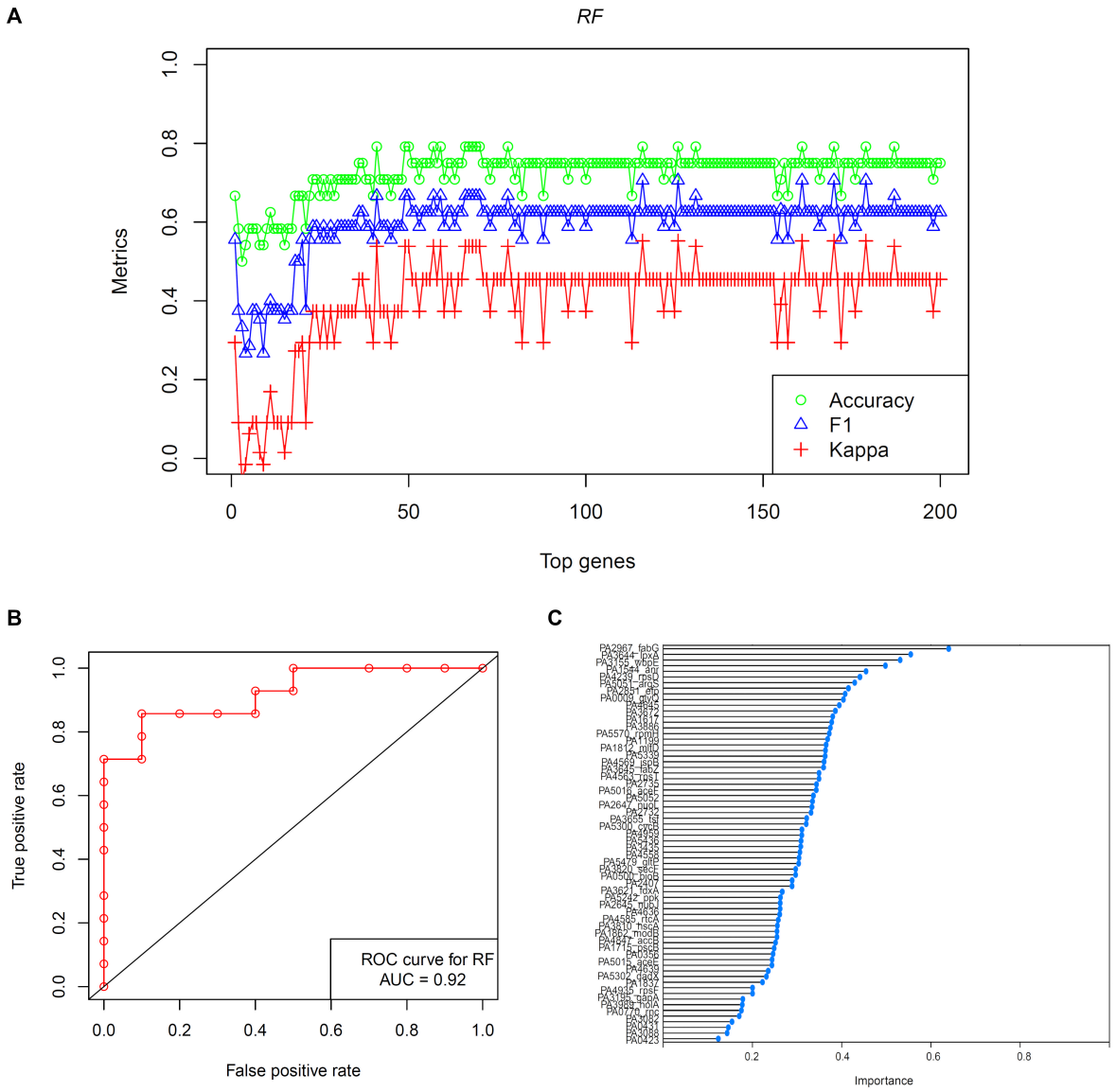
A
RF
Metrics
Top genes
Accuracy
F1
Kappa
B
True positive rate
False positive rate
ROC curve for RF
AUC = 0.92
C
Importance

Figure 4

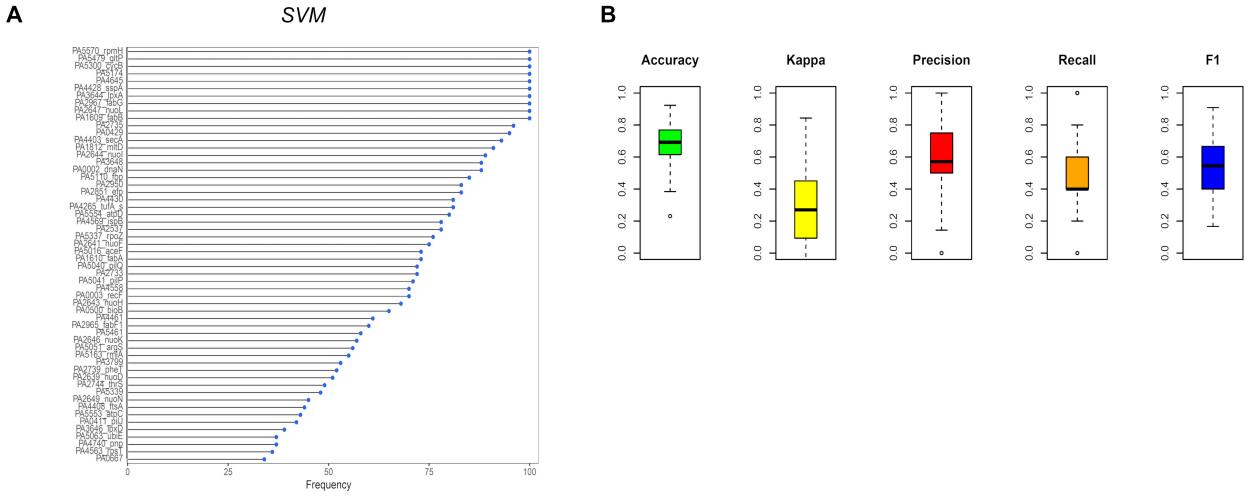
A
SVM
PA5570_rpmH
PA5479_glfP
PA5300_ftsY
PA3114
PA4464_A
PA4428_aspA
PA3844_fpvA
PA2957_fabG
PA2647_nuoL
PA1602_fabB
PA2135
PA0429
PA4403_secA
PA1812_mltD
PA2844_rpoB
PA5648
PA0002_dnaN
PA5114_fis
PA2496_
PA2851_dsbA
PA4265_tuf1_2
PA5584_atpD
PA4568_isbB
PA2312
PA5337_rpoZ
PA2641_nuoF
PA5016_aceF
PA1810_fabA
PA5047_gltQ
PA2733
PA5041_gltP
PA4559
PA0003_recF
PA2643_nuoB
PA5500_
PA2268
PA2965_fabF
PA5461
PA2646_nuoK
PA5015_aceE
PA5163_mtlA
PA3799
PA2739_ghrB
PA2639_nuoD
PA2714_thrC
PA5339
PA2649_nuoN
PA4408_ftsA
PA4553_atpC
PA0411_gcvT
PA3646_tkrB
PA5063_tdcE
PA4740_rnpA
PA4583_0981
PA0667
0
25
50
75
100
Frequency
B
Accuracy
Kappa
Precision
Recall
F1
0.0
0.2
0.4
0.6
0.8
1.0

Figure 5

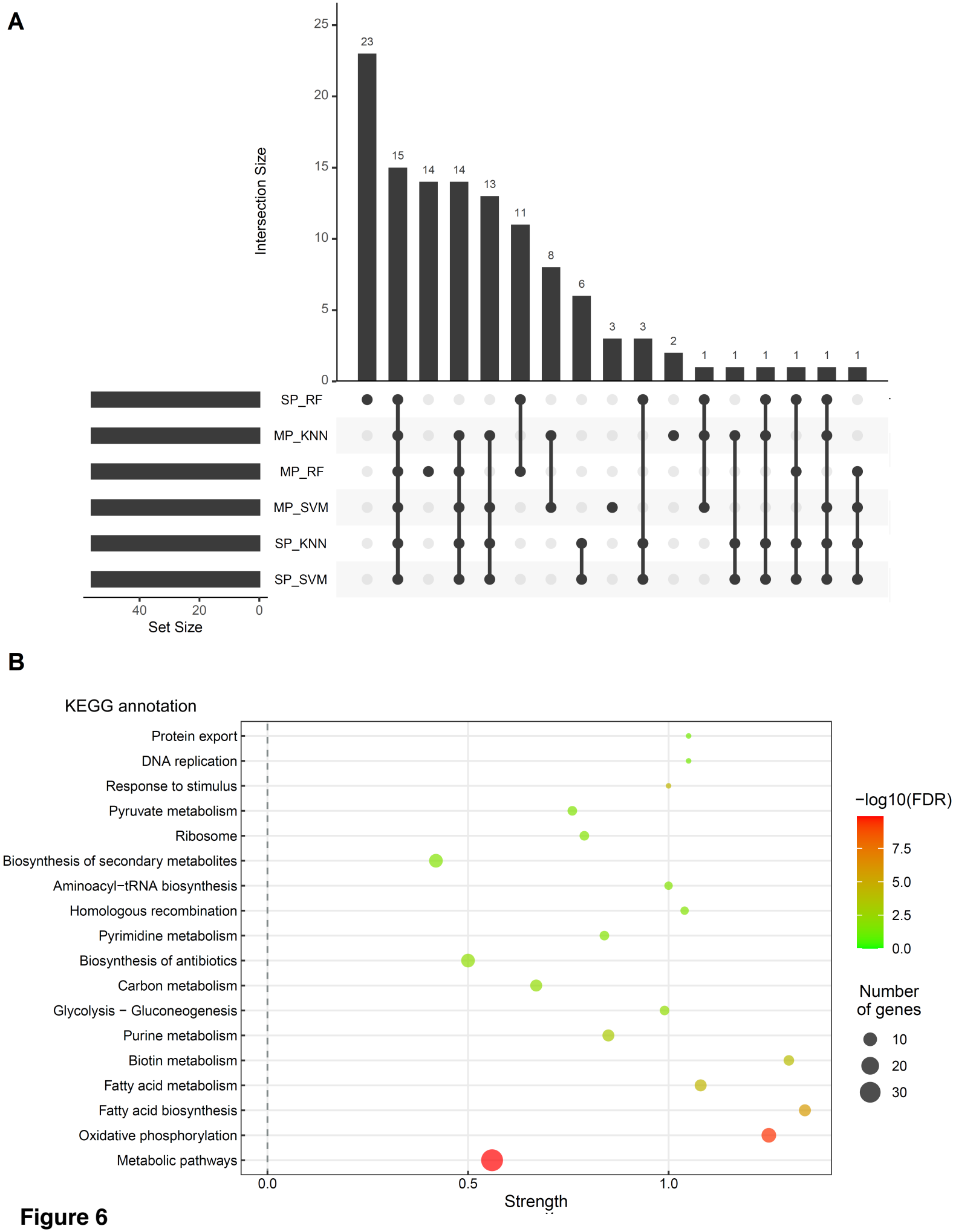
A
Intersection Size
25
20
15
10
5
0
23
15
14
14
13
11
8
6
3
3
2
1
1
1
1
1
1
SP_RF
MP_KNN
MP_RF
MP_SVM
SP_KNN
SP_SVM
40
20
0
Set Size
B
KEGG annotation
Protein export
DNA replication
Response to stimulus
Pyruvate metabolism
Ribosome
Biosynthesis of secondary metabolites
Aminoacyl−tRNA biosynthesis
Homologous recombination
Pyrimidine metabolism
Biosynthesis of antibiotics
Carbon metabolism
Glycolysis − Gluconeogenesis
Purine metabolism
Biotin metabolism
Fatty acid metabolism
Fatty acid biosynthesis
Oxidative phosphorylation
Metabolic pathways
0.0
0.5
1.0
Strength
−log10(FDR)
7.5
5.0
2.5
0.0
Number of genes
10
20
30

Figure 6

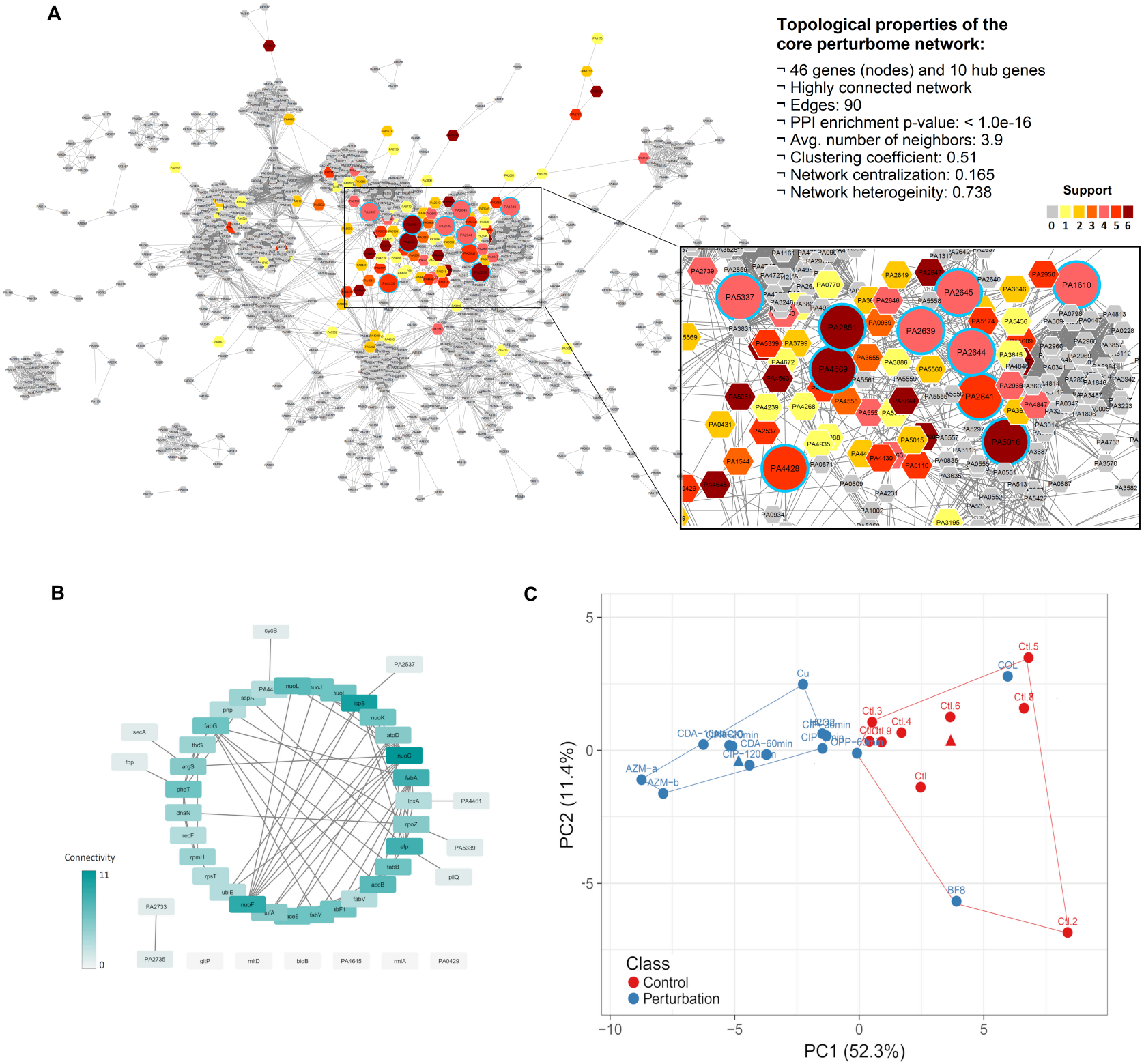
A
Topological properties of the core perturbome network:
¬ 46 genes (nodes) and 10 hub genes
¬ Highly connected network
¬ Edges: 90
¬ PPI enrichment p-value: < 1.0e-16
¬ Avg. number of neighbors: 3.9
¬ Clustering coefficient: 0.51
¬ Network centralization: 0.165
¬ Network heterogeinity: 0.738
Support
0 1 2 3 4 5 6
B
Connectivity
11
0
C
PC2 (11.4%)
PC1 (52.3%)
Class
Control
Perturbation

Figure 7

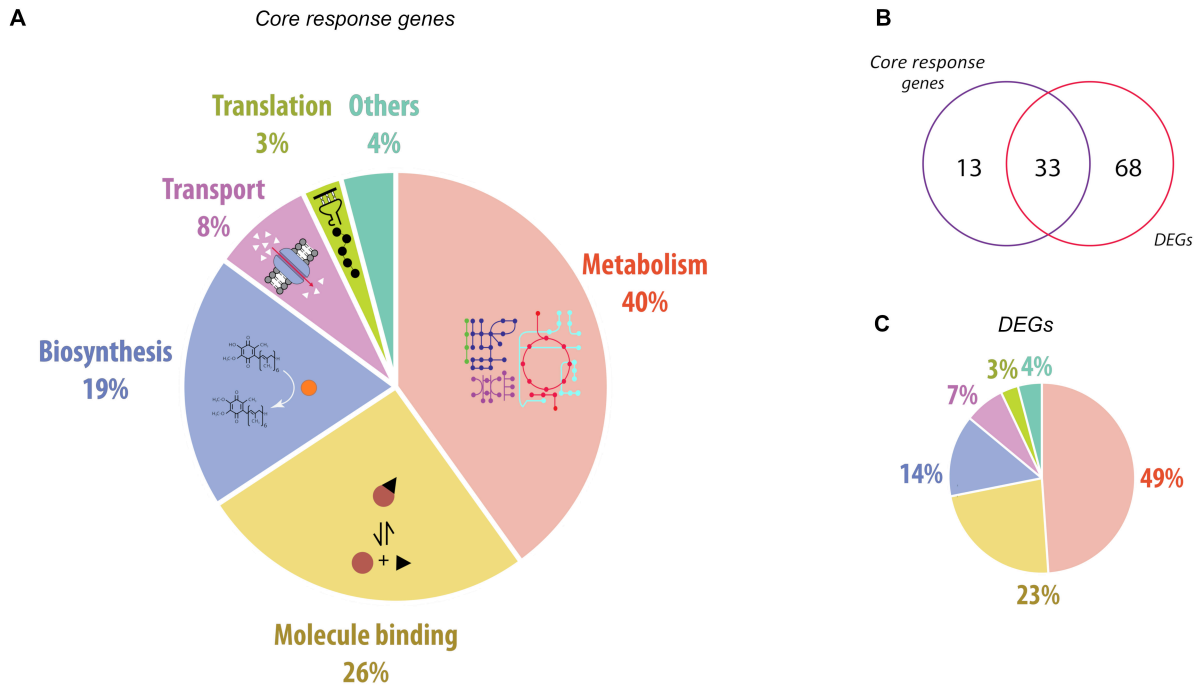
A
Core response genes
Translation
3%
Others
4%
Transport
8%
Metabolism
40%
Biosynthesis
19%
Molecule binding
26%
B
Core response genes
13
33
68
DEGs
C
DEGs
3%
4%
7%
14%
49%
23%

Figure 8

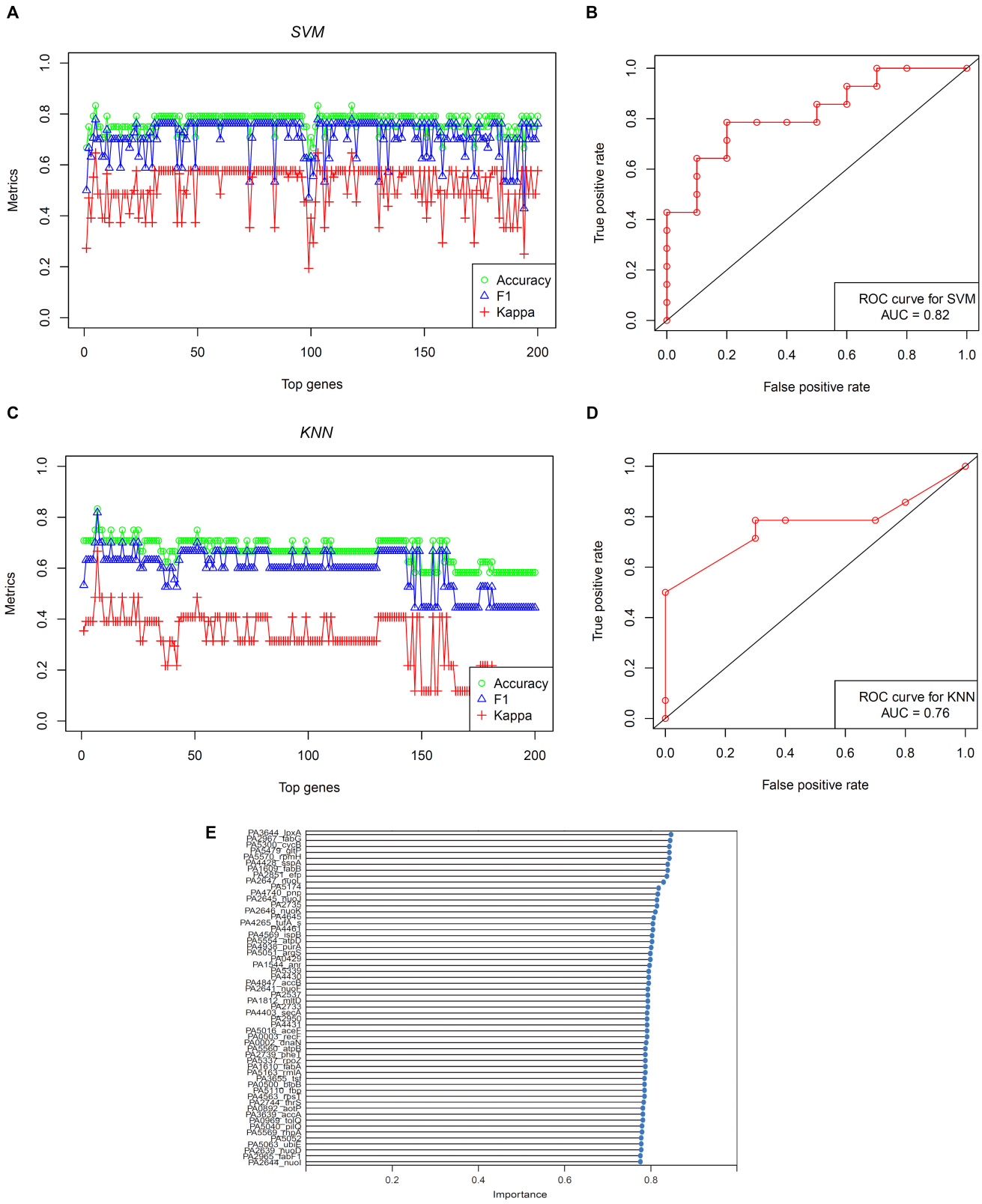
A
SVM
Metrics
Top genes
Accuracy
F1
Kappa
B
True positive rate
False positive rate
ROC curve for SVM
AUC = 0.82
C
KNN
Metrics
Top genes
Accuracy
F1
Kappa
D
True positive rate
False positive rate
ROC curve for KNN
AUC = 0.76
E
Importance

Figure S1

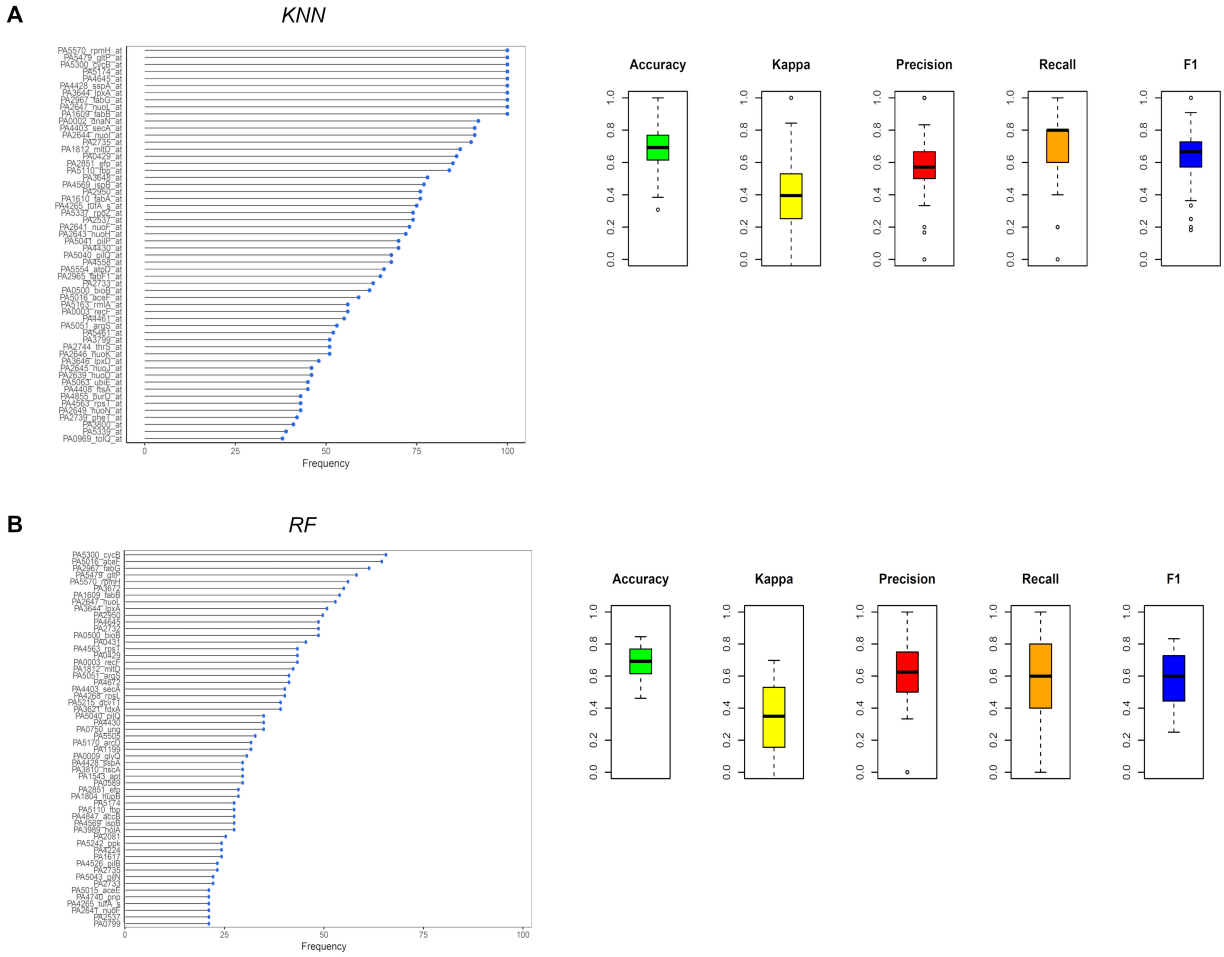
A
KNN
Frequency
Accuracy
Kappa
Precision
Recall
F1
B
RF
Frequency
Accuracy
Kappa
Precision
Recall
F1

Figure S2

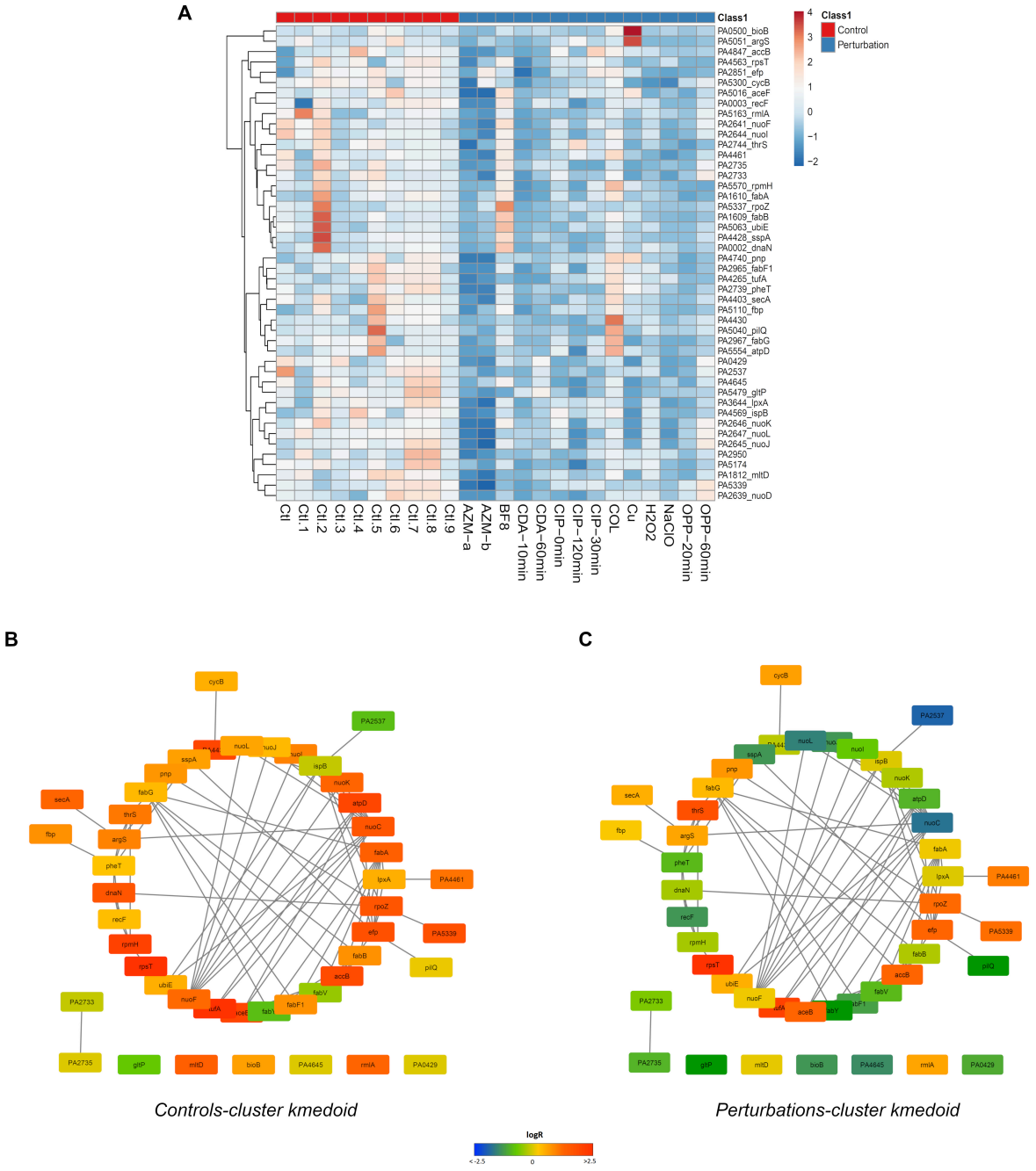
A
Class1
Control
Perturbation
PA0500_bioB
PA5051_argS
PA4847_accB
PA4563_rpsT
PA2851_efp
PA5300_cycB
PA5016_aceF
PA0003_recF
PA5163_rmlA
PA2641_nuoF
PA2644_nuoI
PA2744_thrS
PA4461
PA2735
PA2733
PA5570_rpmH
PA1610_fabA
PA5337_rpoZ
PA1609_fabB
PA5063_ubiE
PA4428_sspA
PA0002_dnaN
PA4740_pnp
PA2965_fabF1
PA4265_tufA
PA2739_pheT
PA4403_secA
PA5110_fbp
PA4430
PA5040_pilQ
PA2967_fabG
PA5554_atpD
PA0429
PA2537
PA4645
PA5479_gltP
PA3644_lpxA
PA4569_ispB
PA2646_nuoK
PA2647_nuoL
PA2645_nuoJ
PA2950
PA5174
PA1812_mltD
PA5339
PA2639_nuoD
Ctl
Ctl.1
Ctl.2
Ctl.3
Ctl.4
Ctl.5
Ctl.6
Ctl.7
Ctl.8
Ctl.9
AZM−a
AZM−b
BF8
CDA−10min
CDA−60min
CIP−0min
CIP−120min
CIP−30min
COL
Cu
H2O2
NaClO
OPP−20min
OPP−60min
B
C
Controls-cluster kmedoid
Perturbations-cluster kmedoid
logR

Figure S3

Table 1. Annotation of genes of the core perturbome of *Pseudomonas aeruginosa*

Support*	Code**	Name	Top annotated functions	DNA damage	DNA repair	Stimulus response	Other stimulus	Differential expression*** (↑ upregulated, ↓downregulated and ↑↓ variable regulation)
6	PA0500	*bioB*	Biotin biosynthesis, catabolic process, diaminopimelate process	x	x	x	heat	↓AZM and ↑↓COL
	PA1812	*mltD*	Microcin transport, cytokinesis, peptidoglycan biosynthesis	-	-	x	drug,heat	↓AZM and ↓COL
	PA2647	*nuoL*	Aerobic respiration, electron/proton transport	x	-	x	response to superoxide, antibiotic	↓NaClO and ↓CIP
	PA2735	-	DNA metabolism, translation, DNA damage	x	x	-		↑Temperature, ↑↓AZM and ↓Cu
	PA2851	***efp***	Peptidyl-prolyl isomerization, DNA damage, ATP (proton)synthesis	x		x	starvation, heat, antibiotic	↑AZM, ↓↑COL and ↓↑CIP
	PA2967	*fabG*	Lipid metabolism, fatty acid and lipid biosynthesis	x	-	x	cold, osmotic stress	↓↑AZM, ↓↑COL and ↓Cu
	PA3644	*lpxA*	Lipid metabolism, fatty acid and peptidoglycan biosynthesis	-	-	x	drug, antibiotic	↓COL, ↓AZM and ↓Cu
	PA4563	*rpsT*	Cold response, ribosomal assembly, transcription antitermination	-	-	x	stress, antibiotic, heat	↓AZM, ↓Cu and ↓↑COL
	PA4569	***ispB***	Ubiquinone biosynthesis, transport, aerobic respiration	x	-	x	stress, antibiotic	↓AZM, ↓↑COL and ↓Cu
	PA4645	-	Enzyme-directed rRNA pseudouridine synthesis, L-serin catabolism, glycine biosynthesis	x	x	-		↑AZM and ↓Cu
	PA5016	***aceF***	Aerobic respiration, electron transport, Krebs cycle	x	-	x	heat, stress	↓AZM, ↓↑CIP, ↓↑COL and ↓Cu
	PA5051	*argS*	DNA damage, antibiotic response, peptidoglycan biosynthesis	x	-	x	heat	↑Cu, ↓COL and ↓AZM
	PA5300	*cycB*	DNA damagage recognition and repair	x	x	-		↓AZM, ↓Cu and ↓COL
	PA5479	*gltP*	DNA damage, transport, fumarate transport	x	-	x	stimulus	↓CIP, ↓Cu and ↓↑COL
	PA5570	*rpmH*	Ribosomal assembly, translation, transcription DNA dependent	x	-	x	drug, cold, antibiotic	↓CIP, ↓Cu and ↓↑AZM, ↓↑COL
5	PA0003	*recF*	Response to DNA damage, SOS response, DNA metabolic process	x	x	x	drug, heat, UV	↓Cu and ↓AZM, ↓↑COL
	PA0429	-	Uncharacterized protein	-	-	-		↓Cu and ↓AZM, ↓↑COL
	PA1609	*fabB*	Fatty acid biosynthesis, lipid metabolism and biosynthesis	x	-	x	drug	↓Cu, ↓CIP, ↓↑AZM and ↓↑COL
	PA2537	-	Phospholipid biosynthesis, wall biogenesis, regulation of endopeptidase	-	-	-		↓Cu and ↑AZM
	PA2641	***nuoF***	Aerobic respiration, electron/proton transport, Krebs cycle	x	x	x	acidity, oxidative stress,	↓AZM, ↓Cu and ↓COL
	PA2733	-	Movement, pilus assembly, histidine biosynthesis	-	-	-		↓↑COL, ↓Cu and ↑AZM
	PA2950	-	Lipid metabolism, fatty acid and lipid biosynthesis	-	x	x	magnesium starvation	↓↑COL, ↓↑Cu and ↓AZM
	PA4265	*tufA*	Heat response, transcription antitermination, DNA damage	x	-	x	drug, antibiotic, cold	↑light, ↓↑CIP and ↓AZM
	PA4403	*secA*	Protein transport, DNA damage, transport	x	-	x	heat, cold, antibiotic,	↓↑COL, ↓Cu, ↓↑CIP and ↓AZM
	PA4428	***sspA***	DNA transcription, outer membrane assembly, lipopolysaccharide transport	-	-	x	drug, oxidative stress,	↓↑COL, ↓Cu, ↓CIP and ↓↑AZM
	PA4430	-	Response to DNA damage, catabolism, phenylacetate catabolism	x	-	x	heat, oxidative stress	↓Cu, ↓↑AZM and ↓NaClO
	PA5040	*pilQ*	Movement, pilus assembly, secretion	x	-	x	antibiotic	↓Cu and ↓AZM
	PA5110	*fbp*	Precursor metabolites and energy, gluconeogenesis, glycolisis	x	-	-		↓↑COL and ↓↑AZM
	PA5174	-	Lipid metabolism, fatty acid and lipid biosynthesis	-	-	-		↓↑COL, ↓AZM and ↓↑Cu
	PA5339	-	DNA damage, nucleobase interconversion, guanosine tetraphosphate metabolism	x	-	-	starvation, oxidative estress	↓↑COL, ↓AZM and ↓Cu
4	PA0002	*dnaN*	DNA damage response, SOS response, DNA metabolic process	x	x	x	heat, stress, UV, drugs	↓↑COL, ↓AZM, ↓↑CIP and ↓Cu
	PA1610	***fabA***	Lipid metabolism, fatty acid and lipid biosynthesis	x	-	x	drug	↓↑COL, ↓AZM and ↓Cu
	PA2639	***nuoD nuoC***	Aerobic respiration , electron/proton transport, DNA damage	x	-	x	heat	↓COL, ↓CIP, ↓AZM, and ↓Cu
	PA2644	***nuoI***	Aerobic respiration, electron/proton transport	x	-	-		↑↓COL, ↓AZM and ↓Cu
	PA2645	***nuoJ***	Aerobic respiration, electron/proton transport			x	oxidative stress	↓↑COL and ↓NaClO
	PA2646	*nuoK*	Aerobic respiration, electron/proton transport	x	x	x	antibiotic, oxidative stress	↓Cu, ↓CIP and ↓AZM and ↓NaClO
	PA2739	*pheT*	Protein transport, ARNt aminocylation for protein transport	-	-	x	cold, heat	↑↓Cu and ↓AZM
	PA2744	*thrS*	Heat response, ATP catabolism, cold response	-	-	x	cold, heat, antibiotic	↑COL, ↓Cu and ↓AZM
	PA2965	*fabF1*	Lipid metabolism, fatty acid and lipid biosynthesis	-	-	x	drug, osmotic stress, Temp,	↓CIP, ↓AZM, and ↓Cu
	PA4461	-	Lipopolysacaridcharid transport, outer memebrane assembly, cold response	-	-	-		↓CIP, ↓AZM, ↑↓COL and ↓Cu
	PA4740	*pnp*	RNA catabolism, transcription antitermination cold response	-	-	x	cold, osmotic stress, heat	↓AZM, ↑↓COL and ↑Cu

	PA4847	*accB*	Fatty acid biosynthesis, lipid metabolism and biosynthesis	x	-	x	drug, heat	↓CIP, ↓↑COL, ↓↑AZM and ↓ low oxigen
	PA5063	*ubiE*	Ubiquinone biosynthesis, aerobic respiration, DNA-dependent transcription	x	x	x	light, oxidative stress	↑↓AZM, ↑↓COL, ↑↓CIP and ↓Cu
	PA5163	*rmlA*	Lipopolysaccharide biosynthesis, DNA damage, antibiotics response	x	x	x	antibiotic	↓AZM, ↓COL, ↓CIP and ↓Cu
	PA5337	***rpoZ***	Regulation of transcription, transcription antitermination	x	-	x	stress, heat, cold, antibiotic	↑↓AZM, ↑↓COL, ↑↓CIP and ↓Cu
	PA5554	*atpD*	Response to DNA damage, aerobic respiration, amino acid metabolism	x	x	x	heat, cold, drug	↑↓AZM, ↓COL, ↑↓CIP and ↓Cu

* Support: number of algorithms (out of six). A gene was included in the core perturbome if the support was ≥4.

** Bold names refer to hub genes of the core perturbome.

*** Expression patterns found in significant works. See Methods for details of the references.

Table S1. Studies of datasets included in this work to identify the core perturbome of *Pseudomonas aeruginosa*

Perturbation (name, acronym and GEO series ID)	**Main results**
Azithromycin AZM-a GSE2430	In this study it was demonstrated that Azithromycin affected the metabolic pathways related to quorum sensing, as an antagonist. In addition, the production of virulence factors decreased, the response to oxidative stress also decreased, the mobility of the bacteria was reduced and the ability to form biofilm was impaired. Additionally, a very strong induction of type II secretion systems was observed (Yusuf Nalca, Lothar Jänsch, Florian Bredenbruch, Robert Geffers, Jan Buer, & Susanne Häussler, 2008).
Hydrogen peroxide H2O2 GSE3090	The results obtained in the study reinforce the conclusion that DNA repair pathways and proteins with catalase activity could be within the antioxidant defense systems that play a more relevant role *in P. aeruginosa*. In the experiment, a sublethal dose of H2O2 was used and a reduction in active and facilitated transport was found. Finally, an increase in the transcription of all the F, R and S type pyocins and at the same time a reduction in the transcription of a cellular immunity protein (pyocin S2) was shown (Chang, Small, Toghrol, & Bentley, 2005).
Copper Cu GSE4152	In the study it was observed that bacteria that were exposed to a copper shock showed a response very similar to that of oxidative stress that was not observed in a population adapted to high concentrations of iron. It is suggested that tolerance to iron could be given by two efflux systems: a P-type ATPase and a two-component regulatory system (Teitzel et al., 2006).
Ciprofloxacin CIP GSE5443 (0, 30 and 120 min)	In this study it was found that the genes involved in virtually all metabolic pathways present a negative regulation in response to ciprofloxacin by knocking down lexA gene, regulator of SOS response. The LexA-controlled SOS regulon identified by microarray analysis includes only 15 genes but does include several genes that encode proteins involved in recombination and replication, including two inducible polymerases known to play a role in mutation and the evolution of antibiotic resistance in other organisms (Cirz, O'Neill, Hammond, Head, & Romesberg, 2006).
Sodium hypoclorite NaClO GSE7402	Exposition to sodium hypochlorite showed a negative regulation in the genes related to metabolism and energy. This regulation suggests that the production of energy in the cell is limited. In addition, another strong negative regulation was shown in the genes linked to small molecule transport proteins. Finally, an increase in the transcription of hypothetical enzymes and membrane proteins was shown (Small, Chang, Toghrol, & Bentley, 2007).
Ortho-Phenylphenol OPP GSE10605 (20 and 60 min)	The results of the study show that after the exposure of Ortho-Phenylphenol, bacteria had a negative regulation in the synthesis of multiple amino acids, in particular the synthesis of enzymes linked to lysine biosynthesis was reduced. Among the genes that showed a positive regulation are genes coding for ribosomal proteins and genes coding for virulence factors. Genes related to phospholipid metabolism and transport and acquisition of iron were inhibited (Nde, Jang, Toghrol, & Bentley, 2008).
Azithromycin AZM-b GSE12738	In this study it was found that in the case of concentrations below the minimum inhibitory concentration, no change in the expression of the genes related to the quorum sensing was observed, however a negative change in the genetic expression was observed on N-acyl homoserine lactones (Kai et al., 2009).
Colistin COL GSE13252	*Pseudomonas* sp faced with a suboptimal dose of colistin, an alteration was obtained in the routes associated with virulence factors and bacterial colonization. Differential expression was obtained for genes related to osmotic stress, biofilm formation, mobility and quorum sensing (Cummins, Reen, Baysse, Mooij, & O'Gara, 2009).
Chlorhexidine diacetate CDA GSE14253 (10 and 60 min)	In this study it is shown that in both times 10 and 60 minutes the genes related to membrane transport, oxidative phosphorylation and electron transport presented a negative regulation. After 10 minutes, inhibition was found in the mechanisms of DNA repair and a stimulation of the oprH gene that causes self-stimulation of antimicrobial absorption. At minute 60, inhibition was found in genes related to flagella, pylons, oxidative phosphorylation, electron transport and outer membrane proteins (Nde, Jang, Toghrol, & Bentley, 2009).
E-4-bromo-5-bromomethylene-3-methylfuran-2-5H-one BF8 GSE36753	In the study it is indicated that the use of BF8 can restore the susceptibility to antibiotics of the *Pseudomonas* strains under study. Interestingly, although BF8 is an inhibitor of quorum sensing in gram-negative bacteria, the data in this study suggest that the activities of BF8 to revert antibiotic tolerance of *P. aeruginosa* PAO1 persister cells is not through QS inhibition and may involve other targets (Pan, Bahar, Syed, & Ren, 2012).

Table S2. Metrics associated with the performance of each method and algorithm

Metrics	SP method*			MP method**		
	SVM	KNN	RF	SVM	KNN	RF
Accuracy	0.79	0.71	0.75	0.66	0.69	0.70
Kappa	0.58	0.39	0.45	0.26	0.37	0.35
McnemarPValue	1.00	1.00	0.22	0.67	0.75	0.78
Sensitivity	0.80	0.60	0.50	0.50	0.69	0.56
Specificity	0.79	0.79	0.93	0.76	0.69	0.80
Positive Predictive Value	0.73	0.67	0.83	0.62	0.59	0.64
Negative Predictive Value	0.85	0.73	0.72	0.72	0.80	0.76
Precision	0.73	0.67	0.83	0.62	0.59	0.64
Recall	0.80	0.60	0.50	0.50	0.69	0.56
F1 score	0.76	0.63	0.63	0.52	0.63	0.59
Prevalence	0.42	0.42	0.42	0.38	0.38	0.38
Balanced Accuracy	0.79	0.69	0.71	0.63	0.69	0.68

* Metrics for top 56 genes, averaged after internal cross validation

** Average after 100 iterations of random partitions and internal cross validation