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

Jose Arturo Molina-Mora[1*], Pablo Montero-Manso[2], Raquel García Batán[1], Rebeca Campos Sánchez[3], Jose Vilar Fernández[2], Fernando García[1].

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

[2] Department of Mathematics, University of A Coruña, Spain

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

* Corresponding author

E-mail: jose.molinamora@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 a 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 public transcriptomic data, 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 systems level were include to describe the perturbome elements.

We were able to select and describe 46 core response genes associated to multiple perturbations in *Pseudomonas 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 normalization analysis, 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.

Introduction

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

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

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

65 To study the central response determinants to different perturbations in a living organism, we used the
66 model *Pseudomonas aeruginosa* PAO1 (reference strain). *P. aeruginosa* is a Gram negative gamma-
67 proteobacterium with a noteworthy metabolic versatility and adaptability enabling colonization of diverse
68 niches infecting plants, animals and humans alike (8,13). In this group, the molecular mechanisms associated
69 to most biological processes remain unclear causing limited action to modulate responses, including the
70 susceptibility to stressors, environmental factors and experimental conditions.

71 Several studies have used machine learning algorithms at the transcriptomic level to recognize gene
72 expression patterns (2,14–16). However, for many common biological contexts, applicability and utility of
73 these machine learning approaches have not been fully explored and utilized (17), for example in the exposure
74 to multiple stressors and the molecular response. To our knowledge, only a few studies have used feature
75 selection methods on biological data to describe the effects of multiple perturbations in complex biological
76 systems (4,6) and so far none in *P. aeruginosa*. A related work in *P. aeruginosa* used a machine learning
77 approach to identify sets of genes that correlate with multiple culture media, but without other conditions
78 (18).

79 The use of microarray and other high throughput technologies data involve challenges for machine
80 learning approaches. These include the curse of dimensionality (19,20), normalization of raw values to
81 compare samples (21,22), data partitions for training and testing models (23,24), and evaluation of
82 performance (21,25). Since comparison between the machine learning algorithms are completely variable
83 (11,17,20,26,27), Support Vector Machine (SVM) (28), K-Nearest neighbor (KNN) (27) and Random Forest (RF)
84 (29) have been successfully used with microarray gene expression data allowing the recognition of emerging
85 patterns (26).

86 Here, we hypothesize that perturbations on living cells will trigger global reprogramming of multiple
87 molecular determinants that can be sensed at the transcriptional level. The initial response after an acute
88 stress will then expand producing the global molecular rearrangement. Then, pleotropic and specific effects
89 on gene networks will be reflected as changes in gene expression profiles and the complexity of molecular
90 regulation at other levels. Therefore, by using a machine learning approach, common molecular features (for
91 all stressors) could be identified as a central or core determinant (see Fig. 1-A-B).

Fig. 1. General pipeline for identifying core response genes in *P. aeruginosa* by a machine learning approach. (A) Schematic representation of hypothesis for identifying core response determinants when bacteria are exposed to multiple perturbations. (B) Workflow of the machine learning approach using microarray data and model fitting by SP and MP methods for identifying and describing core response genes. (C) Representation of data partition methods, SP and MP, including subsamples for testing and an internal 10-fold cross validation for training data set.

Thus, the aims of this study were (i) to implement a machine learning approach to select genes from microarray expression data, and (ii) to identify and describe a subset of genes than can be associated with multiple perturbations or perturbome in *P. aeruginosa,* i.e. the core response components.

Materials and methods

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

Datasets

In order to compare gene expression profiles of strain *P. aeruginosa* PAO1 when exposed to multiple perturbations, GEO database (https://www.ncbi.nlm.nih.gov/geo/) was used for a systematic selection of datasets (details in supplementary Fig. S1-A). Initial evaluation by organism and mRNA profiles by GPL84 platform (Affymetrix *Pseudomonas aeruginosa* PAO1 Array, with all 5549 protein-coding sequences) identified 156 series of datasets with 1310 samples (Date of Access: January 25th 2018). In a second step, data were selected according to experimental design if they included perturbations, leaving only 47 series. Finally, to make datasets comparable by experimental conditions, evaluation and selection were done for series with

119 similar culture conditions (Luria Bertani LB medium and exponential phase when measuring mRNA profile)
120 and if a control condition was available (without any perturbation or treatment). The final dataset was
121 composed of 10 series with 71 samples (Series GSE2430, GSE3090, GSE4152, GSE5443, GSE7402, GSE10605,
122 GSE12738, GSE13252, GSE14253 and GSE36753).

123 Some series included temporal measurements which we considered as separate perturbations, resulting
124 in replicates of 10 controls and 14 perturbations: azithromycin with 2 series (AZM-a and AZM-b) (30,31),
125 Hydrogen peroxide (H_2O_2) (32), copper (Cu) (33), sodium hypoclorite (NaClO) (34), ortho-phenylphenol (OPP)
126 at 20 and 60 minutes (35), colistin (COL) (36), chlorhexidine diacetate (CDA) at 10 and 60 minutes (37), E-4-
127 bromo-5-bromomethylene-3-methylfuran-2-5H-one (BF8) (38) and ciprofloxacin CIP at 0, 30 and 120 minutes
128 (39).

129

130 *Normalization and comparison of global transcriptomic profiles*

131 To compare all the 71 samples, a first normalization step using housekeeping genes was performed, as
132 recommended previously for multiple origin samples (22,40,41). Since *ampC, fabD, proC, pbpA (PBP2), rpoD*
133 and *rpoS* genes have been reported as stable housekeeping genes for *P. aeruginosa* (21), normalization by
134 each one of these genes were compared according to effect on the other. After normalization was done using
135 a specific gene, the other five genes were evaluated by intensity dispersion of controls vrs perturbations and
136 p values were calculated using T-test ($p < 0.001$). Then, comparison was done by calculating variation
137 coefficient for control conditions. These metrics were used for selecting the best normalizer gene (assuming
138 no differences between other housekeeping genes when exposed to stressors).

139 After normalization, clustering algorithms were implemented in order to compare global transcriptomic
140 profiles between perturbations and controls. Principal Component Analysis (PCA) and Hierarchical Clustering
141 (HC) were run with default parameters using the Caret Package (caret.r-forge.r-project.org/) in R software
142 (Version 3.4.3).

143 In order to robustly select a number of genes that could separate experimental conditions (controls and
144 perturbations), reduce noise in microarray data and for finally identifying the core response of *P. aeruginosa*,
145 two approaches of feature selection protocols were implemented, as detailed below.

146

147 *Gene ranking and selection by Single Partition SP method*

148 With the aim of identify genes which expression values were commonly related to multiple
149 perturbations, a first approach was implemented considering a particular partition of dataset for training and
150 testing sets (Fig. 1-C). Single partition was established using the last replicate of each experiment, in both
151 control and perturbation. Because there were 14 perturbations and 10 controls (71 samples including
152 replicates), a total of 24 samples were included in the testing dataset and the remaining 47 samples were
153 included for the training dataset (66% for training and 34% for testing set).

154 Using this partition, ranking of genes was done by a machine learning approach using three classification
155 algorithms: SVM, KNN and RF. A homemade script in R included these functions of the Caret package. For all
156 three algorithms, default parameters were used for training and 10-fold cross validation (Fig. 1-B) was
157 included. After this, variable importance metric was calculated for all genes using the *varImp* function,
158 associating a specific value for each gene. In the case of SVM and KNN, same importance is calculated because
159 function is model-free for these cases (as detailed in Caret Package), resulting in the same list of genes but
160 metrics are specifically calculated for each algorithm.

161 As similarly reported (11), to evaluate the number of genes that should be selected in the top group (the
162 first K ranked genes, as candidates for the core response by each algorithm), multiple classification models
163 were systematically run, starting only with the highest-score ranked gene. In brief, after the training with one
164 gene, model performance was evaluated by calculating metrics using the testing dataset. Then, training was
165 run again when the next ranked gene was added, and new metrics were calculated. This process was repeated
166 up to complete all the ranked genes. Metrics included accuracy (correct classification percentage), kappa
167 value (inter-rater classification agreement), sensitivity, specificity, precision, recall, prevalence and F1 score
168 (harmonic average of the precision and recall).

169 Selection of the K value of top genes was based on the following criteria: (i) the stability of the metrics
170 (priority for accuracy, kappa and F1) when increment of ranked genes was done, as suggested in (11), (ii)
171 consensus value suitable for all the three algorithms (including a 10% of tolerance), and (iii) minimum number
172 of genes as possible. After the selection of the K value, ROC (Receiver-operating characteristic) curve and AUC

173 (Area under the curve) value were calculated for each algorithm. Finally, selection of top K genes between
174 algorithms were compared by metrics and list of genes (see below for details).
175
176 *Gene ranking and selection by Multiple Partitions (MP) method*
177 In order to identify genes related to multiple perturbations independently of a single/specific partition,
178 a second method using multiple random partitions was implemented (Fig. 1-C). To address this, a random
179 data selection for training and testing sets was done using the *createDataPartition* function. Partition was set
180 to 80% (57 samples) for training set and remaining for testing set (14 samples) with experimental conditions
181 equally distributed. Then, protocols with SVM, KNN and RF algorithms (same conditions as previously
182 described in SP method, with 10-fold cross validation) resumed the analysis with a final ranking of genes using
183 the *varImp* function. Using only top K of ranked genes (K value determined using the criteria described in the
184 SP method), new set of training/testing sets were used for evaluating performance of the models and each
185 metric was stored with the list of the K ranked genes. This full process was automatically repeated 100 times
186 using *replicate* function, starting with a new random partition and finishing with the list of the K genes and
187 the metrics associated to that partition. Finally, for each algorithm, full data of all the runs were analyzed for
188 determining frequency of the appearance of genes (*table* function) and calculating average and dispersion of
189 metrics across all the 100 runs. Definitive list of the K more frequent genes was established for each algorithm
190 after this comparison by frequency.
191
192 *Identification of core response genes*
193 After selection of top K genes in each algorithm by SP and MP methods, comparison of genes was done
194 using Venn diagrams in order to identify all the candidate genes using Venn-tool
195 (http://bioinformatics.psb.ugent.be/webtools/Venn/). Genes identified by at least four algorithms were
196 considered part of the perturbome (this guarantees that a gene was identified by the two methods and at
197 least by two different classification algorithms).
198 Gene relationships were represented using molecular networks using a large scale model (using a top-
199 down systems biology approach). The network was built using protein-protein interaction (PPI) graphs in

200 PseudomonasNet database (www.inetbio.org/pseudomonasnet/). Network was downloaded and visualized
201 using the Cytoscape software (https://cytoscape.org/).
202
203 *Description and comparison of core response genes*
204 In order to describe and compare the genes associated with the core response of *P. aeruginosa,* by
205 experimental conditions, four analyzes were established. First, clustering analysis by PCA and HC algorithms
206 were evaluated again but now only considering genes of the core response. Based on distribution in the case
207 of PCA, an additional analysis to cluster the samples into two groups was done using the Kmedoids algorithm
208 in R with fastkmed package (https://rdrr.io/cran/kmed/man/fastkmed.html, default parameters). This
209 allowed not only evaluating the capacity of the core response genes to separate the experimental conditions,
210 but also to calculate the kmedoids for the groups. Additional representation of centroids was done using
211 Kmeans algorithm.
212 Second, using the PseudomonasNet database, a small world network was built and then exported into
213 Cytoscape software with the genes of the core response. The information of topological features (including
214 connectivity) and expression levels of kmedoids were incorporated into different versions of the network.
215 Third, a classic analysis of differentially expressed genes (DEGs, $p<0.05$) was implemented in R with
216 Limma package (https://www.rdocumentation.org/packages/limma/versions/3.28.14) using empirical Bayes
217 moderated t-statistic (eBayes) with Benjamini and Hochberg method for *p* value correction (42). This led us
218 to compare our results with a classical approach for gene expression.
219 Finally, in order to give biological interpretation to the selected genes, to determine levels of expression
220 reported in databases and to study metabolic pathways involved under each perturbation, an exhaustive
221 annotation was made using the databases PseudomonasDB (gene ontology), GEO database and particular
222 literature. This information was integrated with the results obtained by all the analysis and the DEGs in order
223 to fully describe the genes that make up the core response or perturbome of *P. aeruginosa* PAO1.
224

225 # Results

226 *Normalization and comparison of global transcriptomic profiles*

227 In order to compare multi-study microarray data, after data selection (supplementary Fig. S1-A),
228 normalization was performed using six classical housekeeping genes for *P. aeruginosa*, which were evaluated
229 according to two metrics (supplementary Table S1). First, after each normalization by a specific housekeeping
230 gene, intensity dispersion of controls vrs perturbations were compared for other genes and a *p* value was
231 obtained by T-test. For all genes, comparisons between conditions revealed no statistical differences
232 ($p>0.001$). A second criterion was the comparison of variation coefficient for control conditions after each
233 normalization. The lower values of the coefficient were obtained after the normalization with *proC* and with
234 *pbpA.* However, *proC* was selected due it has been reported as a more stable housekeeping gene in *P.*
235 *aeruginosa* (21). Details of the expression levels (between conditions) of the other housekeeping genes after
236 *proC* normalization are shown in supplementary Fig. S1-B. Similarly, average intensity between conditions was
237 compared for each perturbation after *proC* normalization (Fig. 2-A), showing differences in the expression of
238 housekeeping genes for some specific disturbances. Comparison of final dispersion of mean intensities
239 (expression level) of all genes showed similar distribution as expected (Fig. S1-C, case for first replicate of each
240 condition), making samples comparable.

241 Global transcriptomic profiles (all 5549 genes) were compared by both PCA and HC clustering algorithms,
242 revealing a mixed pattern (no separation) between perturbations and controls. Two samples (BF8 and Control
243 8) resulted with extreme global profiles (Fig. 2 B-C).

244

245 **Fig. 2. Normalization and comparison of samples by global profiles using all genes.** (A) Evaluation of
246 housekeeping genes after normalization by *proC* gene, comparing intensity between conditions for each
247 perturbation. Details of gene selection for normalization are shown in Table S1 and Fig. S1. (B-C) Global profiles
248 were compared by both PCA and HC clustering algorithms, showing mixed patterns between classes. More
249 details in supplementary Fig. S1.

250

251 *Perturbome genes of* P. aeruginosa *can be identified by a machine learning approach using SP and MP methods*

252 Two methods using machine learning algorithms (by SP and MP methods) were implemented in order to
253 robustly rank and select genes associated to multiple perturbations in *P. aeruginosa*. In each method, three

classification algorithms were evaluated: SVM, KNN and RF. Metric results associated to RF are shown in Fig. 3, and supplementary Fig. S2 for SVM and KNN.

Fig. 3. **Evaluation of SP method for gene ranking by importance, case for RF algorithm.** (A) Accuracy, F1 and kappa values after iterations of classification with the first top 200 genes (adding genes 1-to-1). (B) ROC plot using selected top 56 genes for evaluation of performance of the algorithm. (C) Ranking and importance value of top 56 genes. Similar results are shown for SVM and KNN algorithms in supplementary Fig. S2.

The first method (SP) implemented a gene ranking based on variable importance using a single/specific data partition. After the ranking was established, multiple classification models were run with the ranked genes (Fig. 3-A and supplementary Fig. S2-A-C). For each classification model, stability of the three metrics were evaluated to select the suitable K value of genes that could be applied to all algorithms at the same time, as detailed in criteria of Material and Methods section. For SVM and RF, stable values of metrics are given with at least the first 51 ranked genes, meanwhile it is 45 genes for KNN. Considering a 10% of tolerance with the highest of these values, K=56 was selected as the number of top genes that were included as preliminary candidates of the core response according to each algorithm. With this value, metrics of each algorithm were compared (Table 1). For example, accuracy was 0.79, 0.71 and 0.75 for SVM, KNN and RF respectively in the SP method. SVM obtained a better performance according to kappa, sensitivity, recall and F1 scores, but higher values of specificity and precision resulted for RF. Also, ROC curve and AUC value were calculated (Fig. 3-B and supplementary Fig. S2-B-D). Best performance was obtained for RF with AUC = 0.92, then 0.82 for SVM and finally 0.76 for KNN. Since *importance* metric for SVM and KNN is the same, they shared same list of top 56 genes. Comparison between implementations showed that 21 genes were identified by the three algorithms at same time, 35 exclusively by RF and same number for SVM/KNN. In total 91 genes were identify by any of the algorithms. List of genes and importance value of top 56 ranked genes for each approach is presented in Fig. 3-C for RF and Fig. S2-E for SVM/KNN.

Table 1. Performance of six different machine learning approaches to identify perturbation-associated genes

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
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

In a second approach, the MP method was implemented using multiple random partitions. Same SVM, KNN and RF algorithms were evaluated by running 100 iterations with different partitions and top-56 more frequent genes for each method were selected. Details of ranking and frequency is shown in Fig. 4-A (SVM) and supplementary Fig. S3 (KNN and RF) and metrics for all 100 iterations are presented in Table 1, Fig. 4-B and supplementary Fig. S3. Accuracies for all the models were 0.66, 0.69 and 0.70 for SVM, KNN and RF respectively. Specific values for kappa, precision, recall and F1 score were found for each algorithm. When comparison of list of genes was done, 29 genes were identified at same time by all implementations, and 23 by both SVM and KNN.

Fig. 4. **Evaluation of MP method for gene ranking by frequency, case for SVM algorithm.** (A) Ranking of top 56 genes by frequency after iterations of 100 data partitions and classification model fitting. (B) Dispersion of metrics across 100 iterations. Similar results are shown for KNN and RF algorithms in supplementary Fig. S3.

With the aim of identifying preliminary core response genes in *P. aeruginosa*, the lists of top 56 genes selected by each algorithm and method were jointly represented using a Venn diagram (Fig. 5-B). Note that the same set was used for SVM and KNN since both procedures lead to the same genes. A total of 118 different genes were identified and 15 genes were simultaneously identified by all the algorithms. Distribution of all

303 118 genes on large scale molecular networks is presented in Fig. 5-A. Results show that selected genes are
304 connected but they do not establish a defined cluster. These genes seem to be associated to different
305 subgroups of highly connected genes.
306
307 **Fig. 5. Identification, systems level description and global profiles comparison given by core response genes.**
308 (A) Distribution and number of algorithms of preliminary 118 selected genes on a basal large scale network of
309 functional associations in *P. aeruginosa.* (B) Set comparison of genes given by classification algorithms of SP
310 and MP methods, with a final selection of 46 genes which were identified by at least 4 algorithms. (C) Small
311 world network showing relationships between the 46 core response genes and connectivity metric for each
312 gene. (D-E) Global profile comparison of samples by both PCA and HC clustering algorithms, showing
313 separation of conditions. For reference, centroids of each cluster were plotted as triangles in each cluster.
314 More details are shown in supplementary Fig. S4.
315
316 Final version of the core perturbome components was established by selecting genes recognized for at
317 least four algorithms, according to the criteria detailed in Materials and Methods. A total of 46 genes were
318 finally associated to core response. In the representation as small world network, relationships between the
319 46 genes revealed different topological patterns of connectivity between molecules, being *nuoC* and *nuoF* the
320 ones with higher connectivity (connection degree). In addition, only six genes had no connections between
321 them (Fig. 5-C).
322
323 *Comparisons of core response genes show separation of global transcriptomic profiles according to*
324 *experimental conditions and biological functions related to tolerance to stress*
325 After the selection of the core response genes, comparisons between controls and perturbations were
326 done to characterize these genes by global profiles. First, clustering algorithms to compare global profiles
327 were run using the 46 genes of the perturbome. In the case of PCA (Fig. 5-D), samples distribution let to
328 differentiate controls and perturbations. A Kmedoid algorithm (k = 2) was able to identify two clusters
329 enriched by samples of each condition: one consisting entirely of perturbation samples (11 samples, blue

330 color) and the other mostly by control samples (10 controls and 3 disturbances, red color). In the case of the
331 blue cluster, the kmedoid was sample CIP-120min, meanwhile control 6 was selected for the another cluster.
332 For the case of HC (Fig. 5-E), the same distribution of samples was obtained. Supplementary analysis of
333 gene expression was included by comparing levels for all core response genes (Fig. S4-A) and comparing
334 expression levels of the kmedoids on the small world network (Fig. S4-B-C).
335 Gene annotation revealed that biological processes related to most of the genes are metabolism,
336 molecule binding and biosynthesis (Fig. 6-A). Specific information about participation in processes of DNA
337 damage response, DNA repairing and response to general/specific stimuli was also searched for each gene,
338 showing that most of genes includes participation in such processes. Literature support shows variable
339 patterns of expression, depending on the disturbance as shown in supplementary Table S2.
340
341 **Fig. 6. Annotation of core response genes and comparison with a differential expression analysis.** (A)
342 General annotation profile of identified genes by core response genes showing associated biological
343 processes. (B) Comparison of identified genes by our machine learning approach and DEGs lists. (C) General
344 annotation of DEGs showing similar profile than our approach. Specific annotation per gene is shown in
345 supplementary Table S2.
346
347 Finally, in order to compare the results of the machine learning strategy with another approach, a
348 differential expression analysis was run. A total of 101 DEGs were identified, which 33 were shared by the
349 core response (Fig. 6-B). This means that 72% of core response genes were also identified by another single
350 and independent method. Annotation profile of DEGs (Fig. 6-C) showed a similar pattern as our machine
351 learning approach.
352
353
354
355
356

357 # Discussion

358 Living organisms face external and internal conditions that compromise cellular functions at molecular,
359 metabolic and structural levels, disrupting their homeostasis (43,44). Cell stress response is crucial for
360 organismal survival and complex networks are usually involved in the molecular mechanism related to
361 tolerance (1). However, few studies have identified central and possible universal regulation of the response
362 to multiple disturbances, a concept termed as perturbome (4,5). Common molecular response was previously
363 reported as a network of common set of genes and pathways that can be generically associated to multiple
364 perturbations in plants (7), pathogenic bacteria (3) or cell lines models (4), and others.

365 In our approach using *P. aeruginosa,* but applicable to other organisms, we hypothesized the existence
366 of a set of core genes regulating the response to stressors in a generic sense of different pathways. *P.*
367 *aeruginosa* has a high proportion (about 5%) of its genome dedicated to regulatory mechanisms, probably
368 explaining its adaptability to such a broad range of growth conditions (25). Since strain *P. aeruginosa* PAO1 is
369 a clinical isolate with a profile of multiresistance to many antibiotics (45), characterization of molecular
370 mechanisms involved in the tolerance to stressors in this strain could eventually help to modulate sensitivity
371 and overcome resistance. Exhaustive integration of multi-omic data and network analysis are required in
372 order to clarify the molecular mechanisms related to stress conditions and eventually use them for modulating
373 cell response.

374

375 *Insights of algorithms to identify core response genes or perturbome*

376 In our study, a first analysis to compare data was done by normalizing intensities using *proC*
377 housekeeping gene. This approach is a prerequisite in qRT-PCR (22), but as reported and exhaustively
378 evaluated in (22), it is also suitable for microarray data, in particular to compare multiple origin samples with
379 divergent sample processing steps (21,25). After normalization, global transcriptomic profiles showed mixed
380 patterns between samples according to their experimental condition. Because its complexity, raw microarray
381 data showed noise and redundant information that can explain the poor resolution of this clustering
382 algorithms to identify classes (20). Thus, a feature (gene) selection analysis was implemented for not only

383 identifying genes associated to multiple perturbations in *P. aeruginosa,* but also to improve performance of
384 predictive/descriptive models capable of separate control and perturbation categories.

385 In our case, these potential patterns were investigated using a robust machine learning approach by
386 implementing six protocols, using SP and MP methods and SVM, KNN and RF algorithms in each case. This was
387 a critical step because gene selection from microarray data is complex *per se*. Feature reduction remains as a
388 challenging task in transcriptomic studies because thousands of genes to select from, and it introduces an
389 additional layer of complexity in the modelling task (12,46). To avoid bias and overfitting, implementations
390 of diverse strategies of data partitioning such as bootstrapping, random partitions and cross validation are
391 recommended. In general, these methodologies can robustly minorate the influence from noise, outliers,
392 absence of ground truth sets, and to reduce variance (2,24,47).

393 The single partition SP method consisted of a particular and invariable data for training (with internal
394 cross validation) and another to test, and it is probably the most common approach used in machine learning.
395 In the case of the multiple partitions MP method, 100 random partitions of dataset were run. MP method had
396 a dual consideration when splitting data (multiple partitions and the internal cross validation). This method
397 can be considered as an *ensemble based on different data partitioning*, as it had been previously proposed
398 (23). Datasets were divided using multiple random partitioning procedures and then genes were ranked. After
399 all runs, a final feature subset is determined by calculating the frequency of features in all the runs (23). A
400 equivalent approach was implemented by Pai and collaborators to classify gene expression data in a cancer
401 model (48).

402 However, one possible problem with MP approach is that cross validation results may depend on
403 similarity of testing and training sets. A classification prediction method is only expected to learn how to
404 predict on unseen samples that are drawn from the same distribution as training samples (24,47), and MP
405 method could violate this assumption. SP method guarantees this because was built always using a replicate
406 for each perturbation and control. Thus, both methods SP and MP are required to robustly select features.

407 On the other hand, many algorithms for dimension reduction have been proposed (6,19,20) but no
408 standard machine learning algorithm can be selected due multiple evaluations results on completely variable
409 metrics associated to performance (26). Many studies have shown that SVM, KNN, and RF generally

410 outperform other traditional supervised classifiers (17,26,49). In our case, a variable pattern was found for
411 different metrics in all evaluated implementations. For example, based on accuracy SVM for SP method and
412 RF for MP method resulted in higher scores, which agree with other studies using machine learning and
413 biological data (26,48).

414 In our subsequent analysis, we were interested in identifying a consensus list of candidate genes for the
415 core response in *P. aeruginosa*, resulting in 46 genes of the perturbome. When global profiles were compared
416 using these genes (PCA and HC), control and perturbation classes (Fig. 5-D) were clearly separated. This gene
417 number seems to be a modest number of elements (less than 1% of all the available dataset with 5549 genes)
418 but it agrees completely with other studies, including machine learning methods (11,17) or other approaches
419 (11,22,50,51). In addition, 72% of core response genes were also identified as DEGs with similar annotation
420 profiles; differences can be explained by significant fluctuations in the differential expression results as
421 previously reported, mainly because it is not a consensus strategy (only based on *p*-value) and it does not
422 incorporate the estimates of the test performance (true positive/negative rates and other metrics) on the
423 results (2).

424

425 *Biological insights of the core response genes in* P. aeruginosa: the perturbome

426 Core response genes or perturbome can be related to a central regulation network, and as convergent
427 point of signal transduction, transcriptional regulation and stress-related pathways, as it has been suggested
428 (2,4,5,44). Annotation of the 46 genes shows that most of them are functionally related to biosynthesis,
429 molecule binding and metabolism (including an important number of hits for lipids), including additional
430 functions for regulation of DNA damage repair, response to stimuli and aerobic respiration. Interestingly,
431 these processes are represented by genes with high connectivity in the small world network. For example,
432 main functions associated to *fabA* are lipid metabolism, fatty acid and lipid biosynthesis, meanwhile for *nuoC*
433 are aerobic respiration, electron/proton transport and DNA damage. Finally, *nuoF* is associated to aerobic
434 respiration, electron/proton transport and Krebs cycle.

435 In this sense, cells are equipped with systems and mechanisms to recover from the environmental stress
436 and stimuli to maintain all necessary physiological functions (52). Other common stimuli such as low ATP, slow

437 growth, and ROS production can also occur before cells express stress specific factors, but mediating a
438 common effect. For example, response to stress includes modulation of energy production and aerobic
439 electron transfer chain components. As it has been reported in *E. coli*, aerobic electron transport chain
440 components are down-regulated in response to growth arrest (53). This corresponds with the global profiles
441 of expression of 46 core response genes. Also, regulation of lipid metabolism is relevant for survival in the
442 wide range of environmental conditions where bacteria thrive (54), even for biofilm-living forms (3) as *P.*
443 *aeruginosa*. Core response genes *fabG*, *lpxA*, and PA5174 could be implied in this process.

444 In the case of DNA damage repair (including the case of *cycB* and *gltP* genes), responses mediated by
445 SOS and *rpoS* help to maintain genome integrity, colonization, and virulence (39,55). These responses are
446 activated under multiple disturbances and modulating a low energy production and shutdown of the
447 metabolism, promotes formation of antibiotic resistance and biofilms (3). Other related pathways for some
448 specific genes included regulation of the transcription during stress by RNA-binding proteins in order to
449 reprogram or shut down translation and to rescue the ribosome stalled by a variety of mechanisms induced
450 (56). Three core response genes (*rpmH*, *tsf* and PA2735) were annotated with these functions.

451 Jointly, the relatively few diversity of metabolic functions and pathways has sense in order to ensure
452 redundancy and robustness in the response to stimuli. Similar results, regarding enriched pathways, have
453 been obtained in other studies with eukaryotic models, including disturbed human cell lines (4,5,44),
454 *Arabidopsis thaliana* under physical and genotoxic stresses (2) or a genome-wide association study of a generic
455 response to stress conditions (6). In the case of prokaryotic organisms, two studies have used *Escherichia coli*
456 as model to identify differentially expressed genes after exposure to stress conditions (3) and to create
457 networks associated with the response to fluctuating environments (7). Differences with other organisms and
458 disturbances can suggest that response cell stress can be organismal specific, although heterogeneity has also
459 been suggested a reasonable explanation because differences in the response in a apparently homogeneous
460 cell population (44,57).

461 As in our case, the response to stimuli and stressors is orchestrated by a pleotropic modulation (58)
462 which can be associated to a central regulation. Alternative mechanisms such as cross stress protection (ability
463 of a stress condition to provide protection against other stressors) (7), role of sigma factors and specific two

464 component systems (3) can contribute to explain this phenomenon. The molecular response can lead to
465 regulate multiple biological activities including metabolism, replication, transcription and translation, changes
466 in membrane composition, motility, modification gene expression, expression of virulence factors, multi-drug
467 resistant phenotypes and biofilm formation, and others (44).

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

473 Further analyses are required to explore potential use of perturbome network to modulate (positively
474 or negatively) the response to disturbances, to model molecular circuits, to identify possible biomarker genes
475 of stressed states, and to experimentally validate our findings. In addition, this approach can be used to model
476 the perturbome in other *P. aeruginosa* strains, as we hope to run soon with a genome we recently described
477 (13), and other organisms.

478

479

480 # Conclusions

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

490 # Abbreviations

491 AUC: Area under the curve

492 AZM: Azithromycin

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

494 CDA: Chlorhexidine diacetate

495 CIP: Ciprofloxacin

496 COL: Colistin

497 Cu: Copper

498 DEGs: Differentially expressed genes

499 DNA: Deoxyribonucleic acid

500 GEO: Gene Expression Omnibus

501 H2O2: Hydrogen peroxide

502 HC: Hierarchical Clustering

503 KNN: K-Nearest Neighbor

504 LB: Luria Bertani

505 mRNA: Messenger RNA

506 NaClO: Sodium hypoclorite

507 OPP: Ortho-phenylphenol

508 PCA: Principal Component Analysis

509 PPI: Protein-protein interaction

510 RF: Random Forest

511 RNA: Ribonucleic acid

512 ROC: Receiver-operating characteristic

513 SVM: Support Vector Machine

514

515

Declarations

Ethics approval and consent to participate

Not Applicable

Availability of data and material

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

Public raw data used in this study: GEO database (https://www.ncbi.nlm.nih.gov/geo/), data Series GSE2430, GSE3090, GSE4152, GSE5443, GSE7402, GSE10605, GSE12738, GSE13252, GSE14253 and GSE36753.

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

Competing interests

No competing interests to declare.

Funding

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

Acknowledgements

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 EF. Prokaryotes : prokaryotic physiology and biochemistry. Springer; 2012.
2. Ma C, Xin M, Feldmann KA, Wang X. Machine Learning-Based Differential Network Analysis: A Study of Stress-Responsive Transcriptomes in Arabidopsis. Plant Cell [Internet]. 2014;26(2):520–37. Available from: http://www.plantcell.org/cgi/doi/10.1105/tpc.113.121913
3. Nagar SD, 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 [Internet]. 2016 May [cited 2018 Nov 16];20(5):310–24. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27195968
4. Caldera M, Müller F, Kaltenbrunner I, Licciardello MP, Lardeau CH, Kubicek S, et al. Mapping the perturbome network of cellular perturbations. Nat Commun [Internet]. 2019;10(1). Available from: http://dx.doi.org/10.1038/s41467-019-13058-9
5. Sadeh S, Clopath C. Theory of Neuronal Perturbome: Linking Connectivity to Coding via Perturbations. bioRxiv [Internet]. 2020 Feb 20 [cited 2020 Apr 29];2020.02.20.954222. Available from: https://www.biorxiv.org/content/10.1101/2020.02.20.954222v1.full
6. Bermingham ML, Pong-Wong R, Spiliopoulou A, Hayward C, Rudan I, Campbell H, et al. Application of high-dimensional feature selection: Evaluation for genomic prediction in man. Sci Rep [Internet]. 2015;5:1–12. Available from: http://dx.doi.org/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 [Internet]. 2014 Apr 28 [cited 2018 Nov 17];9(1):643–643. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23385483
8. Nogales J, Guðmundsson S, Duque E, Ramos JL, Palsson BØ. Expanding the computable reactome in Pseudomonas putida reveals metabolic cycles providing robustness. bioRxiv [Internet]. 2017;139121. Available from: https://www.biorxiv.org/content/early/2017/05/19/139121

570 9. Kc K, Li R, Cui F, Haake AR. GNE: A deep learning framework for gene network inference by aggregating
571 biological information. Bioinformatics [Internet]. 2018;1–9. Available from: https://github.com/kckishan/GNE
572 10. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: protein–protein association
573 networks with increased coverage, supporting functional discovery in genome-wide experimental datasets.
574 Nucleic Acids Res [Internet]. 2019 Jan 8 [cited 2019 Nov 13];47(D1):D607–13. Available from:
575 https://academic.oup.com/nar/article/47/D1/D607/5198476
576 11. Li Y, Wang N, Perkins EJ, Zhang C, Gong P. Identification and optimization of classifier genes from multi-class
577 earthworm microarray dataset. PLoS One. 2010;5(10):1–9.
578 12. Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bioinformatics. Bioinformatics
579 [Internet]. 2007 Oct 1 [cited 2018 Nov 16];23(19):2507–17. Available from:
580 https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btm344
581 13. Molina-Mora J-A, Campos-Sánchez R, Rodríguez C, Shi L, García F. High quality 3C de novo assembly and
582 annotation of a multidrug resistant ST-111 Pseudomonas aeruginosa genome: Benchmark of hybrid and non-
583 hybrid assemblers. Sci Rep [Internet]. 2020 Dec 29 [cited 2020 Jan 29];10(1):1392. Available from:
584 http://www.nature.com/articles/s41598-020-58319-6
585 14. Zhao W, Chen JJ, Perkins R, Wang Y, Liu Z, Hong H, et al. A novel procedure on next generation sequencing data
586 analysis using text mining algorithm. BMC Bioinformatics [Internet]. 2016;17(1):213. Available from:
587 http://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1075-9
588 15. Cornforth DM, Dees JL, Ibberson CB, Huse HK, Mathiesen IH, Kirketerp-Møller K, et al. Pseudomonas aeruginosa
589 transcriptome during human infection. Proc Natl Acad Sci U S A. 2018;115(22).
590 16. Glaab E, Bacardit J, Garibaldi JM, Krasnogor N. Using rule-based machine learning for candidate disease gene
591 prioritization and sample classification of cancer gene expression data. PLoS One. 2012;7(7).
592 17. Leung RKK, Wang Y, Ma RCW, Luk AOY, Lam V, Ng M, et al. Using a multi-staged strategy based on machine
593 learning and mathematical modeling to predict genotype-phenotype risk patterns in diabetic kidney disease: A
594 prospective case-control cohort analysis. BMC Nephrol [Internet]. 2013;14(1):1. Available from: BMC
595 Nephrology
596 18. Tan J, Doing G, Lewis KA, Price CE, Chen KM, Cady KC, et al. Unsupervised Extraction of Stable Expression
597 Signatures from Public Compendia with an Ensemble of Neural Networks. Cell Syst [Internet]. 2017 Jul [cited
598 2020 Apr 30];5(1):63–71.e6. Available from: https://linkinghub.elsevier.com/retrieve/pii/S2405471217302314
599 19. Kar SC. Comparing Prediction Accuracy for Supervised Techniques in Gene Expression Data. Math Theory

600 Model. 2014;4(6):108–16.
601 20. Raza K, Hasan A. A Comprehensive Evaluation of Machine Learning Techniques for Cancer Class Prediction
602 Based on Microarray Data. Int J Bioinform Res Appl [Internet]. 2015;11(5):397–416. Available from:
603 http://arxiv.org/abs/1307.7050%0Ahttp://arxiv.org/abs/1307.7050%5Cnhttp://www.inderscienceonline.com/d
604 oi/abs/10.1504/IJBRA.2015.071940
605 21. Savli H, Karadenizli A, Kolayli F, Gundes S, Ozbek U, Vahaboglu H. Expression stability of six housekeeping genes:
606 a proposal for resistance gene quantification studies of Pseudomonas aeruginosa by real-time quantitative RT-
607 PCR. J Med Microbiol [Internet]. 2003 May 1 [cited 2018 Jan 31];52(5):403–8. Available from:
608 http://www.ncbi.nlm.nih.gov/pubmed/12721316
609 22. Casares FM. A Simple Method for Optimization of Reference Gene Identification and Normalization in DNA
610 Microarray Analysis. Med Sci Monit Basic Res [Internet]. 2016 Apr 28 [cited 2018 Nov 10];22:45–52. Available
611 from: http://www.ncbi.nlm.nih.gov/pubmed/27122237
612 23. Yang P, Zhou BB, Yang JY-H, Zomaya AY. Stability of feature selection algorithms and ensemble feature selection
613 methods in bioinformatics. Biol Knowl Discov Handb Preprocessing, mining, postprocessing Biol data.
614 2013;(January 2014):333–52.
615 24. Tabe-Bordbar S, Emad A, Zhao SD, Sinha S. A closer look at cross-validation for assessing the accuracy of gene
616 regulatory networks and models. Sci Rep [Internet]. 2018;8(1):1–11. Available from:
617 http://dx.doi.org/10.1038/s41598-018-24937-4
618 25. Alqarni B, Colley B, Klebensberger J, McDougald D, Rice SA. Expression stability of 13 housekeeping genes
619 during carbon starvation of Pseudomonas aeruginosa. J Microbiol Methods [Internet]. 2016;127:182–7.
620 Available from: http://dx.doi.org/10.1016/j.mimet.2016.06.008
621 26. Noi PT, Kappas M. Comparison of random forest, k-nearest neighbor, and support vector machine classifiers for
622 land cover classification using sentinel-2 imagery. Sensors (Switzerland). 2018;18(1).
623 27. Li L, Weinberg CR, Darden TA, Pedersen LG. Gene selection for sample classification based on gene expression
624 data: study of sensitivity to choice of parameters of the GA/KNN method. Bioinformatics [Internet]. 2001 Dec
625 [cited 2018 Nov 17];17(12):1131–42. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11751221
626 28. Vapnik VN, Vladimir. Estimation of dependences based on empirical data [Internet]. Springer-Verlag; 1982
627 [cited 2018 Nov 17]. 399 p. Available from: https://dl.acm.org/citation.cfm?id=1098680
628 29. Breiman L. Random Forests. Mach Learn [Internet]. 2001 [cited 2018 Nov 17];45(1):5–32. Available from:
629 http://link.springer.com/10.1023/A:1010933404324

630 30. Yusuf Nalca, Lothar Jänsch, Florian Bredenbruch, Robert Geffers JB, Susanne Häussler. Quorum-sensing
631 antagonistic activities of azithromycin in Pseudomonas aeruginosa PAO1: a global approach. Antimicrob Agents
632 Chemother. 2008;50(5):1680–8.
633 31. Kai T, Tateda K, Kimura S, Ishii Y, Ito H, Yoshida H, et al. A low concentration of azithromycin inhibits the mRNA
634 expression of N-acyl homoserine lactone synthesis enzymes, upstream of lasI or rhlI, in Pseudomonas
635 aeruginosa. Pulm Pharmacol Ther [Internet]. 2009 Dec [cited 2018 Nov 13];22(6):483–6. Available from:
636 http://www.ncbi.nlm.nih.gov/pubmed/19393329
637 32. Chang W, Small DA, Toghrol F, Bentley WE. Microarray analysis of Pseudomonas aeruginosa reveals induction of
638 pyocin genes in response to hydrogen peroxide. BMC Genomics. 2005;6:1–14.
639 33. Teitzel GM, Geddie A, De Long SK, Kirisits MJ, Whiteley M, Parsek MR. Survival and Growth in the Presence of
640 Elevated Copper: Transcriptional Profiling of Copper-Stressed Pseudomonas aeruginosa. J Bacteriol [Internet].
641 2006 Oct 15 [cited 2018 Nov 13];188(20):7242–56. Available from:
642 http://www.ncbi.nlm.nih.gov/pubmed/17015663
643 34. Small DA, Chang W, Toghrol F, Bentley WE. Comparative global transcription analysis of sodium hypochlorite,
644 peracetic acid, and hydrogen peroxide on Pseudomonas aeruginosa. Appl Microbiol Biotechnol.
645 2007;76(5):1093–105.
646 35. Nde CW, Jang H-J, Toghrol F, Bentley WE. Toxicogenomic response of Pseudomonas aeruginosa to ortho-
647 phenylphenol. BMC Genomics. 2008 Oct;9(1):473.
648 36. Cummins J, Reen FJ, Baysse C, Mooij MJ, O'Gara F. Subinhibitory concentrations of the cationic antimicrobial
649 peptide colistin induce the pseudomonas quinolone signal in Pseudomonas aeruginosa. Microbiology. 2009
650 Sep;155(9):2826–37.
651 37. Nde CW, Jang HJ, Toghrol F, Bentley WE. Global transcriptomic response of Pseudomonas aeruginosa to
652 chlorhexidine diacetate. Environ Sci Technol. 2009;43(21):8406–15.
653 38. Pan J, Bahar AA, Syed H, Ren D. Reverting Antibiotic Tolerance of Pseudomonas aeruginosa PAO1 Persister Cells
654 by (Z)-4-bromo-5-(bromomethylene)-3-methylfuran-2(5H)-one. PLoS One. 2012;7(9).
655 39. Cirz RT, O'Neill BM, Hammond JA, Head SR, Romesberg FE. Defining the Pseudomonas aeruginosa SOS response
656 and its role in the global response to the antibiotic ciprofloxacin. J Bacteriol. 2006 Oct;188(20):7101–10.
657 40. Tseng GC. Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and
658 assessment of gene effects. Nucleic Acids Res [Internet]. 2001;29(12):2549–57. Available from:
659 https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/29.12.2549

660 41. Mar J. Applying Data-Driven Normalization Strategies for qPCR Data Using Bioconductor Preprocessing Raw
661 qPCR Data Normalizing qPCR Data. 2010;1–8.
662 42. Smyth GK. Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray
663 Experiments. Stat Appl Genet Mol Biol [Internet]. 2004 Jan 12 [cited 2019 Apr 29];3(1):1–25. Available from:
664 http://www.ncbi.nlm.nih.gov/pubmed/16646809
665 43. Richter K, Haslbeck M, Buchner J. The Heat Shock Response: Life on the Verge of Death. Mol Cell [Internet].
666 2010 Oct 22 [cited 2018 Nov 16];40(2):253–66. Available from:
667 http://www.ncbi.nlm.nih.gov/pubmed/20965420
668 44. Vihervaara A, Duarte FM, Lis JT. Molecular mechanisms driving transcriptional stress responses. Nat Rev Genet
669 [Internet]. 2018;19(6):385–97. Available from: http://dx.doi.org/10.1038/s41576-018-0001-6
670 45. Holloway BW. Genetic Recombination in Pseudomonas aeruginosa. Microbiology [Internet]. 1955 Dec 1 [cited
671 2018 Nov 14];13(3):572–81. Available from: http://www.ncbi.nlm.nih.gov/pubmed/13278508
672 46. Piao J, Sun J, Yang Y, Jin T, Chen L, Lin Z. Target gene screening and evaluation of prognostic values in non-small
673 cell lung cancers by bioinformatics analysis. Gene [Internet]. 2018 Mar 20 [cited 2018 Nov 6];647(1):306–11.
674 Available from: http://www.ncbi.nlm.nih.gov/pubmed/29305979
675 47. Touw WG, Bayjanov JR, Overmars L, Backus L, Boekhorst J, Wels M, et al. Data mining in the Life Sciences with
676 Random Forest: a walk in the park or lost in the jungle? Brief Bioinform [Internet]. 2013 May 1 [cited 2018 Nov
677 18];14(3):315–26. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22786785
678 48. Pai S, Hui S, Isserlin R, Shah MA, Kaka H, Bader GD. netDx: interpretable patient classification using integrated
679 patient similarity networks. Mol Syst Biol [Internet]. 2019 Mar 14 [cited 2019 May 1];15(3):e8497. Available
680 from: http://www.ncbi.nlm.nih.gov/pubmed/30872331
681 49. Park H, Shimamura T, Imoto S, Miyano S. Adaptive NetworkProfiler for Identifying Cancer Characteristic-Specific
682 Gene Regulatory Networks. J Comput Biol [Internet]. 2017;25(2):cmb.2017.0120. Available from:
683 http://online.liebertpub.com/doi/10.1089/cmb.2017.0120
684 50. Falciani F, Diab A, Sabine V, Williams T, Ortega F, George S, et al. Hepatic transcriptomic profiles of European
685 flounder (Platichthys flesus) from field sites and computational approaches to predict site from stress gene
686 responses following exposure to model toxicants. Aquat Toxicol [Internet]. 2008 Nov 11 [cited 2018 Nov
687 18];90(2):92–101. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18823667
688 51. Nota B, Verweij RA, Molenaar D, Ylstra B, van Straalen NM, Roelofs D. Gene Expression Analysis Reveals a Gene
689 Set Discriminatory to Different Metals in Soil. Toxicol Sci [Internet]. 2010 May [cited 2018 Nov 18];115(1):34–

690 40. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20133373
691 52. Krämer R. Bacterial stimulus perception and signal transduction: Response to osmotic stress. Chem Rec
692 [Internet]. 2010 Aug 16 [cited 2018 Nov 23];10(4):217–29. Available from:
693 http://www.ncbi.nlm.nih.gov/pubmed/20607761
694 53. Schurig-Briccio LA, Farías RN, Rodríguez-Montelongo L, Rintoul MR, Rapisarda VA. Protection against oxidative
695 stress in Escherichia coli stationary phase by a phosphate concentration-dependent genes expression. Arch
696 Biochem Biophys [Internet]. 2009 Mar 1 [cited 2018 Nov 22];483(1):106–10. Available from:
697 http://www.ncbi.nlm.nih.gov/pubmed/19138658
698 54. Parsons JB, Rock CO. Bacterial lipids: metabolism and membrane homeostasis. Prog Lipid Res [Internet]. 2013
699 Jul [cited 2018 Nov 22];52(3):249–76. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23500459
700 55. Storvik KAM, Foster PL. RpoS, the stress response sigma factor, plays a dual role in the regulation of Escherichia
701 coli's error-prone DNA polymerase IV. J Bacteriol [Internet]. 2010 Jul [cited 2018 Nov 22];192(14):3639–44.
702 Available from: http://www.ncbi.nlm.nih.gov/pubmed/20472798
703 56. Starosta AL, Lassak J, Jung K, Wilson DN. The bacterial translation stress response. FEMS Microbiol Rev
704 [Internet]. 2014 Nov [cited 2018 Nov 22];38(6):1172–201. Available from:
705 http://www.ncbi.nlm.nih.gov/pubmed/25135187
706 57. Adamson B, Norman TM, Jost M, Cho MY, Nuñez JK, Chen Y, et al. A Multiplexed Single-Cell CRISPR Screening
707 Platform Enables Systematic Dissection of the Unfolded Protein Response. Cell [Internet]. 2016 Dec 15 [cited
708 2018 Nov 16];167(7):1867–1882.e21. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27984733
709 58. Molina-Mora JA, Campos-Sanchez R, Garcia F. Gene Expression Dynamics Induced by Ciprofloxacin and Loss of
710 Lexa Function in Pseudomonas aeruginosa PAO1 Using Data Mining and Network Analysis. In: 2018 IEEE
711 International Work Conference on Bioinspired Intelligence (IWOBI) [Internet]. IEEE; 2018 [cited 2018 Nov 16]. p.
712 1–7. Available from: https://ieeexplore.ieee.org/document/8464130/
713
714
715
716
717
718

Supporting information

Supplementary Figures

Fig. S1. Evaluation of normalization by *proC* gene. (A) Systematic microarray data selection in GEO-NCBI database according to experimental design; (B) Comparison of conditions (control vrs perturbations) for other housekeeping genes after normalization by *proC* gene; (C) Dispersion of intensities of samples, showing similar distribution between samples.

Fig. S2. **Evaluation of SP method for gene ranking by importance, case for SVM and KNN algorithms.** (A-C) Accuracy, F1 and kappa values after iterations of classification with the first top 200 genes (adding genes 1-to-1). (B-D) ROC plot using selected top 56 genes for evaluation of performance of the algorithm. (E) Ranking and importance value of top 56 genes (same importance metric for both algorithms).

Fig. S3. **Evaluation of MP method for gene ranking by frequency, case for KNN and RF algorithms.** (A) KNN evaluation of ranking of top 56 genes by frequency after iterations of 100 data partitions and classification model fitting and dispersion of metrics across 100 iterations. (B) Same evaluation, case for RF algorithm.

Fig. S4. Systems level description and gene expression comparison given by core response genes. (A) Gene expression and clustering of the 46 core response determinants of *P. aeruginosa*. (B-C) Comparison of gene expression (logR) of kmedoids for each cluster using small world networks.

Supplementary tables

Table S1. Evaluation of housekeeping genes (columns) after data normalization with another specific housekeeping gene (rows).

Table S2. Annotation of genes of core response genes of *P. aeruginosa* PAO1

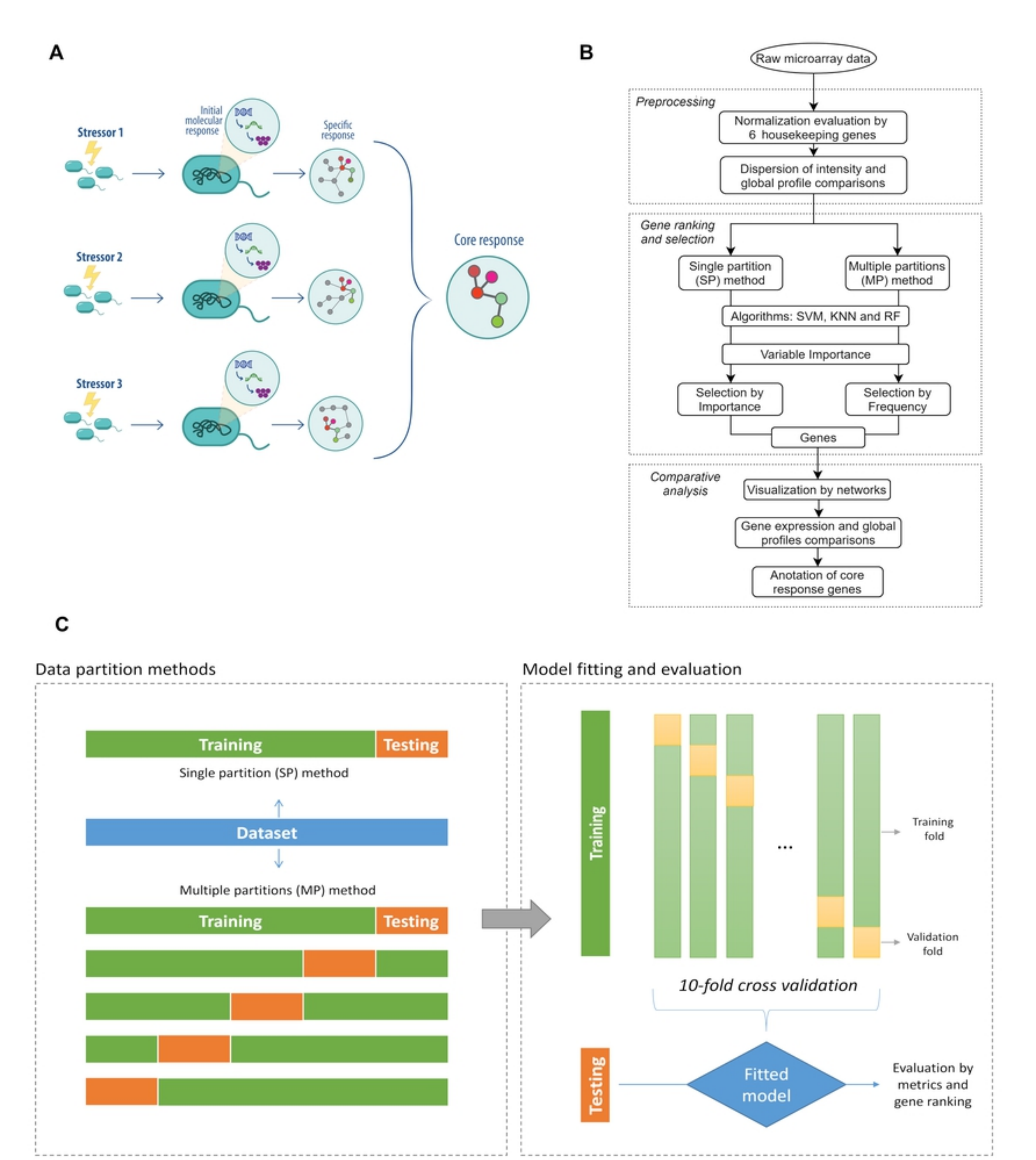
A
Stressor 1
Initial molecular response
Specific response
Stressor 2
Stressor 3
Core response
B
Raw microarray data
Preprocessing
Normalization evaluation by 6 housekeeping genes
Dispersion of intensity and global profile comparisons
Gene ranking and selection
Single partition (SP) method
Multiple partitions (MP) method
Algorithms: SVM, KNN and RF
Variable Importance
Selection by Importance
Selection by Frequency
Genes
Comparative analysis
Visualization by networks
Gene expression and global profiles comparisons
Anotation of core response genes
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

Figure1

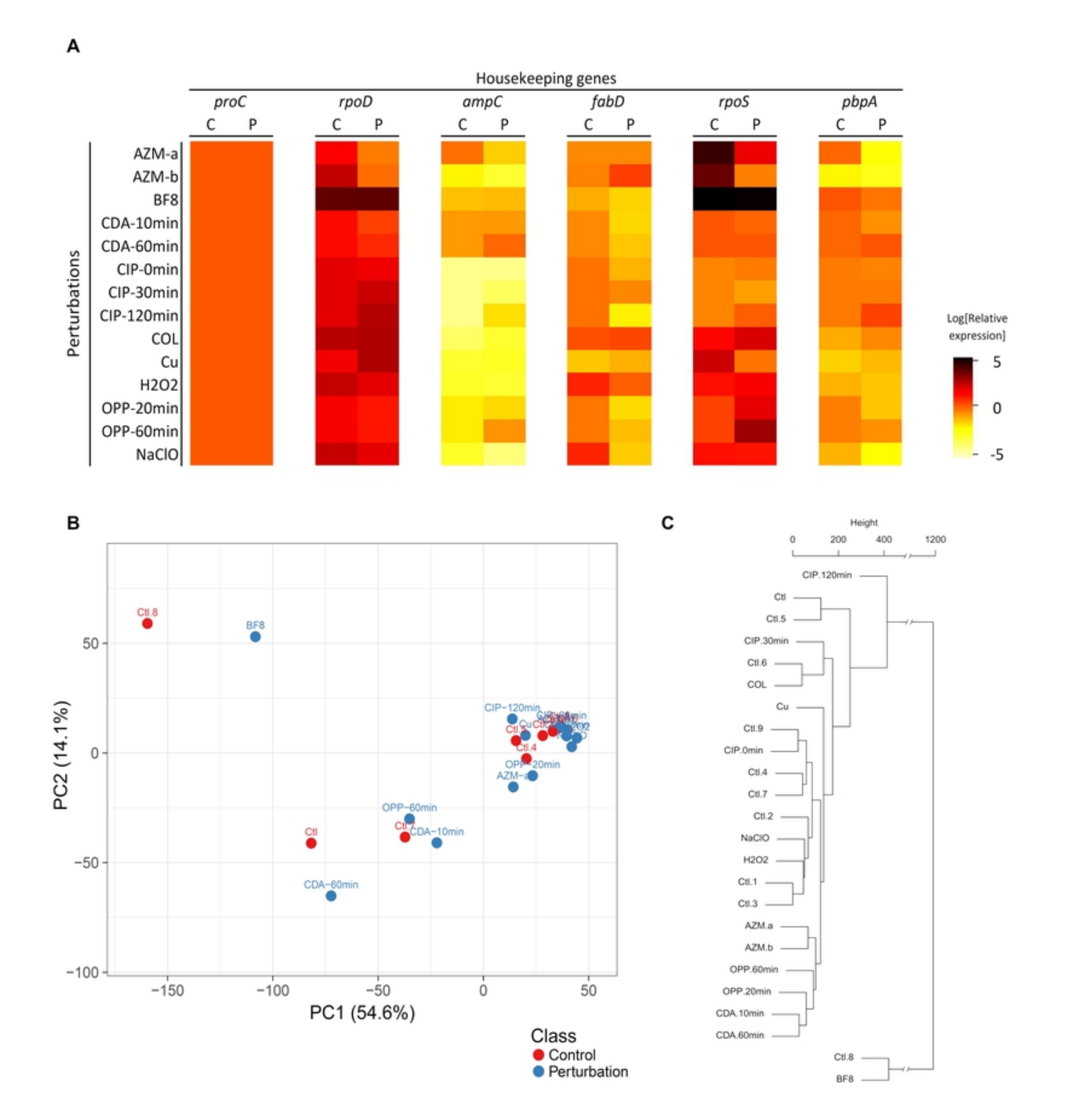
A
Housekeeping genes
proC
rpoD
ampC
fabD
rpoS
pbpA
C
P
Perturbations
AZM-a
AZM-b
BF8
CDA-10min
CDA-60min
CIP-0min
CIP-30min
CIP-120min
COL
Cu
H2O2
OPP-20min
OPP-60min
NaClO
Log[Relative expression]
5
0
-5
B
PC2 (14.1%)
PC1 (54.6%)
Ctl.8
BF8
CIP−120min
Ctl.4
OPP−20min
AZM−a
OPP−60min
Ctl.7
CDA−10min
Ctl
CDA−60min
Class
Control
Perturbation
C
Height
0
200
400
1200
CIP.120min
Ctl
Ctl.5
CIP.30min
Ctl.6
COL
Cu
Ctl.9
CIP.0min
Ctl.4
Ctl.7
Ctl.2
NaClO
H2O2
Ctl.1
Ctl.3
AZM.a
AZM.b
OPP.60min
OPP.20min
CDA.10min
CDA.60min
Ctl.8
BF8

Figure2

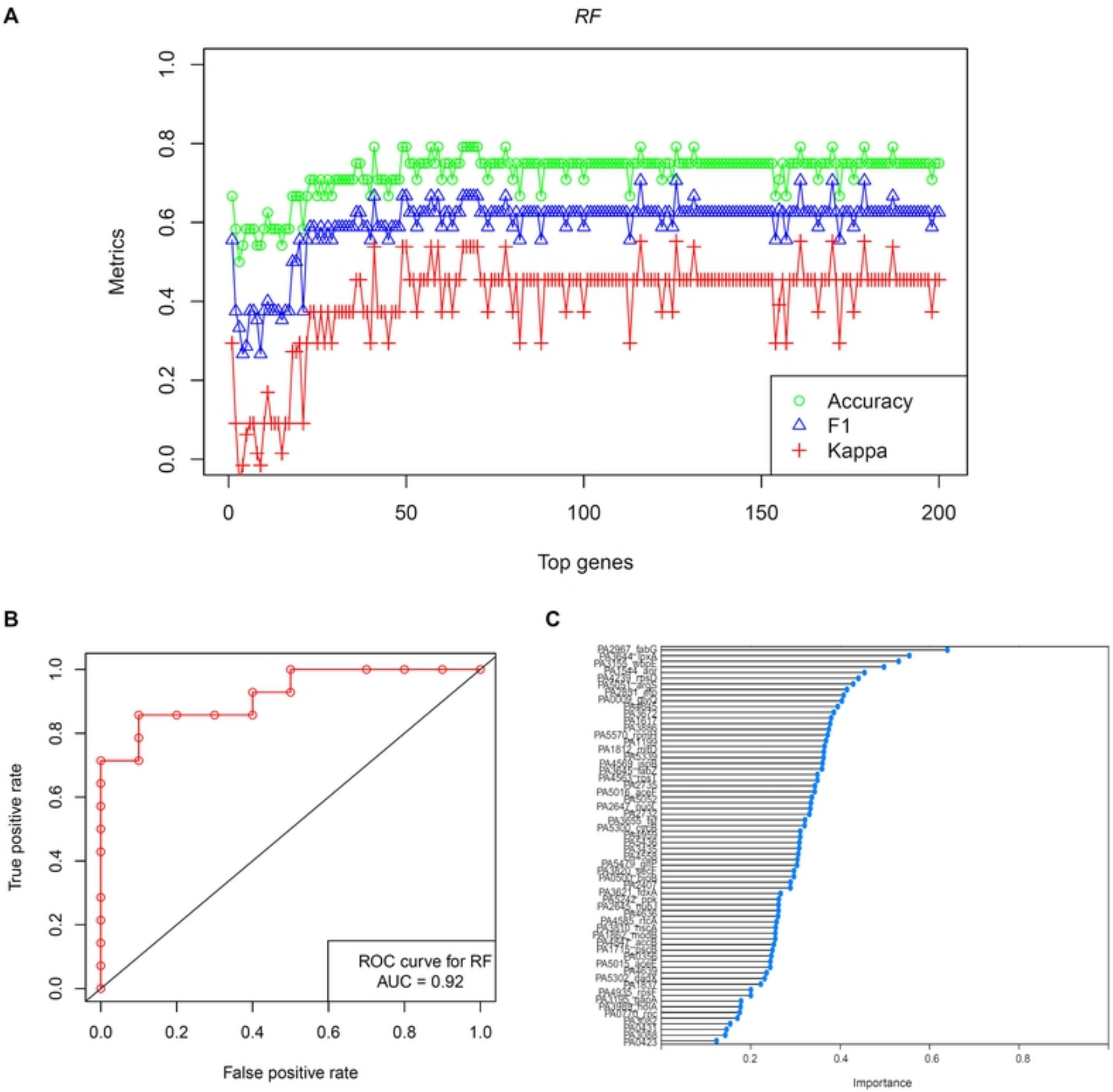
A
RF
Metrics
Top genes
Accuracy
F1
Kappa
B
True positive rate
False positive rate
ROC curve for RF
AUC = 0.92
C
Importance

Figure3

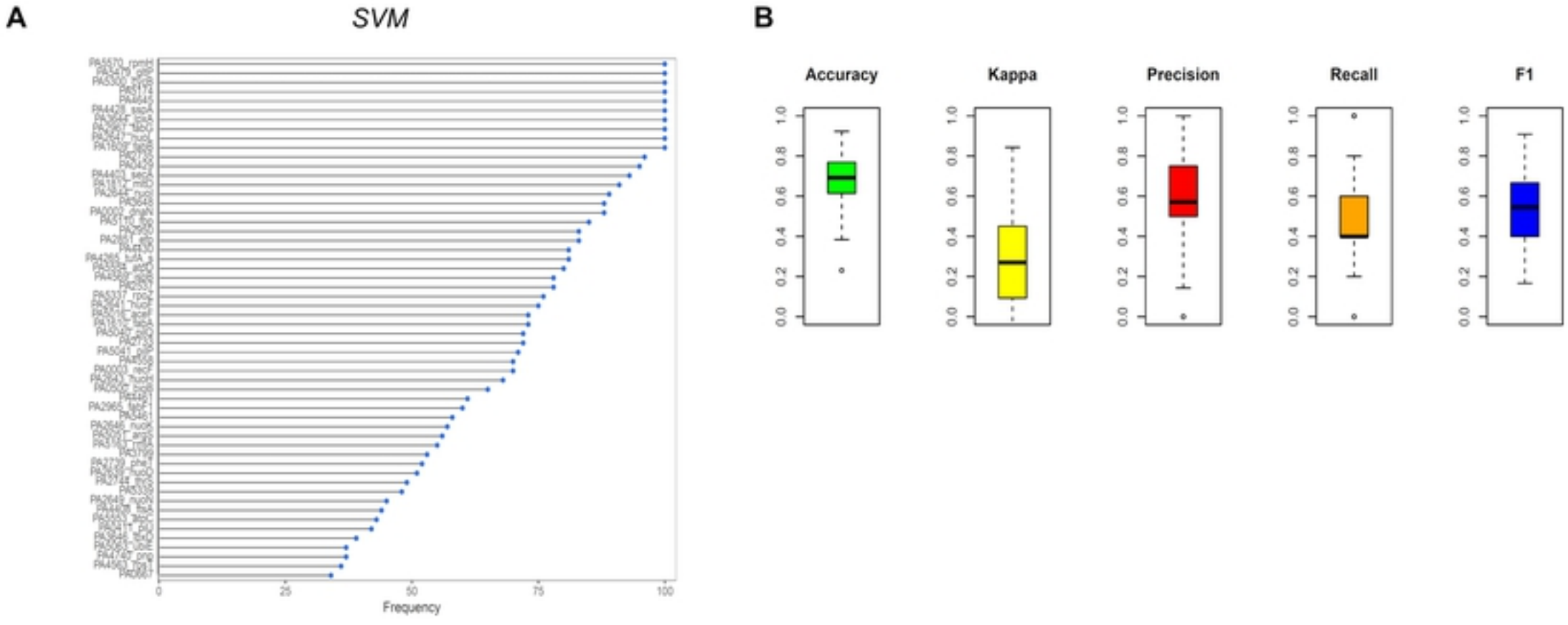
A
SVM
Frequency
0
25
50
75
100
B
Accuracy
Kappa
Precision
Recall
F1
0.0
0.2
0.4
0.6
0.8
1.0

Figure4

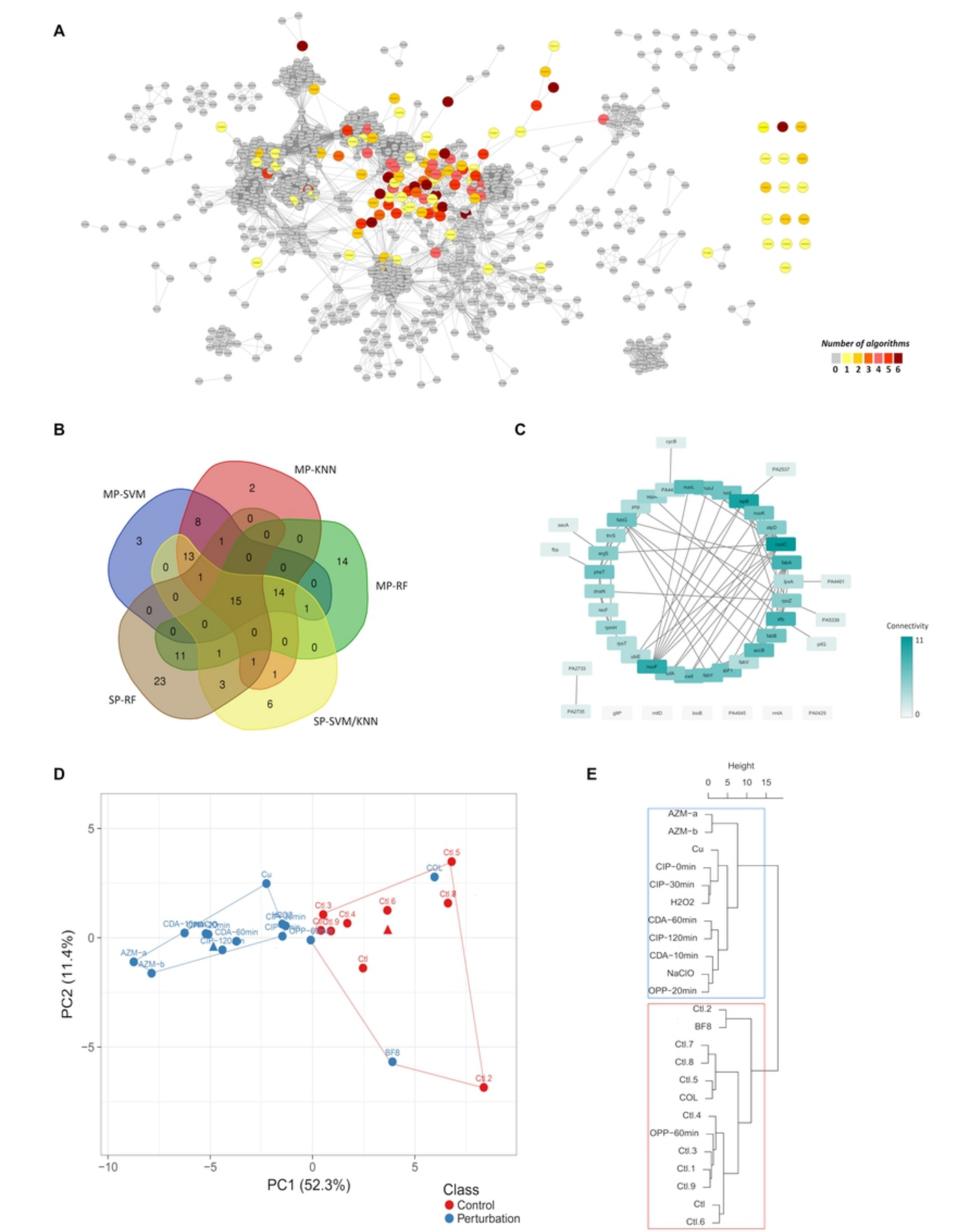
A
Number of algorithms
0 1 2 3 4 5 6
B
MP-SVM
MP-KNN
MP-RF
SP-RF
SP-SVM/KNN
C
Connectivity
11
0
D
PC2 (11.4%)
PC1 (52.3%)
Class
Control
Perturbation
E
Height
0 5 10 15
AZM-a
AZM-b
Cu
CIP-0min
CIP-30min
H2O2
CDA-60min
CIP-120min
CDA-10min
NaClO
OPP-20min
Ctl.2
BF8
Ctl.7
Ctl.8
Ctl.5
COL
Ctl.4
OPP-60min
Ctl.3
Ctl.1
Ctl.9
Ctl
Ctl.6

Figure5

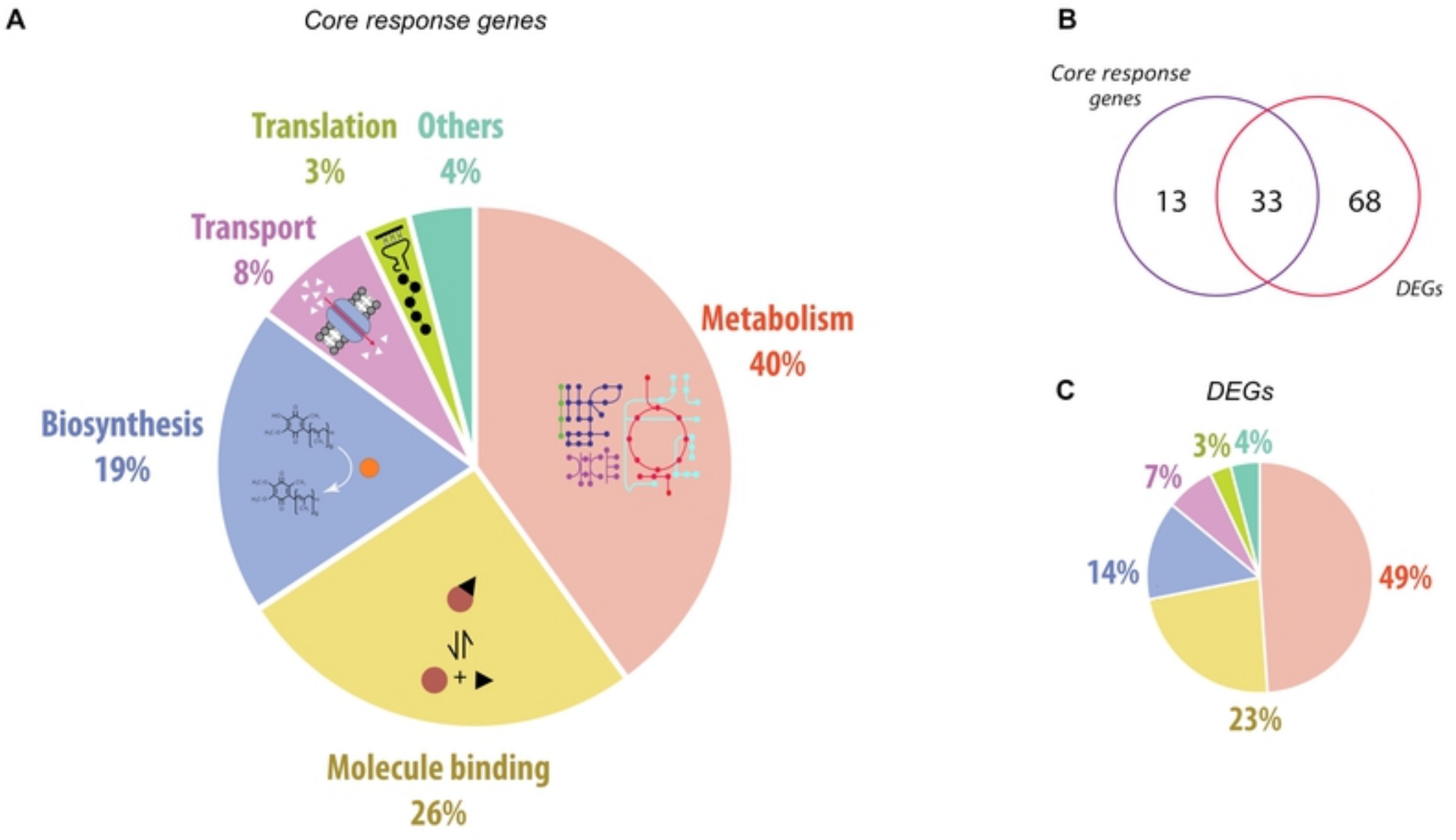
A
Core response genes
Translation
3%
Others
4%
Transport
8%
Metabolism
40%
Biosynthesis
19%
Molecule binding
26%
B
Core response genes
13
33
68
DEGs
C
DEGs
3%
4%
7%
14%
49%
23%

Figure6