UNIVERSIDAD DE COSTA RICA

SISTEMA DE ESTUDIOS DE POSGRADO

MOLECULAR DETERMINANTS OF ANTIBIOTIC TOLERANCE IN THE HIGH-RISK *Pseudomonas aeruginosa* AG1 BY A MULTI-OMICS APPROACH: FROM THE GENOME TO THE TRANSCRIPTOMIC NETWORK IN RESPONSE TO CIPROFLOXACIN.

DETERMINANTES MOLECULARES DE LA TOLERANCIA A LOS ANTIBIÓTICOS EN LA CEPA DE ALTO RIESGO *Pseudomonas aeruginosa* AG1 MEDIANTE UN ENFOQUE MULTI-ÓMICO: DEL GENOMA A LA RED TRANSCRIPTÓMICA EN RESPUESTA A LA CIPROFLOXACINA.

Tesis sometida a la consideración de la Comisión del Programa de Posgrado de Doctorado en Ciencias para optar al grado y título de Doctorado en Ciencias

JOSE ARTURO MOLINA MORA

Ciudad Universitaria Rodrigo Facio, Costa Rica

2020

DEDICATION/DEDICATORIA

I dedicate this work to my family: my parents, brothers and nieces for the love and unconditional support since always, and my best friend Will for the journey we have traveled between happiness and the support in the less fortunate days.

Also to my closest friends, for their encouragement and motivation always.

To all those who believe in dreams and who have always inspired me.

To life, for making everything possible!

Dedico este trabajo a mi familia: mis padres, hermanos y sobrinas por el amor y apoyo incondicional desde siempre, y mi mejor amigo Will por el camino recorrido entre felicidad y el soporte en los días menos afortunados.

También a mis amigos más cercanos, por su aliento y motivación siempre.

A todos que los que creen en los sueños y que me han inspirado siempre.

A la vida, por hacer todo posible!

ACKNOWLEDGMENTS/AGRADECIMIENTOS

I would like to thank all those people who in one way or another have made this possible, including:

To Dr. Fernando García for his guidance, support, scolding and confidence in these four years. To professors Dr. Rodrigo Mora and Dra. Rebeca Campos, who have inspired and received me to work in this field of science.

To colleagues, researchers and collaborators who have participated in the projects associated with this work.

To my friends and professors of the internships at University of A Coruña in Spain and Fudan University in Shanghai-China, who welcomed me with open arms and who made the experience unmatched.

To the academic and administrative staff of the Faculty of Microbiology, Research Center for Tropical Diseases, the Doctoral Program in Sciences and University of Costa Rica for their support in the management processes that allowed me to develop this research.

Quisiera agradecer este trabajo a todas esas personas que de una u otra forma han permitido que esto sea posible, incluyendo:

Al Dr. Fernando García por su guía, apoyo, regaños y confianza en estos cuatro años. A los profesores Dr. Rodrigo Mora y Dra. Rebeca Campos, que me han inspirado y recibido para trabajar en esta área de la ciencia.

A los colegas, investigadores y colaboradores que han participado en los proyectos asociados a este trabajo.

A mis amigos y profesores de mis pasantías en la Universidad de A Coruña en España y Universidad de Fudan en Shanghái-China, que me recibieron con los brazos abiertos y que hicieron que la experiencia fuese inigualable.

Al personal administrativo y académico de la Facultad de Microbiología, Centro de Investigación de Enfermedades Tropicales y Programa de Doctorado en Ciencias por el apoyo en los diversos procesos de gestión que permitieron desarrollar esta investigación.

"Esta tesis fue aceptada por la Comisión del Programa de Estudios de Posgrado de Doctorado en Ciencias de la Universidad de Costa Rica, como requisito parcial para optar al grado y título de **Doctorado en Ciencias**."

Dr. Carlos Chacón Díaz
Representante del Decano
Sistema de Estudios de Posgrado

Fernando

Dr. Fernando García Santamaría
Director de Tesis

Dr. Rodrigo Mora Rodríguez
Asesor

Rebeca

Dra. Rebeca Campos Sánchez
Asesora

Dr. José Guevara Coto
Representante de la Directora del Programa de Posgrado

Jose Arturo Molina Mora
Sustentante

TABLE OF CONTENTS

RESUMEN

La resistencia a los antibióticos es una amenaza importante para la salud pública porque compromete la administración de una terapia antibiótica adecuada. *Pseudomonas aeruginosa* es un patógeno oportunista que causa infecciones entre huéspedes inmunodeprimidos. *P. aeruginosa* AG1 (PaeAG1) es una cepa costarricense con resistencia a múltiples antibióticos como los β-lactámicos (incluidos los carbapenémicos), aminoglucósidos y fluoroquinolonas. PaeAG1 se identificó como el primer aislamiento de *P. aeruginosa* llevando los genes VIM-2 e IMP-18 que codifican las enzimas metalo-β-lactamasas (MBL). Según la Organización Mundial de la Salud (OMS), esta cepa se considera crítica, siendo clasificada en el grupo de Prioridad 1 por su resistencia a los carbapenémicos. PaeAG1 tiene características particulares a niveles genómicos y fenómicos, muchas de ellas relacionadas con la resistencia a los antibióticos. Debido a esto fue de interés estudiar los determinantes moleculares de la tolerancia a los antibióticos en PaeAG1 utilizando un enfoque multi-ómico.

Primero, el ensamblaje del genoma fue el paso inicial para comprender la arquitectura genómica de esta cepa de alto riesgo. Del estudio con 13 enfoques diferentes, la selección del mejor ensamblaje reveló que el genoma de PaeAG1 tiene 57 islas genómicas que albergan seis profagos y dos integrones completos con los genes de las MBL. Además, se encontraron 250 genes de virulencia y 60 genes asociados a la resistencia a los antibióticos.

Segundo, un enfoque genómico comparativo fue implementado para definir y actualizar la relación filogenética entre los genomas completos de *P. aeruginosa*, el contenido de islas genómicas en otras cepas, y la arquitectura de las regiones genómicas alrededor de los dos integrones portadores de MBL. Para el caso del IMP-18, el integrón que lo contiene y la arquitectura alrededor nunca habían sido reportados en la literatura.

Luego, estudiamos el perfil proteómico de PaeAG1 después de la exposición a antibióticos usando electroforesis en gel bidimensional con un protocolo de análisis de imágenes y aprendizaje automático (inteligencia artificial). Los perfiles proteómicos mostraron que ciprofloxacina (CIP) induce un patrón proteico similar al control sin antibióticos, en contraste con otros antibióticos que se agruparon por separado.

En cuarto lugar, para estudiar la respuesta central a múltiples perturbaciones en *P. aeruginosa*, es decir, el perturboma central, un enfoque de aprendizaje automático fue implementado. Utilizando datos transcriptómicos públicos, evaluamos seis enfoques para clasificar y seleccionar genes. La anotación molecular de 46 genes de la respuesta central reveló funciones biológicas relacionadas con la reparación del daño del ADN, metabolismo y la respiración aeróbica en el contexto de la tolerancia al estrés.

Finalmente, para evaluar los efectos de la ciprofloxacina en PaeAG1, realizamos una comparación de curvas de crecimiento, análisis de expresión diferencial usando RNA-Seq y análisis de redes. El análisis transcriptómico mostró una expresión diferencial de 518 genes en el tiempo después del tratamiento con ciprofloxacina, incluyendo genes de fagos residentes que se regularon positivamente. Este último caso se validó a nivel fenómico utilizando ensayos de placa de fagos y que explicó las observaciones fenotípicas en la reducción de las curvas de crecimiento.

En conjunto, utilizando un enfoque multiómico (a niveles genómico, genómico comparativo, perturbómico, transcriptómico, proteómico y fenómico), proporcionamos nuevos conocimientos sobre los determinantes genómicos y transcriptómicos asociados con la tolerancia a antibióticos en PaeAG1. Estos resultados no solo explican en parte la condición de alto riesgo de esta cepa que le permite conquistar ambientes nosocomiales y su perfil de multirresistencia, sino que esta información eventualmente podrá ser usada como parte de las estrategias para combatir a este patógeno.

SUMMARY

Antibiotic resistance is a major threat to public health because it compromises the administration of appropriate antibiotic therapy. *Pseudomonas aeruginosa* is an opportunist pathogen that causes infections among immunocompromised hosts. *P. aeruginosa AG1* (PaeAG1) is a Costa Rican strain with resistance to multiple antibiotics such as β-lactams (including carbapenems), aminoglycosides, and fluoroquinolones. PaeAG1 was identified as the first report of a *P. aeruginosa* isolate carrying both VIM-2 and IMP-18 genes encoding for metallo-β-lactamases (MBL) enzymes. According to the World Health Organization (WHO), this strain is considered critical, being classified as Priority 1 group because of its resistance to carbapenems. PaeAG1 has particular features at genomic and phenomic levels, many of them related to antibiotic resistance. Owing to these traits, we were interested in studying the molecular determinants of antibiotic tolerance in PaeAG1 using a multi-omics approach.

First, the genome assembly was the initial step to understand the genomic architecture of the ST-111 high-risk PaeAG1 strain. From 13 different approaches, the selection of the best assembly revealed that the PaeAG1 genome has 57 genomic islands harboring six prophages, and two complete integrons with the MBLs genes. In addition, 250 genes are anticipated to play a role in virulence and 60 genes in antibiotic resistance.

Second, a comparative genomic approach was implemented to define and update the phylogenetic relationship among complete *P. aeruginosa* genomes, the genomic island content in other strains, and the architecture of genomic regions around the two MBL-carrying integrons. For IMP-18, the integron and the genomic landscape are a unique arrangement, never reported before.

Subsequently, the proteomic profile of PaeAG1 was studied after exposure to antibiotics using 2-dimensional gel electrophoresis data with an image analysis pipeline and a machine learning (artificial intelligence) approach. The proteomic profiles showed that CIP was close to the control (LB medium, without antibiotics), contrasting to other antibiotics that were clustered separately.

Fourth, to study the central response to multiple perturbations in *P. aeruginosa*, i.e. the core perturbome, a machine learning approach was used. Using public transcriptomic data, we evaluated six approaches to rank and select genes. The molecular annotation of 46 core response genes revealed biological functions related to DNA damage repair, metabolism and aerobic respiration in the context of tolerance to stress.

Finally, in order to evaluate the effects of ciprofloxacin on PaeAG1, a growth curves comparison, differential expression analysis using RNA-Seq, and network analysis were performed. It was observed a reduction in the growth curve rate as the sub-inhibitory ciprofloxacin concentrations were increased. The transcriptomic analysis showed a differential expression of 518 genes overtime after ciprofloxacin treatment, including resident-phage genes which were up-regulated. The former was validated at the phenomic level using phage plaque assays, explaining the reduction of the growth curve rate it was observed.

Altogether, using a multi-omics approach (at genomic, comparative genomic, perturbomic, transcriptomic, proteomic and phenomic levels), we provided new insights about the genomic and transcriptomic determinants associated with antibiotic tolerance in PaeAG1. These results not only partially explain the high-risk condition of this strain that enables it to conquer nosocomial environments and its multi-resistance profile, but also this information may eventually be used as part of the strategies to fight this pathogen.

Abbreviations

2D-GE: 2-dimensional gel electrophoresis

CIP: Ciprofloxacin

GIC: Genomic islands cluster

IPM: Imipenem

IMP: Imipenemase

MBL: Metallo-β-lactamase

MLST: Multilocus sequence typing

PaeAG1: *Pseudomonas aeruginosa* AG1

ST: Sequence type

TOB: Tobramycin

VIM: Verona integron-encoded MBLs

WHO: World Health Organization (WHO)

List of Figures

UNIVERSIDAD DE COSTA RICA

SEP Sistema de Estudios de Posgrado

Autorización para digitalización y comunicación pública de Trabajos Finales de Graduación del Sistema de Estudios de Posgrado en el Repositorio Institucional de la Universidad de Costa Rica.

Yo, Jose Arturo Molina Mora, con cédula de identidad 1-1252-0428, en mi condición de autor del TFG titulado Molecular determinants of antibiotic tolerance in the high-risk Pseudomonas aeruginosa AG1 by a multi-omics approach: from the genome to the transcriptomic network in response to ciprofloxacin

Autorizo a la Universidad de Costa Rica para digitalizar y hacer divulgación pública de forma gratuita de dicho TFG a través del Repositorio Institucional u otro medio electrónico, para ser puesto a disposición del público según lo que establezca el Sistema de Estudios de Posgrado. **SI** ☒ **NO** * □

***En caso de la negativa favor indicar el tiempo de restricción: _______________ año (s).**

Este Trabajo Final de Graduación será publicado en formato PDF, o en el formato que en el momento se establezca, de tal forma que el acceso al mismo sea libre, con el fin de permitir la consulta e impresión, pero no su modificación.

Manifiesto que mi Trabajo Final de Graduación fue debidamente subido al sistema digital Kerwá y su contenido corresponde al documento original que sirvió para la obtención de mi título, y que su información no infringe ni violenta ningún derecho a terceros. El TFG además cuenta con el visto bueno de mi Director (a) de Tesis o Tutor (a) y cumplió con lo establecido en la revisión del Formato por parte del Sistema de Estudios de Posgrado.

INFORMACIÓN DEL ESTUDIANTE:

Nombre Completo: Jose Arturo Molina Mora.

Número de Carné: A33282 Número de cédula: 1-1252-0428.

Correo Electrónico: jose.molinamora@ucr.ac.cr.

Fecha: 14-enero-2020. Número de teléfono: 88859445.

Nombre del Director (a) de Tesis o Tutor (a): Dr. Fernando García Santamaría.

FIRMA ESTUDIANTE

Nota: El presente documento constituye una declaración jurada, cuyos alcances aseguran a la Universidad, que su contenido sea tomado como cierto. Su importancia radica en que permite abreviar procedimientos administrativos, y al mismo tiempo genera una responsabilidad legal para que quien declare contrario a la verdad de lo que manifiesta, puede como consecuencia, enfrentar un proceso penal por delito de perjurio, tipificado en el artículo 318 de nuestro Código Penal. Lo anterior implica que el estudiante se vea forzado a realizar su mayor esfuerzo para que no sólo incluya información veraz en la Licencia de Publicación, sino que también realice diligentemente la gestión de subir el documento correcto en la plataforma digital Kerwá.

INTRODUCTION

Biological systems rely on the DNA–RNA–protein information transfer paradigm that determines the phenotype of an organism (O'Donnell, Ross, & Stanton, 2020). The comprehensive or global assessment of a set of molecules, which requires interpretation of molecular intricacy and variations at multiple levels, has been referred to as "-omics sciences" (Subramanian, Verma, Kumar, Jere, & Anamika, 2020). Classical -omics levels refer to genomics, transcriptomics, proteomics, and metabolomics, but the spectrum of –omics has been extended to other biological data such as epigenomics, phenomics, perturbomics, lipidomics, venomics, and many others.

The current high throughput nature of these techniques, as well as their increased accessibility in terms of time and cost, have triggered the volume of information that can be gathered in individual studies including multiple –omics levels, which together are called "multi-omics" (O'Donnell et al., 2020; Subramanian et al., 2020).

Multi-omics can provide a greater understanding of the flow of information in biological systems, from the original biological set-up or condition (genetic, environmental, or developmental) to the functional consequences or relevant interactions (Civelek & Lusis, 2014; Hasin, Seldin, & Lusis, 2017). This makes it possible to draw more comprehensive conclusions on the biological processes in which these data sets must be integrated and analyzed as a holistic system (O'Donnell et al., 2020). Also, integrated approaches that combine individual omics data help to bridge the gap from genotype to phenotype, are considered a promising strategy to understand the complexity of biological systems and unravel the mechanisms underlying the biological condition of interest (Civelek & Lusis, 2014; Subramanian et al., 2020).

In this context, in this work, a comprehensive multi-omics approach was implemented to study molecular determinants of antibiotic tolerance in a model of *Pseudomonas aeruginosa*, including genomics, transcriptomics, perturbomics, proteomics, and phenomics as main –omics levels.

Antibiotic resistance and tolerance

Antimicrobial resistance is the ability of a microbe to grow in an inhibitory concentration of an antibiotic, explained by inherited mechanisms (Berti & Hirsch, 2020; Brauner, Fridman, Gefen, & Balaban, 2016). Tolerance is generally used to describe the ability of microorganisms to survive transient exposure to *bactericidal* antibiotics, which can be inherited or not, with a reduced rate of antimicrobial killing, and often achieved by slowing down the cell growth (Berti & Hirsch, 2020; Brauner et al., 2016).

Antibiotic resistance is a major threat to public health because it compromises the administration of appropriate antibiotic therapy. This reduces the therapeutic options to treat infections, increasing patient morbidity and mortality (Farajzadeh Sheikh et al., 2019; Woodford, Turton, & Livermore, 2011), as well as it causes an increase in the costs of health services. The situation is aggravated by the emergence of strains resistant to multiple antibiotics (Firme, Kular, Lee, & Song, 2010), the knowledge limitation of interactions with pathogens and mechanisms of the action of antimicrobial agents, and the reduced development of new antibiotics (Brazas, Brazas, Hancock, & Hancock, 2005). The use of antibiotics below the minimum inhibitory concentration (MIC) or sub-inhibitory concentration also contributes to antibiotic resistance as it selects pre-existing resistant organisms and allows the strains to continue growing (McVicker et al., 2014). Since sub-inhibitory antibiotic concentrations are found in many natural environments, bacteria can naturally trigger mechanisms of tolerance and resistance (Andersson & Hughes, 2014). However, the fundamental mechanisms of bacterial response to antibiotics have not been fully elucidated (Stewart et al., 2015).

Since in this study we consider not only inherited mechanisms (genomic level, focused on resistance) but also transcriptomic and phenotypical observations using sub-inhibitory antibiotic concentrations (with mechanisms than can be related to tolerance), here we follow the "antibiotic tolerance" definition by Ciofu & Tolker-Nielsen (2019) to refer to all the molecular responses that the bacterial face when exposed to antibiotics.

Pseudomonas aeruginosa AG1 (PaeAG1)

Pseudomonas aeruginosa is an opportunist and versatile pathogen able to survive in a wide variety of environments (Klockgether et al., 2010). With a large genome (6-7.5 Mb), *P. aeruginosa* strains have a large proportion of the genome (>8%) dedicated to regulatory functions (Cabot et al., 2016) resulting in a consequent diversity of metabolic capabilities and responses to stress. Because of these features, *P. aeruginosa* is responsible for infections among immunocompromised hosts (Lu et al., 2016) and nosocomial infections (Fernández, Corral-Lugo, & Krell, 2018). However, the treatment of *P. aeruginosa* infections is challenging due to its many intrinsic and acquired mechanisms of resistance (Toval et al., 2015), resulting in significant morbidity and mortality. According to the World Health Organization (WHO), resistance to carbapenems in *P. aeruginosa*, *Acinetobacter baumannii,* and Enterobacteriaceae family is considered a critical issue in the context of antibiotic resistance, being classified as Priority 1 group (World Health Organization, 2017).

In Costa Rica, the isolation of carbapenem-resistant *P. aeruginosa* strains is relatively common in some major hospitals, up to 63.1% of prevalence, as previously reported (Toval et al., 2015), much higher than the frequencies observed in other countries (Hong et al., 2015). The Costa Rican strain *P. aeruginosa* AG1 (PaeAG1) was identified as the first report of a *P. aeruginosa* isolate carrying both

VIM-2 and IMP-18 genes encoding for Metallo-β-lactamases (MBLs) enzymes, both with carbapenemase activity (Toval et al., 2015). Later, another isolate from the United Kingdom with the same enzymes was reported (Turton et al., 2015).

PaeAG1 was grown from a sputum sample of a patient from the Intensive Care Unit in the San Juan de Dios Hospital (San José, Costa Rica) in 2010. This strain has resistance to multiple antibiotics such as β-lactams (including carbapenems), aminoglycosides, and fluoroquinolones, being only sensible to colistin.

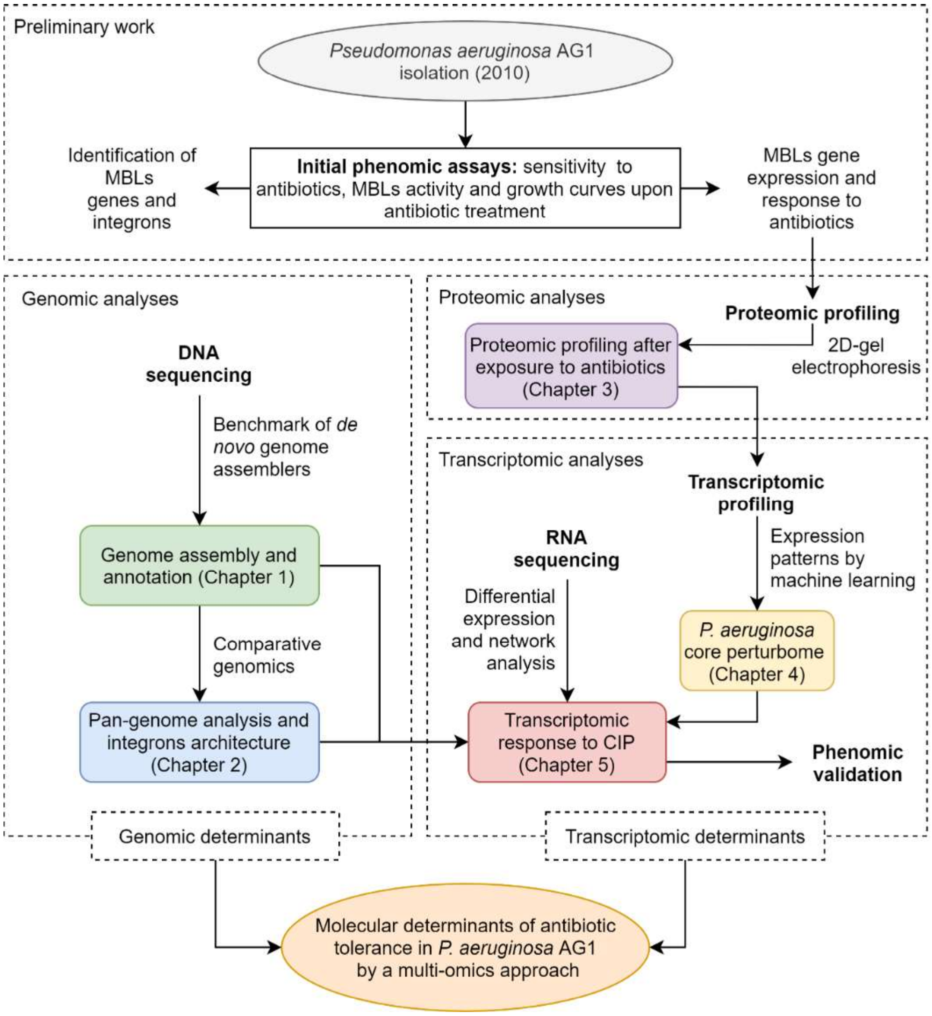

Preliminary work
Pseudomonas aeruginosa AG1 isolation (2010)
Identification of MBLs genes and integrons
Initial phenomic assays: sensitivity to antibiotics, MBLs activity and growth curves upon antibiotic treatment
MBLs gene expression and response to antibiotics
Genomic analyses
DNA sequencing
Benchmark of de novo genome assemblers
Genome assembly and annotation (Chapter 1)
Comparative genomics
Pan-genome analysis and integrons architecture (Chapter 2)
Proteomic analyses
Proteomic profiling
2D-gel electrophoresis
Proteomic profiling after exposure to antibiotics (Chapter 3)
Transcriptomic analyses
Transcriptomic profiling
Expression patterns by machine learning
P. aeruginosa core perturbome (Chapter 4)
RNA sequencing
Differential expression and network analysis
Transcriptomic response to CIP (Chapter 5)
Phenomic validation
Genomic determinants
Transcriptomic determinants
Molecular determinants of antibiotic tolerance in P. aeruginosa AG1 by a multi-omics approach

Figure 1. General workflow to study molecular determinants of antibiotic tolerance in the high-risk *P. aeruginosa* AG1 by a multi-omics approach. This study is based on five main steps: genome assembly and annotation, pan-genome analysis and integrons architecture, proteomic profiling after antibiotics exposure, identification of core perturbome, and the response to ciprofloxacin at transcriptomic and phenomic levels.

The first analysis of the genes in PaeAG1 by Sanger sequencing (primer walking method) confirmed that VIM-2 and IMP-18 genes are encoded in class 1 integrons (NCBI accessions KC907378 and KC907377) (Toval et al., 2015). In addition, at the phenomic level, preliminary comparison to the reference strain (*P. aeruginosa* PAO1) showed that PaeAG1 has particular features after exposure to different antibiotics, including pigment production, biofilm formation, phage plaque induction, and others (Chinchilla, 2018; Toval et al., 2015).

The multi-omics approach to study PaeAG1

In view of the genomic and phenomic features of PaeAG1, we were interested in studying PaeAG1 in-depth using a multi-omics approach. To address this, the strategy was developed in five main steps, each one concretized as a scientific paper and a chapter in this thesis (Figure 1).

First, genome sequencing was done using short and long-read technologies. Although a reference genome is available for the *P. aeruginosa* group (strain PAO1), a *de novo* strategy to assemble (or to build) the PaeAG1 genome was required since it was initially estimated that PaeAG1 has ~ 1.0 Mb additional of DNA sequence in its genome.

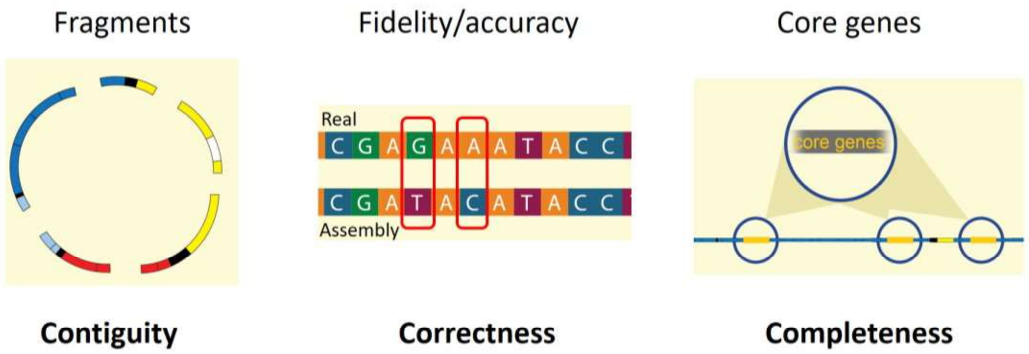


Figure 2. Definition of the 3C criterion: Contiguity, Correction and Completeness. Benchmarking of multiple assemblies can be done using metrics related to the number of pieces obtained vrs expected (contiguity), the fidelity of the assembly compared to the actual sequence (correction), and the ability to construct a minimum set of expected genes, vital to the species (completeness). More details in Chapter 1.

As detailed in Chapter 1, a benchmark of non-hybrid (using a single DNA sequencing technology) and hybrid (using both short and long-read data) assemblers was required to select the optimum model. To make this possible, the 3C criterion (i.e. contiguity, completeness, and correctness) was conceptualized as a set of metrics that can be used to benchmark genome assemblies and select the best approach (Figure 2).

The final assembly (GenBank CP045739), using a hybrid approach, revealed that PaeAG1 has not only the expected gene content for the *P. aeruginosa* group but also specific elements that are absent in the reference genome: 57 genomic islands (corresponding to ~ 1.0 Mb DNA sequence and >1000 genes) harboring the two complete class 1 integrons, six prophages, mobile genetic elements, and some virulence factors (Figure 3). Besides, PaeAG1 has 58 resistance genes, a not functional CRISPR-Cas system (which may explain the high content of genomic islands), and a molecular genotyping profile of a high-risk sequence type 111 (ST-111) strain.

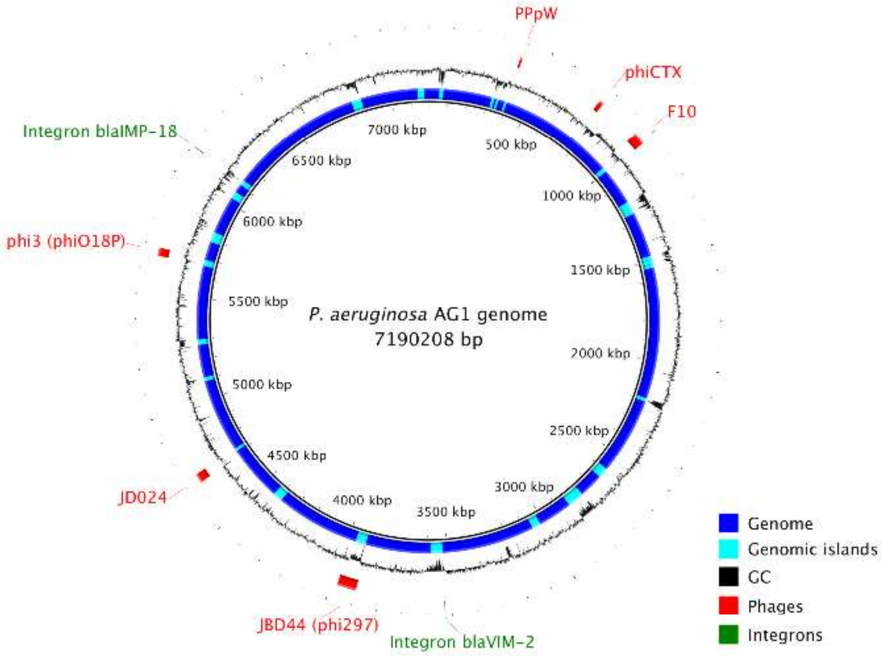


Figure 3. Assembly and annotation of *P. aeruginosa* AG1 genome. Circularized genome showing genomic islands harboring phages, integrons and other elements. Details in Chapter 1 and (J.-A. Molina-Mora, Campos-Sánchez, Rodríguez, Shi, & García, 2020).

These particular results are key components of the multi-omics approach with the subsequent analyses. If a mapping to the reference genome had been selected instead of a *de novo* assembly, the gene content of the extra 1.0 Mb DNA sequence could not have been revealed. In this regard, Chapter 2 focuses on the two PaeAG1 integrons and Chapter 5 reveals the role of phages in the response to ciprofloxacin. Importantly, these integrons and phages are absent in the reference genome.

In order to describe the landscape of the genomic regions associated with the two integrons of PaeAG1, a comparative genomic strategy was performed as a second main step (Chapter 2). It was first demonstrated that VIM-2 and IMP-18 are inducible genes under exposure to carbapenems using RT-qPCR. We then described the phylogenetic relationships among all the complete genomes of *P. aeruginosa* strains using a pan-genome analysis. This led to identify not only the core and the accessory genome for this group, but also other strains sharing the PaeAG1 genomic islands. Phylogenetically related strains were also classified as ST-111 clones, but a variant profile of the PaeAG1 genomic island content was found in other strains. ST-111 is a lineage that belongs to the high-risk group in *P. aeruginosa* (Oliver, Mulet, López-Causapé, & Juan, 2015), which is frequently associated with epidemics where multidrug resistance confounds treatment (Petitjean et al., 2017). Many *P. aeruginosa* high-risk clones carry genomic determinants of antibiotic resistance such as carbapenemases or extended-spectrum β-lactamases (Oliver et al., 2015).

Since PaeAG1 has special genomic features regarding antibiotic multi-resistance, with the carbapenemase activity by the VIM-2 and IMP-18 genes, the profile of genomic island content in phylogenetically related genomes was used to gain insights into the evolution and landscape of genomic regions around the MBL-carrying integrons of PaeAG1. Thus, specific genomic regions associated with the two integrons were reconstructed and characterized to compare the gene content and architecture in close genomes (Figure 4).

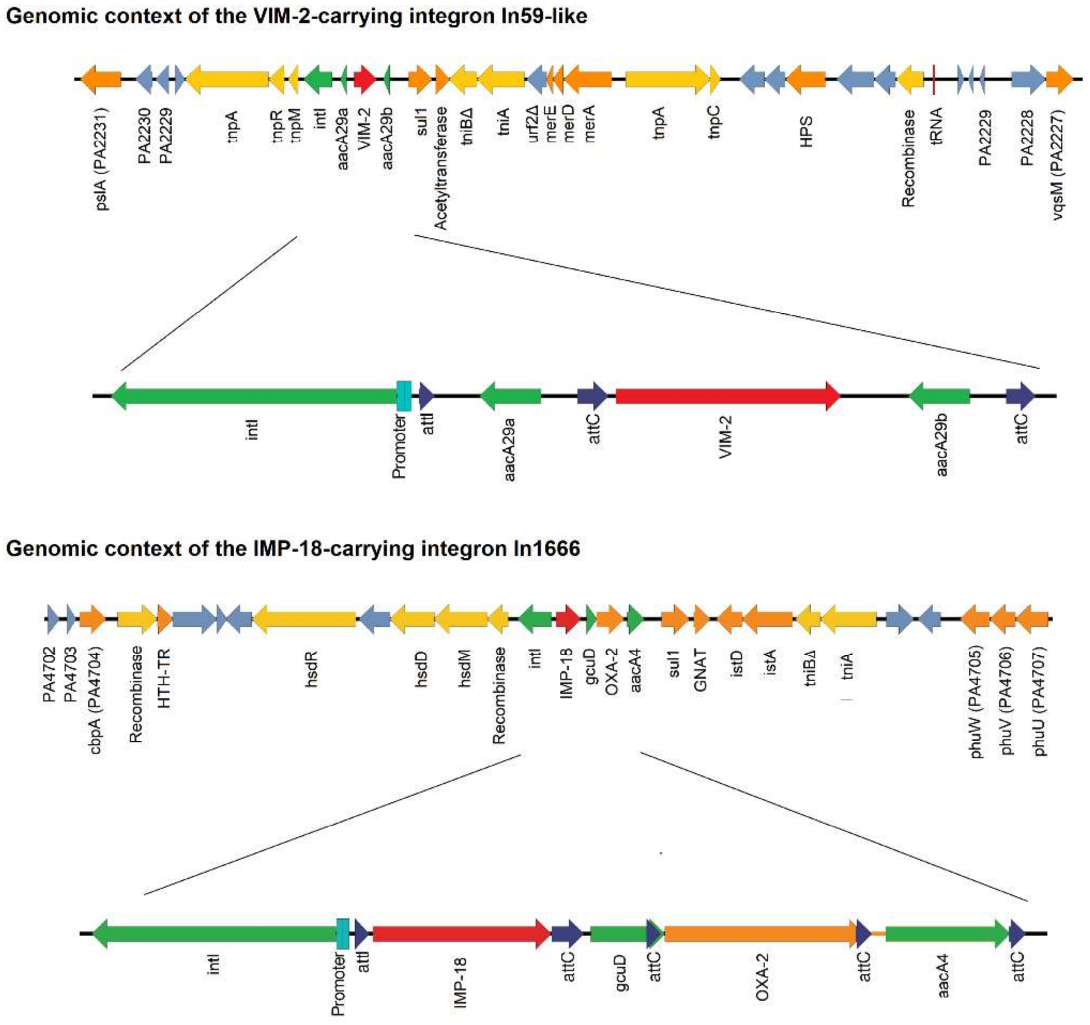


Figure 4. Architecture of the genomic regions containing the MBL-carrying integrons. The genomic region containing the old-acquaintance VIM-2-carrying integron is also present in other ST-111 strains. The architecture of the IMP-18-carrying integron and surrounding regions is shown with an arrangement that is reported here for the first time. Details in Chapter 2 (J.-A. Molina-Mora, Garcia-Batan, & Garcia, 2020).

The genomic region associated with the VIM-2-carrying integron (identified as an In59-like element, INTEGRALL-database http://integrall.bio.ua.pt/) was completely found in the other two ST-111 strains, being considered as an old-acquaintance integron. In the case of the IMP-18-carrying integron, the integron architecture and a surrounding genomic region have never been reported before. The IMP-18-carrying integron was considered as a new element and registered as a mobile element In1666.

Jointly, the chromosome assembly and the comparative genomics were able to define the molecular arsenal of PaeAG1 at the genomic level, including multiple genomic determinants of virulence, mobile elements, and antibiotic resistance genes. On the other hand, in the context of antibiotic resistance, different assays have been performed in PaeAG1 to study its tolerance to antibiotics. Antibiotic susceptibility testing was reported before (Chinchilla, 2018; Toval et al., 2015) and an MBLs differential expression has been tested not only to carbapenems as demonstrated in Chapter 2 but also to other antibiotics (Chinchilla, 2018).

At the proteomic level, the protein content in PaeAG1 under exposure to antibiotics was investigated. 2-dimensional gel electrophoresis (2D-GE) analysis was implemented using different imaging and machine learning algorithms, as presented in Chapter 3. The pipeline to analyze 2D-GE images has been also implemented to study two PaeAG1 subclones C25 and C50, as shown in the Supplementary Material 1 (a proceeding paper): "*Two-Dimensional Gel Electrophoresis Image Analysis of Two* Pseudomonas aeruginosa *Clones*" (José Arturo Molina-Mora, Chinchilla-Montero, Castro-Peña, & García, 2020).

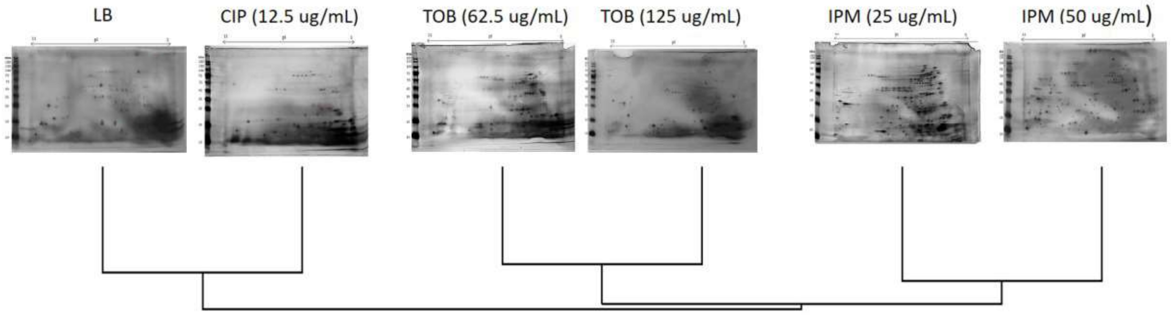


Figure 5. Clustering analysis of the proteomic profiling of PaeAG1 exposed to ciprofloxacin (CIP), imipenem (IPM) and tobramycin (TOB) antibiotics. Under CIP exposure, the proteomic profile after CIP exposure remains close to the control, unlike TOB and IPM. Details in Chapter 3 (Jose Arturo Molina-Mora, Chinchilla-Montero, Castro-Peña, & Garcia, 2020).

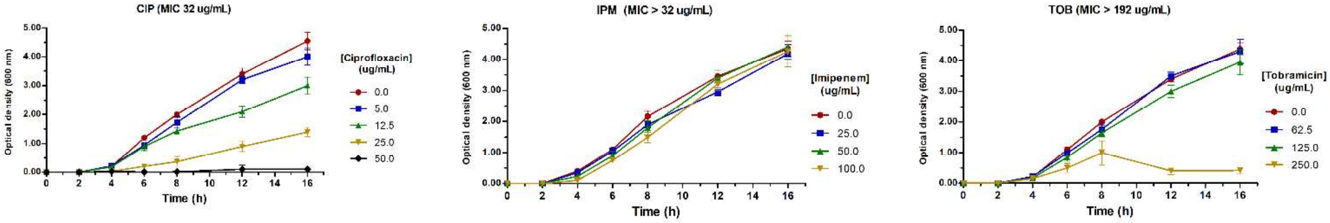


Figure 6. Assessment of PaeAG1 growth curves after treatment with ciprofloxacin (CIP), imipenem (IPM) and tobramycin (TOB) using different concentrations. Concentration-dependent effects were evidenced for CIP but not for the other antibiotics (Jose Arturo Molina-Mora et al., 2020).

For PaeAG1, results reveal that the global proteomic profile after exposure to a sub-inhibitory ciprofloxacin (CIP) concentration remains close to control (LB medium, without antibiotics), contrasting with the results obtained with tobramycin and imipenem, as shown in Figure 5. This means that the effects of ciprofloxacin at the proteomic level are fewer than the changes given by other antibiotics. This is an interesting finding when we compare growth curves. Growth curves showed a particular concentration-effect for PaeAG1 when exposed to sub-inhibitory CIP concentrations, but not to other tobramycin (TOB) or imipenem (IPM) antibiotics (Figure 6) at sub-inhibitory concentrations. Thus, to investigate the association between the PaeAG1 growth and sub-inhibitory CIP concentrations, two main transcriptomic analyses were performed: i) the identification of core perturbome in the *P. aeruginosa* group and ii) transcriptomic profiling of PaeAG1 after exposure to CIP.

As detailed in Chapter 4, the study of the molecular response to diverse perturbations (including CIP), term as perturbome, was carried out for *P. aeruginosa* with the reference strain. This makes it possible to generate the landscape of the central regulatory mechanisms of the stress response at the transcriptomic level in this bacterial group. Tolerance to stress conditions is vital for organismal survival, including bacteria under diverse environmental conditions (including antibiotics) (DeLong, 2012). Thus, to identify the core perturbome of *P. aeruginosa*, a machine learning approach was implemented to recognize gene expression patterns among public transcriptomic data sets, similar to other studies (Cornforth et al., 2018; Glaab, Bacardit, Garibaldi, & Krasnogor, 2012; Ma, Xin, Feldmann, & Wang, 2014; Zhao et al., 2016). In this regard, only a few studies have used machine learning methods on biological data to describe the effects of multiple perturbations in complex biological systems (Bermingham et al., 2015; Caldera et al., 2019) and so far none in *P. aeruginosa*.

In a subsequent analysis, the specific case of CIP exposure was used to standardize a systems biology pipeline to build large-scale molecular networks, as shown in the Supplementary Material 2 (a proceeding paper): "*Gene Expression Dynamics Induced by Ciprofloxacin and Loss of LexA Function in* Pseudomonas aeruginosa *PAO1 Using Data Mining and Network Analysis*" (J.A. Molina-Mora, Campos-Sanchez, & Garcia, 2018).

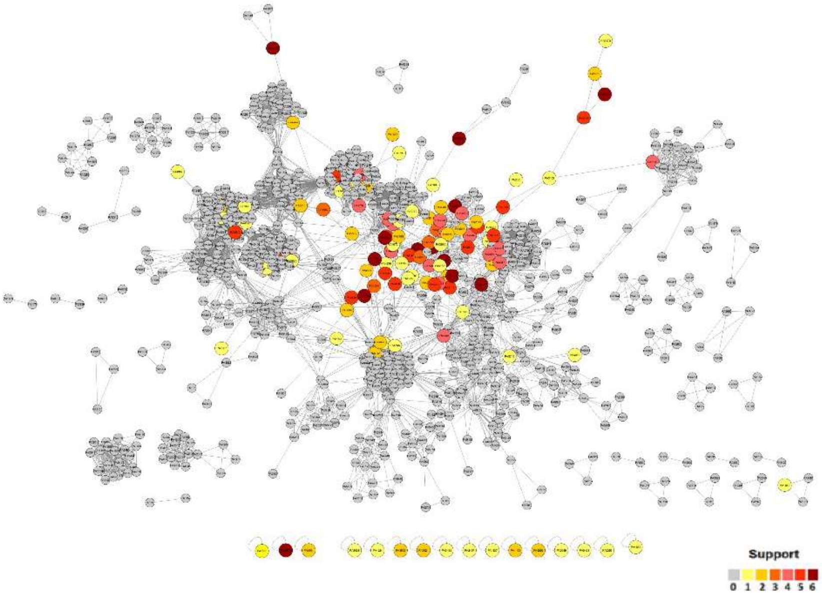


Figure 7. Distribution of core perturbome of *P. aeruginosa* on a basal network of functional associations. Pleiotropic effects are revealed for core perturbome genes. The support indicates the number of algorithms that identified a gene as a relevant element of the perturbome. Details in Chapter 4 (J. Molina-Mora et al., 2020).

The analysis of the central molecular response to perturbations, by both machine learning and large-scale networks, showed that the stress response is pleiotropic in *P. aeruginosa,* composed of at least 118 genes, of which 46 have strong support. Specific effects on gene networks were reflected as changes in gene expression profiles and the complexity of molecular regulation.

With the identification of the landscape of the core perturbome for *P. aeruginosa*, the study was resumed with the particular response to CIP in PaeAG1, as the final main step (Chapter 5). The knowledge of the core perturbome was necessary to differentiate the pathways and responses that are shared by other perturbations, but more importantly, to identify the exclusive responses to CIP in PaeAG1. As detailed before, growth reduction was evidenced for this strain as sub-inhibitory CIP concentrations were increased. Thus, we identified the transcriptomic determinants associated with the response to CIP in PaeAG1. To address this, we used transcriptomic profiling by RNA sequencing and network analysis by applying a top-down systems biology approach.

In order to study in detail the performance of different approaches for transcriptomic data analyses four different pipelines were assessed. Benchmarking of all pipelines was done using bioinformatics and biological criteria according to the genome analysis, phenotypes, and expert knowledge (Figure 8). The pipeline using EDGE-pro was selected as the best one using different criteria according to body coverage and mapping. See Chapter 5 for details. With these pipelines, transcriptomic determinants were identified.

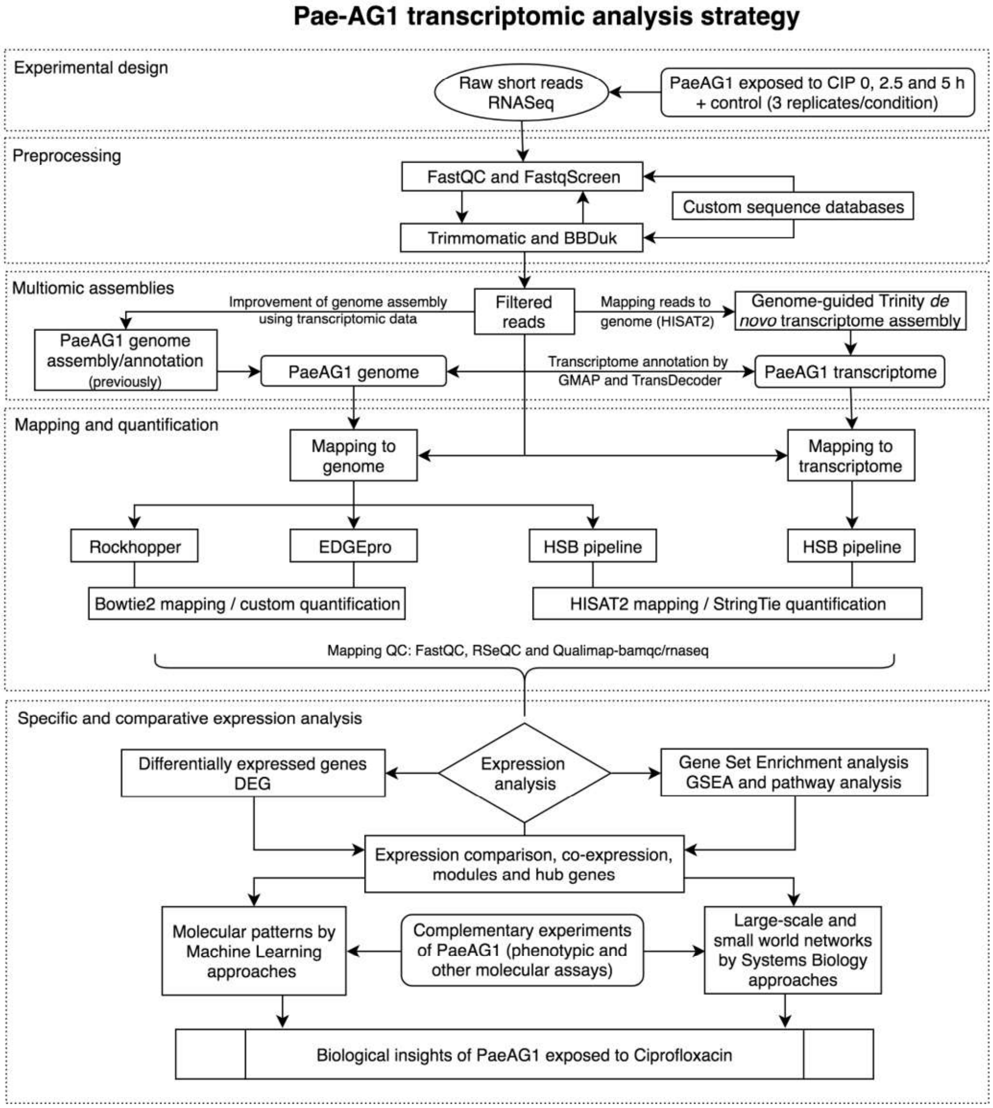

Pae-AG1 transcriptomic analysis strategy
Experimental design
Raw short reads RNASeq
PaeAG1 exposed to CIP 0, 2.5 and 5 h + control (3 replicates/condition)
Preprocessing
FastQC and FastqScreen
Custom sequence databases
Trimmomatic and BBDuk
Multiomic assemblies
Improvement of genome assembly using transcriptomic data
Filtered reads
Mapping reads to genome (HISAT2)
Genome-guided Trinity de novo transcriptome assembly
PaeAG1 genome assembly/annotation (previously)
PaeAG1 genome
Transcriptome annotation by GMAP and TransDecoder
PaeAG1 transcriptome
Mapping and quantification
Mapping to genome
Mapping to transcriptome
Rockhopper
EDGEpro
HSB pipeline
HSB pipeline
Bowtie2 mapping / custom quantification
HISAT2 mapping / StringTie quantification
Mapping QC: FastQC, RSeQC and Qualimap-bamqc/rnaseq
Specific and comparative expression analysis
Differentially expressed genes DEG
Expression analysis
Gene Set Enrichment analysis GSEA and pathway analysis
Expression comparison, co-expression, modules and hub genes
Molecular patterns by Machine Learning approaches
Complementary experiments of PaeAG1 (phenotypic and other molecular assays)
Large-scale and small world networks by Systems Biology approaches
Biological insights of PaeAG1 exposed to Ciprofloxacin

Figure 8. Benchmark of four pipelines for RNA-Seq data analysis to study PaeAG1 after CIP exposure. Pipelines using mapping to the genome or transcriptome with different quantification steps were implemented to identify differentially expressed genes in PaeAG1 after exposure to CIP.

Transcriptomic determinants included classical elements of the core perturbome for *P. aeruginosa* with down-regulation of pathways related to energy metabolism, ribosomal activity, and DNA metabolism, most of them related to bacterial growth reduction. Also, an exclusive feature, the phage induction, was suggested due to the up-regulation of phage genes creating two well-defined clusters at a network level (Figure 9).

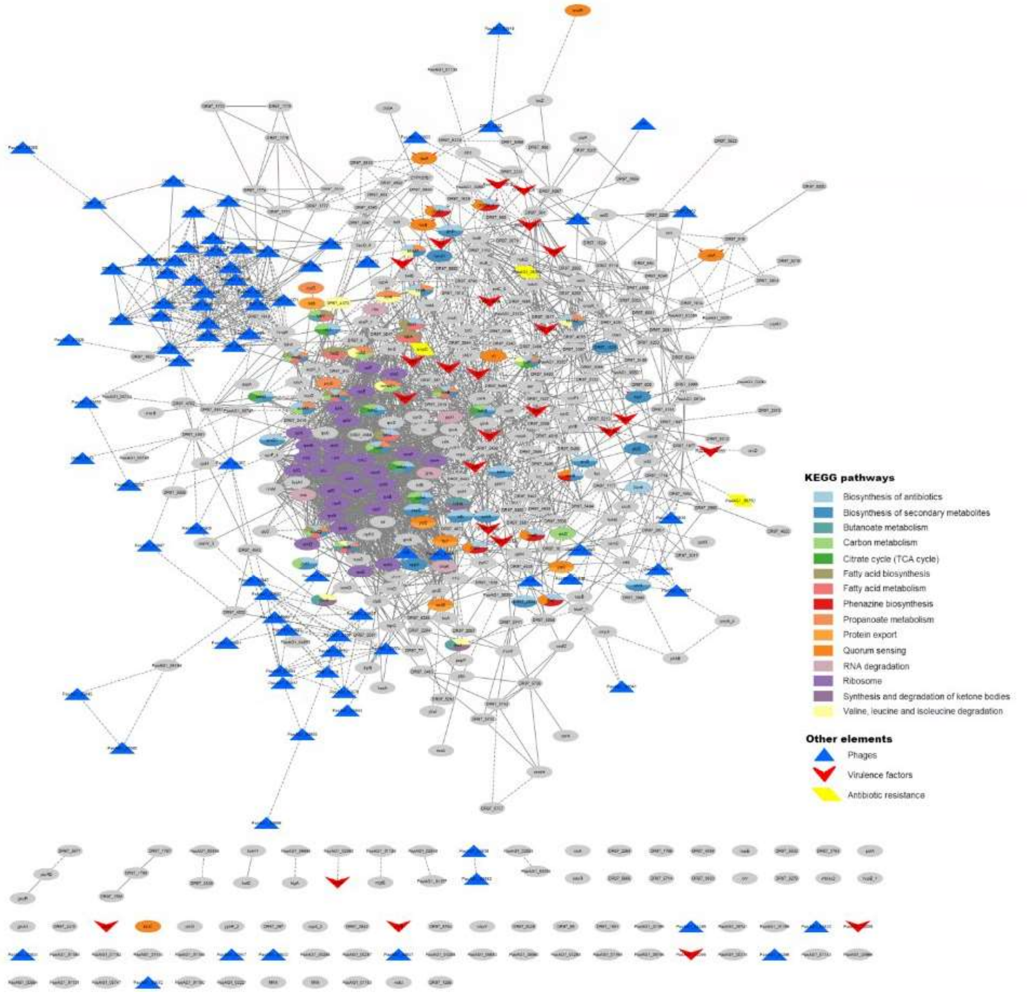


Figure 9. Large-scale network of differentially expressed genes of PaeAG1 after CIP exposure. Multiple elements of virulence, phage, and pathways were found to be modulated by the antibiotic, revealing pleiotropic effects at the transcriptomic level. Details in Chapter 5 (Jose Arturo Molina-Mora et al., 2020).

To validate CIP effects on phage induction, we applied a phage plaque assay (at a phenomic level) that showed an exponential induction as CIP was increased. Since these phages are absent in the reference genome, again, the *de novo* genome assembly was a critical step to obtain biological insights for PaeAG1. Although PaeAG1 is resistant to CIP, a sub-inhibitory concentration of this antibiotic can induce a pleiotropic effect at a transcriptomic level, including pathways of the core perturbome and phage induction. In the last case, with the subsequent bacterial cell lysis, the reduction on the growth curve is explained by CIP in a concentration-dependent manner. This phenomenon is particular to CIP and not found for imipenem or tobramycin, as it was shown in this study.

Phage induction by CIP can be used as a complementary strategy to fight Pseudomonal infections. The fact that PaeAG1 phages are resident elements of the genome and not exogenous elements as in other studies (Fothergill et al., 2011; Kamal & Dennis, 2015), represents an advantage to eventual further implementations. In the context of another study in our group, these results of phage induction were tested in an *in vivo* murine model (Morales-Berrocal, 2016). Very promissory results have been obtained under CIP injection after *P. aeruginosa* infection, in which mortality of infected mice was reduced from 70% to 30% and bacteria quantification dropped-off in organs, but a significant increment in phage counts was evidenced (Figure 10). Specific details will be eventually presented as part of another work. Future studies will also evaluate the modulation of the CIP response using genetic engineering (knock-out, knock-down, and the like), other –omics approaches (proteomics, ChIP-Seq, etc), and other *in vivo* models.

In summary, by using a multi-omics approach, it was able to study molecular determinants of antibiotic tolerance in PaeAG1. Genome assembly using a benchmark strategy led to building a high-quality sequence. A *de novo* approach allowed assembling around 1.0 Mb of sequence that is absent in the reference genome. These exclusive regions are composed of 57 genomic islands harboring two MBL-carrying integrons, phages, and many other genes. Comparison to all available complete sequences showed that the genome could be grouped by MLST profile, including a clear ST-111 cluster containing PaeAG1. In addition, a landscape of genomic regions surrounding integrons was described in which an IMP-18-carrying integron was characterized for the first time. Multi-resistance profile, antibiotic resistance genes, the MLST profile, clusters of the pan-genome analysis, and the architecture of integrons define the genomic determinants of PaeAG1.

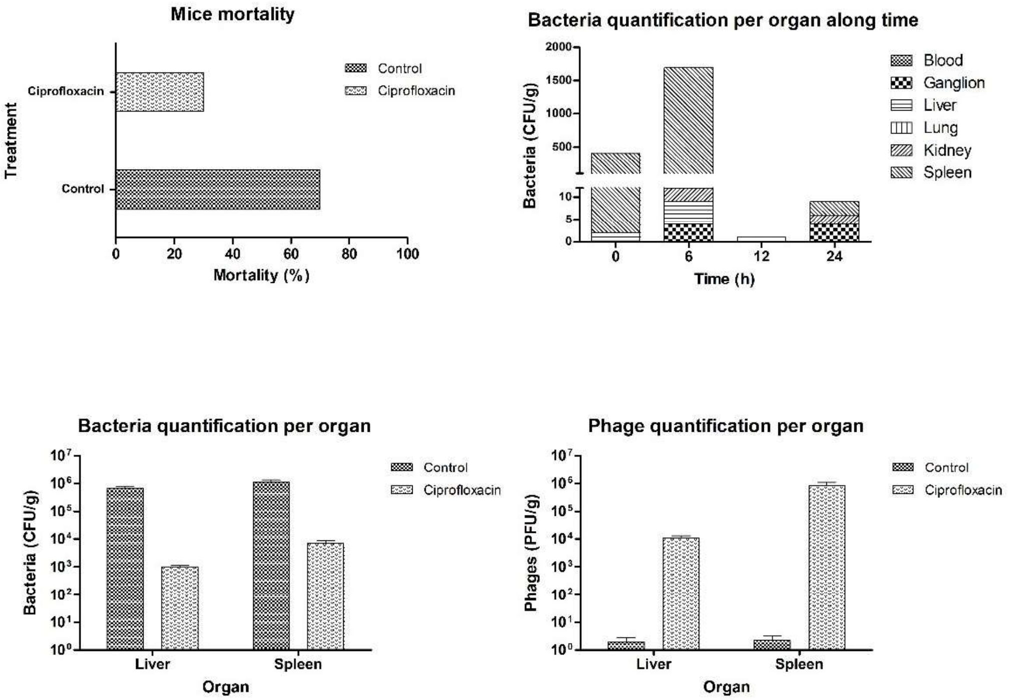


Figure 10. Preliminary results of the in vivo murine model to evaluate phage induction by CIP as a strategy to fight Pseudomonal infections. Upon CIP treatment, the mortality of infected mice was reduced, including reduction of bacteria quantification and increased phage counts in organs.

In order to study the central response to perturbations in the *P. aeruginosa* group, the core perturbome, and to identify gene expression patterns, we used a machine learning approach. Pathways of energy metabolism, ribosomal activity, DNA metabolism, and others were enriched. Similar findings of enriched pathways were obtained for the specific case of PaeAG1 exposed to CIP, but particular genes (absent in the reference strain, such as phage genes) were also identified. Phage induction upon CIP treatment, suggested by phage genes up-regulation, was validated at a phenomic level. Particular key genes, gene clusters, and pathways were recognized as transcriptomic determinants of antibiotic tolerance in PaeAG1.

Together, these genomic and transcriptomic elements are molecular determinants of antibiotic tolerance and resistance in PaeAG1, which in part define the high-risk condition of this strain that enables it to conquer nosocomial environments with a multi-resistance profile.

JUSTIFICATION

This research proposal aims to fill an information gap regarding the molecular determinants associated with tolerance and resistance to antibiotics at the genomic and transcriptomic levels in *P. aeruginosa* AG1. The initial studies determined that this bacterium has high-risk clone characteristics given its success in conquering nosocomial environments and its multi-resistance profile (including resistance to carbapenems). The latter case allows classifying PaeAG1 as a critical and priority 1 organism according to the WHO.

Furthermore, because it was initially estimated that this bacterium contained an additional 1.0 Mb of DNA sequence relative to the reference genome *P. aeruginosa* PAO1, a multi-omics strategy was established to avoid losing genomic information. This is expected to be a crucial point due to the particular PaeAG1 features that differentiate it from the reference strain. To this end, a *de novo* genome assembly and subsequent comparative genomic analyses can identify the genomic determinants associated with tolerance to antibiotics. After proteomic profiling using 2D-GE and comparison of response to antibiotics, the definition of the central response to disturbances or core perturbome in the *P. aeruginosa* group at the transcriptomic level allows identifying the metabolic pathways associated with the stress response. On account of the complexity and amount of data associated with this task, a machine learning strategy was required. For the specific case of PaeAG1 with exposure to CIP, differential expression analyses were performed with RNA sequencing, large-scale molecular network analysis, and experimental validation at the phenomic level. The particular genes, gene clusters, and metabolic pathways of the core perturbome in *P. aeruginosa* and the response to ciprofloxacin in PaeAG1 constitute the transcriptomic determinants of antibiotic tolerance in this strain.

Taken together, these strategies of using a multi-omics approach (at the genomics, transcriptomics, perturbomics, proteomics, and phenomics levels), sequence bioinformatics analyses, machine learning, and systems biology, provided the required approach to identify and characterize the molecular determinants associated with tolerance to antibiotics in PaeAG1.

RESEARCH QUESTION

Which are the general genomic determinants and transcriptomic determinants associated with ciprofloxacin exposure in *P. aeruginosa* AG1 that mediate tolerance to antibiotics?

HYPOTHESIS

Molecular determinants that define antibiotic tolerance in *P. aeruginosa* AG1 can be identified and characterized at the genomic and transcriptomic levels using a multi-omic approach.

RESEARCH OBJECTIVES

General objective

To identify and characterize the genomic and transcriptomic determinants associated with tolerance to antibiotics in *Pseudomonas aeruginosa* AG1 using a multi-omics approach.

Specific objectives

1. To assemble and annotate the *P. aeruginosa* AG1 genome using a benchmarking strategy, in order to characterize the gene content and genomic determinants associated with its multidrug-resistance and other phenotypes.

2. To compare *P. aeruginosa* AG1 genome against other *P. aeruginosa* sequences using comparative genomics to describe pan-genome, phylogenetic relationships, genomic islands content, and architecture of genomic regions associated with the VIM-2- and IMP-18-carrying integrons.

3. To identify genes associated with multiple perturbations in *P. aeruginosa* to describe transcriptomic determinants of the central molecular response (perturbome) using a machine learning approach.

4. To identify transcriptomic determinants using RNA-Seq profiling and network analysis by a top-down systems biology approach to characterize the response to ciprofloxacin in *P. aeruginosa* AG1.

CHAPTER 1

High quality 3C *de novo* assembly and annotation of a multidrug resistant ST-111 *Pseudomonas aeruginosa* genome: Benchmark of hybrid and non-hybrid assemblers

Molina-Mora, J.-A., Campos-Sánchez, R., Rodríguez, C., Shi, L., & García, F. (2020). High quality 3C de novo assembly and annotation of a multidrug resistant ST-111 *Pseudomonas aeruginosa* genome: Benchmark of hybrid and non-hybrid assemblers. Scientific Reports, 10(1), 1392. https://doi.org/10.1038/s41598-020-58319-6

https://www.nature.com/articles/s41598-020-58319-6

Summary

Genotyping methods and genome sequencing are indispensable to reveal genomic structure of bacterial species displaying high level of genome plasticity. However, reconstruction of genome or assembly is not straightforward due to data complexity, including repeats, mobile and accessory genetic elements of bacterial genomes. Moreover, since the solution to this problem is strongly influenced by sequencing technology, bioinformatics pipelines, and selection criteria to assess assemblers, there is no systematic way to select *a priori* the optimal assembler and parameter settings. To assembly the genome of *P. aeruginosa* strain AG1, short reads (Illumina) and long reads (Oxford Nanopore) sequencing data were used in 13 different non-hybrid and hybrid approaches. PaeAG1 is a multiresistant high-risk sequence type 111 (ST-111) clone that was isolated from a Costa Rican hospital and it was the first report of an isolate of *P. aeruginosa* carrying both VIM-2 and IMP-18 genes encoding for metallo-β-lactamases (MBLs) enzymes. To assess the assemblies, multiple metrics regard to contiguity, correctness and completeness (3C criterion, as we define here) were used for benchmarking the 13 approaches and select a definitive assembly. In addition, annotation was done to identify genes (coding and RNA regions) and to describe the genomic content of PaeAG1.

Whereas long reads and hybrid approaches showed better performances in terms of contiguity, higher correctness and completeness metrics were obtained for short read only and hybrid approaches. A manually curated and polished hybrid assembly gave rise to a single circular sequence with 100% of core genes and known regions identified, >98% of reads mapped back, no gaps, and uniform coverage. The strategy followed to obtain this high-quality 3C assembly is detailed in the manuscript and we provide readers with an all-in-one script to replicate our results or to apply it to other troublesome cases.

The final 3C assembly revealed that the PaeAG1 genome has 7,190,208 bp, a 65.7% GC content and 6,709 genes (6,620 coding sequences), many of which are included in multiple mobile genomic elements, such as 57 genomic islands, six prophages, and two complete integrons with VIM-2 and IMP-18 MBL genes. Up to 250 and 60 of the predicted genes are anticipated to play a role in virulence (adherence, quorum sensing and secretion) or antibiotic resistance (β-lactamases, efflux pumps, etc). Altogether, the assembly and annotation of the PaeAG1 genome provide new perspectives to continue studying the genomic diversity and gene content of this important human pathogen.

SCIENTIFIC REPORTS
natureresearch

OPEN

High quality 3C *de novo* assembly and annotation of a multidrug resistant ST-111 *Pseudomonas aeruginosa* genome: Benchmark of hybrid and non-hybrid assemblers

José Arturo Molina-Mora[1]*, Rebeca Campos-Sánchez[2], César Rodríguez[1], Leming Shi[3] & Fernando García[1]

Genotyping methods and genome sequencing are indispensable to reveal genomic structure of bacterial species displaying high level of genome plasticity. However, reconstruction of genome or assembly is not straightforward due to data complexity, including repeats, mobile and accessory genetic elements of bacterial genomes. Moreover, since the solution to this problem is strongly influenced by sequencing technology, bioinformatics pipelines, and selection criteria to assess assemblers, there is no systematic way to select *a priori* the optimal assembler and parameter settings. To assembly the genome of *Pseudomonas aeruginosa* strain AG1 (PaeAG1), short reads (Illumina) and long reads (Oxford Nanopore) sequencing data were used in 13 different non-hybrid and hybrid approaches. PaeAG1 is a multiresistant high-risk sequence type 111 (ST-111) clone that was isolated from a Costa Rican hospital and it was the first report of an isolate of *P. aeruginosa* carrying both blaVIM-2 and blaIMP-18 genes encoding for metallo-β-lactamases (MBL) enzymes. To assess the assemblies, multiple metrics regard to contiguity, correctness and completeness (3C criterion, as we define here) were used for benchmarking the 13 approaches and select a definitive assembly. In addition, annotation was done to identify genes (coding and RNA regions) and to describe the genomic content of PaeAG1. Whereas long reads and hybrid approaches showed better performances in terms of contiguity, higher correctness and completeness metrics were obtained for short read only and hybrid approaches. A manually curated and polished hybrid assembly gave rise to a single circular sequence with 100% of core genes and known regions identified, >98% of reads mapped back, no gaps, and uniform coverage. The strategy followed to obtain this high-quality 3C assembly is detailed in the manuscript and we provide readers with an all-in-one script to replicate our results or to apply it to other troublesome cases. The final 3C assembly revealed that the PaeAG1 genome has 7,190,208 bp, a 65.7% GC content and 6,709 genes (6,620 coding sequences), many of which are included in multiple mobile genomic elements, such as 57 genomic islands, six prophages, and two complete integrons with blaVIM-2 and blaIMP-18 MBL genes. Up to 250 and 60 of the predicted genes are anticipated to play a role in virulence (adherence, quorum sensing and secretion) or antibiotic resistance (β-lactamases, efflux pumps, etc). Altogether, the assembly and annotation of the PaeAG1 genome provide new perspectives to continue studying the genomic diversity and gene content of this important human pathogen.

Genotyping methods and genome sequencing are indispensable to reveal genomic structure and evolution of bacterial clones with high resolution[1]. In this sense, production of large amounts of short sequencing data from genomes (reads) has been facilitated by continuous advances in Next Generation Sequencing (NGS) technologies.

[1]Centro de Investigación en Enfermedades Tropicales, Facultad de Microbiología, Universidad de Costa Rica, San José, Costa Rica. [2]Centro de Investigación en Biología Celular y Molecular, Facultad de Microbiología, Universidad de Costa Rica, San José, Costa Rica. [3]Human Phenome Institute of Fudan University, Shanghai, China. *email: jose.molinamora@ucr.ac.cr

This includes short read sequencing technologies (a few hundred bp read length) such as Illumina and long read sequencing technologies (several hundred kb read length) such as Pacific Biosciences (PacBio) single-molecule real-time (SMRT) and Oxford Nanopore Technology (ONT)[2].

Using sequencing data, it is expectable that full-length chromosomes could be produced when the genome is fully sequenced and assembled[3]. However, reconstruction of genome or assembly is not straightforward due data complexity. This is a challenging problem that requires time and expertise[4]. If a reference genome is available, an assembly can be made by comparison or direct mapping, otherwise, a *de novo* assembly, in which only the information obtained from reads is used to reconstruct the genome, without prior knowledge of its organization[5]. In *de novo* assembly, sequences (reads) are grouped into contigs using graph based algorithms such as Overlap-Layout-Consensus, *de Bruijn* and greedy approaches[5,6]. Then contigs are assembled into scaffolds (supercontigs or metacontigs). Alternatively, some *de novo* assemblers use reference genomes to solve specific inconsistencies or for scaffolding[5].

Reconstruction can be favored by some previous information, such as expected genome size, GC content and repetitive region content, as they help choose the best strategy to follow. Even though many algorithms to assemble genome by *de novo* approaches are available, performance is completely dependent on data (short or long reads, instruments, technology), genomic complexity (repeats, number of chromosomes or plasmids) and complementary algorithms (pre-processing, databases, annotations, etc)[7]. Therefore, for a specific genome and dataset, selection of the optimal assembly strategy to use is not a trivial task because there is no systematic way to determine which assembler and what parameter settings must be selected[8].

Since a key first requirement in the study of genomes is accuracy[9], short reads technologies are preferred because they produce high fidelity reads[10]. Also, the low cost and high accuracy of Illumina sequencing makes it well suited to high-throughput bacterial genomics[10]. However, genomes present complex repeat structures difficult to solve by different assemblers. As reported, if the repeats are longer than the reads, genomic regions sharing perfect repeats can be indistinguishable[6]. With this, resolving a full genome is a challenging issue for short reads approaches. Consequently, most available bacterial genomes are incomplete[11], highly fragmented, and of low quality[3].

Long reads, by contrast, can exceed the length of repeats in a typical bacterial genome, facilitating genome assembly[10]. Long reads technology offers an important advantage for complex genomes with high level of repetitive elements or genome duplication[7]. Thus, use of long reads data has shown improvements in the context of *de novo* genome assemblies, rising contiguity, solving fragmented regions, and closing gaps[12]. However, these third generation sequencing methods deal with relatively high sequencing error[8], which has been estimated up to 15% of random but also systematic errors[10,12]. In addition, long reads sequencing has a higher cost per base than that with Illumina platforms[11].

Combination of reads of different length and from different sequencing platforms in so-called hybrid approaches often counterbalances the drawbacks of each method[4]. The growing interest in hybrid assemblies is justified by the popularity, cost and accuracy of short reads sequencing, plus the resolution capacity of repetitive regions and genomic structures of long reads[10]. In some cases, a hybrid approach is sufficient to produce a single and closed sequence of the microbial genome[13]. However, to accurately assemble a genome, neither the optimum combination and coverage of long and short reads, nor the minimum required length of long-reads are known a priori[9]. Due to this, hybrid and non-hybrid assembly must be individually evaluated with regard to select the best assembly conditions, and different metrics and tools are available for this purpose. However, no single or completely useful strategy is considered as universal and sufficient to benchmark assemblies[3,14].

Benchmark of assemblies can be achieved using metrics related to contigs and scaffolds (contiguity), ability to complete the whole structure of the genome (completeness), and the accuracy of the assembly (correctness). Although most of studies of assemblies exploit these parameters to evaluate the performance of assemblers[3,8,10,15–17], here we define the general assessment by "3C criterion" as all metrics required to evaluate and benchmark genome assemblies using contiguity, completeness and correctness metrics, as detailed:

- Contiguity: It evaluates the assembly in terms of number and size of contigs and scaffolds[6], the pieces found in an assembly. Metrics includes statistics related to maximum length, average length, combined total length, and contig N50 (length-weighted median of ordered contigs or scaffolds)[2]. However, contiguity metrics thereof need to be interpreted with caution due they do not contain information on assembly accuracy and completeness[4].
- Correctness: it refers to how well those pieces accurately represent the genome sequenced[16] and, in general is acceptable that it is essential to prioritize correctness rather than contiguity[12]. However, correctness is difficult to evaluate if a preliminary reference genome is not available, which is a particular problem for *de novo* assembly[6]. Mapping and comparison to reference or draft genome (or a consensus sequence) can be used to detect misassemblies, including mismatches, indels, and misjoins[8].
- Completeness: it assesses how much of the genome is represented by the pieces of the assembly[16]. This implies the evaluation of ability to assembly not only all the genes, but also to solve all complicated regions, including repetitive sequences and, if it is expected, circularization of genome. The most important metric for this case is the "completeness score", calculated by the examination of single-copy orthologs conserved genes[18]. In addition, information of known sequences, unexpected variations in coverage, and remapping of reads allows to analyze the consistency of the genome and identification of potentially poorly assembled regions[5,19].

Thus, to develop a strategy to assembly a bacterial genome using the non-hybrid and hybrid approaches as well as the 3C criterion, we used a ST-111 strain of *Pseudomonas aeruginosa*. *P. aeruginosa* is Gram-negative bacterium and a well-known opportunistic pathogen[20]. It is responsible for acute and chronic nosocomial and community infections in immune-compromised patients[21]. However, the treatment of *P. aeruginosa* infections is challenging due many intrinsic and acquired mechanisms of resistance[22], including the production of to β-lactamases antibiotic modifying enzymes and target alteration.

Multi-resistance in *P. aeruginosa* is becoming more and more serious, not only due resistance to classical β-lactams, aminoglycosides and fluoroquinolones, but also to resistance last resort treatments including carbapenems (β-lactams) and colistin, which causes great difficulties in clinical treatment[23,24] and resistant to these antibiotics emerge as a final level of fight of bacteria which compromises infections treatments[24]. Many bacterial clones with carbapenemase-producing features are recognized as high-risk clones[25]. A high-risk clone is a multidrug-resistant clone with highly efficient transmission and/or maintenance among humans or animals[26], playing a major role in the spread of resistance in the hospital and other environments[27] and a flexible ability to accumulate and switch resistance[28]. However, the term high-risk is not necessarily associated with severity[26]. A limited number of *Pseudomonas aeruginosa* genotypes (mainly ST-111, ST-175, and ST-235) are recognized as high-risk clones, and they are responsible for epidemics of nosocomial infections by multidrug-resistant or extensively drug-resistant strains worldwide[29].

In Costa Rica, isolation of carbapenem resistant *P. aeruginosa* strains is relatively common in some major hospitals as we reported before[22], most of them carrying one blaVIM and one blaIMP allele carbapenemases and up to 63.1% of prevalence[22], much higher than the frequencies observed in other countries[30].

The Costa Rican multi-resistant strain *P. aeruginosa* AG1 (PaeAG1) was isolated from a sputum sample of a patient with pneumonia from the Intensive Care Unit of the San Juan de Dios Hospital (San José, Costa Rica) in 2010. PaeAG1 has a resistance phenotype to β-lactam (including carbapenems), aminoglycosides and fluoroquinolones, showing susceptibility only to colistin. In addition, PaeAG1 was identified as the first report worldwide of a strain carrying both blaIMP-18 (or IMP-18) and blaVIM-2 (VIM-2) genes, coding for metallo-β-lactamases (MBL) with carbapenemase activity[22].

PaeAG1 is a high-risk clone with a genotyping profile ST-111, which includes strains with a phenotype extremely resistant to antibiotics, responsible for various types of infections in hospitals and rapid spread between the individuals[29,31]. Sanger sequencing confirmed that the blaVIM-2 and blaIMP-18 genes of strain AG1 (Accessions KC907377 and KC907378) are encoded in class 1 integrons, likely in two different structures[22]. In addition, preliminary experimental assays suggested no existence of plasmids[22].

We were interested in assembling and annotating the genome of the clinical isolate PaeAG1 due to its importance as a high-risk clone with multi-resistance to antibiotics and to identify molecular determinants related to the ability to conquer nosocomial environments, virulence and other phenotypes. Thus, the aims of our study were: (i) to assemble the PaeAG1 genome using short and long reads data by hybrid and non-hybrid multiple approaches, (ii) to benchmark assemblers and select the best genome assembly approach using the 3C criterion, and (iii) to annotate the PaeAG1 genome to characterize and identify general gene content and genomic determinants associated with its multidrug-resistance and virulence phenotypes.

Methods

The general pipeline followed to assembly the PaeAG1 genome by hybrid and non-hybrid approaches is shown in Fig. 1. Complete details of settings of implemented algorithms are shown in supplementary material "Scripts for bioinformatics analysis".

Bacterial isolate.

The Costa Rican PaeAG1 strain was isolated in 2010 from a sputum sample of a patient with pneumonia from the Intensive Care Unit of the San Juan de Dios Hospital (San José, Costa Rica). This isolate has phenotypic resistance (AST-GN cards, bioMeriux Vitek) to β-lactams, aminoglycosides and fluoroquinolones, shows susceptibility only to colistin and expresses metallo-β-lactamase activity (E-test MBL strips, AB Biodisk), as reported[22].

Bacterial growth and DNA isolation.

PaeAG1 cells were grown overnight in Luria-Bertani broth (LB) medium at 37 °C with shaking. Then, cells were collected by centrifugation and genomic DNA was isolated with the QIAGEN DNeasy Kit (QIAGEN, UK) following the manufacturer's instructions.

The yield of genomic DNA obtained was determined using a Nanodrop (Nanodrop 2000, Thermo Scientific, UK) and by Qubit Fluorometric Quantitation (Qubit 3.0 Fluorometer, Thermo Scientific). DNA integrity was verified by electrophoresis using 0.7% agarose gels.

Whole genome sequencing using short reads.

Genomic DNA was sequenced using Illumina technology (Illumina Inc.) at Macrogen, Korea. The sequencing library was prepared using TruSeq DNA Sample Prep kit with the standard Illumina DNA shotgun library preparation protocol. DNA fragmentation was achieved by ultrasonication, and then adapter ligation and PCR enrichment were done. Paired end reads of 101 bp were generated using a HiSeq. 2000 sequencing instrument. Sequence files were evaluated using FastQC v0.11.7[32] before and after trimming. Reads were trimmed (including adapters removal) using Trimmomatic v0.38[33] to discard sequences with per base sequence quality score <30. After selection, 7.4 Gb of sequences were kept, with a 14 million of pairs of reads and mean coverage >400X according to expected genome size (approx. 7 Mb).

Whole genome sequencing using long reads.

Long reads from genomic DNA was sequenced using Oxford Nanopore technology by NextOmics, Wuhan-China. Sequencing libraries were prepared according to the ONT 1D ligation library protocolSQK-LSK109. FLO-MIN-106 flowcell and the standard 48-hour run script with active channel selection enabled were used to sequence reads in a GridION instrument. Poretools v0.6.0[34] was used to extract and evaluate reads by quality before and after trimming. Adapters were removed using Porechop v0.2.3 (github.com/rrwick/Porechop) and trimming was done using Filtlong v0.2.0 (github.com/rrwick/Filtlong). Reads with mean quality weight <10 and/or shorter than 1 kb were discarded. The final dataset consisted of

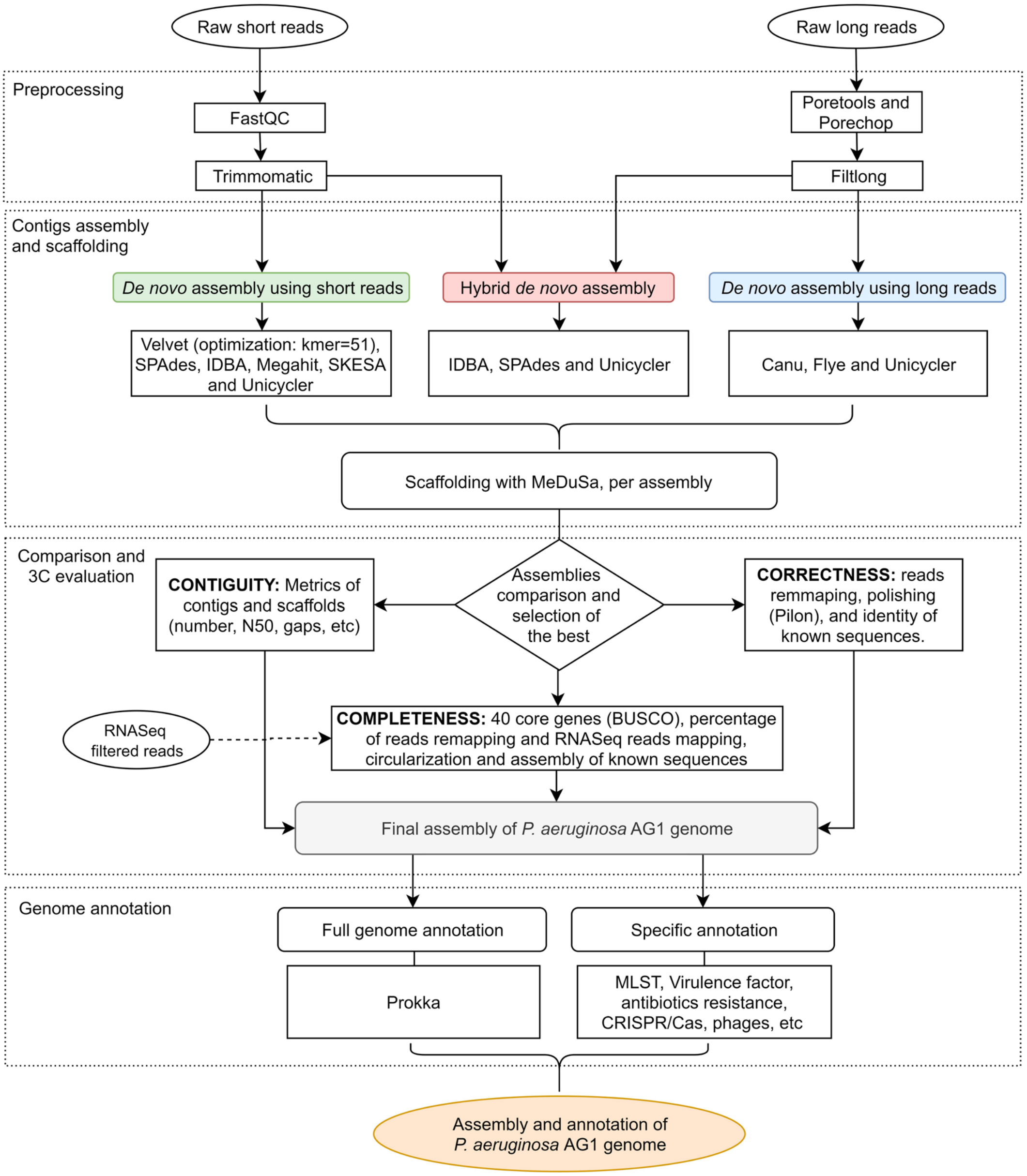

Raw short reads
Raw long reads
Preprocessing
FastQC
Poretools and Porechop
Trimmomatic
Filtlong
Contigs assembly and scaffolding
De novo assembly using short reads
Hybrid *de novo* assembly
De novo assembly using long reads
Velvet (optimization: kmer=51), SPAdes, IDBA, Megahit, SKESA and Unicycler
IDBA, SPAdes and Unicycler
Canu, Flye and Unicycler
Scaffolding with MeDuSa, per assembly
Comparison and 3C evaluation
CONTIGUITY: Metrics of contigs and scaffolds (number, N50, gaps, etc)
Assemblies comparison and selection of the best
CORRECTNESS: reads remmaping, polishing (Pilon), and identity of known sequences.
RNASeq filtered reads
COMPLETENESS: 40 core genes (BUSCO), percentage of reads remapping and RNASeq reads mapping, circularization and assembly of known sequences
Final assembly of *P. aeruginosa* AG1 genome
Genome annotation
Full genome annotation
Specific annotation
Prokka
MLST, Virulence factor, antibiotics resistance, CRISPR/Cas, phages, etc
Assembly and annotation of *P. aeruginosa* AG1 genome

Figure 1. General bioinformatic pipeline to assemble, compare and annotate the *Pseudomonas aeruginosa* AG1 genome using short and long reads as well as hybrid approaches.

4.5 Gb of sequence, with 259,491 reads in total, a read mean length of 17,343 bp, a longest read of 201,659 bp, and a final mean coverage >560X.

Short reads genome assembly. Six *de Bruijn* graph based assemblers were used with default parameters and without reference guided option, if applicable. The classical assemblers included in the study were Velvet v1.2.10[35], SPAdes v3.13.0[36], IDBA v1.1.3[37], and Megahit v1.1.3[38]. Two newer assemblers were also included: SKESA v2.3.0[39] and Unicycler v0.4.7[11]. To estimate the best k-mer length for genome *de novo* assembly for Velvet, KmerGenie 1.7051 was implemented[40]. Other algorithms selected best k-mer length values automatically, if needed. Assembly sequences were kept at contig level with minimum size of 1,000 bp.

Long reads genome assembly. Three graph-based long read assemblers were used: Canu 1.8[41], Flye 2.3.7[42] and Unicycler v0.4.7[11]. Default parameters and no reference genome nor alternative sequencing data were considered. Only contigs with size higher than 1,000 bp were kept.

Hybrid genome assembly. Three graph-based hybrid approaches were applied. Default parameters without reference sequence were used to run IDBA-hyb v1.1.1 (https://github.com/loneknightpy/idba), Unicycler v0.4.7[11] and SPAdes v3.13.0[43]. Only contigs with size higher than 1,000 bp were kept.

Scaffolding. Prior the final version of the genome assembly of PaeAG1, BLASTn (blast.ncbi.nlm.nih.gov/Blast.cgi) was used to search closest genome according to contig sequences. All assemblies at contig level were assembled into scaffolds using the closest genome as reference sequences (*P. aeruginosa* strain RIVM-EMC2982, more details in Results) using MeDuSa v1.6[44]. When final version was achieved, scaffolding and benchmarking was done using the definitive version of the PaeAG1 genome with same scaffolder.

3C Benchmark of approaches and selection of best assembly. Benchmark of all assemblers were done according to 3C criterion, as follow:

Contiguity. Genome assembly statistics about quality and contiguity were assessed using QUAST 5.0.1[14] at both contig and scaffold levels. Assembler outputs were compared with regards to total assembly length (expected: around 7 Mb), number of contigs/scaffolds (one sequence expected), N50 (expected: as large as possible, close to genome size), NG50 (as large as possible), and others.

Completeness. Four strategies were implemented to assess completeness. First, single copy ortholog gene sets were searched (expected: 100%) in the assemblies using the BUSCO tool[45] within the gVolante plataform (https://gvolante.riken.jp)[18] and comparing gene content against 40 genes of the bacteria database (available at https://busco.ezlab.org/v1/). We also checked the ability of the assemblers to reproduce the complete sequences of the two class I integrons of PaeAG1 previously obtained by Sanger sequencing (KC907377 and KC907378). The third analysis used Circlator[19] to assess the replicon circularization achieved by assemblers that gave rise to single sequences (expected: a circular sequence). A last approach calculated the percentage of genomic and transcriptomic reads mapping to each genome reconstruction (expected: >95% mapping). To this end, short and long reads were remapped to the assemblies using BWA 0.7.17[46]. In addition, 12 reads files from a RNASeq experiment (triplicates of same strain under four experimental conditions with or without ciprofloxacin) were mapped to the assemblies using HISAT2 v2.1.0[47]. Qualimap v2.2.2[48] was used to calculate coverage and percentage of mapped reads, and comparison was done in a single report using MultiQC v1.7[49].

Correctness. Two strategies were used to evaluate correctness. The first one was to estimate error rates, check for uniform coverage, and detect false variants of short reads that mapped to the polished genome (see below, expected: 0% errors). This was done using Qualimap results. The second strategy was to calculate the percentage of identity of local alignments between known Sanger sequences (integrons, expected: 100% identity) of PaeAG1 and the final assembly (BLASTn).

All above criteria were considered to select the best assembly. This draft genome was polished and curated (next section) and the new version was included as extra 13th assembly.

We used all quantitative data to run a Principal Components Analysis (PCA), which was implemented in R software v3.5.1 (www.r-project.org/) using the Carret package (caret.r-forge.r-project.org/). This let to compare global profiles and performance given by assemblers. The final version of genome assembly was also included as an independent unit.

De novo assembly graphs were visualized using Bandage v0.8.1[50]. Finally, assembled sequences were visualized and compared against the final assembly using the BLAST Ring Image Generator (BRIG) tool v0.95[51].

Curation and polishing of the definitive genome assembly. Final adjustments of selected genome assembly were made manually based on the assembly graph, read coverage and distribution. Pilon 1.23[52] with BWA-mapped reads were implemented to automatically polish the assemblies. After this, a final polished assembly was obtained. Remapping of short and long reads, as well as all metrics calculations and 3C criterion evaluations were done again.

Comparative genome analysis. BLASTn of complete sequence was run again to find the closest genome, which jointly with the genome of the reference strain *P. aeruginosa* PAO1 were compared using Mauve v2.4.0[53] to determine the level of synteny and to describe global genomic structure.

Also, in order to compare the PaeAG1 genome with other ST-111 strains, a phylogenetic analysis was done using all the available complete sequences of ST-111 *P. aeruginosa* genomes. The reference strain *P. aeruginosa* PAO1 was also included. All the records were retrieved from Pseudomonas Genomes Database (PGDB, pseudomonas.com), and Roary program v3.12.0[54] was run with default parameters to establish relationships between strains using gene content by a pan-genome analysis. Scripts supplied with the program were used to create plots.

Whole genome annotation. For all assemblies, gene prediction and gene annotation was achieved using Prokka v1.13.3[55] and a custom database created with the genome of *P. aeruginosa* PAO1 and closest annotated strain to PaeAG1 as primary sources for annotation, or the default bacterial database provided with the software distribution. Also, Clusters of Orthologous Groups (COG), Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway were searched using EggNOG (http://eggnogdb.embl.de/)[56] for all coding sequences (CDS).

Specific genome annotation. Specific annotation and searching for specific genomic determinants was only done for the definitive final assembly. Default parameters were used in all cases. *In silico* serotyping was done using Past v1.0 (https://cge.cbs.dtu.dk/services/PAst-1.0/) and multilocus sequence typing[57] using MLST v2.0 (https://cge.cbs.dtu.dk/services/MLST/). Antimicrobial resistance genes were detected using RGI tool v5.1.0 (Resistance Gene Identifier, https://card.mcmaster.ca/analyze/rgi) and ResFinder v3.2 (https://cge.cbs.dtu.dk/services/ResFinder/). CRISPR-Cas arrays were investigated using CRISPRCasFinder v1.1.2 (https://crisprcas.i2bc.

3C Criterion	Level and metrics		Short reads only approaches					Long reads only approaches			Hybrid approaches				
			Velvet	SPAdes	IDBA	Megahit	SKESA	Unicycler	Canu	Flye	Unicycler	IDBA	SPAdes	Unicycler	Final assembly
Contiguity	Contigs assembly	Contigs	*227*	89	127	125	217	113	2	**1**	5	121	16	**1**	**1**
		Total length	*7027785*	7094145	7090598	7103650	7047434	7074438	7121028	**7209472**	**7465726**	7092836	7188777	7189601	**7190208**
		GC (%)	65.79	65.73	65.74	65.73	65.77	65.77	65.66	65.59	65.64	65.74	65.68	65.71	65.71
		N50	*65258*	223421	170948	168521	68375	151417	4329427	**7209472**	7178173	141288	1593634	**7189601**	**7190208**
		L50	33	11	14	14	34	15	1	1	1	15	2	1	1
	Scaffolding	Scaffolds	1	10	10	10	2	1	1	1	1	10	10	1	1
		N50 & NG50	7039385	7078855	7079244	7091835	7056837	7080238	7121028	7209472	7465826	7082290	7171429	7189601	7190208
		Genome fraction (%)	*97.714*	98.362	98.293	98.484	98.054	98.382	99.381	99.991	**100**	98.356	99.717	**99.992**	**100**
		NA50	177145	375326	491929	478607	708585	709611	4328063	7207242	7177177	477586	3956502	7189601	7190208
		LA50	12	6	5	5	4	4	1	1	1	4	1	1	1
		N's per 100 kbp	*217.06*	52.13	77.51	75.96	151.6	81.92	**0**	**0**	1.34	74.67	5.56	**0**	**0**
Correctness		Misassemblies	*81*	22	37	33	24	19	1	**0**	4	26	2	**0**	**0**
		Unaligned mis. contigs	0	0	0	0	0	0	0	0	0	0	0	0	0
		Mismatches per 100 kbp	6.56	2.42	4.88	1.61	1.84	***0.48***	*35.94*	*28.01*	*101.21*	3.68	11.33	**0.07**	**0**
		Indels per 100 kbp	6.49	0.41	0.67	0.28	1.79	**0.34**	*324.66*	*284.54*	*186.53*	1	1.14	**0**	**0**
Completeness	40 core genes (BUSCO)	Fragmented genes	0	0	0	0	0	0	4	9	9	0	0	0	0
		Intact genes	40	40	40	40	40	40	20	13	23	40	40	40	40
		Lost genes	0	0	0	0	0	0	*16*	*18*	*8*	0	0	0	0
		Completeness score (strict, %)	100	100	100	100	100	100	*50*	*32.5*	*57.7*	100	100	100	100
	Whole genome annotation	CDS	6574	6554	6543	6565	6540	6567	11229	9565	9089	6559	6605	6621	6620
		Contigs	1	10	10	10	2	1	1	1	1	10	10	1	1
		rRNA	2	5	5	5	3	3	12	12	12	4	14	12	12
		tmRNA	1	1	1	1	1	1	0	1	0	1	1	1	1
		tRNA	70	62	69	70	61	70	72	65	75	69	76	76	76
Completeness & correctness		Mean length of CDS (bp)	938.34	957.54	956.28	954.9	950.19	953.49	*499.35*	*607.14*	*664.14*	955.11	963.51	961.89	961.86
	Integron blaVIM-2	Identity (%)	100.0	99.5	99.8	100.0	100.0	99.7	99.488	99.257	99.843	99.753	99.778	99.778	100
		Coverage	0.5	0.7	0.6	*0.4*	0.5	0.6	**1.0**	**1.0**	1.0	0.6	**0.9**	**0.9**	**1.0**
	Integron blaIMP-18	Identity (%)	100.0	100.0	100.0	100.0	100.0	100.0	99.515	98.744	99.728	100	100	100	100
		Coverage	*0.6*	0.9	0.9	0.9	0.8	0.8	**1.0**	**1.0**	1.0	0.8	**1.0**	**1.0**	**1.0**

Table 1. Comparison of contiguity and annotation of *P. aeruginosa* AG1 genome assembly by different approaches*. *For some metrics, best and worst values are marked as bold or italics, respectively.

paris-saclay.fr/CrisprCasFinder/Index). Virulome was identified using Virulence Factor DataBase (VFDB, http://www.mgc.ac.cn/VFs/).

For mobilome delimitation, genomic islands were identified using IslandViewer v4 (www.pathogenomics.sfu.ca/islandviewer/). PHASTER was used to find prophages (phaster.ca/)[58] and integrons were searched using IntegronFinder v2.0[59]. The results of this series of searches were visualized in the genome using BRIG.

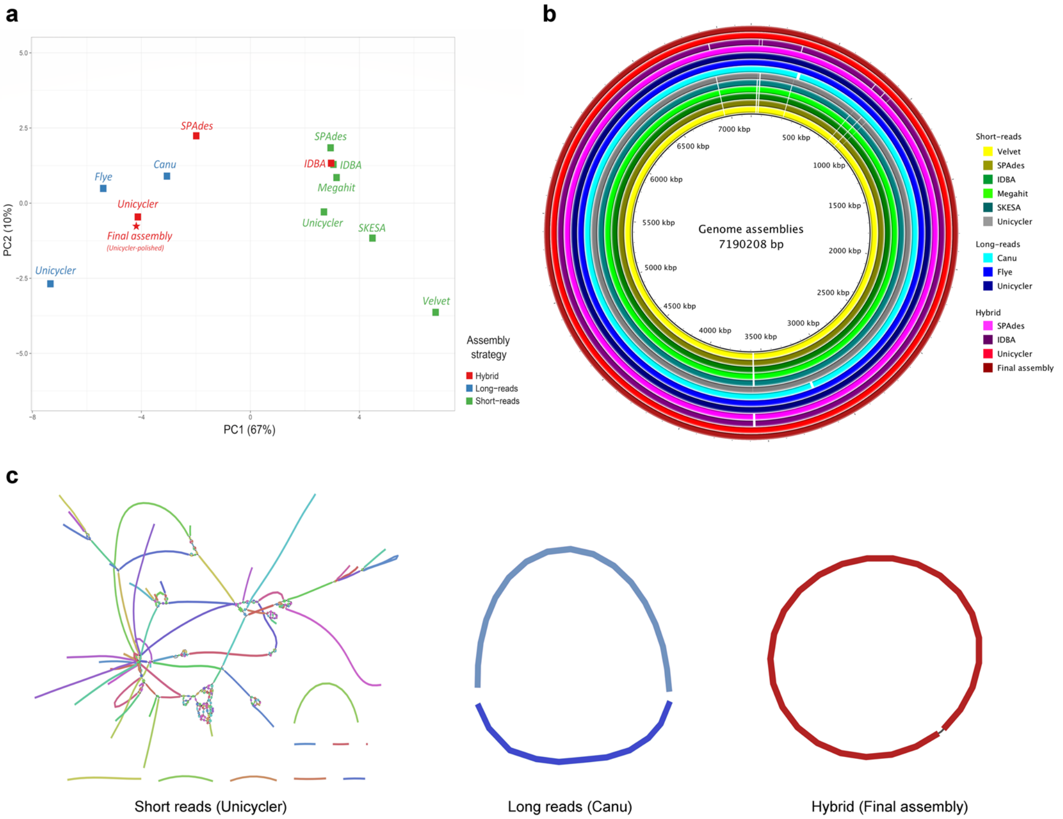
Results

In order to assembly the genome of *P. aeruginosa* AG1, an exhaustive workflow was implemented using hybrid and non-hybrid approaches, using Illumina short reads sequencing and Oxford Nanopore long reads sequencing data. General protocol is presented in Fig. 1. After sequencing and four bioinformatic steps, a single circular sequence was achieved and it was also annotated.

Benchmarking of hybrid and non-hybrid assemblers: a winner?.

Using different approaches, the PaeAG1 genome assembly was evaluated using the 3C criterion. The final version was presented as a last case, cured and polished. The contiguity and completeness criteria were initially the most important for the selection of the draft assembly, and then, a final polishing strategy focused on ensuring correctness (see next section). A summary of the most important metrics related to these criteria is presented in Table 1. Metrics related to scaffolding were obtained using the final assembly as reference, although various attempts to create scaffolds were made with closely related genomes.

According to results of contiguity, the use of short reads only approaches shows a lower performance (89 to 227 contigs and 1–10 scaffolds) compared to other approaches that exploit long reads (1 to 5 contigs and one scaffold for all cases) or hybrid methods (1–121 contigs and 1–10 scaffolds). Performance profiles between assemblers are compared in Table 1 and Fig. 2. Short reads assemblies are similar to each other according to Table 1 and PCA



Short reads (Unicycler)

Long reads (Canu)

Hybrid (Final assembly)

Figure 2. General comparison of *P. aeruginosa* AG1 genome assemblies. (**a**) Relationship between different assemblers by PCA using contiguity and annotation features. (**b**) Completeness evaluation and comparison for all different approaches using the final assembly as reference. (**c**) *De novo* assembly graph of three different approaches by short reads, long reads or hybrid assemblers. More details in Supplementary Fig. S1.

(Fig. 2a). In the case of long reads approaches, hybrid or not, the performance was also similar to each other at this contiguity level. Differences depending on technology and assembly strategy are recognized according to metrics and global profiles in PCA, gaps in the assembly and graphs (Fig. 2).

Only two assemblers generated a single contig. One is a long reads only approach (Flye) and the other one is a hybrid assembler (Unicycler). The hybrid assembler IDBA obtained metrics equivalent to the mode without the use of long reads (short reads only with 127 contigs and 121 contigs for hybrid approach), and also similar to Megahit (125 contigs and other metrics). Velvet and SKESA had the higher contigs values, 227 and 217 respectively.

The anticipated total genome length was similar among the 13 assemblers (7–7.2 Mb for all cases, except for long read only Unicycler with 7.4 Mb), while the N50 value tended to be much shorter for short reads assemblies (65–171 kb) compared to long reads (4.3–7.2 Mb). However, at the scaffold level N50 values were comparable among all cases (>7.0 Mb). At this same level, all assemblies covered virtually the entire final genome, although the lower performance was obtained for short reads only approaches (>97%).

As to correctness, long reads only were linked to high rates of mismatches (28–101 per 100 kb) and indels (186–324 per 100 kb), which were not solved by posterior polishing steps (as in Unicycler). Better values were obtained for other approaches using short reads, hybrid (0–11 mismatches and 1–1.14 indels) or not (0.48–6.6 mismatches and 0.3–6.5 indels). In addition, although long reads only assemblies generated sequences of approximately the same length as the other approaches, their annotations revealed high CDS numbers (9,089–11,229, which contrast with the 6,550–6,600 for short reads and hybrid approaches). Specific analysis of sequences showed a low median CDS size (average <600 bp) from long reads only assemblers compared to other cases with short reads only or hybrid (average 955 bp, which is an expected value for PaeAG1), suggesting fragmentation of CDS in the long reads assemblies.

Evaluation of 40 core genes using BUSCO tool and completeness score showed a 100% performance for short reads only and hybrid assemblers. However, in long reads only approaches it was possible to identify 13 to 23 core genes only (32.5–57.7%).

Regarding the PaeAG1 integron sequences obtained by Sanger sequencing, with a length greater than 2,500 bp and 3,000 bp, the assemblies of short reads only had low coverage (0.4–0.9), specifically in regions with repetitions. On the other hand, models with long reads had the best performance (1.0 in all cases), and their use in the hybrid approaches improved the assembly of the aforementioned repetitive zones (0.9–1.0 for all cases, except IDBA with 0.6–0.8).

Using all information, global profiles were compared the samples using a PCA. The full table used for PCA and the components values are provided in the "Supplementary Material PCA data". As presented in Fig. 2a, these profiles show a separation between the profiles of the short reads only (green color) and the others, creating two clusters. Also, unpolished and polished Unicycler assemblies kept close, as might be expected.

Enhancing the winner: polishing of hybrid unicycler assembly. The assembly directly obtained from the hybrid Unicycler approach was selected as the winner for its better fulfilled the 3C criteria, and it was used for downstream analyses. However, a review of the assembly was required in evidence of: (i) missing coverage for one of the known integrons sequences (Table 1) and (ii) presence of a zone with irregular/non-uniform distribution in the remapping of long reads (Supplementary Fig. S1a -left). Due to this, a manual curation was required. Curation was carried out with the help of the known sequences of the integrons, assembly graphs, and the assemblies of long reads only (because long reads could assemble that region). A detailed explanation of the curation is provided in the "Supplementary Material Manual curation" file, including a graphical representation.

After curation with short reads, a final polishing step was carried out to guarantee completeness. Only 5 bases were modified, which is reflected in the mismatches rate (per 100 kbp) of Unicycler hybrid of 5/7,190,208*100 kb = 0.07 (Table 1). When remapping of reads was done, regular and uniform coverage was detected, even in the conflictive zone (Supplementary Fig. S1a-right). Furthermore, the known integron sequences showed complete identity and coverage (Table 1, last column).

With this improved version of the assembly, in addition to the PCA comparison, an alignment of all assemblies was done against the final assembly to highlight the problematic regions to assemble. As shown in Fig. 2b some gaps were evident in all assemblies that were derived from short reads only and these gaps were not always compensated through the use of hybrid approaches. However, for most assemblers, the use of long reads only or hybrid improved those regions. Benchmark of all assemblers in a specific conflictive region is presented in Supplementary Fig. S1b. The assembly graphs of three cases are presented in Fig. 2c, showing the variable ability of assemblers to solve the *de novo* assembly problem.

3C assessment of PaeAG1 final genome assembly. To assess the final assembly of PaeAG1 genome, 3C criterion was re-evaluated:

Contiguity. The final assembly was built with hybrid Unicycler, with curation and polishing steps, but without the need for a reference genome. Full contiguity was achieved. A single and circular sequence was obtained.

Completeness. With all the elements evaluated, maximum completeness is considered. This includes circularization of sequence, 100% identity and coverage of known sequences of the integrons and 100% completeness scores in 40 expected genes (single copy orthologs set). Regarding the remapping of genomic reads, 99.85% of the short reads were mapped with an average coverage of 403X (See coverage graph in Supplementary Fig. S1c left). About long reads, 97.81% were mapped to the genome with an average coverage of 560X (Supplementary Fig. S1c right). Additional data from the same strain PaeAG1 using RNASeq technology achieved a mapping of 98.6% of read sequences.

Correctness. The polishing rounds that Unicycler includes and the additional polishing after curation using short reads guarantee the maximum accuracy of the genome assembly.

Thus, circular assembled genome was built according to 3C criterion: high contiguity, completeness and correctness was achieved.

Annotation of PaeAG1 genome. The PaeAG1 genome is composed of a single and circular sequence of 7,190,208 bp, with 65.71% GC content (Fig. 3a). A total of 6,620 CDS, 12 rRNA, 76 tRNA and 1 tmRNA (6,709 genes in total) were determined (Table 1). In addition, 2,197 genes were associated with Gene ontology terms, 5,537 related to defined COGs, and 3,060 to KEGG when orthologous groups and functional annotation were analyzed.

As shown in Fig. 3b, specific annotation of different genomic determinants was done, including antibiotic resistance genes, mobilome, virulence factors and others. Regarding antibiotic resistance gene profiling, genetic determinants of resistance to β-lactams, aminoglycosides, and fluoroquinolones, fosfomycin, phenicol and sulphonamide were found. By mechanism, 60 resistance associated genes were identified, including 44 efflux pumps and 8 associated with drug inactivation, including blaVIM-2 and blaIMP-18 gene alleles. Also, six determinant of target alteration and two of target replacement were identified. More details are shown in the Supplementary Table S1.

In the case of virulence factors, *P. aeruginosa* AG1 has more than 250 genomic determinants for 11 classes or enriched groups, including adherence (flagella, type IV pili biosynthesis and motility), antimicrobial activity (phenazines biosynthesis), antiphagocytosis (alginate production), iron uptake (pyochelin and pyoverdine), enzymes (phospholipases), biosurfactant (rhamnolipid biosynthesis), quorum sensing, proteases, regulation of two component system, type three secretion systems (T3SS) and toxins (exotoxin-A). More details are shown in the Supplementary Table S2.

In the study of the mobilome, diversity of elements were identified. At the genomic islands level, a total of 57 laterally acquired regions (size >10 kb) were identified (light blue in Fig. 3a), which correspond to drastic changes in the average GC composition. Six prophages (including two intact) were identified. The two complete integrons already described were also found. In correspondence to this diversity of mobile elements, no complete/functional CRISPR-Cas systems were recognized.

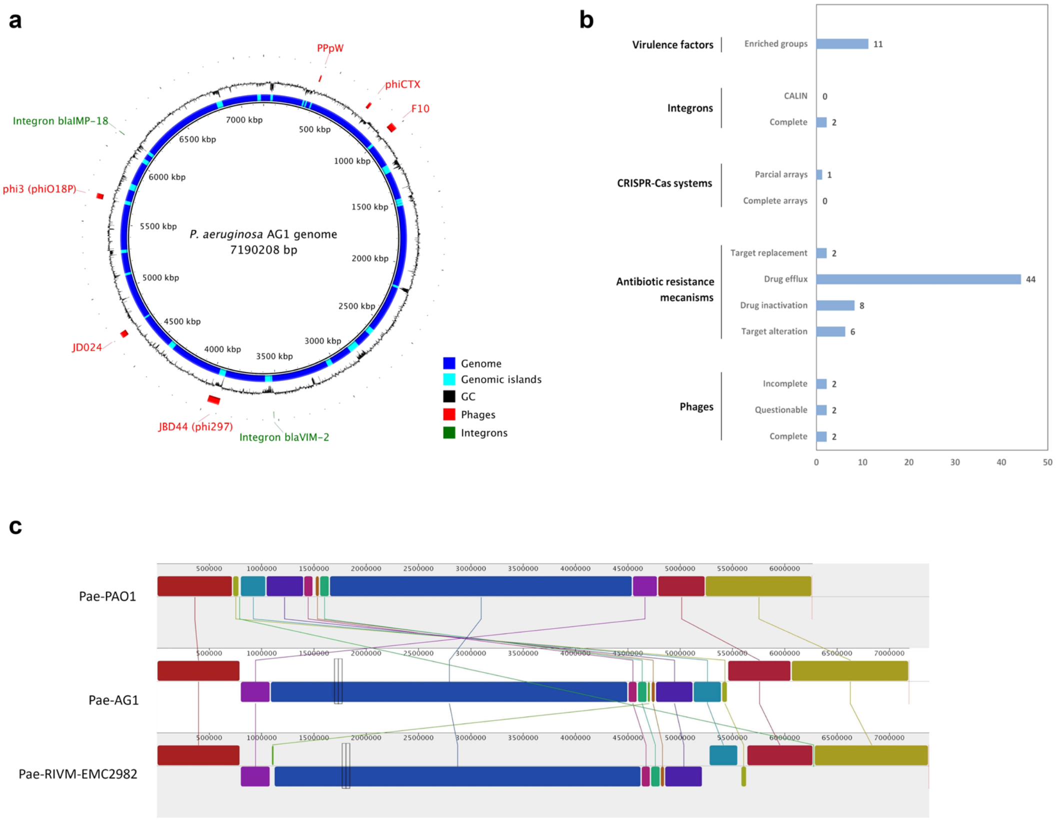


Figure 3. Annotation of *P. aeruginosa* AG1 genome. (**a**) Circularized genome showing phages and integrons locations. (**b**) Specific annotation of different genomic determinants including number of elements. (**c**) Genome synteny comparison among three strains of *P. aeruginosa*: PAO1 (general reference), AG1 (our assembly) and RIVM-EMC2982 (closest one to PaeAG1 according to BLAST analysis).

Using BLASTn, RIVM-EMC2982 (Accession CP016955.1; 7,380,063 bp, 65.7% GC content and Prokka annotation: 6,871 CDS, 76 tRNAs, 1 tmRNA and 12 rRNA; ST-111 and blaVIM-2+) was identified as the closest genome to PaeAG1 (Query cover 99%, identity 100%), which is a ST-111 and blaVIM-2 carrying strain. Both strains have same number of RNAs genes. Synteny comparison of the nucleotide sequences of both strain revealed 99% identity and 92% of coverage comparing PaeAG1 strain against RIVM-EMC2982. In addition, comparison of genome of PaeAG1 (genome size of 7.2 Mb) was done against strains PAO1 (6.3 Mb) and RIVM-EMC2982 (7.4 Mb). As shown in Fig. 3c, genomic blocks contrast with the general reference of the *P. aeruginosa* group, PAO1, which has almost 1 MB of difference of the genome size and around 1 000 genes. In the case of comparison with RIVM-EMC2982, general profile by blocks found similar arrays between both strains, congruent with genome sizes and content of mobile determinants in both strains.

In addition, comparison of gene content of ST-111 strains was used for phylogenetic analysis. A total of 9 complete genomes were available in PGDB, all with variable genome size (6.7–7.3 Mb) and gene content (6,200–7,400 genes). Pan-genome analysis revealed a total of 10,637 genes, which can separate strains in two clusters, one of them including PaeAG1 and *P. aeruginosa* RIVM-EMC2982 (Fig. 4a). The reference strain PAO1 was found to be completely separated from the group. Regarding core-genome, 4,783 genes (45% of total genes) were identified (present in at least 10 of the 11 sequences). A third part of genes were identified in only one of the strains. More details are shown in Fig. 4b,c. Interestingly, PaeAG1 is the only isolate which carries blaIMP-18 gene, in contrast to blaVIM-2 which was present in most of the strains.

Discussion

P. aeruginosa is an opportunistic pathogen able to adapt to different environments and it causes a variety of acute and chronic infections. PaeAG1 is a clinical isolate from a Costa Rican hospital with a profile of multi-resistance to antibiotics. In this context, concern over the increasing prevalence in hospitals of high-risk clones, including *Pseudomonas aeruginosa*, has prompted the use of typing methods and sequencing strategies to study the genomic epidemiology of bacterial clones at high resolution[1]. Interested in the assembly and annotation of PaeAG1 genome, we implemented different approaches using short and long reads and we benchmark them using the 3C criterion.

Benchmark of hybrid and non-hybrid assemblies. Of the more than 50 assemblies we run for pipeline standardization (considering different pre-processing, assembly and annotation steps), best cases per assembly

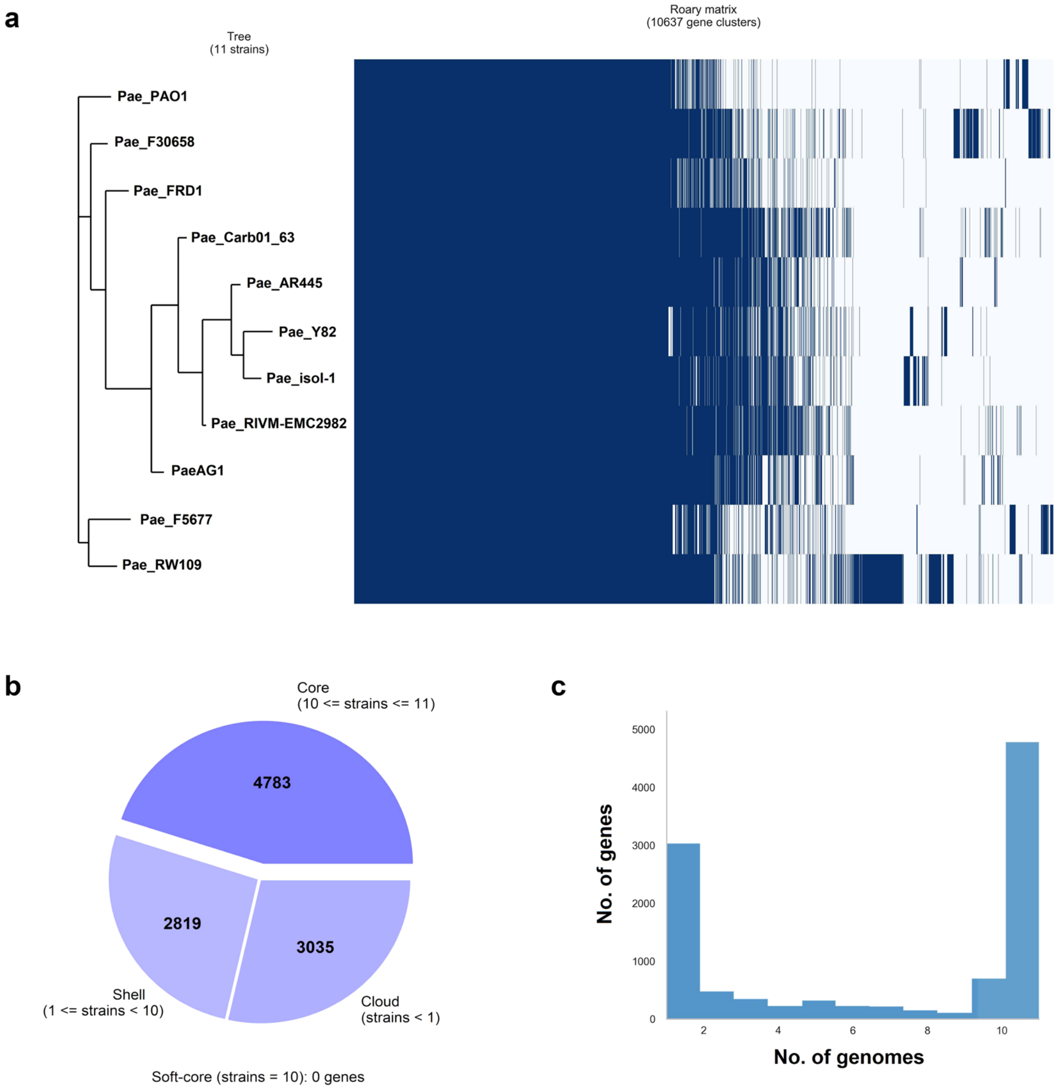


Figure 4. Pan-genome analysis of ST-111 *P. aeruginosa* strains. (**a**) Clustering according to strains profile by gene content. A total of 10,637 genes were identified. (**b**) Distribution of the gene content in all the strains, including that the core genome is composed of 4,783 (45% of total genes). Distribution of genes number by number of genomes is presented in (**c**).

were compared. In total 12 approaches were presented, and the best one was included as a 13th case after polishing and curation. According to the global profiles given by metrics and 3C benchmark, variable results were obtained (Table 1 and Fig. 2).

Regarding contiguity, fewer contigs were assembled using long reads or hybrid approaches in comparison to short reads. As reported, assembly continuity and genome size seems not to be correlated[60]. This is verified in our case, and dependency on technology seems more evident. Also, dependency on algorithms showed different contiguity, even for same type of approach. Use of long reads (non-hybrid or hybrid method) improved contiguity metrics, solving most of conflictive regions that short reads could not assemble.

In the case of correctness, long reads only approaches presented critical problems in accuracy. As in our study, in a recent study error rates for short reads and hybrid assemblies were similar but were much higher for long reads assemblies using Unicycler in all cases[1]. Even though we had ultra-deep coverage for both sequencing technologies, this could be no enough to correct error in long reads only assemblies. This is probably due to systematic errors that have been detected in long reads sequencers, without compensation even increased sequencing depth[10]. In addition, our results using long reads only assemblers tended to have larger assemblies (total length) and duplication in different contigs was recognized. This is has been previously reported for long read assemblers[10] and it could be a major obstacle for polishing the genome[12] and compromising accuracy.

To assess completeness, we implemented an analysis using expected gene content by searching single-copy orthologs[61]. Short reads only and hybrid approaches achieved the assembly of 100% of core genes, but long reads only had a poor performance. Also, despite the larger number of CDS for long reads, incomplete assembly of genes was evidenced. Fragmentation of genes was confirmed by comparing the average size of all those elements.

In long reads only assemblies the CDS average size was <600 bp, but for all other approaches this value was around 955 bp (Table 1). The CDS average size of the closest genome to PaeAG1, RIVM-EMC2982, is 955 bp, meanwhile for PAO1 strain is 1000 bp. This appreciation has been briefly reported before[62]. The incompleteness of genome assembly will not matter if genome structure is not the focus of a study[9], but it is not the case of PaeAG1, where genomic events reconstruction would be crucial to understand the special features of this strain.

When all features of assemblies are included in the PCA analysis, general profiles of short reads approaches define a separated cluster, and another one for long reads and hybrid methods (Fig. 2a). Considering all the metrics of the 3C criterion, definitively SPAdes and Unicycler hybrid approaches outperformed non-hybrids methods. This can be explained due reference-free genomes assembly is feasible using best features of both short and long reads technologies[9]. IDBA assembler is a particular case which remains as the same using the hybrid or non-hybrid approach.

About other works related to the algorithms we evaluated, different results have been found depending on data and genome complexity. However, since introduction of Unicycler assembler, a last generation algorithm, most studies have suggested that Unicycler outperforms other approaches[1,10,11,63]. In the case of IDBA and Velvet, performance was comparable to SPAdes when it was introduced[36]. For Megahit, an assembler for metagenomes but also working for single genomes[38], it has been also used in recent studies, mainly related to microbial communities or particular strains[64]. More restricted works using SKESA are reported, but performance seem to be better than SPAdes and Megahit for some cases[39].

For short reads only or hybrid assemblies, SPAdes is still used to aseembly genomes[36,39,65]. In a recent study, SPAdes had better results when compared to others, where Unicycler was not included[3].

For long reads, Canu has been successfully implemented in different studies[10,12,41], showing well performance when benchmark is done (but most of them without Unicycler assember). For Flye, it has been used in recent studies[66,67], including a case where Canu, Flye and Unicycler (using long reads only and hybrid approaches) had very similar performance[68]. Comparison between Unycicler, SPAdes and Canu has shown that in some cases Canu and SPAdes are not able to circularize the final assembly, unlike Unicycler[11]. In another study with long reads only, Canu was the best ranked assembler using *Escherichia coli* genome[12].

All this variable results of assemblers (in our benchmark and the literature) are congruent with several reports about the diversity of assemblers, which have been developed to generate high quality *de novo* assemblies, but their output is very different because of algorithmic differences, data source and genomic complexity[2]. This complicates selection of appropriate strategy. Thus, the need for more capable assemblers is still mandatory in terms of capabilities, accuracy and the way to deal with genomic features[3].

Regarding the differences in cost for both technologies (only considering sequencing step and no other complementary costs) Illumina short reads sequencing cost ($1500) was around three times more expensive than ONT ($500) sequencing. In our case, the hybrid approach has a cost of around $2000 for both technologies. Although we had ultra-deep sequencing data for both platforms, the minimal coverage requirements for PaeAG1 genome assembly are not known, which could significantly reduce the sequencing price. This cost is higher than other studies but with hundreds of sequenced samples[69,70], in contrast with our case in that a single genome was sequenced (increasing costs).

In the case of conflictive regions, each assembler implements slightly different heuristics to deal with repetitions in the genome, uneven coverage, sequencing errors and chimeric reads[8]. Efforts to generate complete genome sequences with repetitive regions has been hampered by dramatic expansion of mobile elements, especially when short read sequencing methodologies are used[13]. In PaeAG1 genome assembly, different complicated regions were identified when short reads only approaches (all methods) and hybrid IDBA were used, creating gaps in an incomplete assembly (Fig. 2b). Although the PaeAG1 has not really a repeat-dense genome, mobile elements add repetitive sequences. This has complicated the assembly of its genome using short reads only approaches. All this regions were apparently solved by long reads only and for hybrid SPAdes and hybrid Unicycler. This results are expectable according to previous reports and the differences in each technology. Use of long reads technologies achieve repeat regions spanning[63] and it permits bridging of repetitive sequences[65].

However, evaluation of remapping of reads with the selected assembly (hybrid Unicycler according to 3C criterion) revealed a variation in the coverage in one specific region, as shown in Supplementary Fig. S1a (left), with an irregular and non-uniform distribution of reads. This conflictive region was preliminary annotated as a flanking repetitive sequence of one of the integrons (containing blaVIM-2 gene). This is a common phenomenon in regions carrying antimicrobial resistance determinants, which are often flanked by repetitive insertion sequences, and it can be difficult to assemble using short reads because are very short compared to the repetitions[10]. In our case, the conflictive region is part of the known region of the integron (approx. 2,500 bp, sequenced using Sanger method), and 100% of short reads had a size of 101 bp. Although this region was identified in a hybrid approach, this problem is an in force limitation of the algorithms[11] and curation step was required.

No resolution of repetitive region made that short reads were mapped incorrectly[9], evidenced as a coverage peak of reads in the remnant conflictive region of PaeAG1 genome assembly. In addition, this is congruent with the alignment of known sequence against the assembly. At least a 12% of the blaVIM-2 carrying integron sequence was lost in the hybrid approaches, including hybrid Unicycler (Table 1). We can conclude that those identical flanking regions of integrons were not well assembled using short reads. Long reads approaches were able to coverage both regions completely. The compromised ability of the Unicycler algorithm to assemble this conflictive region in the hybrid mode is related to the approach. In general, hybrid assembly can be accomplished with either a short-read-first or long-read-first approach. In the short-read-first method, contigs are assembled using short reads followed by a scaffolding is addresses using long reads[11]. Drawbacks of this approach include scaffolding mistakes and structural errors (misassemblies) in the sequence[71]. This could be the reason of our case in the conflictive region due Unicycler in hybrid mode is a short-read-first approach. In this context, the genome assembly problem is an open issue due is a NP-hard problem, and no universal solution to find the optimal route in graph-based approaches is available, in particular which is aggravated by repetitive regions. To deal with repetitive sequences in the genome, Unicycler determine the occurrence (multiplicity) of contigs in the assembly using both depth and connectivity using a greedy algorithm, and a bridging step is used to connect contigs and solve repeats using paired-end short reads[11]. However, due the algorithm used by Unicycler is a greedy approach, optimal solution is not warranted, and assembly errors can be induced. Thus, additional steps, as the manual curation, are required.

In this sense, manual curation is a common practice to finish genome due complexity of genomic data which algorithms not always can deal with[9,10]. In a case, by comparing long reads only and hybrid assemblies, this manual curation it implied recovery of lost sequences up to 18 kbp for some assemblies in another study[10]. Same situation was presented in another ST-111 *P. aeruginosa* strain, where flanking regions of blaVIM-2 gene was broken during assembly[72]. In other studies, no polishing strategy improves the completeness of assemblies[65].

To improve the genome assembly of PaeAG1, curation was done with the help of the known sequences of PaeAG1 (Sanger sequencing), assembly graphs and the assemblies of long reads only. After this polishing step, remapping showed a uniform distribution of reads (Supplementary Fig. S1a right) and complete matching (100% identity and coverage) of the known sequences of the integrons, as expected.

At graph assembly level, when topological structure of assembly is analyzed for short reads assemblies (Fig. 2c, short reads), a collapsed graph is evidenced, where sequences are shown as cycles due the repeats or small shared sequences in many reads at same time. This means that there is insufficient information to disambiguate the repeat or shared sequences in the graph. This problem was solved when long reads were implemented, showing no cycles for long reads approaches (although shown case had two contigs), and a complete circularized genome for the final hybrid assembly.

Assessment of the genome assembly of PaeAG1. Based on best overall quality statistics and polishing, hybrid approach using Unicycler was selected as the final assembly of PaeAG1 genome using 3C criterion.

In our initial efforts to assembly the genome, using only short reads, most of assemblers generated more than 100 contigs, and using RIVM-EMC2982 strain (which was selected after doing a full genome BLASTn of contigs), scaffolding finished with 1 sequence for the case of Unicycler and 22 gaps. In order to improve the genome assembly, ONT technology was used to produce long reads and new evaluations were made using both, long read only or hybrid methods.

On the other hand, notwithstanding all the three contiguity, completeness and correctness evaluation are frequently evaluated in genome assembly studies[3,8,12,15–17], no explicit conceptualization of “3C criterion” has been achieved. Here we emphasized its use to referrer to the classical metrics and comparisons.

The final assessment of the definitive assembly of PaeAG1 genome accomplished an ultra-deep coverage for both, short (>400X) and long reads (>560X) technologies. Also it achieved high performance according to 3C criterion: (i) full contiguity with a single and circular genome without gaps; (ii) correctness based on short reads remapping and polishing, achieving full accuracy (including known sequences of the strain); and (iii) completeness according to identification of 100% of expected core gene set and percentage of remapping of genomic reads as well mapping of reads from RNASeq technology.

Altogether, the use of a hybrid strategy allowed the PaeAG1 genome to be inferred by a *de novo* or reference-free assembly approach, which it represent a key element in the study of this strain due its exclusive genomic features[9]. To our knowledge, this is the first genome assembly of a ST-111 *P. aeruginosa* strain using a hybrid approach.

The first hybrid assemblies for other-class *P. aeruginosa* strains were published recently[23,73,74]. In order to evaluate our pipeline in these publicly available sequencing data, we implemented our hybrid approach to the two cases with Illumina and ONT sequencing technologies. For the case of the *P. aeruginosa* strain Houston-1[73], we were able to reproduce the assembly of the chromosome and the plasmid with our approach. For the *P. aeruginosa* strain CRPA[23], the published draft genome was composed of three contigs, and with our approach we were able to finish into two contigs, representing an improvement in the assembly. More details of the assemblies of these two strains are shown at the end of the Supplementary Material Manual curation.

Annotation of the PaeAG1 genome and epidemiological insights. In order to identify main features of the PaeAG1 genome, including its architecture, composition and functions, genome characterization and annotation was done. The PaeAG1 chromosome is a large and circular sequence of 7,190,208 bp, larger than reference strain PAO1 and similar to other ST-111 strains size[31,75]. Same pattern was found for the GC content of 65.7%. This relatively large genome in *P. aeruginosa* has been associated to thrive in a repertoire of hosts and environments[21].

The general annotation of genome revealed that PaeAG1, contain 6,709 genes (including 6,620 CDS), which are related to 2,197 Gene ontology terms, 3,060 elements in KEGG and 5,537 COGs. In similar way as reported in first whole genome sequencing of a *P. aeruginosa* strain[76], genome analysis of PaeAG1 shows determinants associated to versatility and successful ability to conquer multiple niches in nature. For example broad capabilities to transport and metabolize organic substances, presence of chemotaxis systems, biofilms production and efflux systems have been described and all of them were annotated for PaeAG1.

Genome sequence analysis using molecular typing methods showed that PaeAG1 has a ST-111 profile and O12 serotype. ST-111 is a lineage that belongs to the O12 serotype, which has been associated with multidrug resistance and expansion in hospitals for decades[28,72,75]. Thus, emergence of high-risk clones, including the ST-111 clones of *P. aeruginosa*, undermines the available therapeutic strategies and therefore, compromises public health. The presence of this kind of high-risk clones in Costa Rican hospitals is a nationwide concern because MBL and particular virulence factors producing isolates cause serious infections that are difficult to treat[77]. This same

ST-111 profile has been identified in most of MBL producing *P. aeruginosa* strains in the United Kingdom[75] thus as in Netherlands[77].

Annotation of virulence factors found classical elements in *P. aeruginosa* group[78], including elements related to adherence, antiphagocytosis, iron uptake, phospholipases, biosurfactant, quorum sensing, proteases, regulation, secretion systems, and toxins. Some particular virulence factors of PaeAG1 are substrate for type I protein secretion system T1SS (alkaline protease aprA), T2SS (elastases LasA and LasB, exotoxin-A and phospholipases PlcH, PlcN, and PlcB) and T3SS (ExoS, ExoT, and ExoY)[78]. It has been reported that secretion of ExoS is predominantly identified in invasive *P. aeruginosa* strains[78]. Recently, this determinant was identified in two blaVIM-2 carrying strains, one serotype O12 and ST-111 isolate (*P. aeruginosa* Carb01 63) and another O11 strain of ST-446 (*P. aeruginosa* S04 90) in Netherlands[31]. In PaeAG1, a potential invasive role of this strain can be related to the presence of this element.

In the context of mobile genetic elements, large number of determinants were identified in the chromosome of PaeAG1, including multiple genomic islands, six prophages and two integrons. Comparison of PaeAG1 against the reference of the *P. aeruginosa* group PAO1 and the closest strain to PaeAG1, RIVM-EMC2982, is consistent with genome size and mobile elements content. In the case of strain PAO1, this reference has a 6.3 Mb genome, meanwhile PaeAG1 has almost 1 Mb more of bases pairs (around 1,000 genes). This difference is congruent with high content of genomic island and other mobile elements in PaeAG1 but it is compromised in PAO1 strain. In the case of RIVM-EMC2982 (ST-111 and blaVIM-2+), this strain was identified as the closest to PaeAG1 and similar profile by genomic blocks were recognized (Fig. 3c). Meticulous analysis showed some different genomic arrangements, including differences in composition of mobile elements and absence of blaIMP-18 in RIVM-EMC2982.

In the case of the six prophages, all of them are also found in RIVM-EMC2982 genome (ten prophages in total) in same conditions of integrity. However, there are variable results of prophage presence in many ST-111 strains, which has been discussed as difficult to interpret, due transient nature of phages or the more methodological issues[72]. In addition, these high numbers of prophages might be related to the absence of CRISPR-Cas systems in the genome[31], as the case of PaeAG1. Reports of compromised CRISPR-Cas defense systems are associated to better ability to acquire mobile element carrying antibiotic resistance genes in *P. aeruginosa* and other organisms[79].

Regarding the integrons of PaeAG1, identification of genes *intl*1, *sul*1 and *qacE*Δ1 for class I integrons, suggested two integron-like structures carrying the VIM-2 and IMP-18 genes[22]. This was confirm when Sanger method was used for sequencing both integrons. In our assembly, these two complete integrons and same structure were found, one carrying blaVIM-2 and another one including blaIMP-18. This is congruent with previous studies showing that these two genes are regularly identified in integrons in *P. aeruginosa*[30,31,80].

In more detail, VIM (Verona integron-encoded metallo-β-lactamase) enzymes have same hydrolytic spectrum than the IMP-type enzymes, and specifically blaVIM-2 is responsible of multiple outbreaks being the most widespread MBL in *P. aeruginosa*[30]. Multiple strains carrying VIM-2 have been identified in different latitudes around the world[75,80–83]. In United Kingdom, a study with 87 ST-111 *P. aeruginosa* strains found that 73 isolates carried VIM-2 and others carried different IMPs and one isolate had both VIM-2 and IMP-18, the second report of a clone carrying both MBL[75]. In a Netherlands outbreak, another strain (Carb01–63 strain, isolated from drains and sinks in a hospital) had a ST-111 profile and it was closely related to same RIVM-EMC2982[31]. All the three strains (PaeAG1, Carb01–63 and RIVM-EMC2982, in the same group according to phylogenetic analysis) are resistant to multiple antibiotics and carry blaVIM-2 allele.

In the case of imipenemases coded by blaIMP-18 gene, outbreaks reports and genetic context is limited in *P. aeruginosa*, including some cases in United States[84], México[85], France[81] and Puerto Rico[86].

For other antibiotic resistance determinants, annotation also included serine- and metallo-β-lactamases (PDC-3, OXA-2, as well as VIM-2 and IMP-18), porins and efflux pumps (including mexAB–oprM, mexCD–oprJ, mexEF–oprN, mexHI–opmD operons). All of them may contribute to the multi-resistance phenotype in PaeAG1.

As it was revealed by pan-genome analysis of ST-111 members, variable composition of gene content separate strains in relatively independent groups. The strains (including PaeAG1) belongs to the O12 serotype, which has been associated with multidrug resistance and nosocomial expansion[28,29]. PaeAG1 was close to the main group with 5 isolates, including the *P. aeruginosa* RIVM-EMC2982 (the closest to PaeAG1 by BLAST analysis) and Carb01–63 strains. Although all the strains (except the reference) are part of same group, differences in gene content is a remarkable feature, including that PaeAG1 was the only strain carrying blaIMP-18 genes. In contrast, ST-111 strains has been frequently associated with blaVIM-2, as mentioned before[28,75]. Other less commonly associated lactamases genes include VIM-4 or other IMP-type enzymes, but also only with extended-spectrum β-lactamases without carbapenemase activity (such as VEB-1 and OXA)[75].

Due differences in size of the genome (6.7–7.3 Mb) and gene content, as well as the particular genomic features of this strains (genomic island composition and evolution, mobile elements, integrons, phages and others), further analysis are required to describe high plasticity in this group.

Conclusions

Advances in sequencing technology play an increasing and determinant role in infection investigations and tracking evolution of international lineage of high-risk bacterial clones in clinical context over long times and in great detail[87]. However, genome assembly is not obvious and it is challenged by sequencing technology, genomic features and all bioinformatic algorithms, making it an open problem. Exhaustive comparison of different strategies to assembly the genome and it assessment gives a better way to get close to the real genome sequence. Benchmarking using the 3C criterion is a consensus approach that includes different levels and aims of comparison for the robust selection of a final assembly.

In our case, a hybrid assembly was the best approach to achieve a single circular sequence with high quality 3C for the case of the genome of a high-risk *P. aeruginosa* strain. Thus, best features of short and long reads sequencing technologies are included and their drawbacks are compensated.

The case of PaeAG1 genome assembly is a first and important step to understand the genomic architecture of an ST-111 high-risk strain. Annotation could reveal all the genomic content and molecular determinants related to phenotypes, which for PaeAG1 are related to multi-resistance and virulence mainly. This highlighting the need for more studies using epidemiological information and both high throughput technologies and conventional methods to understand the molecular mechanisms and phenotypes, make decisions at clinical level and to fight, and hopefully, overcome the antibiotic multi-resistance problem.

Data availability

Data input and output data for PCA are provided as Supplementary material PCA data. The details of the approach for the manual curation are available in the Supplementary Material Manual Curation.

Scripts for bioinformatics analysis are provided as a supplementary material, but also available at https://github.com/josemolina6/PaeAG1_genome/blob/master/Script_for_bioinformatic_analysis.sh.

To specifically run the analysis of the 3C criterion, access a simplified Script at: https://github.com/josemolina6/PaeAG1_genome/blob/master/Script_3C_evaluation.sh.

The annotated final assembly of the PaeAG1 chromosome was deposited in GenBank under the accession number CP045739. Short reads and long reads raw data were uploaded to the NCBI Sequence Read Archive (SRA) and it is available under the accessions numbers SRX7088413 and SRX7088414, respectively. A full table of all the details of the genome annotation is provided as a Supplementary material, and it is also available at: https://github.com/josemolina6/PaeAG1_genome. Files of the annotation in different formats as well as the fasta files of all the assemblies are available in the same link.

Received: 14 November 2019; Accepted: 6 January 2020;
Published online: 29 January 2020

References

1. Gonzales Decano, A. *et al*. Complete Assembly of Escherichia coli Sequence Type 131 Genomes Using Long Reads Demonstrates Antibiotic Resistance Gene Variation within Diverse Plasmid and Chromosomal Contexts. *mSphere* **4** (2019).
2. Kwon, D., Lee, J. & Kim, J. GMASS: A novel measure for genome assembly structural similarity. *BMC Bioinformatics* **20**, 1–9 (2019).
3. Yahav, T. & Privman, E. A comparative analysis of methods for de novo assembly of hymenopteran genomes using either haploid or diploid samples. *Sci. Rep.* **9**, 1–10 (2019).
4. Ekblom, R. & Wolf, J. B. W. A field guide to whole-genome sequencing, assembly and annotation. *Evol. Appl.* **7**, 1026–1042 (2014).
5. Aguilar-Bultet, L. & Falquet, L. Secuenciación y ensamblaje de novo de genomas bacterianos: una alternativa para el estudio de nuevos patógenos. *Rev. Salud Anim.* **37**, 125–132 (2015).
6. Miller, J. R., Koren, S. & Sutton, G. Assembly algorithm for Next-Generation Sequencing data. *Genomics* **95**, 315–327 (2010).
7. Bellec, A., Courtial, A., Cauet, S. & Rodde, N. Long Read Sequencing Technology to Solve Complex Genomic Regions Assembly in Plants. *J. Next Gener. Seq. Appl.* **3** (2016).
8. Alhakami, H., Mirebrahim, H. & Lonardi, S. A comparative evaluation of genome assembly reconciliation tools. *Genome Biol.* **18**, 1–14 (2017).
9. Wang, W. *et al*. Assembly of chloroplast genomes with long- and short-read data: A comparison of approaches using Eucalyptus pauciflora as a test case. *BMC Genomics* **19**, 1–15 (2018).
10. Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Completing bacterial genome assemblies with multiplex MinION sequencing. *Microb. Genomics* **3** (2017).
11. Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput. Biol.* **13**, 1–22 (2017).
12. Jayakumar, V. & Sakakibara, Y. Comprehensive evaluation of non-hybrid genome assembly tools for third-generation PacBio long-read sequence data. *Brief. Bioinform.* **20**, 866–876 (2019).
13. Batty, E. M. *et al*. Long-read whole genome sequencing and comparative analysis of six strains of the human pathogen Orientia tsutsugamushi. *PLoS Negl. Trop. Dis.* **12**, 1–17 (2018).
14. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).
15. Michael, T. P. *et al*. High contiguity Arabidopsis thaliana genome assembly with a single nanopore flow cell. *Nat. Commun.* **9**, 1–8 (2018).
16. Broad Institute. GAEMR. Available at: http://software.broadinstitute.org/software/gaemr/ (Accessed: 30th July 2019) (2019).
17. Nadalin, F., Vezzi, F. & Policriti, A. GapFiller: a de novo assembly approach to fill the gap within paired reads. *BMC Bioinformatics* **13**, S8 (2012).
18. Nishimura, O., Hara, Y. & Kuraku, S. gVolante for standardizing completeness assessment of genome and transcriptome assemblies. *Bioinformatics* **33**, 3635–3637 (2017).
19. Liao, Y. C. *et al*. Completing bacterial genome assemblies: strategy and performance comparisons Oxford Nanopore MinION sequencing and genome assembly Circlator: automated circularization of genome assemblies using long sequencing reads Versatile genome assembly evaluation. 2016–2017 (2019).
20. Duan, J., Jiang, W., Cheng, Z., Heikkila, J. J. & Glick, B. R. The Complete Genome Sequence of the Plant Growth-Promoting Bacterium Pseudomonas sp. UW4. *PLoS One* **8** (2013).
21. Freschi, L. *et al*. The Pseudomonas aeruginosa Pan-Genome Provides New Insights on Its Population Structure, Horizontal Gene Transfer, and Pathogenicity. *Genome Biol. Evol.* **11**, 109–120 (2019).
22. Toval, F. *et al*. Predominance of carbapenem-resistant Pseudomonas aeruginosa isolates carrying blaIMP and blaVIM metallo-β-lactamases in a major hospital in Costa Rica. *J. Med. Microbiol.* **64**, 37–43 (2015).
23. Yu, X. *et al*. Long-read Nanopore Sequencing-based Draft Genome of a Carbapenem-resistant Pseudomonas aeruginosa. *J. Glob. Antimicrob. Resist.* https://doi.org/10.1016/j.jgar.2019.05.023 (2019).
24. Farajzadeh Sheikh, A. *et al*. Molecular epidemiology of colistin-resistant Pseudomonas aeruginosa producing NDM-1 from hospitalized patients in Iran. *Iran. J. Basic Med. Sci.* **22**, 38–42 (2019).
25. Miriagou, V. *et al*. Acquired carbapenemases in Gram-negative bacterial pathogens: detection and surveillance issues. *Clin. Microbiol. Infect.* **16**, 112–22 (2010).

26. Baquero, F., Coque, T. M. & Cruz, Fdela Ecology and Evolution as Targets: the Need for Novel Eco-Evo Drugs and Strategies To Fight Antibiotic Resistance. *Antimicrob. Agents Chemother.* **55**, 3649–3660 (2011).
27. Willems, R. J. L., Hanage, W. P., Bessen, D. E. & Feil, E. J. Population biology of Gram-positive pathogens: high-risk clones for dissemination of antibiotic resistance. *FEMS Microbiol. Rev.* **35**, 872–900 (2011).
28. Woodford, N., Turton, J. F. & Livermore, D. M. Multiresistant Gram-negative bacteria: the role of high-risk clones in the dissemination of antibiotic resistance. *FEMS Microbiol. Rev.* **35**, 736–755 (2011).
29. Mulet, X. *et al*. Biological markers of Pseudomonas aeruginosa epidemic high-risk clones. *Antimicrob. Agents Chemother.* **57**, 5527–5535 (2013).
30. Hong, D. J. *et al*. Epidemiology and characteristics of metallo-ß-lactamase-producing Pseudomonas aeruginosa. *Infect. Chemother.* **47**, 81–97 (2015).
31. van der Zee, A. *et al*. Spread of carbapenem resistance by transposition and conjugation among Pseudomonas aeruginosa. *Front. Microbiol.* **9**, 1–11 (2018).
32. Andrews, S. FastQC A Quality Control tool for High Throughput Sequence Data. Available at, https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (Accessed: 10th April 2018) (2010).
33. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
34. Loman, N. J. & Quinlan, A. R. Poretools: a toolkit for analyzing nanopore sequence data. *Bioinformatics* **30**, 3399–3401 (2014).
35. Zerbino, D. R. & Birney, E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821–829 (2008).
36. Bankevich, A. *et al*. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–77 (2012).
37. Peng, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. IDBA – A Practical Iterative de Bruijn Graph De Novo Assembler. In 426–440, https://doi.org/10.1007/978-3-642-12683-3_28 (Springer, Berlin, Heidelberg, 2010).
38. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).
39. Souvorov, A., Agarwala, R. & Lipman, D. J. SKESA: strategic k-mer extension for scrupulous assemblies. *Genome Biol.* **19**, 153 (2018).
40. Chikhi, R. & Medvedev, P. Informed and automated k-mer size selection for genome assembly. *Bioinformatics* **30**, 31–37 (2014).
41. Koren, S. *et al*. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
42. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546 (2019).
43. Antipov, D., Korobeynikov, A., McLean, J. S. & Pevzner, P. A. hybridSPAdes: an algorithm for hybrid assembly of short and long reads. *Bioinformatics* **32**, 1009–1015 (2016).
44. Bosi, E. *et al*. MeDuSa: a multi-draft based scaffolder. *Bioinformatics* **31**, 2443–2451 (2015).
45. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
46. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–60 (2009).
47. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
48. Okonechnikov, K., Conesa, A. & García-Alcalde, F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* **32**, 292–4 (2016).
49. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).
50. Wick, R. R., Schultz, M. B., Zobel, J. & Holt, K. E. Bandage: interactive visualization of *de novo* genome assemblies: Fig. 1. *Bioinformatics* **31**, 3350–3352 (2015).
51. Alikhan, N.-F., Petty, N. K., Ben Zakour, N. L. & Beatson, S. A. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics* **12**, 402 (2011).
52. Walker, B. J. *et al*. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLoS One* **9**, e112963 (2014).
53. Darling, A. C. E., Mau, B., Blattner, F. R. & Perna, N. T. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* **14**, 1394–403 (2004).
54. Page, A. J. *et al*. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691–3693 (2015).
55. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
56. Huerta-Cepas, J. *et al*. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, D309–D314 (2019).
57. Larsen, M. V. *et al*. Multilocus sequence typing of total-genome-sequenced bacteria. *J. Clin. Microbiol.* **50**, 1355–61 (2012).
58. Arndt, D. *et al*. PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res.* **44**, W16–W21 (2016).
59. Cury, J., Jové, T., Touchon, M., Néron, B. & Rocha, E. P. Identification and analysis of integrons and cassette arrays in bacterial genomes. *Nucleic Acids Res.* **44**, 4539–4550 (2016).
60. Koren, S. *et al*. Reducing assembly complexity of microbial genomes with single-molecule sequencing. *Genome Biol.* **14**, R101 (2013).
61. Wang, W. *et al*. Data descriptor: The sequence and de novo assembly of hog deer genome. *Sci. Data* **6**, 4–11 (2019).
62. Kirkegaard, R. What is a good genome assembly? – Albertsen Lab. Available at, https://albertsenlab.org/what-is-a-good-genome-assembly/ (Accessed: 9th August 2019) (2019).
63. Peter, S. *et al*. Tracking of antibiotic resistance transfer and rapid plasmid evolution in a hospital setting by Nanopore sequencing. *bioRxiv* 639609, https://doi.org/10.1101/639609 (2019)
64. Learman, D. R. *et al*. Comparative genomics of 16 Microbacterium spp. that tolerate multiple heavy metals and antibiotics. *PeerJ* **6**, e6258 (2019).
65. Nicholls, S. M., Quick, J. C., Tang, S. & Loman, N. J. Ultra-deep, long-read nanopore sequencing of mock microbial community standards. *Gigascience* **8**, 1–9 (2019).
66. Schmid, M. *et al*. Pushing the limits of *de novo* genome assembly for complex prokaryotic genomes harboring very long, near identical repeats. *Nucleic Acids Res.* **46**, 8953–8965 (2018).
67. Bishara, A. *et al*. High-quality genome sequences of uncultured microbes by assembly of read clouds. *Nat. Biotechnol*, https://doi.org/10.1038/nbt.4266 (2018)
68. Ring, N. *et al*. Resolving the complex Bordetella pertussis genome using barcoded nanopore sequencing. *Microb. genomics* **4** (2018).
69. De Maio, N. *et al*. Comparison of long-read sequencing technologies in the hybrid assembly of complex bacterial genomes. *Microb. Genomics* **5**, e000294 (2019).
70. Risse, J. *et al*. A single chromosome assembly of Bacteroides fragilis strain BE1 from Illumina and MinION nanopore sequencing data. *Gigascience* **4**, 60 (2015).
71. Koren, S. *et al*. Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat. Biotechnol.* **30**, 693–700 (2012).

72. Witney, A. A. *et al*. Genome sequencing and characterization of an extensively drug-resistant sequence type 111 serotype O12 hospital outbreak strain of Pseudomonas aeruginosa. *Clin. Microbiol. Infect.* **20**, O609–O618 (2014).
73. Spinler, J. K., Raza, S., Runge, J. K. & Luna, R. A. Complete Genome Sequence of the Multidrug-Resistant Pseudomonas aeruginosa Endemic Houston-1 Strain, Isolated from a Pediatric Patient with Cystic Fibrosis and Assembled Using Oxford Nanopore and Illumina Sequencing. *Microbiol. Resour. Announc.* **8** (2019).
74. Magalhães, B., Senn, L. & Blanc, D. S. High-Quality Complete Genome Sequences of Three *Pseudomonas aeruginosa* Isolates Retrieved from Patients Hospitalized in Intensive Care Units. *Microbiol. Resour. Announc.* **8** (2019).
75. Turton, J. F. *et al*. High-resolution analysis by whole-genome sequencing of an international lineage (Sequence Type 111) of pseudomonas aeruginosa associated with metallo-carbapenemases in the United Kingdom. *J. Clin. Microbiol.* **53**, 2622–2631 (2015).
76. Olson, M. V. *et al*. Complete genome sequence of Pseudomonas aeruginosa PAO1, an opportunisticpathogen. *Nature* **406**, 959–964 (2000).
77. Van der Bij, A. K. *et al*. Metallo-β-lactamase-producing Pseudomonas aeruginosa in the Netherlands: the nationwide emergence of a single sequence type. *Clin. Microbiol. Infect.* **18**, E369–E372 (2012).
78. Bleves, S. *et al*. Protein secretion systems in Pseudomonas aeruginosa: A wealth of pathogenic weapons. *Int. J. Med. Microbiol.* **300**, 534–543 (2010).
79. Pawluk, A., Bondy-Denomy, J., Cheung, V. H. W., Maxwell, K. L. & Davidson, A. R. A new group of phage anti-CRISPR genes inhibits the type I-E CRISPR-Cas system of Pseudomonas aeruginosa. *MBio* **5**, e00896 (2014).
80. Giakkoupi, P. *et al*. Spread of Integron-Associated VIM-Type Metallo-β-Lactamase Genes among Imipenem-Nonsusceptible Pseudomonas aeruginosa Strains in Greek Hospitals. *J. Clin. Microbiol.* **41**, 822 (2003).
81. Hocquet, D. *et al*. Nationwide investigation of extended-spectrum beta-lactamases, metallo-beta-lactamases, and extended-spectrum oxacillinases produced by ceftazidime-resistant Pseudomonas aeruginosa strains in France. *Antimicrob. Agents Chemother.* **54**, 3512–5 (2010).
82. Poirel, L. *et al*. Characterization of VIM-2, a carbapenem-hydrolyzing metallo-beta-lactamase and its plasmid- and integron-borne gene from a Pseudomonas aeruginosa clinical isolate in France. *Antimicrob. Agents Chemother.* **44**, 891–7 (2000).
83. Poirel, L. *et al*. Characterization of Class 1 Integrons from Pseudomonas aeruginosa That Contain the blaVIM-2 Carbapenem-Hydrolyzing -Lactamase Gene and of Two Novel Aminoglycoside Resistance Gene Cassettes. *Antimicrob. Agents Chemother.* **45**, 546–552 (2001).
84. Borgianni, L. *et al*. Genetic Context and Biochemical Characterization of the IMP-18 Metallo-β-Lactamase Identified in a *Pseudomonas aeruginosa* Isolate from the United States. *Antimicrob. Agents Chemother.* **55**, 140–145 (2011).
85. Garza-Ramos, U. *et al*. Metallo-β-lactamase IMP-18 is located in a class 1 integron (In96) in a clinical isolate of Pseudomonas aeruginosa from Mexico. *Int. J. Antimicrob. Agents* **31**, 78–80 (2008).
86. Martínez, T., Vazquez, G. J., Aquino, E. E., Goering, R. V. & Robledo, I. E. Two novel class I integron arrays containing IMP-18 metallo-β-lactamase gene in Pseudomonas aeruginosa clinical isolates from Puerto Rico. *Antimicrob. Agents Chemother.* **56**, 2119–21 (2012).
87. Dößelmann, B. *et al*. Rapid and Consistent Evolution of Colistin Resistance in Extensively Drug-Resistant Pseudomonas aeruginosa during Morbidostat Culture. *Antimicrob. Agents Chemother.* **61**, e00043–17 (2017).

Acknowledgements

We thank to all members of both Centro de Investigación en Enfermedades Tropicales (Universidad de Costa Rica Costa Rica) and PGx group of The Human Phenome Institute (Fudan University, Shanghai-China) for their support in the experimental and bioinformatics analysis, respectively. This work was funded by project "B8114 Definición de la red transcriptómica y de las alteraciones genómicas inducidas por ciprofloxacina en *Pseudomonas aeruginosa* AG1", Vicerrectoría de Investigación, Universidad de Costa Rica (period 2017–2019).

Author contributions

J.M.M., C.R. and F.G. participated in the conception, design of the study and data selection. JMM implemented the bioinformatics analysis. J.M.M., R.C.S., C.R. and L.S. participated in the interpretation of bioinformatics results. J.M.M., C.R. and F.G. participated in the interpretation of the data in the biological context. J.M.M. drafted the manuscript and all authors were involved in its revision. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-58319-6.

Correspondence and requests for materials should be addressed to J.A.M.-M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

CHAPTER 2

Genomic context of the two integrons of ST-111 *Pseudomonas aeruginosa* AG1: a VIM-2-carrying old-acquaintance and a novel IMP-18-carrying integron

Molina-Mora, J.-A., Garcia-Batan, R., & Garcia, F. (2020). Genomic context of the two integrons of ST-111 *Pseudomonas aeruginosa* AG1: A VIM-2-carrying old-acquaintance and a novel IMP-18-carrying integron. Research Square (Pre-Print). https://doi.org/10.21203/RS.3.RS-41474/V1

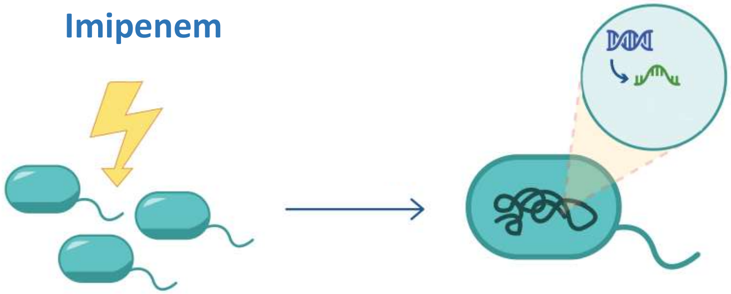
Summary

P. aeruginosa AG1 is a high-risk ST-111 strain with resistance to multiple antibiotics, including carbapenems by the activity of VIM-2 and IMP-18 metallo-β-lactamases. These genes are harbored in two class 1 integrons, belonging to genomic islands. However, the genomic context related to these determinants in PaeAG1 is unclear. Thus, we implemented a comparative genomic approach to define and up-date the phylogenetic relationship among complete *P. aeruginosa* genomes and genotyping profiles using a pan-genome analysis. We also studied the PaeAG1 genomic islands content in other strains and the architecture of genomic regions around the integrons.

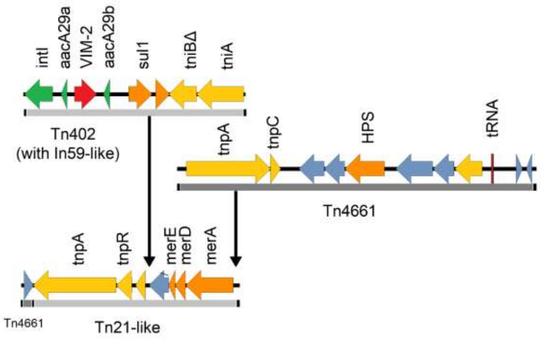
With 211 strains, the pan-genome analysis revealed that complete genome sequences are able to separate clones by MLST, including a ST-111 cluster with PaeAG1. The PaeAG1 genomic islands were found to define a diverse presence/absence pattern among related genomes, but content was related to phylogenetic relationships. Finally, landscape reconstruction of specific genomic regions showed that VIM-2-carrying integron (In59-like) is an old-acquaintance element harbored in a known genomic region completely found in other two ST-111 strains. In addition, PaeAG1 has an exclusive genomic region containing a novel IMP-18-carrying integron (registered as In1666), with an arrangement never reported before. Altogether, we provide new insights about the genomic determinants associated with the resistance to carbapenems in this high-risk *P. aeruginosa* using comparative genomics.

Graphical Abstracts

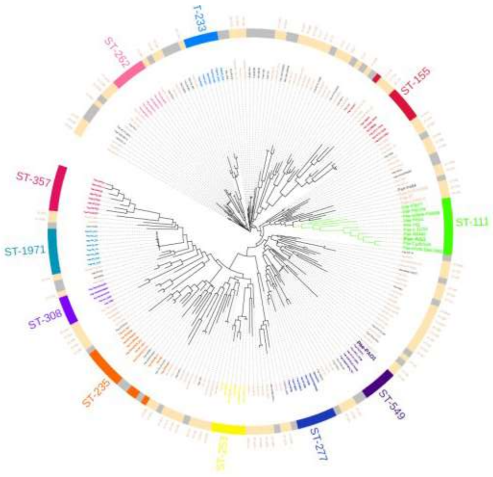
Genomic context of the two integrons of ST-111 *Pseudomonas aeruginosa* AG1: a VIM-2-carrying old-acquaintance and a novel IMP-18-carrying integron



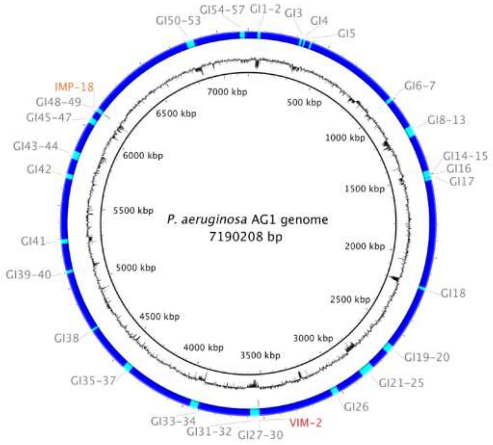
VIM-2 and IMP-18 expression
after imipenem exposure (RT-qPCR)



Reconstruction of genomic regions associated with
VIM-2- and IMP-18-carrying integrons



Pan-genome analysis:
selection of related genomes



PaeAG1 genomic islands in
other realted strains

Genomic context of the two integrons of ST-111 *Pseudomonas aeruginosa* AG1: a VIM-2-carrying old-acquaintance and a novel IMP-18-carrying integron

Highlights

- *Pseudomonas aeruginosa* AG1 (PaeAG1) carries VIM-2 and IMP-18 genes, which are induced by carbapenems
- Pan-genome analysis is able to separate strains by MLST profile
- Few PaeAG1 genomic islands were found in other related genomes
- The VIM-2-carrying integron (In59-like) is an old-acquaintance element
- A novel IMP-18-carrying integron (registered as In1666) was described for the first time

Genomic context of the two integrons of ST-111 *Pseudomonas aeruginosa* AG1: a VIM-2-carrying old-acquaintance and a novel IMP-18-carrying integron

Authors:

Jose Arturo Molina Mora, M.Sc.*

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

Email: jose.molinamora@ucr.ac.cr

*** Corresponding author**

Diana Chinchilla-Montero, M.Sc.

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

Email: dchinchilla@inciensa.sa.cr

Raquel García Batán, M.D.

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

Email: raquel.garcia@ucr.ac.cr

Fernando García, Ph.D.

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

Email: fernando.garcia@ucr.ac.cr

Abstract

Pseudomonas aeruginosa is an opportunist and versatile organism responsible for infections among immunocompromised hosts. This pathogen has high intrinsic resistance to most antimicrobials. *P. aeruginosa* AG1 (PaeAG1) is a Costa Rican high-risk ST-111 strain with resistance to multiple antibiotics, including carbapenems due to the activity of both VIM-2 and IMP-18 metallo-β-lactamases (MBLs). These genes are harbored in two class 1 integrons, belonging to one out of the 57 PaeAG1 genomic islands. However, the genomic context related to these determinants in PaeAG1 and other *P. aeruginosa* strains is unclear. Thus, we first assessed the transcriptional activity of VIM-2 and IMP-18 genes when exposed to imipenem (a carbapenem) by RT-qPCR. To select related genomes to PaeAG1, we then implemented pan-genome analysis to define and up-date the phylogenetic relationship among complete *P. aeruginosa* genomes. We also studied the PaeAG1 genomic islands content in the related strains and finally we described the architecture and possible evolutionary steps of the genomic regions around the VIM-2- and IMP-18-carrying integrons.

Expression of VIM-2 and IMP-18 genes was demonstrated to be induced after imipenem exposure. In a subsequent comparative genomics analysis with 211 strains, the *P. aeruginosa* pan-genome revealed that complete genome sequences are able to separate clones by MLST profile, including a clear ST-111 cluster with PaeAG1. The PaeAG1 genomic islands were found to define a diverse presence/absence pattern among related genomes. Finally, landscape reconstruction of genomic regions showed that VIM-2-carrying integron (In59-like) is an old-acquaintance element harbored in a known region completely found in other two ST-111 strains. In addition, PaeAG1 has an exclusive genomic region containing a novel IMP-18-carrying integron (registered as In1666), with an arrangement never reported before. Altogether, we provide new insights about the genomic determinants associated with the resistance to carbapenems in this high-risk *P. aeruginosa* using comparative genomics.

Keywords: *Pseudomonas aeruginosa* AG1, ST-111, IMP-18, VIM-2, pan-genome, genomic islands

Abbreviations

FDR: False Discovery Rate

GI: Genomic island

GIC: Genomic islands cluster

IMP: Imipenemase

KEGG: Kyoto Encyclopedia of Genes and Genomes

MBLs: Metallo-β-lactamases

MLST: Multilocus sequence typing

PaeAG1: *Pseudomonas aeruginosa* AG1

ST: Sequence type

VIM: Verona integron-encoded MBLs

WHO: World Health Organization (WHO)

1. INTRODUCTION

Pseudomonas aeruginosa is an opportunist and versatile pathogen able to survive in a wide variety of environments (Klockgether et al., 2010). With a large genome (6-7.5 Mb), *P. aeruginosa* strains have a the large proportion of the genome (>8%) dedicated to regulatory functions (Cabot et al., 2016) resulting in a consequent diversity of metabolic capabilities and responses to stress studied (J. A. Molina-Mora et al., 2020; J. Molina-Mora et al., 2020). Due to these features, *P. aeruginosa* is responsible for infections among immunocompromised hosts (Lu et al., 2016) and nosocomial infections (Fernández, Corral-Lugo, & Krell, 2018). This pathogen has high intrinsic resistance to most antimicrobials used in therapeutic practice (Brazas, Brazas, Hancock, & Hancock, 2005), many of them by multidrug-resistant or extensively drug-resistant strains (Oliver, Mulet, López-Causapé, & Juan, 2015). This severely compromises the selection of appropriate treatments (X. Mulet et al., 2013) causing significant morbidity and mortality. According to World Health Organization (WHO) resistance to carbapenems in *P. aeruginosa*, *Acinetobacter baumannii* and Enterobacteriaceae family is considered a critical issue in the context of antibiotic resistance, being classified as Priority 1 group (World Health Organization, 2017).

P. aeruginosa AG1 (PaeAG1) is a particular *P. aeruginosa* strain isolated from an immunocompromised patient in a Costa Rican hospital in 2010 (Toval et al., 2015). This strain has resistance to multiple antibiotics such as β-lactams (including carbapenems), aminoglycosides, and fluoroquinolones, being only sensible to colistin. This strain was the first report of a *P. aeruginosa* isolate carrying both VIM-2 and IMP-18 genes encoding for metallo-β-lactamases (MBLs) enzymes, both with carbapenemase activity (Toval et al., 2015). As shown in our previous works, including the genome assembly (GenBank CP045739) (J.-A. Molina-Mora, Campos-Sánchez, Rodríguez, Shi, & García, 2020), these genes belong to two independent class 1 integrons, each contained in one out the 57 predicted genomic islands of PaeAG1 (J.-A. Molina-Mora et al., 2020; Toval et al., 2015). Other elements such as six phages, mobile genetic elements and some virulence factors are also harbored in genomic islands. Ciprofloxacin exposure in PaeAG1 induces phage activity with a very complex activity, affecting the growth despite the strain is sensitive to this antibiotic (J. A. Molina-Mora et al., 2020). In addition, PaeAG1 has a not functional CRISPR-Cas system and molecular genotyping by multilocus sequence type (MLST) classifies PaeAG1 as a high-risk sequence type 111 (ST-111) strain.

ST-111 is a lineage that belongs to the O12 serotype, including a multi-resistance profile and the ability to colonize nosocomial environments (X. Mulet et al., 2013; Turton et al., 2015; Witney et al., 2014; Woodford, Turton, & Livermore, 2011). Jointly with ST-235 and ST-175 genotypes, ST-111 belong to the high-risk group in *P. aeruginosa* (Oliver et al., 2015). High-risk clones are frequently associated with epidemics where multidrug resistance confounds treatment (Petitjean et al., 2017).

In this context, it is considered that *P. aeruginosa* high-risk clones are part of a non-clonal epidemic population structure (Oliver et al., 2015; Petitjean et al., 2017), many carrying genomic determinants such as carbapenemases or extended-spectrum β-lactamases (Oliver et al., 2015). Carbapenemases include Ambler class A enzymes such as KPC and GES variants, Ambler class B MBLs (IMP, VIM, SPM, GIM, NDM and FIM type), and Amber class D (OXA variants) enzymes (Farajzadeh Sheikh et al., 2019; Hong et al., 2015). In Costa Rica, isolation of carbapenem resistant *P. aeruginosa* strains is relatively common in some major hospitals as we reported, most of them carrying VIM or IMP alleles and up to 63.1% prevalence (Toval et al., 2015). This is much higher than the frequencies observed in other countries (Hong et al., 2015).

VIM and IMP genes, as well as other MBLs, are frequently found as part of gene cassettes carried by integrons (Walsh, 2005; Zhao & Hu, 2011), leading to the dissemination of multidrug resistance among Gram negative bacteria (Jones-Dias et al., 2016). Thus, there is a growing interest in the reconstruction of the genomic context of mobile elements (in particular for integrons) to gain insights into bacterial evolution and its association with human activities, as well as to identify possible ways to mitigate antibiotic resistance (Ghaly, Chow, Asher, Waldron, & Gillings, 2017).

However, the genomic context of *P. aeruginosa* high-risk clones associated with integrons has been studied in some few studies (Chowdhury et al., 2016).

In this sense, comparative genomic strategies can provide insights not only about gene content, architecture and evolutionary details, but also dynamics of mobile genetic elements, pathogenicity determinants, and others (Peter et al., 2019). Several studies at genomic level have been implemented to describe the molecular diversity in *P. aeruginosa* (including high-risk clones) using different comparative approaches (Xavier Mulet et al., 2013; Petitjean et al., 2017; Turton et al., 2015).

Since PaeAG1 has special genomic features regarding antibiotic multi-resistance, including VIM-2 and IMP-18 genes with carbapenemase activity, 57 genomic islands and a ST-111 profile, we hypothesized that the comparative genomics can reveal insights about the evolution and landscape of genomic regions around the MBLs-carrying integrons of PaeAG1. Thus, the aim of the study was to compare PaeAG1 genome against other *P. aeruginosa* sequences using comparative genomics to describe phylogenetic relationships, genomic islands content and architecture of genomic regions associated with the VIM-2- and IMP-18-carrying integrons of PaeAG1. We first demonstrated that VIM-2 and IMP-18 are functional genes that can be induced after treatment with imipenem (a carbapenem antibiotic). We then analyzed all the complete *P. aeruginosa* genomes using a pan-genome analysis approach to identify related genomes to PaeAG1, revealing that whole genome sequences are able to separate clones by MLST profile (ST). Afterward, PaeAG1 genomic islands were searched in the related genomes, including all the ST-111 genomes, and diverse presence/absence patterns were found in related genomes. Finally, specific genomic regions associated with the two integrons were reconstructed and characterized to compare the gene content and architecture in close genomes. Genomic region associated with the VIM-2-carrying integron (In59-like) was completely found in other two ST-111 strains (i.e. it is an old-acquaintance integron), but an IMP-18-carrying integron (registered as In1666), with an architecture never reported before, was found when the landscape of the related genomic region was described.

2. MATERIALS AND METHODS

2.1 Bacterial isolate

The PaeAG1 strain is a Costa Rican isolate with resistance to β-lactams (including carbapenems, $\mathrm{MIC}_{\mathrm{Imipenem}}$ >32 µg/mL), aminoglycosides, and fluoroquinolones, being only sensitive to colistin. We recently assembled and annotated the PaeAG1 genome (J.-A. Molina-Mora et al., 2020) and data is available in Genbank under accession CP045739 (Bioproject PRJNA587210).

2.2 RT-qPCR for VIM-2 and IMP-18 expression after imipenem exposure

In order to study the expression of VIM-2 and IMP-18 genes by imipenem exposure in PaeAG1, experiments of growth curves and RT-qPCR were performed.

Growth curves assay: Three aliquots of pre-cultured PaeAG1 cells were added to fresh Lysogenic Broth (LB) broth to an initial optical density ($\mathrm{OD}_{600\mathrm{nm}}$) of 0.01. Each aliquot was treated with 0.0 (control), 25.0 or 50.0 µg/mL of imipenem. Growth was monitored at times 0, 2, 4, 6, 8, 12 and 16 hours. The assay was performed in triplicates. Two specific aliquots at times 6 and 12 hours were taken for RT-qPCR assay, as follows.

*RNA isolation***:** Aliquots at times 6 and 12 hours after imipenem exposure were preserved using the RNA protect reagent (QIAGEN). Total RNA was extracted using the RNeasy Mini kit (QIAGEN, UK) following the manufacturer´s instructions. Subsequently, RNA was transcribed into cDNA with the Maxima H Minus First Strand cDNA Synthesis kit (Thermo Scientific™ Inc.). In the different steps, quality and quantity of extracted RNA or cDNA were determined using a Nanodrop (Nanodrop 2000, Thermo Scientific™ Inc.).

Primers sequences: Primers sequences for target VIM-2 and IMP-18 genes and the reference gene *rpoD* were found from literature (Kim, Kim, & Choi, 2003; Mendes et al., 2007; Savli et al., 2003) . See Table 1 for details. Primers were manufactured by Thermo Scientific™ Inc.

RT-qPCR: The standard curve method was implemented to quantify expression of target and reference genes. Each reaction mixture contained 12.5 µL of SYBR green Master mix (Thermo Scientific™ Inc.), 0.25 µL of each primer, 10 µL of PCR-grade water, and 2 µL of cDNA. Thermocycling was performed on the StepOnePlus Real-Time PCR System (Thermo Scientific™ Inc.). For VIM-2 and IMP-18 genes, assay was run with a denaturation at 95°C (10 min), 35 amplification cycles of 94°C (20 s), 53°C (45 s), and 72°C (30 s), with data acquisition at 72°C. For *rpoD* gene, conditions were denaturation at 95°C (10 min), 45 amplification cycles of 95°C (15 s), 20°C (10 s), and 72°C (15 s), with data acquisition at 72°C. Melt curve data were used to determine whether only the correct product had been amplified.

Relative gene expression analysis: Gene expression of VIM-2 and IMP-18 in the experimental conditions (0, 25 and 50 µg/mL imipenem) were normalized using the *rpoD* housekeeping gene. The data was analyzed using the delta-delta Ct method (12). The change in gene expression within samples (time and antibiotic concentration) was calculated with reference to the control (0 µg/mL imipenem) and a two-way ANOVA test was performed between conditions (95% confidence level).

2.3 Datasets of complete *P. aeruginosa* genome sequences

In order to compare all the complete genomic sequences of *P. aeruginosa* by a pan-genome analysis, metadata (including strain names, alternative ID, gene content, MLST profile, and others), genome and protein sequences (".fasta" format), and annotation (genbank ".gbk" and ".tab" formats) files were retrieved from Pseudomonas Genomes Database (PGDB, https://pseudomonas.com).

2.4 Comparative genomic analysis by a pan-genome approach

Since differences in annotation were identified for many sequences, even in exactly the same genomic regions, we decided to identify and annotate genes from the complete genomic sequences using the same approach. To achieve this, gene prediction and annotation was done using Prokka v1.13.3 (with --genus Pseudomonas --species aeruginosa and other parameters by default configuration) (Seemann, 2014). The Prokka annotation files (in ".gbk" format) were used to run the phylogenetic analysis by a pan-genome approach based on gene content in the Roary program v3.12.0 (Page et al., 2015) with default parameters. The phylogenetic tree (".newick" file) was visualized using Interactive Tree Of Life Tool (iTOL, https://itol.embl.de/) v5 (Letunic & Bork, 2019), and strain names and MLST profiles were incorporated for each strain. For strains with unknown MLST, the profile was verified using the complete genome sequence approach (Larsen et al., 2012) in the MLST tool v2.0 (https://cge.cbs.dtu.dk/services/MLST/). For a functional analysis for all core-genes, STRINGdb (https://string-db.org/) was used to identify significantly enriched KEGG pathways (cutoff of false discovery rate FDR < 0.05).

2.5 Comparative analysis of the presence of PaeAG1 genomic islands in other strains

The 57 PaeAG1 genomic islands were previously identified using IslandViewer v4 (www.pathogenomics.sfu.ca/islandviewer/), as we reported recently (J.-A. Molina-Mora et al., 2020). Genomic islands regions (".bed" file) were downloaded from the same platform and sequences (".fasta" format) were obtained using the *getfasta* function in bedtools software v2.29.2 (Quinlan & Hall, 2010). Distribution of genomic islands along the genome was visualized using the BLAST Ring Image Generator BRIG tool v0.95 (Alikhan, Petty, Ben Zakour, & Beatson, 2011).

In order to determinate the presence and frequency of these genomic islands in other strains, a comparative analysis based on sequence alignment was done. Thus, we implemented a BLASTn pipeline to align PaeAG1 genomic island sequences and the complete genome sequences of all strains. A minimum length for coverage of 95% (overlap between query and subject sequences) and 80% of minimum sequence identity between sequences were used to define that a specific genomic island was present in a strain, otherwise, it was considered absent. Final comparison of presence/absence of genomic islands was done for selected strains (see Results) using a small phylogenetic tree and a heatmap, which were visualized using *phylo.heatmap* function from *phytools* package v0.7-20 (https://www.rdocumentation.org/packages/phytools), in the R software (https://www.r-project.org/).

2.6 Landscape of genomic regions associated with the two class 1 integrons of PaeAG1

Two complete and independent class 1 integrons were previously identified in PaeAG1, one carrying the VIM-2 gene and another harboring the IMP-18 gene (J.-A. Molina-Mora et al., 2020). To better understand the possible evolutionary history of these integrons and its potential for lateral transfer, we reconstructed the genetic landscape of the genomic regions around these elements. Identity of the integrons was investigated using INTEGRALL database (http://integrall.bio.ua.pt). For the new integron (see Results), the same database was used for the registry and the integron number assignment.

Since the two integrons are absent in the reference strain Pae-PAO1, an alignment of the genomic regions (BLASTn) and another of amino acid (AA) sequences (BLASTp) were used to identify the limits of the complete inserted region in PaeAG1. The two specific inserted regions were composed of two or more genomic islands in a row, as obtained in our previous study (grouped or with overlapping regions) (J.-A. Molina-Mora et al., 2020). Thus, regions were called $GIC_{VIM\text{-}2}$ (genomic island cluster containing VIM-2-carrying integron) and $GIC_{IMP\text{-}18}$ (genomic island cluster harboring the IMP-18-carrying integron).

Once the insertions were delimited in PaeAG1 and the insertion point in the reference genome was identified, we expanded the loci up to cover three coding genes on each side. A final alignment (BLASTn) of the expanded regions of $GIC_{VIM\text{-}2}$ and $GIC_{IMP\text{-}18}$ was done against selected genomes. Genomes selection was done based on the phylogenetic relationships of strains close to PaeAG1 (pan-genome analysis) and the profile of presence/absence of the PaeAG1 genomic islands in other strains. All the syntenic regions of selected strains were compared using annotation files in Easyfig software v2.2.3 (Sullivan, Petty, & Beatson, 2011), leading to visualize alignments, gene content and identity, exclusive/shared elements by strain and possible evolutionary steps, and others.

3. RESULTS

3.1 Expression of VIM-2 and IMP-18 genes is induced after imipenem treatment in PaeAG1

In order to assess the functional activity of VIM-2 and IMP-18 genes, a RT-qPCR was performed. Exposition to imipenem had no effects on the growth curves of PaeAG1 (Fig. 1-A). Evaluation of gene expression after exposition to imipenem (Fig. 1-B-C) showed that VIM-2 and IMP-18 increased its expression at least by a 1.7-fold (respect to control) at 6 hours, but only 1.1-fold at 12 hours. This observation was independent of the imipenem concentration (25 or 50 μg/mL), as supported by the statistical analysis in which changes in the relative expression by time but not by concentration were significant for each gene.

3.3 Pan-genome analysis with the complete genome sequences defines P. aeruginosa *clusters which correlates with the MLST genotyping profile*

To select related genomes to PaeAG1, a total of 211 strains were selected to compare the genomic composition (including PaeAG1). Supplementary file 1 *All_strains_information.xlsx* contains the list of all the selected genomes, ID, strain, MLST profile, and others. Gene content comparison was done based on a pan-genome approach. A total of 2726 genes were identified as part of the core-genome (present > 99% strains). More details of results and complementary plots are provided in the Supplementary file 2 *Pan-genome analysis results.xlsx*.

Enrichment analysis of KEGG pathways for all core genes (Table S1) found 42 biological processes implicated in several metabolism routes related to energy (carbon, fatty acids, amino acids), DNA and RNA, ribosomal activity, protein synthesis, and others.

As shown in Fig. 2, similarity in the genomic composition by pan-genome analysis defines a phylogenetic tree able to separate groups that can be described in turn by the MLST genotyping profile. Although we identified a total of 67 different MLST profiles (and unknown cases), many of them resulted with low frequency. For example, 35 different ST classes had only a single strain (35 strains, 17% of all genomes) and 88 strains (42%) belonged to the 56 ST profiles with less than five genomes. In addition, 44 strains (21%) had an allelic composition with an unknown ST profile. On the other hand, a total of 79 (37%) genomes corresponded to 11 ST classes with five or more strains. The last were evidenced using different colors by ST profile (as showed in the Fig. 2), meanwhile strains belonging to low frequency ST profiles were colored in the same way. Representative genomes such as the reference strain Pae-PAO1 (ST-549, purple cluster) and Pae-UCBPP-PA14 (ST-253, yellow group) were identified in the main ST groups.

Regarding PaeAG1, this strain was located in the same group with the other nine ST-111 strains in a clearly separated cluster (green). Other two ST profiles (low frequency ST-234 and ST-654) and one unknown case (Pae-Pa84 strain) kept close to this group. The whole group of these related strains, and the reference strain Pae-PAO1, were used for subsequent analysis, including their phylogenetic relationships. For other high-risk clones, a single ST-175 genome was identified, and a clear cluster was found for the ten ST-235 genomes (including other genomes with unknown profile).

3.3 Varying profiles of the presence/absence of the 57 PaeAG1 genomic islands are found in the ST-111 strains and related genomes

A comparative analysis based on sequence alignment was run in order to determinate the presence and frequency of the PaeAG1 genomic islands in other phylogenetically related strains. Genomic islands locus were previously predicted (J.-A. Molina-Mora et al., 2020). We first represented the distribution of the genomic islands along the PaeAG1 genome, as presented in Fig. 3. Many of the islands kept together, including overlapping regions or an arrangement in a row. Thus, we termed this as a genomic islands cluster (GIC) to refer to this group of islands. In Fig. 3, GICs correspond to the genomic regions labeled as joined names of the genomic islands, for example “GI48-49” represents the genomic region of islands GI48 and GI49. In some cases each genomic island in the cluster can be differentially distributed in the genomes (for example GI48 is present in PaeAG1 and Pae-97, but GI49 is only found in PaeAG1, Fig. 4). For this reason, we do not re-define the locus neither joined the islands.

Analysis of the presence/absence of PaeAG1 genomic islands in other ST-111 strains and related genomes is shown in Fig. 4. Profiles for all the 211 is available in the Supplementary file 1 *All_strains_information.xlsx,* including total counts of strains by genomic islands, and total genomic islands per genome. The closest genomes to PaeAG1 (Pae-RIVM-EMC2982 and Pae-Carb0163) had the most similar profiles in the genomic islands content (carrying 41 genomic islands), but different patterns are obtained for other ST-111 strains. None of the islands is present in the reference genome Pae-PAO1, and other few genomic islands are rarely present in other non ST-111 strains.

On the other hand, two particular genomic islands were particularly recognized due to they carry the two PaeAG1 integrons. GI27 genomic island harbors the VIM-2-carrying integron, while IMP-18-carrying integron belongs to GI49. As shown in Fig. 4, GI27 (red) is present in PaeAG1 and two other ST-111 strains, and it is also absent in the rest of the 208 genomes. GI49 (blue) is unique to PaeAG1 and it is not it is present in none of the other 210 strains in the study.

Additionally, both genomic island are associated with a GIC, GI27-30 and GI48-49 (Fig. 3) respectively. Since the importance of these genomic regions to study the integrons, we specifically called them $\mathrm{GIC}_{\mathrm{VIM-2}}$ (genomic island cluster containing VIM-2-carrying integron) and $\mathrm{GIC}_{\mathrm{IMP-18}}$ (genomic island cluster harboring the IMP-18-carrying integron).

Based on phylogenetic relationships, ST profile and genomic islands content, we selected specific genomes to compare the GICs associated with the integrons. As shown in Fig. 4, the four genomic islands of $\mathrm{GIC}_{\mathrm{VIM-2}}$ (GI27-30) are differentially present in the genomes. For example, GI28 and G29 are present in eight strains, but GI27 in three and G30 in four. To specifically compare the genomic regions of $\mathrm{GIC}_{\mathrm{VIM-2}}$, we used the reference Pae-PAO1, Pae-RIVM-EMC2982 (with the four genomic islands), and Pae-AR445 (with three of the genomic islands). For the case of GIC$_{IMP-18}$, the two islands GI48 and GI49 are absent in other ST-111 strains, but GI48 is present in Pae-97. Except for this case, no other strains in all 211 genomes were identified harboring both islands. To compare the genomic regions, the reference genome Pae-PAO1, Pae-RIVM-EMC2982 as a closest genome, and Pae-97 (the only genome sharing a section of the GIC) were used.

3.4 GIC$_{VIM-2}$ is a known region containing the old-acquaintance VIM-2-carrying integron in PaeAG1

With the aim of describing the possible evolutionary history of the VIM-2-carrying integron in PaeAG1, we described the architecture of the genomic regions delimited by the GIC$_{VIM-2}$ (including three extreme genes on each side: 35 798 bp and 32 protein-coding genes). Using Pae-PAO1 as reference, we found that genomic insertion occurred in the middle of the PA2229 gene, as shown in the top of Fig. 5. The insertion resulted mostly present in Pae-AR445 (coverage 94% and identity 99.97% of the PaeAG1 region), but without most of the integron (integrase *intI1* and *sul1* are present, unlike the gene cassette including VIM-2). However, a full coverage region was identified in Pae-RIV-EMC2982, with a 100% coverage and identity 99.99%. The only two variants identified in the full region were non-synonymous mutations, with an amino-acid change in PaeAG1_03254 (transcriptional regulator *merD*, 99.0% identity) and PaeAG1_03255 (mercuric reductase merA, 99.8% identity). See Table 2 and supplementary Table S2 for more details. Although not shown in Fig. 5, alignment was also done for Pae-Carb0163, which has the same profile of genomic islands content as Pae-RIV-EMC2982. In this case, a 100% coverage and identity 99.87% (45 variants) were obtained in the GIC$_{VIM-2}$ region; most of the variants resulted in a change in the amino-acid sequence in PaeAG1_03245 (aacA29a, part of the integron with a 95.8% identity resulting in aacA29e allele), but also affecting other three proteins (mercuric reductase, integrase *IntI* and a transposase). See supplementary Table S2 for more details.

Regarding the gene content (Table 2), this genomic insertion contains the complete integron carrying VIM-2 gene. Composition of this integron is described in Fig. 5 (bottom), containing classical elements *int1*, *attI, sul1* and the gene cassette (with aacA29a-b and VIM-2) of a class 1 integron, being classified as In59-like. Furthermore, GIC_{VIM-2} has at least other mobile genetic elements, including transposases and recombinases modules. Other coding modules are associated with mercury metabolism or they remain unknown (hypothetical proteins). Details of the protein alignment of PaeAG1 against four genomes is also provided (supplementary Table S2). Reconstruction of the evolutionary steps related to the conformation of this genomic region include participation of four transposons (Tn402, Tn*21*-like, a disrupted and another complete Tn*4661*) as shown in Fig. 7-A. See details in Discussion.

Considering the full coverage and very high identity in at least two genomes, Pae-RIVM-EMC2982 and Pae-Carb0163, GIC_{VIM-2} can be considered a genomic region present in two well-known VIM-2+ strains, being this gene located in an old-acquaintance class 1 integron (In59-like).

3.5 GIC_{IMP-18} is a PaeAG1 exclusive genomic region harboring a new IMP-18-carrying integron

In a similar way as before, we compared four genomes to described the architecture of the genomic regions delimited by the GIC_{IMP-18} (including three extreme genes on each side: 30 258 bp and 29 protein-coding genes). Using Pae-PAO1 as reference, we found that genomic insertion occurred between the genes PA4704 and PA4705, as shown in the top of Fig. 6. Genomic islands GI48-49 are absent in Pae-RIV-EMC2982 and Pae-Carb0163 genomes (the last not shown in the Fig.).

BLAST of GIC_{IMP-18} identified the highest scored sequence in Pae-97 genome (ST-234). Thus, since Pae-97 carries GI48, syntenic comparison was done using this genome (Fig. 6). Analysis revealed a 77% coverage with identity 99.92%. The Pae-97 integron also contains *Int1*, aacA genes and another allele of the IMP gene (IMP-1), all with a different arrangement.

Regarding gene content (Table 3), this genomic insertion contains the complete integron carrying IMP-18 gene. Composition of this integron is described in Fig. 6 (bottom), containing *int1, attI, sul1* and the gene cassette (IMP-18, gcuD, OXA-2 and aacA4). GIC_{IMP-18} also has genes coding for endonucleases and recombinases, or hypothetical proteins. Details of the protein alignment of PaeAG1 against the four genomes are also provided (see supplementary Table S3).

Considering the absence of the complete region in other genomes and the first report of the architecture of this integron, GIC_{IMP-18} can be considered a PaeAG1 exclusive region harboring a new IMP-18-carrying integron. This integron was registered as In1666 in INTEGRALL database.

Conformation of GIC_{IMP-18} region seems to include the participation of at least three mobile elements (the new integron In1666, insertion sequence IS1326 and transposon TnAs3) as shown in Fig. 7-B. However, a lack of information about the role of other elements (regions without matching sequences) makes difficult to complete the possible evolutionary steps related to this genomic region.

In summary, the pan-genome analysis lead us to identify that the genomic content can separate groups according to the ST profile (MLST genotyping). All the ST-111 strains, including PaeAG1, resulted in the same phylogenetic group but different presence/absence profiles of PaeAG1 genomic islands were identified in other strains, even for grouped genomic islands, the GICs. Analysis of the landscape of regions GIC_{VIM-2} and GIC_{IMP-18} revealed one known and another new arrangement of genomic sequences in PaeAG1, harboring two independent MBLs-carrying integrons. The IMP-18-carrying integron has a unique and exclusive composition, reported here for the first time.

4. DISCUSSION

Antibiotic multi-resistance is a major threat to public health because continuous emergence, worldwide spread, and increasing prevalence (Hong et al., 2015). With a high-risk ST-111 profile, PaeAG1 is a critical organism due to its resistance to multiple antibiotics but in particular the resistance to carbapenems (World Health Organization, 2017). In our study, we first demonstrated that expression of VIM-2 and IMP-18 genes (with carbapenemase activity) are induced after imipenem exposure, evidencing that are functional genes. To describe the genomic context associated with theses MBLs, we performed a pan-genome analysis, a comparison of genomic islands between representative strains and the reconstruction of the surrounding genomic regions.

In the pan-genome analysis, we were able not only to reveal that whole genome sequences could separate clones by ST profile (MLST), but also identification of core and accessory genes was achieved. Other pan-genome analysis in *P. aeruginosa* also found clusters than could be identified by the ST profile (Aguilar-Rodea et al., 2017; Weiser et al., 2019). While multiple comparative genomic analyses (many using a pan-genome approach) have been reported for *P. aeruginosa* (Aguilar-Rodea et al., 2017; Chowdhury et al., 2016; Freschi et al., 2019; Gomila, Peña, Mulet, Lalucat, & García-Valdés, 2015; Hilker et al., 2015; Mosquera-Rendón et al., 2016; Ozer, Allen, & Hauser, 2014; Poulsen et al., 2019; Valot et al., 2015; Weiser et al., 2019; Wendt & Heo, 2016), most of them include incomplete, fragmented or draft genomes, or sequences of few genes. In 2015, complete genomes were used in a similar approach, but only 17 genomes were available (NCBI), which only three corresponded to high-risk clones (Valot et al., 2015). Thus, our analysis provides an up-date of the general status of relationships of the 211 available complete genomes by pan-genome analysis.

In relation to gene content among all strains, we identified a total of 2726 genes as part of the core-genome (>99% strains), similar to another similar approach (Mosquera-Rendón et al., 2016). Other studies have suggested a higher number of core genes (4000-5300) (Hilker et al., 2015; Ozer et al., 2014; Valot et al., 2015; Weiser et al., 2019). The relatively high number of conserved genes in the core-genome can be associated with the ability to conquer multiple environments and to facilitate infectious capability towards a large set of hosts (Valot et al., 2015). According to functional analysis, 42 KEGG pathways (energy metabolism, nucleic acids, amino acids, ribosomal activity, and many others) were found as part of the enriched routes for all the core genes, with functions that are in line with other similar pan-genome studies (Mosquera-Rendón et al., 2016; Valot et al., 2015).

P. aeruginosa genome is composed of a mosaic structure including the large core-genome (Valot et al., 2015), into which regions of genomic plasticity lead to the insertion of block of genes belonging to the accessory genome (Mathee et al., 2008). In the case of PaeAG1 and other ST-111 strains, genome sequence is around 1.0 Mb longer that the reference genome Pae-PAO1, difference that is reflect as genomic islands distributed along the genome.

Pae-RIVM-EMC2982 and Pae-Carb0163 (closest genomes to PaeAG1) had the most similar profiles carrying 41 out the genomic islands. As highlighted in Results, many genomic islands formed clusters (GICs, Fig. 3 and 3), including the genomic islands clusters harboring the two integrons (GIC_{VIM-2} and GIC_{IMP-18}). Genomic islands groups have been reported before as integrative and conjugative elements or ICEs (Petitjean et al., 2017), but ICEs in PaeAG1 (using ICEberg 2.0 platform, https://db-mml.sjtu.edu.cn/ICEfinder/ICEfinder.html) overlap with other GICs but none with GIC_{VIM-2} and GIC_{IMP-18}. Since size of the core-genome and its content is not well known (Valot et al., 2015), prediction methods are required to define accessory regions, but outcome depends on algorithms (Ozer et al., 2014), which could explain differences and the GICs**.**

On the other hand, this prominent number of genomic islands in PaeAG1 and other ST-111 strains can be explained due to the absence of a functional CRISPR-Cas system (bacterial defense system against foreign DNA) and consequent high number of successful events of horizontal gene transfer (Petitjean et al., 2017). This genome plasticity of individual strains represents an advantage for *P. aeruginosa* to fit the needs for survival in virtually any environment (Mathee et al., 2008).

In the context of carbapenems resistance, genes encoding for MBLs are usually found as gene cassettes in class 1 integrons (Jones-Dias et al., 2016; Walsh, 2005). This allows a rapid dissemination in the clinical setting due to the selective pressure by the use of antibiotics (Sánchez-Martinez et al., 2010), which is aggravated due to this antibiotic represents the last therapeutic source to treat *P. aeruginosa* infections (Toval et al., 2015). While multiple studies correlate antibiotic resistance and the presence of integrons, genetic context surrounding class 1 integrons is often not investigated in *P. aeruginosa,* as remarked before (Chowdhury et al., 2016).

Carbapenem resistance in PaeAG1 was demonstrated to be explained by activity of two MBLs (VIM-2 and IMP-18) (Toval et al., 2015), each gene harbored in two independent class 1 integrons (J.-A. Molina-Mora et al., 2020; Toval et al., 2015).

Evaluation of the sequence showed that GIC_{VIM-2} is also present in Pae-RIVM-EMC2982 (100% coverage and 99.99% identity) and Pae-Carb0163 (100% coverage and 99.87% identity) at chromosomal level. However, a study including these strains showed that VIM-2-carrying integron and surrounding regions (~30 Kb, equivalent to GIC_{VIM-2}) were shared with a plasmid of ST-446 *P. aeruginosa* S04-90 with 99% identity. Based on identity, mobilization of the fragment between plasmids and chromosomes may have occurred recently (van der Zee et al., 2018).

In the same study, analysis of genome landscape showed that the regions (equivalent to GIC_{VIM-2}) corresponded to a DNA segment acting as a composite transposon, composed of four different transposons (Tn402, Tn*21*-like, a disrupted and another complete Tn*4661*). The class 1 integron carrying VIM-2 is contained in the Tn402 transposon (Gillings, 2017; van der Zee et al., 2018). Evolutionary details are completely explained in (van der Zee et al., 2018). GIC_{VIM-2} carries the genes involved in its own transposition module (transposases such as TniB and TnpA) and mercury resistance module, as described in other similar transposons and insertion sequences (Chowdhury et al., 2016; Ghaly et al., 2017; Jones-Dias et al., 2016; Liebert, Hall, & Summers, 1999; van der Zee et al., 2018). Presence of gene cassettes unrelated to the antibiotic resistance can be result of anthropogenic settings (Ghaly et al., 2017) and selection pressures in environments polluted with heavy metals and other substances such as mercury, arsenic and disinfectants (Gillings et al., 2015).

Regarding the VIM-2-carrying integron, this element is an In59-like integron. In59 was first reported two decades ago in France (Poirel et al., 2001) and then worldwide (Gillings, 2017; Samuelsen et al., 2010; Toval et al., 2015; van der Zee et al., 2018). Among all the 211 strains in our study, VIM-2 was only present into PaeAG1 and the two closest genomes (all ST-111). Differences in aacA29 genes defined the aacA29e allele found in Pae-Carb0163 (van der Zee et al., 2018), in contrast to aacA29a-b in PaeAG1, all coding for aminoglycoside acetyltransferases. Since GIC_{VIM-2} sequence and architecture is completely found in two VIM-2+/ST-111 strains, VIM-2-carrying integron (In59-like) can be considered old-acquaintance element in a well-known genomic context.

Additionally, genomic context defined by GIC_{IMP-18} was also analyzed. Using Pae-PAO1 as reference, it is shown that GIC_{IMP-18} insertion occurred in a specific point (*prrH*) between PA4704 and PA4705 (Fig. 6). This region contains three genes for regulatory small RNAs (*prrF1, prrH* and *prrF2*) are found, which are involved in iron homeostasis under iron-depleted conditions (Reinhart et al., 2017) or to avoid iron toxicity (Reinhart et al., 2015).

While complete GIC_{IMP-18} (composed of GI48-GI49 genomic islands) was not found in none of other strains, GI48 section was found in Pae-97 strain (ST-234, with a class 1 integron), a genome close to ST-111 group (Fig. 2 and 3). Sequences comparison of GIC_{IMP-18} and Pae-97 showed 77% coverage and 99.92% identity. Gene composition of GIC_{IMP-18} includes endonucleases and recombinases module, the class1 integron, transposase TniB and hypothetical proteins.

In relation to the integron harbored in GIC_{IMP-18}, the IMP-18-carrying element is composed of the *intI1*, the gene cassette (carrying IMP-18, *gcuD* and OXA-2), aacA4 and *sul1*. In another strain, similar genes with another arrangement (orderly IMP-18, a disrupted aacA43, OXA-2 and *gcuD*) were reported for the first time in the In706 integron in 2012 (Martínez, Vazquez, Aquino, Goering, & Robledo, 2012). Pae-97 contains a class 1 integron, but with a different arrangement with IMP-1 allele (without OXA nor *gcuD* genes). Other studies found multiple strains carrying both IMP-18 and OXA-2 (without *gcuD* nor aacA4) in Mexican isolates as part of In169 (Sánchez-Martinez et al., 2010) and In1215 (López-García et al., 2018) integrons, including some located in plasmids.

Since there is a lack of information about the genomic context of many IMP-carrying integrons (such as region GIC_{IMP-18}, unlike GIC_{VIM-2}), and the particular architecture of the class 1 integron in PaeAG1 with the gene cassette IMP-18/*gcuD*/OXA-2/aacA4, we consider that this IMP-18-carrying integron (registered as In1666) is a novel element that we report here for the first time.

In the partial reconstruction of the evolutionary steps related to the $\mathrm{GIC}_{\mathrm{IMP\text{-}18}}$ region, the integron In1666, the insertion sequence IS1326 and the transposon TnAs3 seem to play a key role in the current state of this genomic region. Both IS1326 and TnAs3 have been reported in different integrons and high plasticity regions (He et al., 2016; Jones-Dias et al., 2016; Liebert et al., 1999; Szuplewska, Czarnecki, & Bartosik, 2014). Further analyses are required to complete the evolutionary steps which have defined this genomic region as well as the implications of multiresistant in PaeAG1.

Jointly, identification of the landscape of the genomic context defined by $\mathrm{GIC}_{\mathrm{VIM\text{-}2}}$ and $\mathrm{GIC}_{\mathrm{IMP\text{-}18}}$, provides insights about the dissemination and evolution of mobile elements, in this particular case for integrons carrying MBLs. Since MBL-producing *P. aeruginosa* is able to produce epidemic outbreaks and responsible for the dissemination of carbapenemase resistance worldwide (Castanheira, Deshpande, Costello, Davies, & Jones, 2014), it is worrisome that strains such as PaeAG1 are able to circulate among Costa Rican hospitals. This can be correlated with the high prevalence of carbapenem resistant strains in Costa Rica, many carrying VIM or IMP genes (Toval et al., 2015). Future works are necessary to trigger the surveillance system in order to evaluate if other circulating strains carry these two elements, to identify its possible dissemination and hence carry out an adequate infection control program in medical centers.

5. CONCLUSIONS

PaeAG1 is a high-risk and a critical organism due to its resistance to carbapenems by the activity of VIM-2 and IMP-18 enzymes, both harbored in two class 1 integrons. To describe the genomic context associated with these integrons, we first verified the functionality of VIM-2 and IMP-18 after imipenem exposure. We then analyzed 211 complete genome sequences using a pan-genome analysis, separating strains by MLST profile. Analysis of the 57 PaeAG1 genomic islands showed a varying pattern of the presence/absence among all the strains, in particular for closest genomes to PaeAG1. Two selected genomic islands clusters, $GIC_{VIM\text{-}2}$ and $GIC_{IMP\text{-}18}$, were studied in-depth. $GIC_{VIM\text{-}2}$ sequence was completely found in other two known ST-111 strains, which contained the VIM-2-carrying integron as an old-acquaintance In59-like element. $GIC_{IMP\text{-}18}$ was partially found in another genome, but the IMP-18-carrying integron has an architecture never reported before, being considered as a novel In1666 integron. We provide new insights about the genomic determinants associated with this high-risk *P. aeruginosa* clone and its resistance to carbapenems using comparative genomics.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

All the strains we used in this study were obtained from NCBI. All the IDs are available in the "Supplementary_file 1 All_strains_information". For PaeAG1, the genome sequence and annotation files are available from our previous work at https://github.com/josemolina6/PaeAG1_genome, and in GenBank under the accession number CP045739.

Declaration of Competing Interest

The authors declare that there is no conflict of interest.

Acknowledgements

This work was funded by projects “B8114 Definición de la red transcriptómica y de las alteraciones genómicas inducidas por 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”, Vicerrectoría de Investigación, Universidad de Costa Rica (period 2017-2020).

We thank all members of Centro de Investigación en Enfermedades Tropicales (Universidad de Costa, Costa Rica) for their support in the activities associated with the projects.

REFERENCES

Aguilar-Rodea, P., Zúñiga, G., Rodríguez-Espino, B. A., Cervantes, A. L. O., Arroyo, A. E. G., Moreno-Espinosa, S., … Velázquez-Guadarrama, N. (2017). Identification of extensive drug resistant Pseudomonas aeruginosa strains: New clone ST1725 and high-risk clone ST233. *PLoS ONE*, *12*(3), 2007–2013. https://doi.org/10.1371/journal.pone.0172882

Alikhan, N.-F., Petty, N. K., Ben Zakour, N. L., & Beatson, S. A. (2011). BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics*, *12*(1), 402. https://doi.org/10.1186/1471-2164-12-402

Brazas, M. D., Brazas, M. D., Hancock, R. E. W., & Hancock, R. E. W. (2005). Ciprofloxacin Induction of a Susceptibility Determinant in Pseudomonas aeruginosa. *Antimicrobial Agents and Chemotherapy*, *49*(8), 3222–3227. https://doi.org/10.1128/AAC.49.8.3222

Cabot, G., Zamorano, L., Moyà, B., Juan, C., Navas, A., Blázquez, J., & Oliver, A. (2016). Evolution of Pseudomonas aeruginosa antimicrobial resistance and fitness under low and high mutation rates. *Antimicrobial Agents and Chemotherapy*, *60*(3), 1767–1778. https://doi.org/10.1128/AAC.02676-15.Address

Castanheira, M., Deshpande, L. M., Costello, A., Davies, T. A., & Jones, R. N. (2014). Epidemiology and carbapenem resistance mechanisms of carbapenem-non-susceptible Pseudomonas aeruginosa collected during 2009-11 in 14 European and Mediterranean countries. *Journal of Antimicrobial Chemotherapy*, *69*(7), 1804–1814. https://doi.org/10.1093/jac/dku048

Chowdhury, P. R., Scott, M., Worden, P., Huntington, P., Hudson, B., Karagiannis, T., … Djordjevic, S. P. (2016). Genomic islands 1 and 2 play key roles in the evolution of extensively drug-resistant ST235 isolates of Pseudomonas aeruginosa. *Open Biology*, *6*(3). https://doi.org/10.1098/rsob.150175

Farajzadeh Sheikh, A., Shahin, M., Shokoohizadeh, L., Halaji, M., Shahcheraghi, F., & Ghanbari, F. (2019). Molecular epidemiology of colistin-resistant Pseudomonas aeruginosa producing NDM-1 from hospitalized patients in Iran. *Iranian Journal of Basic Medical Sciences*, *22*(1), 38–42. https://doi.org/10.22038/ijbms.2018.29264.7096

Fernández, M., Corral-Lugo, A., & Krell, T. (2018). The plant compound rosmarinic acid induces a broad quorum sensing response in Pseudomonas aeruginosa PAO1. *Environmental Microbiology*, *20*(12), 4230–4244. https://doi.org/10.1111/1462-2920.14301

Freschi, L., Vincent, A. T., Jeukens, J., Emond-Rheault, J. G., Kukavica-Ibrulj, I., Dupont, M. J., … Levesque, R. C. (2019). The Pseudomonas aeruginosa Pan-Genome Provides New Insights on Its Population Structure, Horizontal Gene Transfer, and Pathogenicity. *Genome Biology and Evolution*, *11*(1), 109–120. https://doi.org/10.1093/gbe/evy259

Ghaly, T. M., Chow, L., Asher, A. J., Waldron, L. S., & Gillings, M. R. (2017). Evolution of class 1 integrons: Mobilization and dispersal via food-borne bacteria. *PLoS ONE*, *12*(6), 1–11. https://doi.org/10.1371/journal.pone.0179169

Gillings, M. R. (2017). Class 1 integrons as invasive species. *Current Opinion in Microbiology*, *38*, 10–15. https://doi.org/10.1016/j.mib.2017.03.002

Gillings, M. R., Gaze, W. H., Pruden, A., Smalla, K., Tiedje, J. M., & Zhu, Y.-G. (2015). Using the class 1 integron-integrase gene as a proxy for anthropogenic pollution. *The ISME Journal*, *9*(6), 1269–1279. https://doi.org/10.1038/ismej.2014.226

Gomila, M., Peña, A., Mulet, M., Lalucat, J., & García-Valdés, E. (2015). Phylogenomics and systematics in Pseudomonas. *Frontiers in Microbiology*, *6*(MAR), 1–13. https://doi.org/10.3389/fmicb.2015.00214

He, S., Chandler, M., Varani, A. M., Hickman, A. B., Dekker, J. P., & Dyda, F. (2016). Mechanisms of evolution in high-consequence drug resistance plasmids. *MBio*, *7*(6), 1987–2003. https://doi.org/10.1128/mBio.01987-16

Hilker, R., Munder, A., Klockgether, J., Losada, P. M., Chouvarine, P., Cramer, N., … Tümmler, B. (2015). Interclonal gradient of virulence in the *P seudomonas aeruginosa* pangenome from disease and environment. *Environmental Microbiology*, *17*(1), 29–46. https://doi.org/10.1111/1462-2920.12606

Hong, D. J., Bae, I. K., Jang, I. H., Jeong, S. H., Kang, H. K., & Lee, K. (2015). Epidemiology and characteristics of metallo-ß-lactamase-producing Pseudomonas aeruginosa. *Infection and Chemotherapy*, *47*(2), 81–97. https://doi.org/10.3947/ic.2015.47.2.81

Jones-Dias, D., Manageiro, V., Ferreira, E., Barreiro, P., Vieira, L., Moura, I. B., & Caniça, M. (2016). Architecture of class 1, 2, and 3 integrons from gram negative bacteria recovered among fruits and vegetables. *Frontiers in Microbiology*, *7*(SEP), 1–13. https://doi.org/10.3389/fmicb.2016.01400

Kim, S.-M., Kim, E.-C., & Choi, S.-Y. (2003). Typing by Pulsed Field Gel Electrophoresis and Detection of Metallo-β-lactamase Gene Against Acinetobacter baumannii from Clinical Specimens. *Korean J Clin Lab Sci*, *35*(2), 90–98. Retrieved from http://www.kjcls.org/journal/view.html?spage=90&volume=35&number=2

Klockgether, J., Munder, A., Neugebauer, J., Davenport, C. F., Stanke, F., Larbig, K. D., … Tümmler, B. (2010). Genome diversity of Pseudomonas aeruginosa PAO1 laboratory strains. *Journal of Bacteriology*, *192*(4), 1113–1121. https://doi.org/10.1128/JB.01515-09

Larsen, M. V, Cosentino, S., Rasmussen, S., Friis, C., Hasman, H., Marvig, R. L., … Lund, O. (2012). Multilocus sequence typing of total-genome-sequenced bacteria. *Journal of Clinical Microbiology*, *50*(4), 1355–1361. https://doi.org/10.1128/JCM.06094-11

Letunic, I., & Bork, P. (2019). Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Research*, *47*(W1), W256–W259. https://doi.org/10.1093/nar/gkz239

Liebert, C. A., Hall, R. M., & Summers, A. O. (1999). Transposon Tn21, Flagship of the Floating Genome. *Microbiology and Molecular Biology Reviews*, *63*(3), 507–522. https://doi.org/10.1128/mmbr.63.3.507-522.1999

López-García, A., Rocha-Gracia, R. del C., Bello-López, E., Juárez-Zelocualtecalt, C., Sáenz, Y., Castañeda-Lucio, M., … Lozano-Zarain, P. (2018). Characterization of antimicrobial resistance mechanisms in carbapenem-resistant Pseudomonas aeruginosa carrying IMP variants recovered from a Mexican Hospital. *Infection and Drug Resistance*, *11*, 1523. https://doi.org/10.2147/IDR.S173455

Lu, P., Wang, Y., Zhang, Y., Hu, Y., Thompson, K. M., & Chen, S. (2016). RpoS-dependent sRNA RgsA regulates Fis and AcpP in Pseudomonas aeruginosa. *Molecular Microbiology*, *102*(2), 244–259. https://doi.org/10.1111/mmi.13458

Martínez, T., Vazquez, G. J., Aquino, E. E., Goering, R. V, & Robledo, I. E. (2012). Two novel class I integron arrays containing IMP-18 metallo-β-lactamase gene in Pseudomonas aeruginosa clinical isolates from Puerto Rico. *Antimicrobial Agents and Chemotherapy*, *56*(4), 2119–2121. https://doi.org/10.1128/AAC.05758-11

Mathee, K., Narasimhan, G., Valdes, C., Qiu, X., Matewish, J. M., Koehrsen, M., … Lory, S. (2008). Dynamics of Pseudomonas aeruginosa genome evolution. *Proceedings of the National Academy of Sciences*, *105*(8), 3100–3105. https://doi.org/10.1073/PNAS.0711982105

Mendes, R. E., Kiyota, K. A., Monteiro, J., Castanheira, M., Andrade, S. S., Gales, A. C., … Tufik, S. (2007). Rapid detection and identification of metallo-β-lactamase-encoding genes by multiplex real-time PCR assay and melt curve analysis. *Journal of Clinical Microbiology*, *45*(2), 544–547. https://doi.org/10.1128/JCM.01728-06

Molina-Mora, J.-A., Campos-Sánchez, R., Rodríguez, C., Shi, L., & García, F. (2020). High quality

3C de novo assembly and annotation of a multidrug resistant ST-111 Pseudomonas aeruginosa genome: Benchmark of hybrid and non-hybrid assemblers. *Scientific Reports*, *10*(1), 1392. https://doi.org/10.1038/s41598-020-58319-6

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

Molina-Mora, J., Montero-Manso, P., Batán, R. G., Sánchez, R. C., Fernández, J. V., & García, F. (2020). A first Pseudomonas aeruginosa perturbome: Identification of core genes related to multiple perturbations by a machine learning approach. *BioRxiv*, 2020.05.05.078477. https://doi.org/10.1101/2020.05.05.078477

Mosquera-Rendón, J., Rada-Bravo, A. M., Cárdenas-Brito, S., Corredor, M., Restrepo-Pineda, E., & Benítez-Páez, A. (2016). Pangenome-wide and molecular evolution analyses of the Pseudomonas aeruginosa species. *BMC Genomics*, *17*(1), 1–14. https://doi.org/10.1186/s12864-016-2364-4

Mulet, X., Cabot, G., Ocampo-Sosa, A. A., Dominguez, M. A., Zamorano, L., Juan, C., … Spanish Network for Research in Infectious Diseases (REIPI). (2013). Biological Markers of Pseudomonas aeruginosa Epidemic High-Risk Clones. *Antimicrobial Agents and Chemotherapy*, *57*(11), 5527–5535. https://doi.org/10.1128/AAC.01481-13

Mulet, Xavier, Cabot, G., Ocampo-Sosa, A. A., Dominguez, M. A., Zamorano, L., Juan, C., … Oliver, A. (2013). Biological markers of Pseudomonas aeruginosa epidemic high-risk clones. *Antimicrobial Agents and Chemotherapy*, *57*(11), 5527–5535. https://doi.org/10.1128/AAC.01481-13

Oliver, A., Mulet, X., López-Causapé, C., & Juan, C. (2015). The increasing threat of Pseudomonas aeruginosa high-risk clones. *Drug Resistance Updates*, *21–22*, 41–59. https://doi.org/10.1016/j.drup.2015.08.002

Ozer, E. A., Allen, J. P., & Hauser, A. R. (2014). Characterization of the core and accessory genomes of Pseudomonas aeruginosa using bioinformatic tools Spine and AGEnt. *BMC Genomics*, *15*(1), 737. https://doi.org/10.1186/1471-2164-15-737

Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T. G., … Parkhill, J. (2015). Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*, *31*(22), 3691–3693. https://doi.org/10.1093/bioinformatics/btv421

Peter, S., Bosio, M., Gross, C., Bezdan, D., Gutierrez, J., Oberhettinger, P., … Ossowski, S. (2019). Tracking of antibiotic resistance transfer and rapid plasmid evolution in a hospital setting by Nanopore sequencing. *BioRxiv*, 639609. https://doi.org/10.1101/639609

Petitjean, M., Martak, D., Silvant, A., Bertrand, X., Valot, B., & Hocquet, D. (2017). Genomic characterization of a local epidemic Pseudomonas aeruginosa reveals specific features of the widespread clone ST395. *Microbial Genomics*, *3*(10), e000129. https://doi.org/10.1099/mgen.0.000129

Poirel, L., Lambert, T., Turkoglu, S., Ronco, E., Gaillard, J., & Nordmann, P. (2001). Characterization of Class 1 Integrons from Pseudomonas aeruginosa That Contain the blaVIM-2 Carbapenem-Hydrolyzing -Lactamase Gene and of Two Novel Aminoglycoside Resistance Gene Cassettes. *Antimicrobial Agents and Chemotherapy*, *45*(2), 546–552. https://doi.org/10.1128/AAC.45.2.546-552.2001

Poulsen, B. E., Yang, R., Clatworthy, A. E., White, T., Osmulski, S. J., Li, L., … Hung, D. T. (2019). Defining the core essential genome of Pseudomonas aeruginosa. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(20), 10072–10080. https://doi.org/10.1073/pnas.1900570116

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

Reinhart, A. A., Nguyen, A. T., Brewer, L. K., Bevere, J., Jones, J. W., Kane, M. A., … Oglesby-Sherrouse, A. G. (2017). The Pseudomonas aeruginosa PrrF Small Acute Murine Lung

Infection. *Infection and Immunity*, *85*(5), 1–15. https://doi.org/10.1128/IAI.00764-16

Reinhart, A. A., Powell, D. A., Nguyen, A. T., O'Neill, M., Djapgne, L., Wilks, A., … Oglesby-Sherrouse, A. G. (2015). The prrF-encoded small regulatory RNAs are required for iron homeostasis and virulence of Pseudomonas aeruginosa. *Infection and Immunity*, *83*(3), 863–875. https://doi.org/10.1128/IAI.02707-14

Samuelsen, Ø., Toleman, M. A., Sundsfjord, A., Rydberg, J., Leegaard, T. M., Walder, M., … Giske, C. G. (2010). Molecular epidemiology of metallo-β-lactamase-producing Pseudomonas aeruginosa isolates from Norway and Sweden shows import of international clones and local clonal expansion. *Antimicrobial Agents and Chemotherapy*, *54*(1), 346–352. https://doi.org/10.1128/AAC.00824-09

Sánchez-Martinez, G., Garza-Ramos, U. J., Reyna-Flores, F. L., Gaytán-Martínez, J., Lorenzo-Bautista, I. G., & Silva-Sanchez, J. (2010). In169, A New Class 1 Integron that Encoded blaIMP-18 in a Multidrug-Resistant Pseudomonas aeruginosa Isolate from Mexico. *Archives of Medical Research*, *41*(4), 235–239. https://doi.org/10.1016/j.arcmed.2010.05.006

Savli, H., Karadenizli, A., Kolayli, F., Gundes, S., Ozbek, U., & Vahaboglu, H. (2003). Expression stability of six housekeeping genes: a proposal for resistance gene quantification studies of Pseudomonas aeruginosa by real-time quantitative RT-PCR. *Journal of Medical Microbiology*, *52*(5), 403–408. https://doi.org/10.1099/jmm.0.05132-0

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, *30*(14), 2068–2069. https://doi.org/10.1093/bioinformatics/btu153

Sullivan, M. J., Petty, N. K., & Beatson, S. A. (2011). Easyfig: a genome comparison visualizer. *Bioinformatics*, *27*(7), 1009. https://doi.org/10.1093/BIOINFORMATICS/BTR039

Szuplewska, M., Czarnecki, J., & Bartosik, D. (2014). Autonomous and non-autonomous Tn 3 -family transposons and their role in the evolution of mobile genetic elements . *Mobile Genetic Elements*, *4*(6), 1–4. https://doi.org/10.1080/2159256x.2014.998537

Toval, F., Guzmán-Marte, A., Madriz, V., Somogyi, T., Rodríguez, C., & García, F. (2015).

Predominance of carbapenem-resistant Pseudomonas aeruginosa isolates carrying blaIMP and blaVIM metallo-β-lactamases in a major hospital in Costa Rica. *Journal of Medical Microbiology*, *64*(1), 37–43. https://doi.org/10.1099/jmm.0.081802-0

Turton, J. F., Wright, L., Underwood, A., Witney, A. A., Chan, Y. T., Al-Shahib, A., … Woodford, N. (2015). High-resolution analysis by whole-genome sequencing of an international lineage (Sequence Type 111) of pseudomonas aeruginosa associated with metallo-carbapenemases in the United Kingdom. *Journal of Clinical Microbiology*, *53*(8), 2622–2631. https://doi.org/10.1128/JCM.00505-15

Valot, B., Guyeux, C., Rolland, J. Y., Mazouzi, K., Bertrand, X., & Hocquet, D. (2015). What It Takes to Be a Pseudomonas aeruginosa? The Core Genome of the Opportunistic Pathogen Updated. *PLOS ONE*, *10*(5), e0126468. https://doi.org/10.1371/journal.pone.0126468

van der Zee, A., Kraak, W. B., Burggraaf, A., Goessens, W. H. F., Pirovano, W., Ossewaarde, J. M., & Tommassen, J. (2018). Spread of carbapenem resistance by transposition and conjugation among Pseudomonas aeruginosa. *Frontiers in Microbiology*, *9*(SEP), 1–11. https://doi.org/10.3389/fmicb.2018.02057

Walsh, T. R. (2005). The emergence and implications of metallo-β-lactamases in Gram-negative bacteria. *Clinical Microbiology and Infection, Supplement*, *11*(6), 2–9. https://doi.org/10.1111/j.1469-0691.2005.01264.x

Weiser, R., Green, A. E., Bull, M. J., Cunningham-Oakes, E., Jolley, K. A., Maiden, M. C. J., … Mahenthiralingam, E. (2019). Not all pseudomonas aeruginosa are equal: Strains from industrial sources possess uniquely large multireplicon genomes. *Microbial Genomics*, *5*(7). https://doi.org/10.1099/mgen.0.000276

Wendt, M., & Heo, G.-J. (2016). Multilocus sequence typing analysis of Pseudomonas aeruginosa isolated from pet Chinese stripe-necked turtles (Ocadia sinensis) . *Laboratory Animal Research*, *32*(4), 208. https://doi.org/10.5625/lar.2016.32.4.208

Witney, A. A., Gould, K. A., Pope, C. F., Bolt, F., Stoker, N. G., Cubbon, M. D., … Hinds, J.

(2014). Genome sequencing and characterization of an extensively drug-resistant sequence type 111 serotype O12 hospital outbreak strain of Pseudomonas aeruginosa. *Clinical Microbiology and Infection*, *20*(10), O609–O618. https://doi.org/10.1111/1469-0691.12528

Woodford, N., Turton, J. F., & Livermore, D. M. (2011). Multiresistant Gram-negative bacteria: the role of high-risk clones in the dissemination of antibiotic resistance. *FEMS Microbiology Reviews*, *35*(5), 736–755. https://doi.org/10.1111/j.1574-6976.2011.00268.x

World Health Organization. (2017). *Guidelines for the prevention and control of carbapenem-resistant Enterobacteriaceae, Acinetobacter baumannii and Pseudomonas aeruginosa in health care facilities*. Geneva. Retrieved from https://apps.who.int/iris/bitstream/handle/10665/259462/9789241550178-eng.pdf?sequence=1&ua=1

Zhao, W. H., & Hu, Z. Q. (2011). IMP-type metallo-β-lactamases in Gram-negative bacilli: Distribution, phylogeny, and association with integrons. *Critical Reviews in Microbiology*, *37*(3), 214–226. https://doi.org/10.3109/1040841X.2011.559944

FIGURES AND TABLES LEGENDS:

Fig. 1. VIM-2 and IMP-18 expression after imipenem exposure. A RT-qPCR was performed to assess the transcriptomic activity of VIM-2 and IMP-18 genes. PaeAG1 was exposed to two imipenem concentrations, showing no effects on the growth curves (A). Relative gene expression showed that higher induction occurs at 6 hours after exposure, not only for VIM-1 (B), but also for IMP-18 (C). Relative expression was statistically different by time but not by concentration ($p<0.05$).

Fig. 2. Comparative genomic analysis of 211 *P. aeruginosa* strains. By a pan-genome analysis strategy, the complete genomes were compared and the gene composition defined groups that can

be described in turn by the MLST genotyping profile. ST groups with a low frequency of less than 5 strains are shown in beige and cases with unknown ST were represented in gray. ST groups with 5 or more strains were represented with colors. The Pae-AG1 strain and all the other ST-111 strains are located in a clearly separated cluster, as shown in green.

Fig. 3. Distribution of genomic islands of PaeAG1 along the genome. The 57 predicted genomic islands are distributed along the PaeAG1 genome, and most of them forming groups with two or more islands in a row (genomic islands clusters, GIC), which are jointly named in a single label.

Fig. 4. Comparative analysis of the presence/absence of PaeAG1 genomic islands in other ST-111 strains and representative genomes. The 57 genomic islands were searched in the genomes of the other ST-111 strains, the reference strain PAO1 (ST-549) and three other strains close to the ST-111 group (see Fig. 2). The GI27 genomic island includes the VIM-2-carrying integron and it is present in PaeAG1 and two other ST-111 strains, while the GI49 (blue) harboring IMP-18-carrying integron is unique to PaeAG1 and is not it is present in none of the other 210 strains in the study. Other genomic islands linked to $\mathrm{GIC}_{\mathrm{VIM\text{-}2}}$ and $\mathrm{GIC}_{\mathrm{IMP\text{-}18}}$ have a different pattern of occurrence between strains.

Fig. 5. Description of the architecture of the genomic region $\mathbf{GIC}_{\mathbf{VIM\text{-}2}}$ containing the old-acquaintance VIM-2-carrying integron. The genomic region $\mathrm{GIC}_{\mathrm{VIM\text{-}2}}$ is absent in the reference sequence Pae-PAO1, meanwhile it is mostly present in Pae-AR445, but without most of the integron. Full coverage of the region was identified in Pae-RIV-EMC2982. The architecture of the VIM-2-carrying integron is shown.

Fig. 6. Description of the architecture of the exclusive genomic region $\mathbf{GIC}_{\mathbf{IMP\text{-}18}}$ containing the new IMP-18-carrying integron. The genomic region $\mathrm{GIC}_{\mathrm{IMP\text{-}18}}$ is absent in the reference sequence

Pae-PAO1 and Pae-RIV-EMC2982 strains, meanwhile it is partially present in Pae-97. The architecture of the IMP-18-carrying integron is shown with an arrangement that is reported here for the first time.

Fig. 7. Possible evolutionary steps associated with the genomic regions of the VIM-2- and IMP-18-carrying integrons. Different mobile elements are involved in the current state of the genomic region, being completely described for $\mathrm{GIC}_{\mathrm{VIM\text{-}2}}$ (A) and partially for $\mathrm{GIC}_{\mathrm{IMP\text{-}18}}$ (B).

Table 1. Primer sequences used for RT-qPCR experiments.

Table 2. Annotation of protein-coding genes of the genomic region $\mathrm{GIC}_{\mathrm{VIM\text{-}2}}$ associated with the VIM-2-carrying integron.

Table 3. Annotation of protein-coding genes of the genomic region $\mathrm{GIC}_{\mathrm{IMP\text{-}18}}$ associated with the IMP-18-carrying integron.

SUPPLEMENTARY MATERIAL

Supplementary Table S1

Supplementary Table S2

Supplementary Table S3

Supplementary_file 1 All_strains_information

Supplementary_file 2 Pan-genome analysis results

Table 1

Table 1. Primer sequences used for RT-qPCR experiments.

Gene	Primer	Sequence (5´→ 3´)	Final concentration	Amplicon length
IMP-18	Forward	GAATAG(A/G)(A/G)TGGCTTAA(C/T)TCTC	1 µM	188 pb
	Reverse	CCAAAC(C/T)ACTA(G/C)GTTATC		
VIM-2	Forward	CCGCGTCTATCATGGCTATT	0.1 µM	181 pb
	Reverse	ATGAGACCATTGGACGGGTA		
rpoD	Forward	GGGCGAAGAAGGAAATGGTC	1 µM	178 pb
	Reverse	CAGGTGGCGTAGGTGGAGAA		

Table 2

Table 2. Annotation of protein-coding genes of the genomic region GIC$_{\text{VIM-2}}$ associated with the VIM-2-carrying integron

PaeAG1	**Pae-RIV-EMC2982**		**Annotation**	**Name**	**RefSeq**
	Gene*	**Identity**			
PaeAG1_03237	EMC2982_03491	100.0	PslA, psl cluster plays a role in cell-cell and/or cell-surface interaction in biofilm formation	pslA (PA2231)	NP_250921.1; WP_003111160.1
PaeAG1_03238	EMC2982_03490	100.0	Hypothetical protein PA2230	PA2230	NP_250920.1; WP_003122761.1
PaeAG1_03239	EMC2982_03489	100.0	Hypothetical protein PA2229	PA2229	NP_250919.1 ; WP_003113716.1
PaeAG1_03240	EMC2982_03488	100.0	Hypothetical protein	HP	WP_034066849.1
PaeAG1_03241	EMC2982_03487	100.0	Transposase TnpA	tnpA	WP_003460108.1
PaeAG1_03242	EMC2982_03486	100.0	Transposase TnpR	tnpR	WP_000147567.1; YP_005211182.1
PaeAG1_03243	EMC2982_03485	100.0	Transposase TnpM	tnpM	WP_004217866.1
PaeAG1_03244	EMC2982_03484	100.0	Class I integron integrase IntI	**intI**	YP_005221021.1
PaeAG1_03245	EMC2982_03483	100.0	6'-N-aminoglycoside acetyltransferase type I aacA29a	**aacA29a**	WP_032490447.1
PaeAG1_03246	EMC2982_03482	100.0	Carbapenem-hydrolyzing metallo-beta-lactamase VIM-2	**VIM-2**	WP_032491390.1
PaeAG1_03247	EMC2982_03481	100.0	6'-N-aminoglycoside acetyltransferase type I aacA29b	**aacA29b**	WP_032490447.1
PaeAG1_03248	EMC2982_03480	100.0	Sulfonamide-resistant dihydropteroate synthase Sul1	**sul1**	WP_000259031.1
PaeAG1_03249	EMC2982_03479	100.0	Acetyltransferase	Acetyltransferase	WP_000376623.1

PaeAG1_03250	EMC2982_03478	100.0	Transposase TniB	tniBΔ	WP_003107582.1; WP_021264342.1
PaeAG1_03251	EMC2982_03477	100.0	Transposase TniA	tniA	WP_000179844.1; YP_008766137.1
PaeAG1_03252	EMC2982_03476	100.0	Hypothetical protein	urf2Δ	WP_000204520.1
PaeAG1_03253	EMC2982_03475	100.0	Mercury resistance protein merE	merE	WP_000993386.1; YP_789372.1
PaeAG1_03254	EMC2982_03474	**99.0**	Transcriptional regulator merD	merD	WP_001277456.1; YP_789373.1
PaeAG1_03255	EMC2982_03473	**99.8**	Mercuric reductase merA	merA	WP_000105636.1; YP_789374.1
PaeAG1_03256	EMC2982_03472	100.0	Transposase	tnpA	WP_003111042.1; WP_003460108.1
PaeAG1_03257	EMC2982_03471	100.0	TpnA repressor protein	tnpC	WP_003111043.1; NP_745109.1
PaeAG1_03258	EMC2982_03470	100.0	Hypothetical protein	HP	WP_003111045.1
PaeAG1_03259	EMC2982_03469	100.0	Hypothetical protein	HP	WP_003111046.1
PaeAG1_03260	EMC2982_03468	100.0	Homospermidine synthase (HPS)	HPS	WP_003111047.1
PaeAG1_03261	EMC2982_03467	100.0	Hypothetical protein	HP	WP_003111048.1
PaeAG1_03262	EMC2982_03466	100.0	Hypothetical protein	HP	WP_003111049.1
PaeAG1_03263**	EMC2982_03465	100.0	Recombinase	Recombinase	WP_003111050.1
PaeAG1_03265**	EMC2982_03463	100.0	Hypothetical protein	HP	WP_010792965.1
PaeAG1_03266	EMC2982_03462	100.0	Hypothetical protein	HP	WP_003092560.1
PaeAG1_03267	EMC2982_03461	100.0	Hypothetical protein PA2229	PA2229	NP_250919.1 ; WP_003113716.1
PaeAG1_03268	EMC2982_03460	100.0	Hypothetical protein PA2228	PA2228	NP_250918.1 ; WP_003113715.1
PaeAG1_03269	EMC2982_03459	100.0	AraC-type transcriptional regulator VqsM	vqsM (PA2227)	NP_250917.1 ; WP_003113714.1

Notes:

* "EMC2982_" is the locus with our annotation (see Methods). See Supplementary Table S1 for locus in PGDB annotation file and amino-acid comparison against other genomes.

**PaeAG1_03264 is a tRNA, i.e. not included here.

Table 3

Table 3. Annotation of protein-coding genes of the genomic region GIC$_{IMP-18}$ associated with the IMP-18-carrying integron

PaeAG1	Pae-97		Annotation	Name	RefSeq
	Gene*	Identity			
PaeAG1_05736	Pa97_05533	100.0	Hypothetical protein PA4702	PA4702	NP_253390.1 ; WP_003095090.1
PaeAG1_05737	Pa97_05534	100.0	Hypothetical protein PA4703	PA4703	NP_253391.1 ; WP_003095094.1
PaeAG1_05738	Pa97_05535	100.0	cAMP-binding protein A PA4704 , cbpA	cbpA (PA4704)	NP_253392.1 ; WP_003095096.1
PaeAG1_05739	Pa97_05536	100.0	Recombinase	Recombinase	WP_023442562.1
PaeAG1_05740	Pa97_05537	100.0	helix-turn-helix transcriptional regulator (HTH-TR)	HTH-TR	WP_003148665.1
PaeAG1_05741	Pa97_05538	**99.8**	Hypothetical protein	HP	WP_137462639.1
PaeAG1_05742	Pa97_05539	100.0	Hypothetical protein	HP	WP_071567699.1
PaeAG1_05743	Pa97_05540	100.0	Hypothetical protein	HP	WP_042855636.1
PaeAG1_05744	Pa97_05541	100.0	Type I restriction endonuclease subunit R	hsdR	WP_042855635.1; YP_005974822.1
PaeAG1_05745	Pa97_05542	100.0	Hypothetical protein	HP	WP_003148682.1
PaeAG1_05746	Pa97_05543	100.0	restriction endonuclease subunit S	hsdS	WP_079393399.1; YP_005974824.1
PaeAG1_05747	Pa97_05544	100.0	type I restriction-modification system (RMS) subunit M	hsdM	WP_003148685.1; YP_005974823.1
PaeAG1_05748	Pa97_05545	100.0	recombinase family protein	Recombinase	WP_003148687.1
PaeAG1_05749	Pa97_05546	100.0	class 1 integron integrase IntI1	**intI**	YP_005221021.1
PaeAG1_05750	Pa97_05548 (IMP-1)	**80.5**	subclass B1 metallo-beta-lactamase IMP-18	**IMP-18**	WP_060614779.1
PaeAG1_05750.1	CP913_RS21750	**36.4**	DUF1010 domain-containing protein gcuD	**gcuD**	WP_001336345.1

PaeAG1_05751	Pa97_05547	**36.4**	oxacillin-hydrolyzing class D beta-lactamase OXA-2	**OXA-2**	WP_034033256.1
PaeAG1_05751.1	CP913_RS28765	**99.4**	Aminoglycoside N(6')-acetyltransferase type 1 aacA4	**aacA4**	WP_003159191.1
PaeAG1_05752	Pa97_04840	100.0	sulfonamide-resistant dihydropteroate synthase Sul1	**sul1**	WP_000259031.1
PaeAG1_05753	Pa97_04839	100.0	GNAT family N-acetyltransferase	GNAT	WP_000376623.1
PaeAG1_05754	Pa97_05603	44.4	ATP-binding protein, protease istD	istD	WP_000983249.1
PaeAG1_05755	Pa97_04622	44.1	Transposase istA	istA	WP_001324342.1; WP_000996451.1
PaeAG1_05756	Pa97_05551	100.0	Transposase TniB	tniBΔ	WP_003107582.1; WP_021264342.1
PaeAG1_05757	Pa97_05552	100.0	Transposase TniA	tniA	WP_000179844.1; YP_008766137.1
PaeAG1_05758	Pa97_05553	100.0	Hypothetical protein	HP	WP_003157545.1
PaeAG1_05759	Pa97_05554	99.6	Hypothetical protein	HP	WP_003157546.1
PaeAG1_05760	Pa97_05555	97.6	iron(III) ABC transporter PhuW	phuW	NP_253393.1 ; WP_003113451.1
PaeAG1_05761	Pa97_05556	99.6	heme ABC transporter ATP-binding protein PhuV	phuV	NP_253394.1 ; WP_003095098.1
PaeAG1_05762	Pa97_05557	100.0	iron ABC transporter permease PhuU	phuU	NP_253395.1 ; WP_003121063.1

Notes:

* “Pa97_” is the locus with our annotation (see Methods). See Supplementary Table S2 for locus in PGDB annotation file and amino-acid comparison against other genomes. Cases with “CP913_” locus refers to the PGDB annotation file with a better score due to annotation algorithms differences.

Figure-1

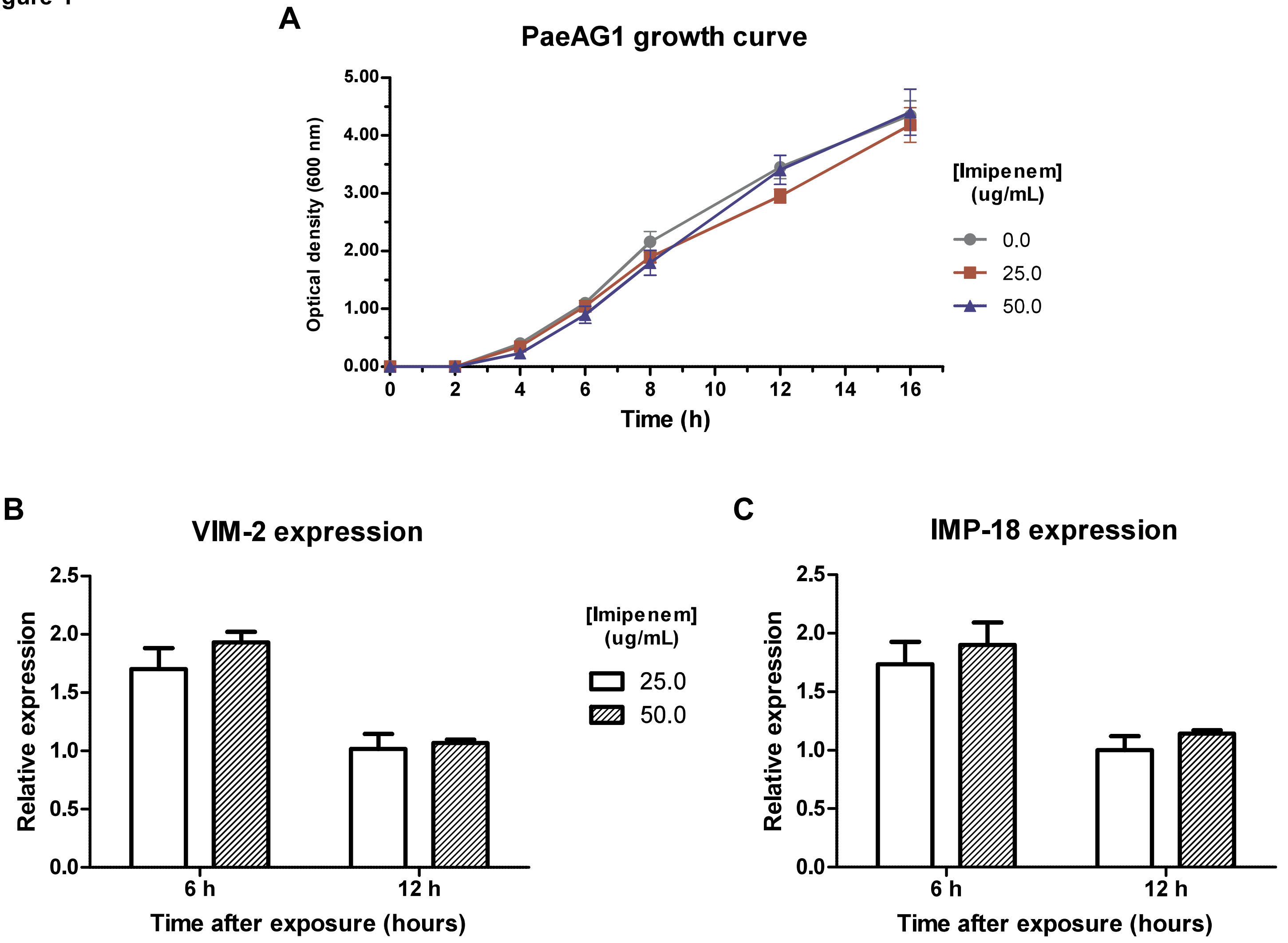

A
PaeAG1 growth curve
Optical density (600 nm)
0.00
1.00
2.00
3.00
4.00
5.00
0
2
4
6
8
10
12
14
16
Time (h)
[Imipenem]
(ug/mL)
0.0
25.0
50.0
B
VIM-2 expression
Relative expression
0.0
0.5
1.0
1.5
2.0
2.5
6 h
12 h
Time after exposure (hours)
[Imipenem]
(ug/mL)
25.0
50.0
C
IMP-18 expression
Relative expression
0.0
0.5
1.0
1.5
2.0
2.5
6 h
12 h
Time after exposure (hours)

Figure-2

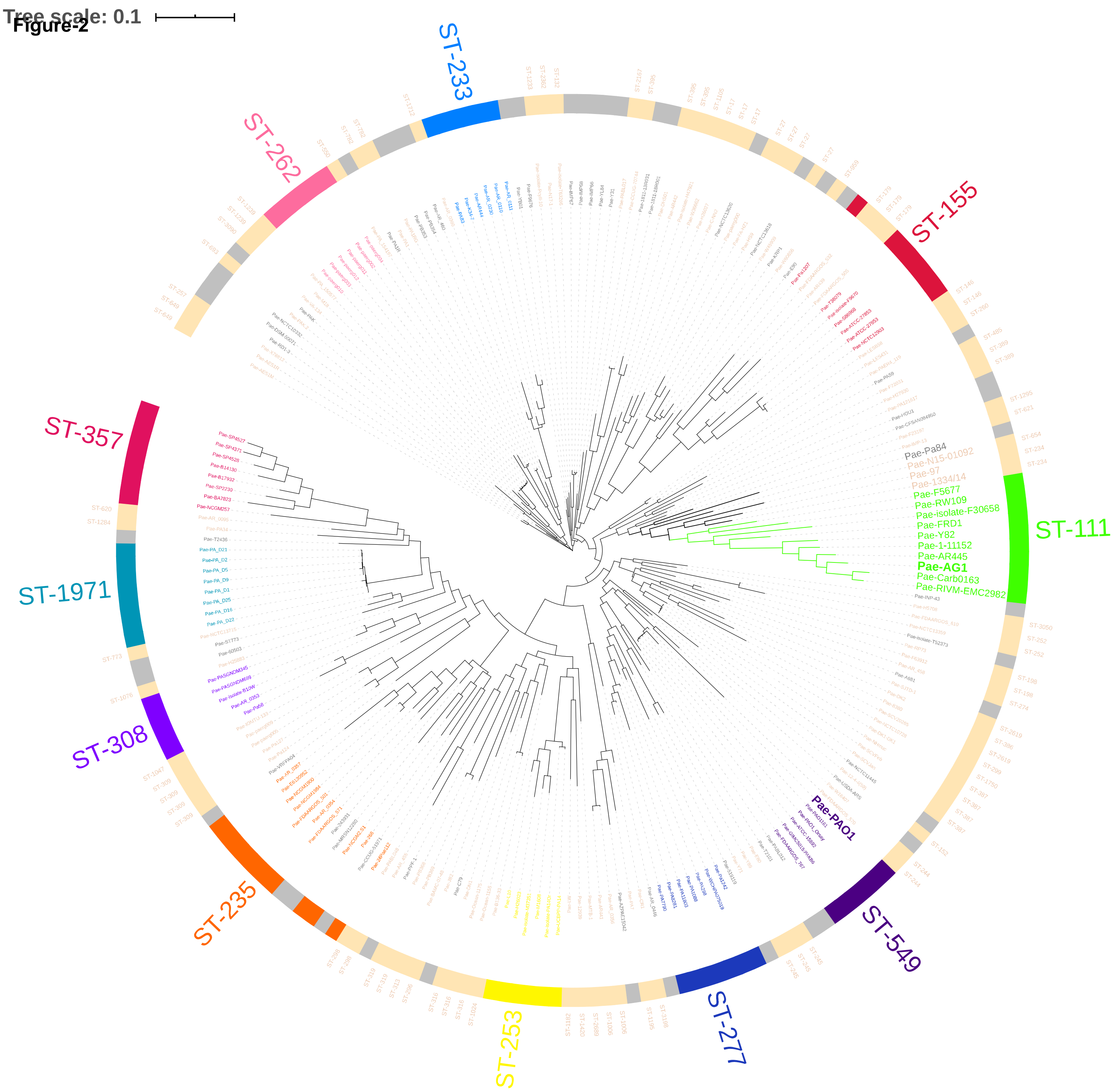

Tree scale: 0.1
ST-233
ST-262
ST-155
ST-111
ST-357
ST-1971
ST-308
ST-235
ST-253
ST-277
ST-549
Pae-Pa84
Pae-N15-01092
Pae-97
Pae-1334/14
Pae-F5677
Pae-RW109
Pae-isolate-F30658
Pae-FRD1
Pae-Y82
Pae-1-11152
Pae-AR445
Pae-AG1
Pae-Carb0163
Pae-RIVM-EMC2982
Pae-PAO1
Pae-SP4527
Pae-SP4371
Pae-SP4528
Pae-B14130
Pae-B17932
Pae-SP2230
Pae-BA7823
Pae-NCGM257
Pae-PA_D21
Pae-PA_D2
Pae-PA_D5
Pae-PA_D9
Pae-PA_D1
Pae-PA_D25
Pae-PA_D16
Pae-PA_D22

Figure-3

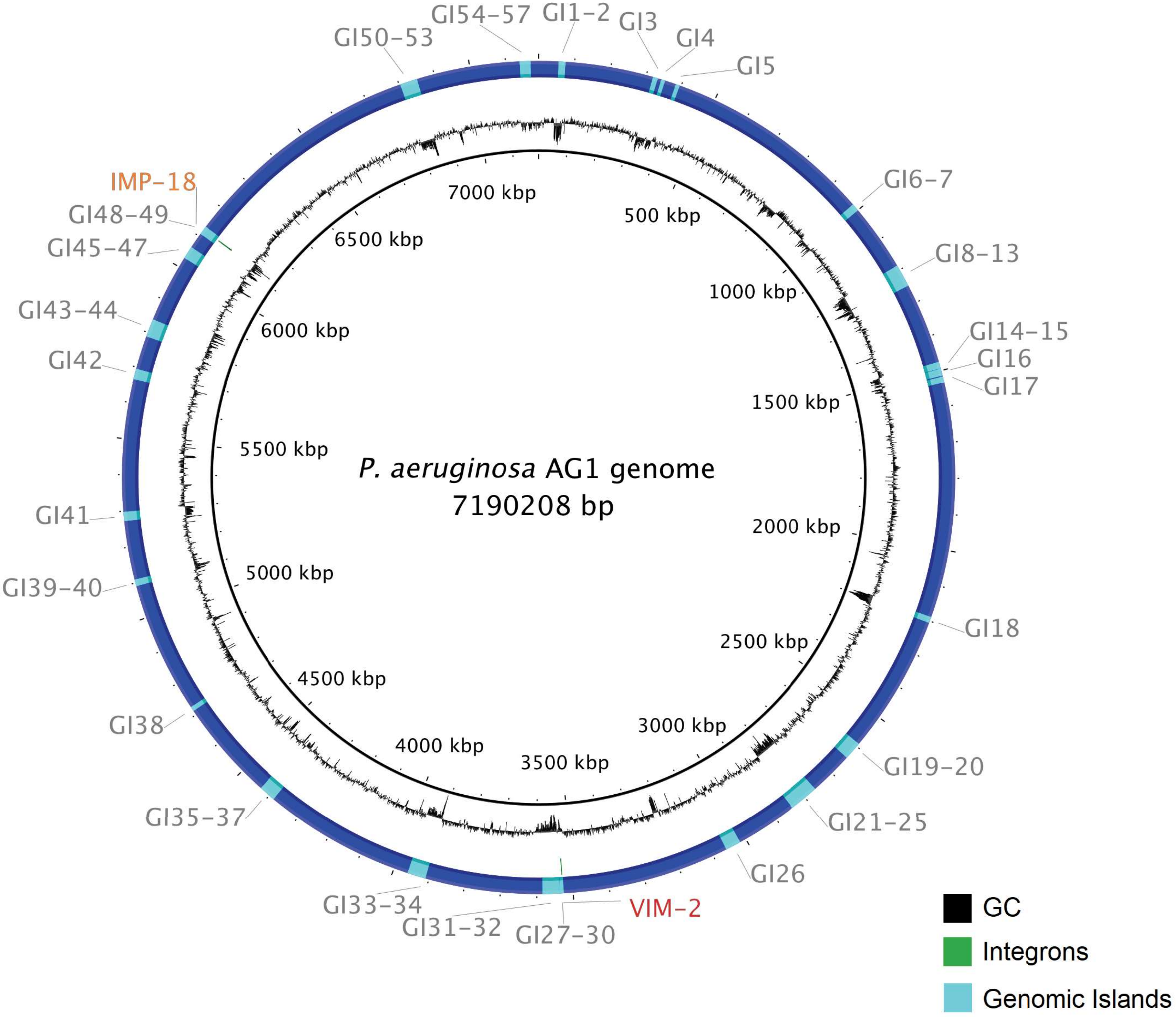

GI54-57
GI1-2
GI3
GI4
GI5
GI50-53
GI6-7
GI8-13
GI14-15
GI16
GI17
GI18
GI19-20
GI21-25
GI26
VIM-2
GI27-30
GI31-32
GI33-34
GI35-37
GI38
GI39-40
GI41
GI42
GI43-44
GI45-47
GI48-49
IMP-18
7000 kbp
500 kbp
1000 kbp
1500 kbp
2000 kbp
2500 kbp
3000 kbp
3500 kbp
4000 kbp
4500 kbp
5000 kbp
5500 kbp
6000 kbp
6500 kbp
P. aeruginosa AG1 genome
7190208 bp
GC
Integrons
Genomic Islands

Figure-4

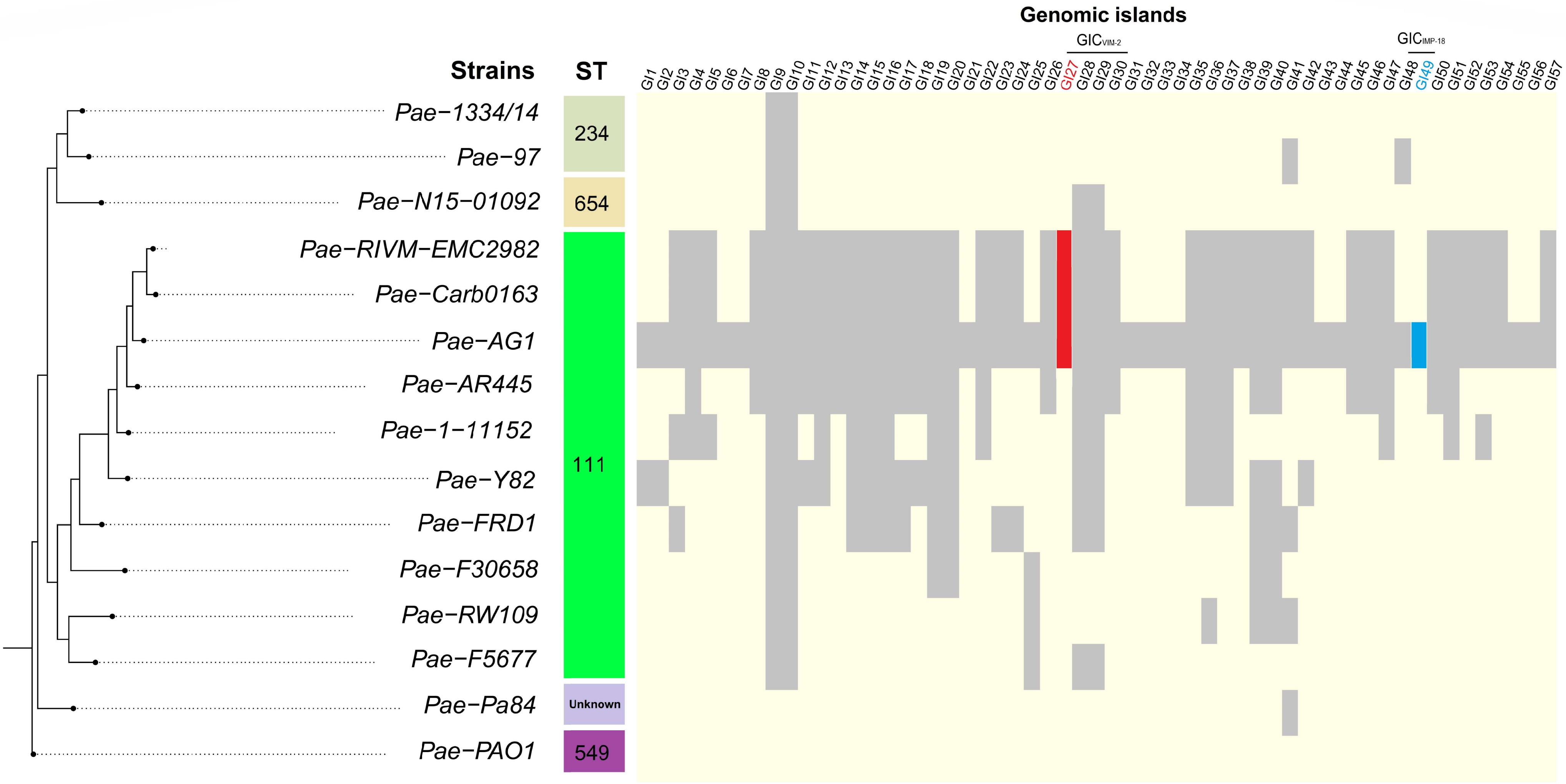
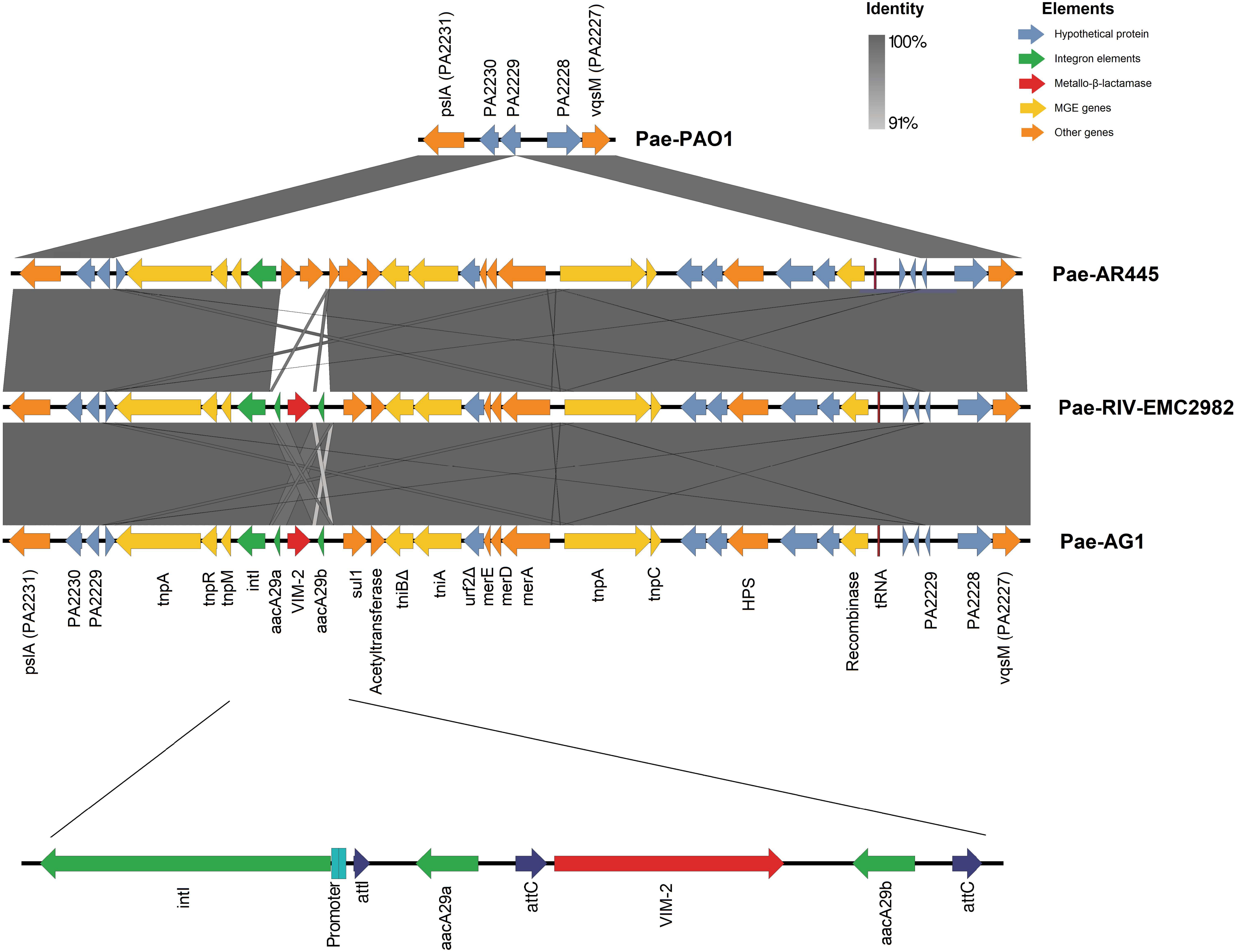

Genomic islands
GICVIM-2
GICIMP-18
Strains
ST
GI1
GI2
GI3
GI4
GI5
GI6
GI7
GI8
GI9
GI10
GI11
GI12
GI13
GI14
GI15
GI16
GI17
GI18
GI19
GI20
GI21
GI22
GI23
GI24
GI25
GI26
GI27
GI28
GI29
GI30
GI31
GI32
GI33
GI34
GI35
GI36
GI37
GI38
GI39
GI40
GI41
GI42
GI43
GI44
GI45
GI46
GI47
GI48
GI49
GI50
GI51
GI52
GI53
GI54
GI55
GI56
GI57
Pae−1334/14
Pae−97
234
Pae−N15−01092
654
Pae−RIVM−EMC2982
Pae−Carb0163
Pae−AG1
Pae−AR445
Pae−1−11152
Pae−Y82
Pae−FRD1
Pae−F30658
Pae−RW109
Pae−F5677
111
Pae−Pa84
Unknown
Pae−PAO1
549

Figure-5



psIA (PA2231)
PA2230
PA2229
PA2228
vqsM (PA2227)
Pae-PAO1
Identity
100%
91%
Elements
Hypothetical protein
Integron elements
Metallo-β-lactamase
MGE genes
Other genes
Pae-AR445
Pae-RIV-EMC2982
Pae-AG1
psIA (PA2231)
PA2230
PA2229
tnpA
tnpR
tnpM
intI
aacA29a
VIM-2
aacA29b
sul1
Acetyltransferase
tniBΔ
tniA
urf2Δ
merE
merD
merA
tnpA
tnpC
HPS
Recombinase
tRNA
PA2229
PA2228
vqsM (PA2227)
intI
Promoter
attI
aacA29a
attC
VIM-2
aacA29b
attC

Figure-6

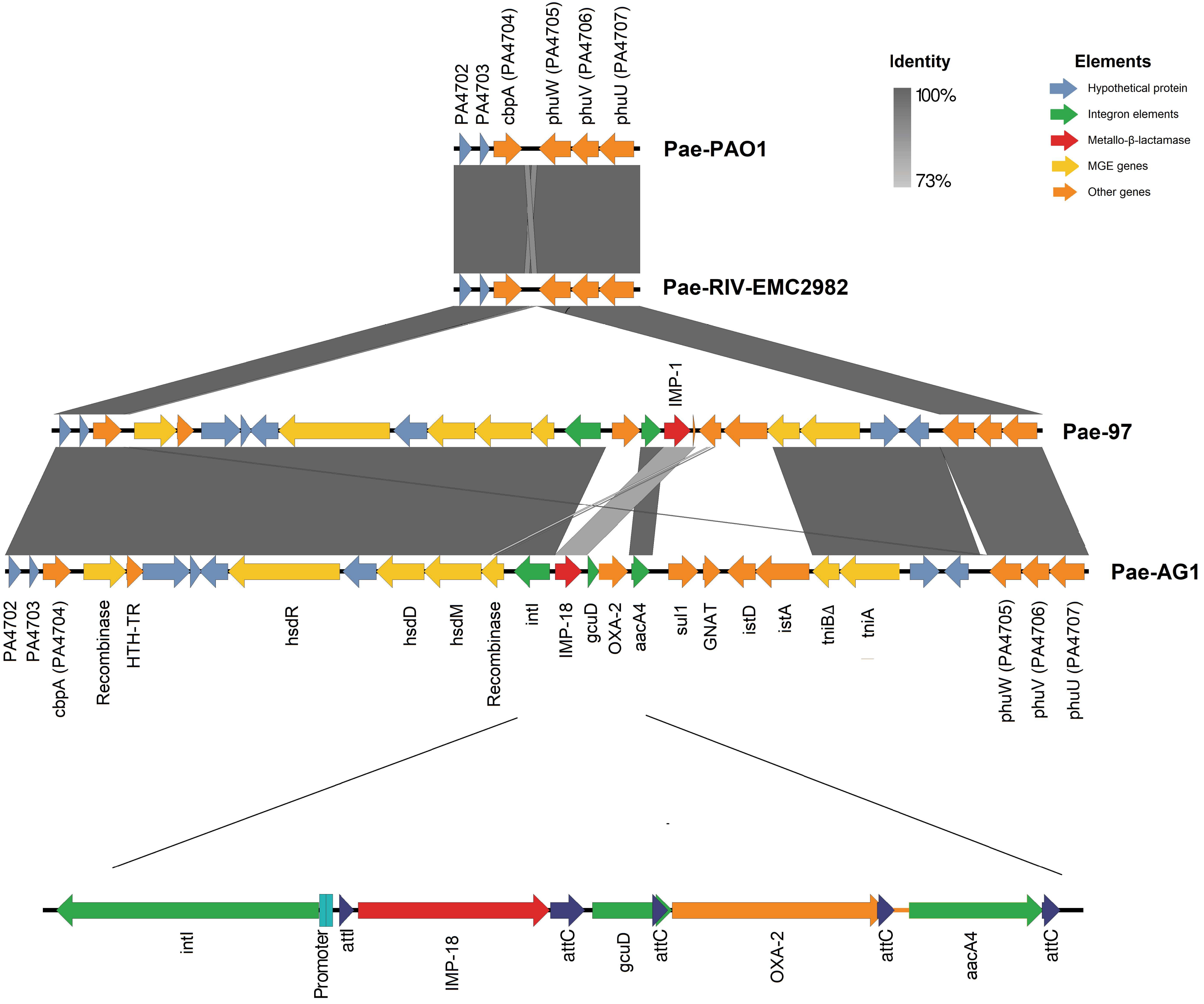
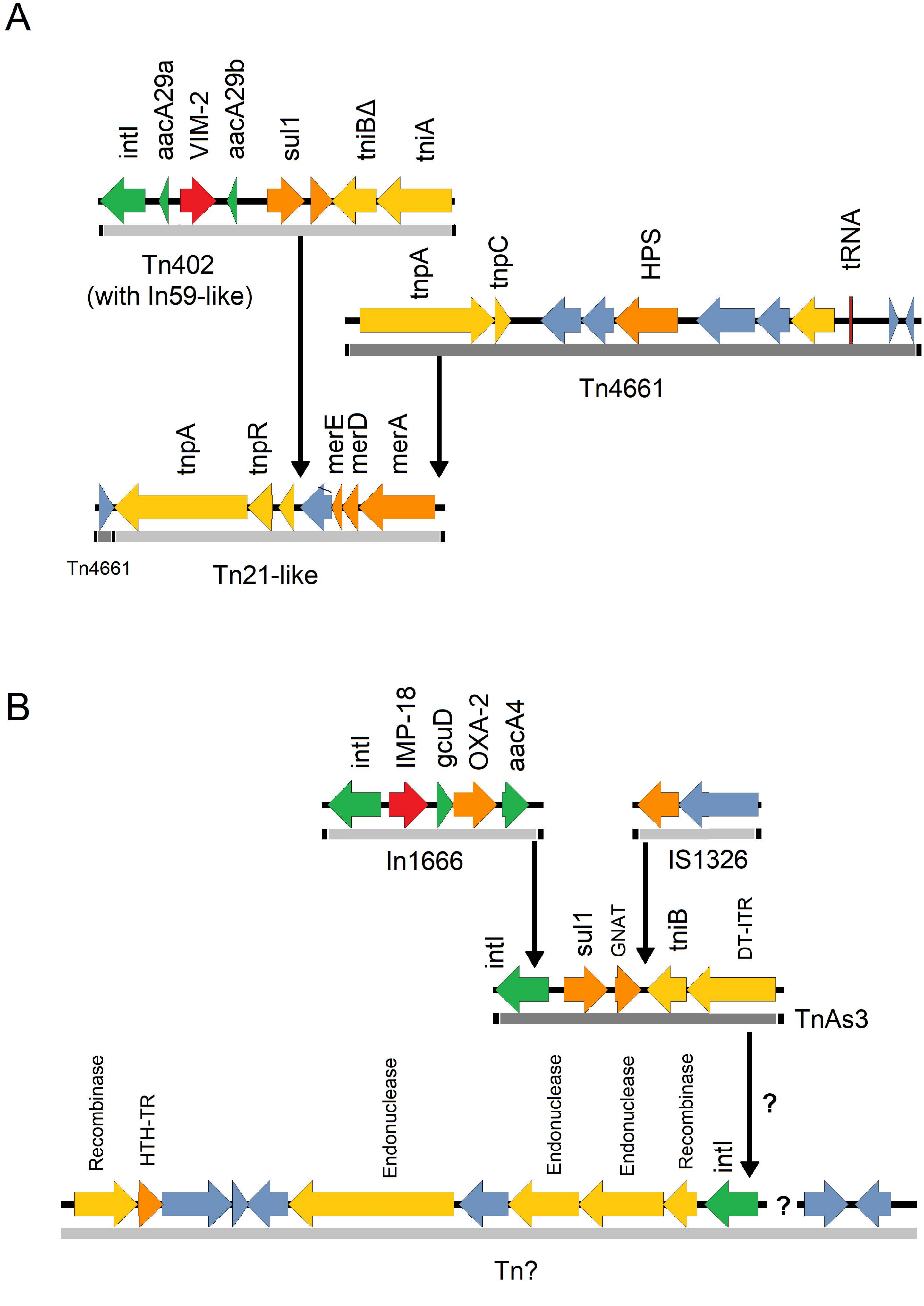

PA4702
PA4703
cbpA (PA4704)
phuW (PA4705)
phuV (PA4706)
phuU (PA4707)
Pae-PAO1
Identity
100%
73%
Elements
Hypothetical protein
Integron elements
Metallo-β-lactamase
MGE genes
Other genes
Pae-RIV-EMC2982
IMP-1
Pae-97
Pae-AG1
PA4702
PA4703
cbpA (PA4704)
Recombinase
HTH-TR
hsdR
hsdD
hsdM
Recombinase
intI
IMP-18
gcuD
OXA-2
aacA4
sul1
GNAT
istD
istA
tniBΔ
tniA
phuW (PA4705)
phuV (PA4706)
phuU (PA4707)
intI
Promoter
attI
IMP-18
attC
gcuD
attC
OXA-2
attC
aacA4
attC

Figure-7


A
intI
aacA29a
VIM-2
aacA29b
sul1
tniBΔ
tniA
Tn402
(with In59-like)
tnpA
tnpC
HPS
tRNA
Tn4661
tnpA
tnpR
merE
merD
merA
Tn4661
Tn21-like
B
intI
IMP-18
gcuD
OXA-2
aacA4
In1666
IS1326
intI
sul1
GNAT
tniB
DT-ITR
TnAs3
?
Recombinase
HTH-TR
Endonuclease
Endonuclease
Endonuclease
Recombinase
intI
?
Tn?

CHAPTER 3

Two-dimensional gel electrophoresis (2D-GE) image analysis based on CellProfiler: *Pseudomonas aeruginosa* AG1 as model

Molina-Mora, J. A., Chinchilla-Montero, D., Castro-Peña, C., & Garcia, F. (2020). Two-dimensional gel electrophoresis (2D-GE) image analysis based on CellProfiler: *Pseudomonas aeruginosa* AG1 as model. Medicine, IN-PRESS.

Summary

Using the bacterial strain *Pseudomonas aeruginosa* AG1 as a model, we obtained images from Two-dimensional gel electrophoresis (2D-GE) of periplasmic protein profiles when the strain was exposed to multiple antibiotics. As reported, 2D-GE is an indispensable technique for the study of proteomes of biological systems, providing an assessment of changes in protein abundance under various experimental conditions. However, due to the complexity of 2D-GE gels, there is no systematic, automatic and reproducible protocol for image analysis and specific implementations are required for each context. In addition, practically all available solutions are commercial, which implies high cost and little flexibility to modulate the parameters of the algorithms. Then we proceeded to implement and evaluate an image analysis protocol with an open-source software, CellProfiler. First, a preprocessing step included a bUnwarpJ-Image pipeline for aligning 2D-GE images. Then, using CellProfiler we standardized two pipelines for spots identification. Total spots recognition was achieved using segmentation by intensity, whose performance was evaluated when compared with a reference protocol. In a second pipeline with the same program, differential identification of spots was addressed when comparing pairs of protein profiles. Due to the characteristics of the programs used, our workflow can automatically analyze a large number of images and it is parallelizable, which is an advantage with respect to other implementations. Finally, we compared six experimental conditions of bacterial strain in the presence or absence of antibiotics, determining protein profiles relationships by applying clustering algorithms PCA (Principal Components Analysis) and HC (Hierarchical Clustering). Results revealed that global proteomic profile after exposure to a sub-inhibitory ciprofloxacin (CIP) concentration remains close to control (LB medium, without antibiotics), contrasting with the results obtained with tobramycin and imipenem. This means that the effects of ciprofloxacin at the proteomic level are fewer than the changes given by other antibiotics.

Quality Improvement Study

Medicine®

OPEN

Two-dimensional gel electrophoresis (2D-GE) image analysis based on CellProfiler

Pseudomonas aeruginosa AG1 as model

Jose Arturo Molina-Mora, MSc*, Diana Chinchilla-Montero, MSc, Carolina Castro-Peña, MSc, Fernando García, PhD

Abstract

Two-dimensional gel electrophoresis (2D-GE) is an indispensable technique for the study of proteomes of biological systems, providing an assessment of changes in protein abundance under various experimental conditions. However, due to the complexity of 2D-GE gels, there is no systematic, automatic, and reproducible protocol for image analysis and specific implementations are required for each context. In addition, practically all available solutions are commercial, which implies high cost and little flexibility to modulate the parameters of the algorithms. Using the bacterial strain, *Pseudomonas aeruginosa*AG1 as a model, we obtained images from 2D-GE of periplasmic protein profiles when the strain was exposed to multiple conditions, including antibiotics. Then, we proceeded to implement and evaluate an image analysis protocol with open-source software, CellProfiler. First, a preprocessing step included a bUnwarpJ-Image pipeline for aligning 2D-GE images. Then, using CellProfiler, we standardized two pipelines for spots identification. Total spots recognition was achieved using segmentation by intensity, whose performance was evaluated when compared with a reference protocol. In a second pipeline with the same program, differential identification of spots was addressed when comparing pairs of protein profiles. Due to the characteristics of the programs used, our workflow can automatically analyze a large number of images and it is parallelizable, which is an advantage with respect to other implementations. Finally, we compared six experimental conditions of bacterial strain in the presence or absence of antibiotics, determining protein profiles relationships by applying clustering algorithms PCA (Principal Components Analysis) and HC (Hierarchical Clustering).

Abbreviations: 2D-GE = two-dimensional gel electrophoresis, ANOVA = analysis of variance, CIP = Ciprofloxacin, FDR = false discovery rate, HC = hierarchical clustering, IMP = Imipenem, PCA = Principal Component Analysis, pI = isoelectric point, TOB = Tobramycin.

Keywords: 2D-GE, bUnwarpJ, CellProfiler, image analysis, proteomics, *Pseudomonas aeruginosa*

1. Introduction

Proteomics is a field of study of the omic sciences that focuses on the analysis of the complete set of proteins produced in a cell, tissue, or organism at a given moment, that is, proteomes. The evaluation of protein profiles of biological samples, either by the presence or absence of proteins, or the measurement of their relative abundance, can help to understand the cellular processes, including associated to pathologies, particular biological conditions or to understand molecular mechanisms of biological relevance.[1] However, since cells can produce thousands of proteins, the processing of protein information is complex.

Editor: Duane R. Hospenthal.

Ethics approval and consent to participate: Not Applicable. Since, our study is focused on a bacterial strain, an ethical approval was not necessary, as it was evaluated by the scientific committee of the Research Center in Tropical Diseases (CIET), University of Costa Rica, Costa Rica. Neither humans nor animals were used in this study.

This work was funded by project B7124 Analysis of the extracellular proteome of Pseudomonas aeruginosa AG1, registered in the University of Costa Rica (period 2017–2019).

The authors have no conflicts of interest to disclose.

Data availability: Image analysis pipelines for both "Identification of total spots" and "Differential identification of spots" are available at: https://github.com/josemolina6/2D-GE. Both need to be opened using CellProfiler software.

The datasets generated during and/or analyzed during the present study are publicly available.

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

** Correspondence: Jose Arturo Molina-Mora, Research Center in Tropical Diseases (CIET), University of Costa Rica, San Pedro, Costa Rica (e-mail: jose.molinamora@ucr.ac.cr).*

How to cite this article: Molina-Mora JA, Chinchilla-Montero D, Castro-Peña C, García F. Two-dimensional gel electrophoresis (2D-GE) image analysis based on CellProfiler: Pseudomonas aeruginosa AG1 as model. Medicine 2020;99:49(e23373).

Received: 25 October 2019 / Received in final form: 20 October 2020 / Accepted: 27 October 2020

http://dx.doi.org/10.1097/MD.0000000000023373

In this sense, two-dimensional gel electrophoresis (2D-GE) has become a method of choice for proteomic studies since its introduction more than 40 years ago.[2] Its current use in part is explained due to its high performance in terms of the separation of complex protein mixtures.[3] The use of 2D-GE gels allows the comparison of complex protein profiles, first separating them by isoelectric point (pI) and then by molecular weight.[4] With this, the proteins are separated as spots, which are revealed with stains such as Coomassie blue or silver stain, to then capture images of the gel. These images are then analyzed to identify the points and study the protein content, as well as continue with subsequent proteomic studies by other strategies.[1]

However, due to the anomalies present in the images of 2D-GE gels, there is still no reliable, automatic and highly reproducible pipeline for 2D-GE image analysis.[4] At a strictly experimental level, the challenges of this type of technique include experimental variation (reagents, running conditions, etc), particular mobility of the proteins, deformation of the gel and the high probability of finding several proteins in the same space of the plane of the gel.[5] At the level of image analysis, the difficulties are greater, including anomalies such as the presence of vertical and horizontal stripes, noise around protein spots, diffuse spots and background noise, fusions of spots, artifacts due to the presence of dust or bubbles, saturation of certain spots and lack of linear intensity of protein spots.[1,3]

At the preprocessing level, one of the basic tasks is the alignment of images, in which one of the images is intentionally deformed to match the spots with the other image. This is done with a transformation that optimizes the measure of similarity and in turn quantifies the quality of alignment.[6] Then, algorithms are implemented to detect protein spots, that is, the recognition of objects by segmentation to define the limits of each spot, many of them with methods based on intensity, form, or hybrid strategies.[3] In a subsequent step, the quantification of the level of protein expression is performed according to the intensity and the number of pixels.[1] If required, a differential expression analysis can be performed by comparing conditions, in which multivariate statistical criteria are used, including analysis of variance (ANOVA) according to the size and intensity of the spot, strategies of correction of *P* values such as FDR (false discovery rate) or machine learning algorithms for clustering or classifying protein profiles.[5]

For the implementation of these analysis modules, there are software packages, practically all commercially available. This has the disadvantage that many are for a particular proteomics market, subject to purchase of equipment and that makes it even more expensive. Within these commercial solutions are PDQuest, ImageMaster2D, ProteomeWeaver, ProteinMine, Delta2D and Melanie, among others,[1] which generally contain modules that include the alignment of images to be compared, automatic identification and edition of spots, counting, quantification of intensity, and area calculation by spot. Within the options of free software, ImageJ[7] has been widely used for analysis of images of biological origin, but automation is limited, given that its approach is of individual analysis, as has been described.[8,9] In the approach of Natele and collaborators, a protocol was implemented with ImageJ for the study of spots in 2D-GE gels, applicable to pairs of images but with a strategy of limited scalability to large sets of images.[4]

Thus, due to all above, the aim of this work was to implement and evaluate an image analysis protocol with open-source software for identifying spots in 2D-GE gels images, which also includes the potential application to automatically analyze a large number of images and, due to the computational requirements, that is potentially parallelizable. For this, we standardized experimental protocols for the study of the periplasmic proteins of *Pseudomonas aeruginosa* AG1 under various conditions of exposure to antibiotics. This bacterium is an opportunistic pathogen that survives diversity of environments, including hospital environments.[10] Specifically, our study model is the strain *P aeruginosa* AG1, a Costa Rican isolate[11] with a multiresistance profile to antibiotics and with clonal MLST (https://pubmlst.org/) categorized as ST111, which implies a high risk for public health because of its resistance to therapy and association with nosocomial infections.

With this bacterial model, from the experimental assays, separation of the proteins was achieved using 2D-GE gels and it was revealed with silver staining. After capturing the respective images, we implemented a pre/processing step that included an initial phase of image alignment using the script bUnwarpJ[12] in the program ImageJ[7,13]; this package has the ability to align hundreds of images to the same reference in one step. Subsequently, we made the spots identification with two protocols using the program CellProfiler.[8,14] A first protocol was established to identify total spots in the images of the gels, and that was contrasted with a reference analysis with the commercial program Melanie (https://2d-gel-analysis.com/). In a second implementation, spots differential identification in experimental conditions was made, separating the common spots from the exclusive ones. Finally, a comparison of several experimental conditions was carried out with two clustering algorithms, showing the similarity of protein profiles of *P aeruginosa* AG1 exposed to antibiotics. To the best of our knowledge, CellProfiler program has not been used for the identification of spots on 2D-GE gels, although it has been implemented to for recognizing biological objects (cells, complete organisms, tumors, colonies of microorganisms and others) in hundreds of images, making this implementation as promising for the analysis of hundreds of gels in proteomics studies.

2. Methods

2.1. Experimental assays for 2D-GE gels

For the extraction and analysis of periplasmic proteins of *P aeruginosa* AG1, cultures were used at exponential phase in LB medium (Luria Bertani, 2 clones) and LB medium added with subinhibitory concentrations of antibiotics ciprofloxacin (CIP, 12.5 μg/mL), tobramycin (TOB, 62.5 and 125 μg/mL), and imipenem (IMP, 25 and 50 μg/mL). The marker "IEF 3–10 SERVA liquid mix" (with proteins of size and known isoelectric point) was used as migration control. After pre-cultivation for 16 h under the corresponding conditions, the bacteria were cultured for 6 h at 37°C under agitation. After verifying their exponential growth by optical density, the samples were centrifuged at 10,000 rpm for 30 min and the supernatant was discarded.

For the extraction of periplasmic proteins with chloroform, pellets were washed with sterile PBS 1× and then 0.01 M Tris–hydrochloride pH 8.0 filtered and chloroform were added. After an incubation, the sample was centrifuged and the supernatant stored at −80°C. For protein precipitation, the supernatants were treated with methanol and chloroform. After vigorous stirring and a strong centrifugation, the separation was achieved in

2 phases, an upper one of methanol/water and a lower one of chloroform. The periplasmic protein fraction was found in the middle of both phases, which was finally precipitated with more alcohol and centrifugation. After the supernatant was removed, the protein pellet was dried and resuspended in 0.05% SDS lysis buffer, obtaining the protein extract of interest. Modified protocol of Ames et al.[15]

Finally, the protein separation in two-dimensional gel was performed by adding the proteins to Isoelectric Focusing (IEF) strips and hydrated for 24 h at room temperature. Then, the proteins were separated using a non-linear 3 to 11 pH gradient, following the manufacturer's instructions (GE HealthCare Immobiline Dry Strip GelsTM). For the second dimension (molecular weight), the IEF strips were incubated in equilibrium buffer (50 mM Tris–HCl, 6 M Urea, 30% glycerol and 2% SDS) with 4-dithiothreitol (DTT), for 10 min, before separation into a SDS-GE gradient of 4% to 20% for 90 min at 150 V. PageRuler Protein Ladder (Fermentas) was used as a molecular weight marker. All gels were visualized with silver stain. The bands were observed in the ChemiDoc photo viewer (BioRad).

2.2. Preprocessing of 2D-GE images by alignment

Due to the conditions inherent in the assembly of 2D-GE gels, the images require preprocessing alignment (Fig. 1). Thus, the detailed protocol was implemented by Natale and collaborators[4] using the bUnwarpJ package in the ImageJ program.[12] Using 5 reference points, with spots known as common between the images, we proceeded to the deformation of the larger images to align with the spots of the smaller image, using the parameter of "degree of deformation" as fine. After the deformation, the aligned images were saved for the following analysis steps.

2.3. Identification of total spots

In order to identify the totality of visible protein spots in the gels, an image analysis protocol was implemented using the CellProfiler program (https://CellProfiler.org/). As detailed in Figure 1 (middle-left) the protocol consisted of 5 steps:

1. the inversion of the images to enable recognition,
2. the implementation of an object recognition, evaluating different parameters and recognition algorithms and segmentation,
3. improving the identification by manual editing,
4. calculating different metrics by object and, finally,
5. visualizing the recognition in the images.

Similarly, the automatic protocol of a specialized program for 2D gels, Melanie (https://2d-gel-analysis.com/), was used to compare the performance of our protocol, contrasting the number of recognized elements and the intensity measured with a linear regression.

2.4. Differential identification of spots

To compare the differential expression of proteins between experimental conditions, we proceeded to implement an analysis of pairs of images (Fig. 1 middle-right, also see Figure 4A for case of two clones of control condition). The steps for this process included:

1. the inversion of aligned images,
2. creating a new image of spots commonly shared by the images, preserving the minimum value of pixels in the same location,
3. automatic identification and manual edition of primary objects (same as protocol of total spots), and
4. the elimination of common spots of each image.

With this, we obtained images of gels with common spots eliminated, so in a next step we performed

5. the identification and edition of primary objects of the exclusive spots of each gel,
6. calculation of metrics for each spot, and finally,
7. the representation of common and exclusive spots for each image.

With this, each image of each condition identified spots present in both conditions (configured to be marked in red), or, exclusive of each gel (blue or green colors in each image).

2.5. Comparison of gels from multiple experimental conditions

In order to compare different profiles of periplasmic proteins in various conditions of antibiotic exposure in *P aeruginosa* AG1, we proceeded to run two machine learning algorithms for clustering: a Principal Component Analysis (PCA) and a Hierarchical Clustering (HC) analysis (Fig. 1 down). To address this, the images were first aligned (as previously described) and then the images were divided into 121 sectors (11 × 11 quadrants) and, given that the location was in coordinates, the counting of spots was made for each of the zones. This information was used to implement the clustering algorithms, which used Euclidean distance for the dissimilarity and default parameters of the Caret package (http://caret.r-forge.r-project.org/) in the R program (https://www.r-project.org/).

3. Results

In order to establish an automatic procedure for the identification of spots of proteins in 2D-GE gels, we first proceeded with the generation of images from experimental assays with the periplasmic proteins of *P aeruginosa* AG1, in conditions with or without antibiotics. Then, we proceeded with the analysis of images, including alignment, identification of total spots and validation, differential identification when comparing pairs of conditions and finally analysis by clustering, as summarized in Figure 1.

To align and compare the protein migration profile in 2D-GE gels, the bUnwarpJ package was used to deform the larger images and align them to a reference. In the case presented in Figure 2A, which starts with two images of different sizes (two clones of the strain in control condition), five points of reference or common denominator are established between the images, which are used by the algorithm to optimize the alignment by calculating a field and network of deformation (Fig. 2B). With this, the larger image is reduced to align and make the spots comparable between conditions (the image was cropped to visualize the distribution, Fig. 2C).

Using the CellProfiler software, two spots recognition protocols were implemented. In the first one, with the identification of total spots, it was established that the optimal conditions were the use of a global algorithm (assuming relatively homogeneous background pixels and other parameters with default values),

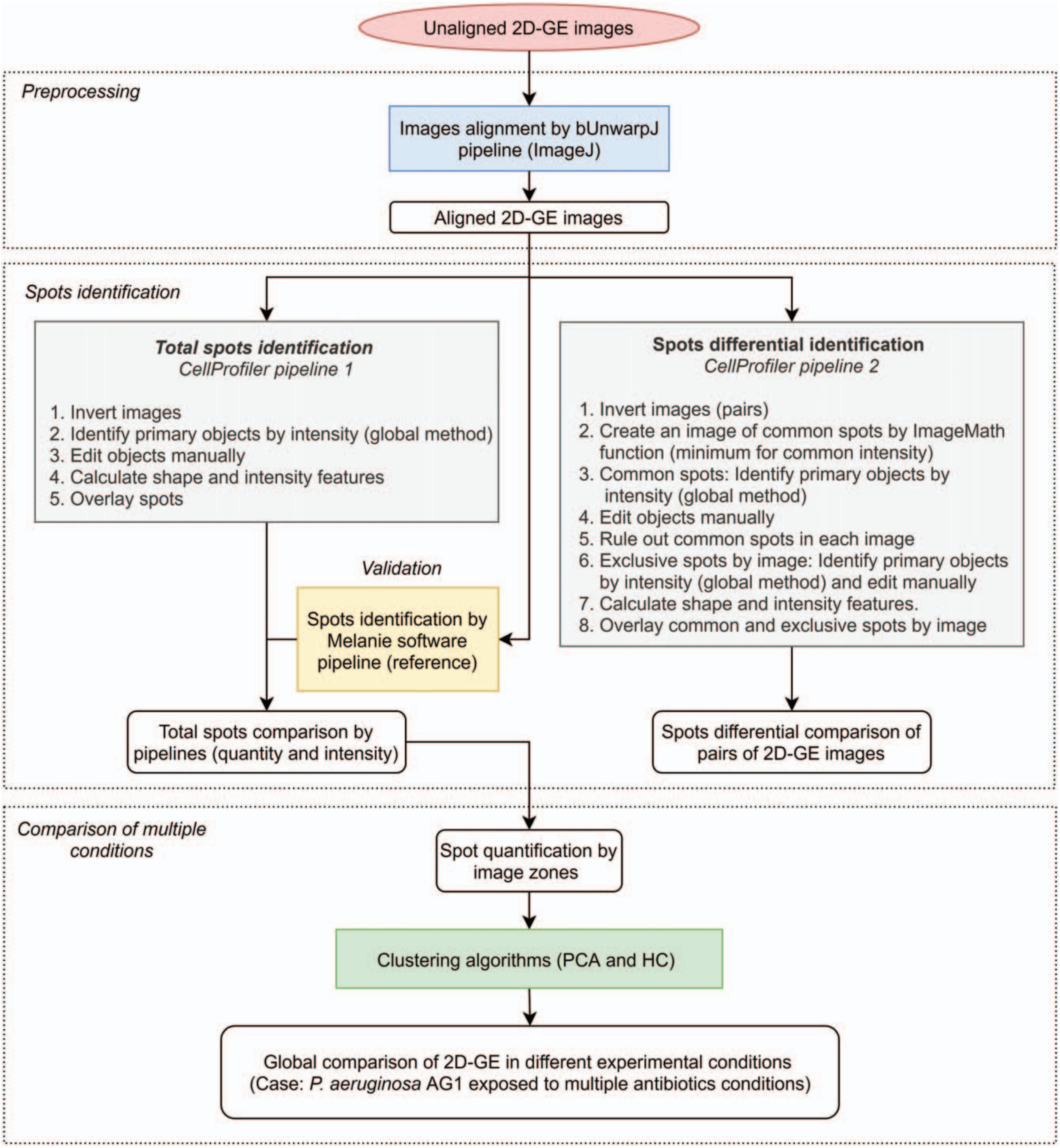

Unaligned 2D-GE images
Preprocessing
Images alignment by bUnwarpJ pipeline (ImageJ)
Aligned 2D-GE images
Spots identification
Total spots identification
CellProfiler pipeline 1
1. Invert images
2. Identify primary objects by intensity (global method)
3. Edit objects manually
4. Calculate shape and intensity features
5. Overlay spots
Spots differential identification
CellProfiler pipeline 2
1. Invert images (pairs)
2. Create an image of common spots by ImageMath function (minimum for common intensity)
3. Common spots: Identify primary objects by intensity (global method)
4. Edit objects manually
5. Rule out common spots in each image
6. Exclusive spots by image: Identify primary objects by intensity (global method) and edit manually
7. Calculate shape and intensity features.
8. Overlay common and exclusive spots by image
Validation
Spots identification by Melanie software pipeline (reference)
Total spots comparison by pipelines (quantity and intensity)
Spots differential comparison of pairs of 2D-GE images
Comparison of multiple conditions
Spot quantification by image zones
Clustering algorithms (PCA and HC)
Global comparison of 2D-GE in different experimental conditions (Case: P. aeruginosa AG1 exposed to multiple antibiotics conditions)

Figure 1. General workflow by image analysis for identifying and comparing spots in 2D-GE gel images.

sizes of 40 to 100 pixels for the objects and the use of intensity to recognize and segment objects. Thus, after the inversion of the image and the recognition of objects, the recognized objects were presented on the original image (Fig. 3A left). When performing the comparison with an automatic protocol with the Melanie program (used as a reference for validation), it was verified that the resolution capacity of the protocol we implemented had the same ability to identify spots (Fig. 3A, right). The number of spots was counted in 124 for both protocols (this value was controlled with the manual edition available both in our protocol and in Melanie pipeline and that includes cases of proteins grouped as a single spots in cases of large spots). Given that the boundaries or edges of recognition of an object varied between protocols, we proceeded to perform a linear regression between the intensity values, determining that the intensity behavior between the algorithms is linear (Fig. 3B).

In a second protocol with the same program, we proceeded with the differential identification of spots when comparing pairs of gels, obtained from two clones of the same strain *P aeruginosa* AG1 in LB medium condition. The identification of objects was

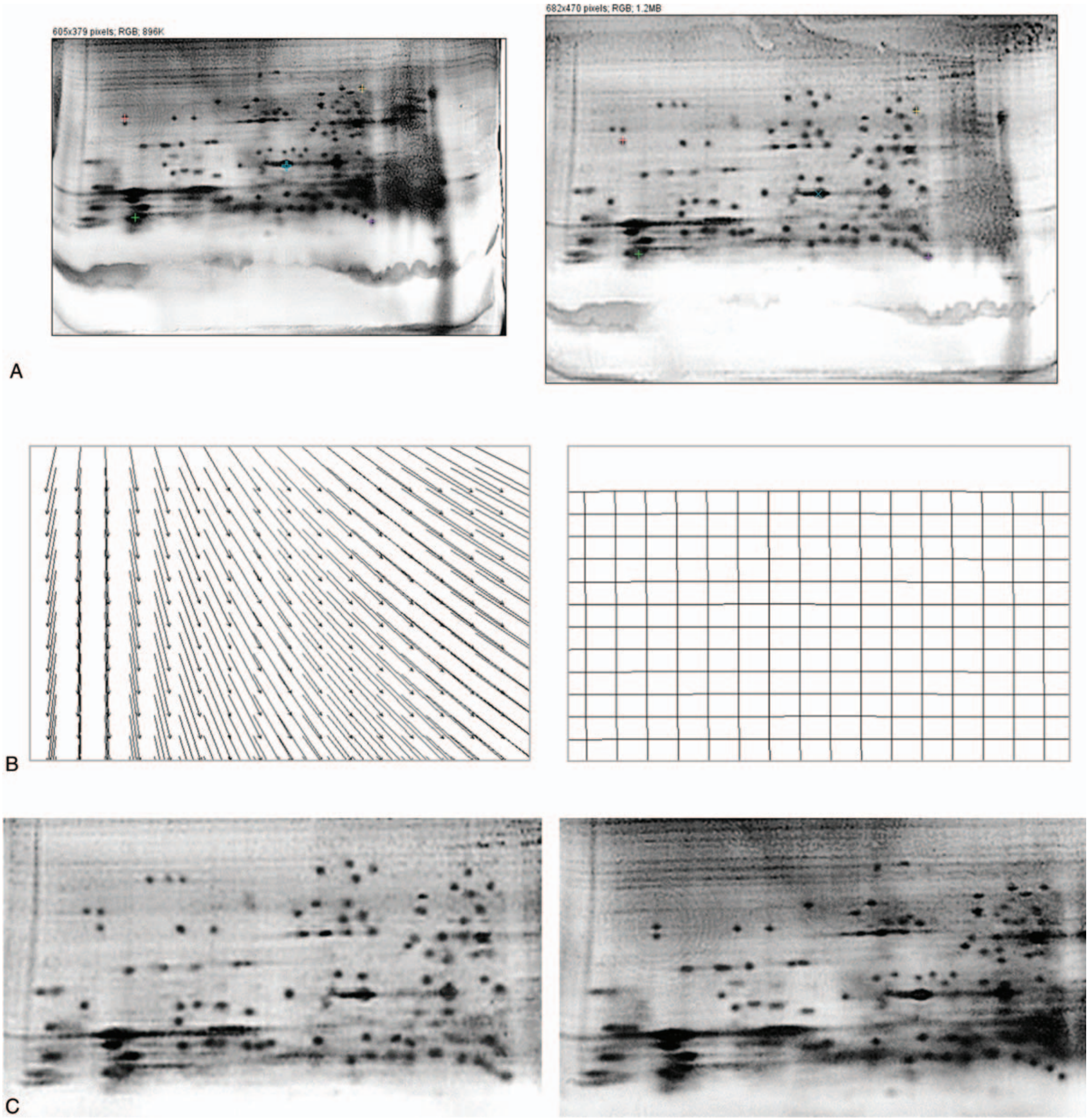


Figure 2. Alignment of 2D-GE gel images by warping method with bUnwarpJ pipeline (gels of proteic profiles from two clones from same strain *Pseudomonas aeruginosa* AG1). (A) Raw images showing differences by size and scale. Color marks define the reference points for warping. (B) Deformation field (left) and deformation grid (right) of larger image to align to the small one. (C) Aligned 2D-GE gel images after warping and cropping (two clones).

done with the algorithm and intensity conditions described for the previous case, both for common spots and exclusive spots. Obtaining common spots was achieved by creating a new image, preserving the lower pixel value for the two images (so if a dot was present in both conditions, the image created would have a high value). Then, the spots were identified and they were labeled as proteins common to both conditions. Using the MaskImage function, the elimination of these common objects was achieved and, in a new recognition for each image, it was possible to identify the exclusive elements of each gel. Using colors, each type of object, common spots (red) or exclusive spots (green or blue) were marked on the images, showing that for this case the majority of proteins were shared by the two clones of the bacteria (Fig. 4B and C).

Finally, with the identification of spots made for each gel in different conditions including antibiotics, we proceeded to the comparison of the protein profiles. First, a division of the images into zones was carried out, and the number of spots was counted. Then, the PCA and HC clustering algorithms were evaluated, obtaining that the profiles given by different antibiotics generate more differences than the concentration of the antibiotic. In the case of the PCA (Fig. 5A), using first two components (with a cumulative variation between both >60%), they show a similar relationship between the control with LB medium and the case of ciprofloxacin. This relationship is maintained when evaluating HC (Fig. 5B), but the relationship between imipenem and its two concentrations shows minor differences. In addition, for this same case, the division by zones shows the sectors of gels with

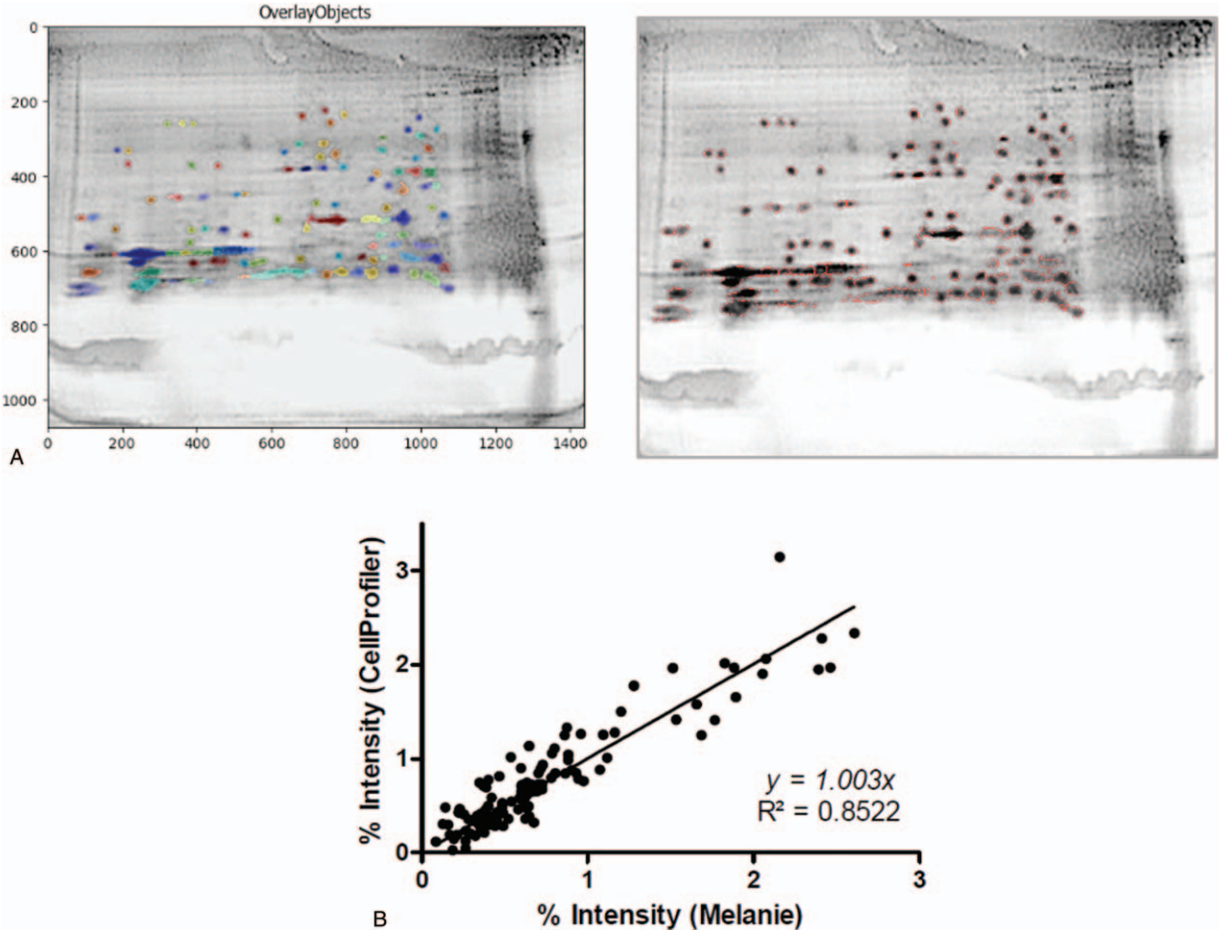


Figure 3. Total spots identification by a CellProfiler pipeline and comparison with Melanie pipeline. (A) CellProfiler pipeline (left) vrs Melanie software (right) for segmentation of objects and final identification after manual edition. (B) Comparison of spots intensity using the CellProfiler pipeline and Melanie software.

similar or very different compartment (potentially useful to select zones for subsequent analysis, see discussion). The HC results are shown with the respective gels in Figure 5C.

4. Discussion

Proteomics is considered an essential field for the systematic analysis of biological systems, an assessment of changes in the abundance of proteins that occur in living organisms and that can be studied at various levels.[3] The two-dimensional gel electrophoresis 2D-GE, separating the proteins according to their isoelectric point and molecular weight, is still used in proteomics laboratories due to the relative ease of implementation in terms of execution and cost, the capacity of solve and visualize miles of proteins in a single run and it is compatible with other high-performance protein techniques, such as mass spectrometry.[1] 2D-GE and subsequent strategies have been implemented in recent studies using bacterial models, including application of protein phosphorylation (phosphoproteomics) in *Bacillus anthracis*[16,17] or biotechnological applications in *Xanthomonas campestris*.[18]

After the experimental phase, the visualization of the proteins is done with the particular stains and gel images are captured, which must be analyzed qualitatively and quantitatively for the extraction of biologically relevant protein information. Of the existing implementations, although there are some investigations in methods of analysis of gels 2D-GE work directly at the level of pixels, most focus on recognizing spots on gel to describe the abundance in each condition.[3] Despite this, there is no protocol for universal or consensus analysis, and multiple limitations are reported in various processing steps.[4] At commercial field, the available programs have additional drawbacks of having a high cost, in addition to many of them are for sale with hardware equipment, which restricts the possibilities of use. In addition, due to its nature, the private code of the implementations is not available, which prevents knowing the details of strategy at the level of algorithms and makes the modification impossible for specific applications. In addition, some limitations of commercial or open access programs include the limited number of images to analyze.

With the aim of implementing and evaluating an image analysis protocol for the recognition of spots in 2D-GE gels images, using open-source software, parallelizable, and applicable to hundreds of images, we obtained experimental data of protein profiles of *P aeruginosa* AG1 under standardized conditions with or without antibiotics. The general protocol was presented in Figure 1. Although it is possible to find variations between runs for the same sample, in our work, we used data from different samples but the same run. Comparison of other protein concentrations, experimental conditions, or

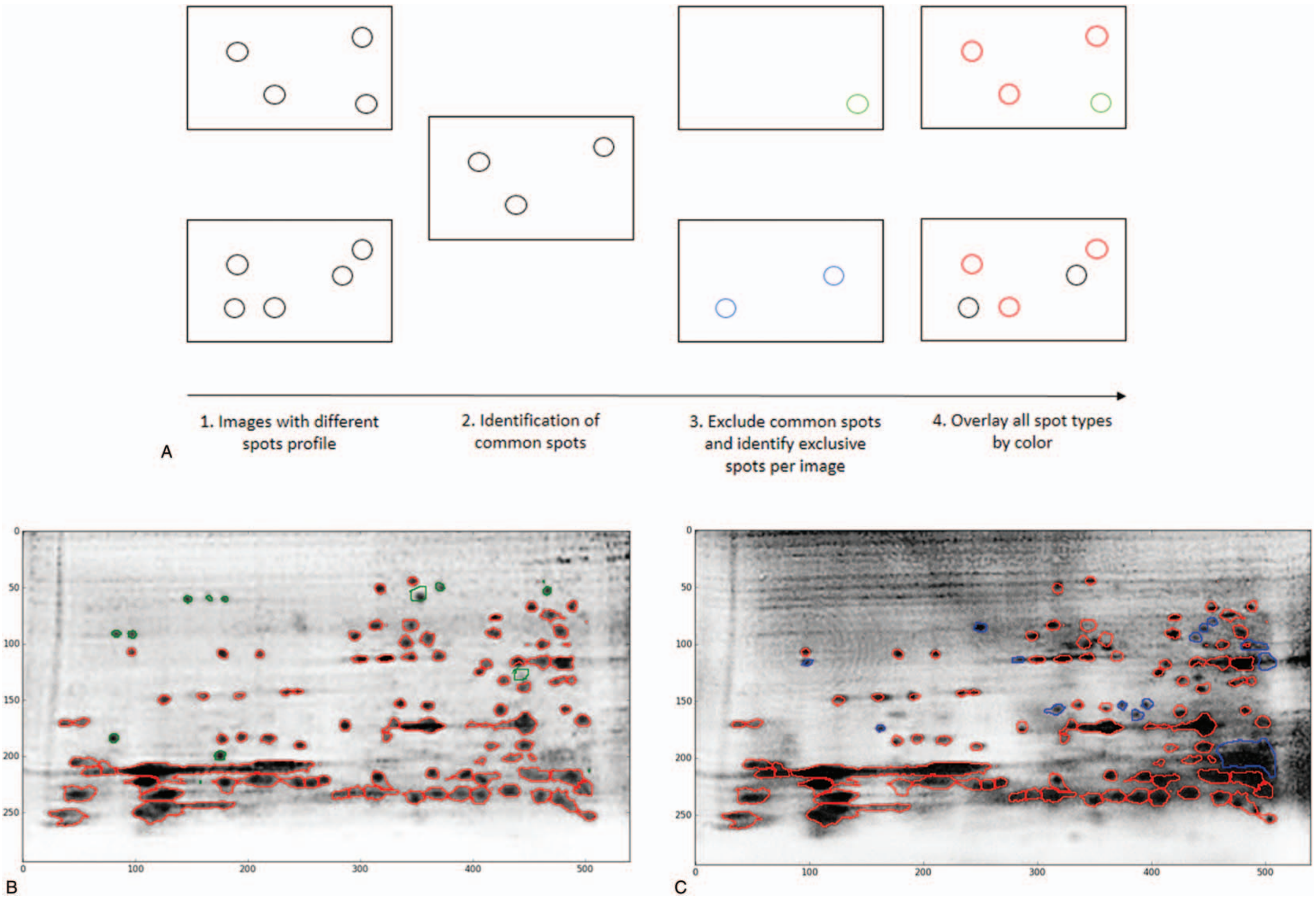

1. Images with different spots profile
2. Identification of common spots
3. Exclude common spots and identify exclusive spots per image
4. Overlay all spot types by color
A
B
C
0
50
100
150
200
250
0
100
200
300
400
500

Figure 4. Spots differential identification and comparison of 2D-GE gel images from two experimental conditions (clones from same strain). (A) General pipeline for identifying common (red, 124 spots) and exclusive spots (blue or green), which was applied to two different proteomic profiles, Clone 1 with 11 exclusive spots (B) or Clone 2 with 14 exclusive spots (C), respectively.

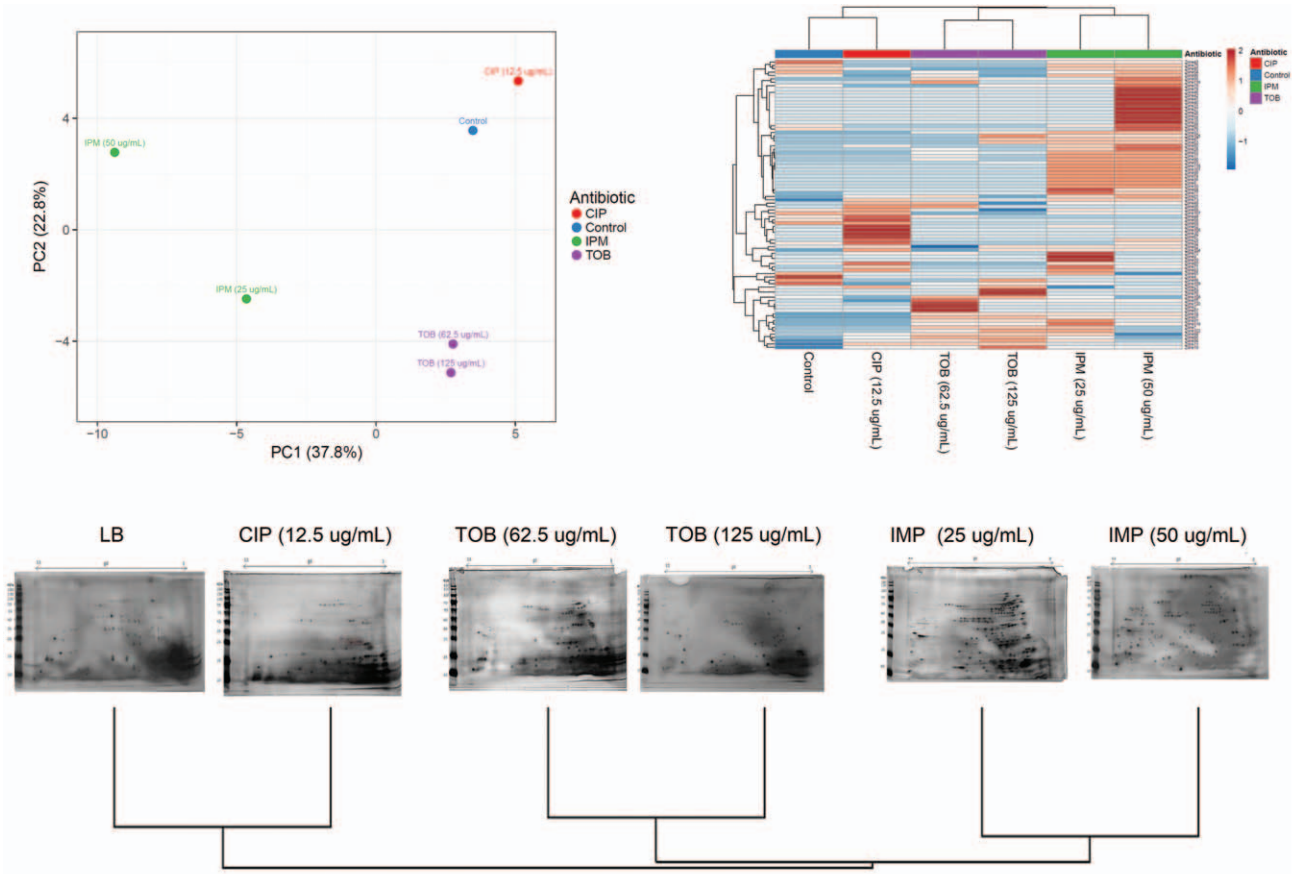


Figure 5. Machine learning approach of clustering analysis for comparing 2D-GE gel images from multiple experimental conditions. (A) PCA algorithm, (B) HC analysis showing zones and spots count, and (C) HC showing images.

replicates are known to produce changes in the proteomic profile, and further analyses are required to study these effects and the performance of our pipeline considering this.

Images were aligned with the bUnwarpJ package in the ImageJ program. This step is required as preprocessing of data since the final performance depends to a great extent on the quality of the images to be processed. This processing includes the alignment of images to match the corresponding protein points of different conditions.[1] In our case, the larger image was adjusted to the smaller one and as an example the case of two protein profiles of two clones of the bacteria was presented in the control condition with LB culture medium (Fig. 2). Although in our final implementation we use 6 images when aligning, the alignment of hundreds or thousands of images is possible using a single reference, as we did in another application with data of cell cultures followed over time, aligning 600 images to the initial image (unpublished data), showing the potential of using this package for the analysis of multiple gel images. Other applications with other types of images show this fact.[9,19,20]

After the preprocessing, we carried out the implementation of two protocols with CellProfiler software. Particular features of this software are discussed below. In a first approach, we recognized total spots (Fig. 3), allowing the counting of spots and the quantification of the area and intensity integrated by each one. Additionally, we compared the performance of this protocol with a pipeline of the commercial software Melanie, showing an equivalent performance when comparing the intensity obtained per object. Due to the fact that in both protocols a module of manual editing of the identification is implemented, the count of elements was intentionally controlled according to expert criteria, for a total of 124 spots. Similar results in performance have been previously reported when an analysis with ImageJ was compared with Melanie,[4] but as mentioned before, with limited number of images to be processed. In the case of CellProfiler, automation is an essential component from its design, as well as the option to parallelize in computer clusters.[14]

In a second protocol (Fig. 4), we implemented a procedure to differentially recognize the expression of proteins in pairs of experimental conditions, allowing us to identify common and exclusive spots of experimental conditions. To do this, our strategy was based on the construction of a new image using the minimum value of pixels of the two images aligned and inverted, using the MathImage function of the program. In this way common spots were preserved. The recognition by segmentation based on intensity allowed the identification of objects, which were later excluded in each. In a second phase, the remaining spots were recognized in each image, to then differentially represent the edges of the shared and exclusive spots.

To the best of our knowledge, there are no approximations that allow the display of common and exclusive spots automatically, given that it is regularly done manually. This information is used to identify proteins differentially expressed in the conditions studied. However, our approach is very robust considering only the presence or absence of spots, and true cases of differential expression with significant changes in intensity are not contemplated, so we consider that this protocol allows the differential identification of spots, but not properly the differential protein expression analysis. This last type of analysis is carried out by commercially available packages, but they are mainly based on the intensity and area, and due to the preprocessing of the image in terms of image contrast, dimensions and other modifications, the normalization and transformation of data it remains a challenge.[1]

Regarding the CellProfiler program and its convenience for this implementation, this software offers the management of hundreds of thousands of images, freely available and with an open and flexible code platform to share, test, and develop new methods by experts in image analysis. In addition, it offers an easy-to-use interface and the possibility of implementing in computational clusters.[8] In addition, due to its nature of automation, the program is capable of handling hundreds of thousands of images, which high performance infrastructure is required for massive analyzes, such as those implemented at the omics level. Although many of the applications of the CellProfiler program are formulated for cells, other applications have been implemented at the level of recognition of complete organism in images, such as the parasite *Caenorhabditis elegans*,[21] or complete tumors, colonies of yeast or bacteria, and other images of biological origin, as evidenced in the Educational Modules section of the web page (https://CellProfiler.org/outreach/).

In contrast, we finally carried out the implementation of a machine learning analysis to compare protein profiles from gel images using PCA and HC clustering algorithms. This type of strategy has been previously implemented with PCA and heuristic clustering algorithms,[22–24] as well as supervised classification algorithms to separate conditions, including Support Vector Machine.[23] Other approaches have implemented comparison modules using directly the properties of intensity, brightness, and contrast of images to contrast with databases,[25] or, other levels of proteomic analysis, such as mass spectrometry.[26] Regarding the methodology used in our case of division by zones and grouping of regions with a similar profile, this strategy can be used to make subsequent decisions of work in proteomics laboratories, where the task after the gels is the selection of spots and continue with identification applications with techniques such as HPLC or mass spectrometry.[1]

In the biological aspect according to the results obtained when comparing 6 experimental conditions with *P aeruginosa* AG1 bacteria with or without antibiotic, it was possible to identify the relationships between the total protein expression profiles. Both with the results of the analysis by PCA and by HC, it is concluded that there is greater similarity between the profiles obtained for the same antibiotic at different concentrations, and that they are separated from the conditions of other antibiotics, congruent according to the mechanisms of action of each type of antibiotic. In the case of ciprofloxacin, its profile was separated to a greater degree from the other antibiotics and was grouped with the control with LB medium.

Because the bacterial strain *P aeruginosa* AG1 is resistant to those antibiotics, this information and subsequent analysis at the proteomic level, together with other genomic, transcriptomic, and phenomic analyzes that we are conducting, will allow us to obtain new findings of the biological relationships to molecular level that provide insights to begin to explain the mechanisms of tolerance to antibiotics and the modulation of biological processes in response to cellular stress.

5. Conclusions

In the context of proteomics and its importance for the study of different biological conditions, our implementation of the image analysis of gels 2D-GE offers an opportunity to continue with studies of analysis of protein profiles. Using the open-source software, CellProfiler (and bUnwarpJ for preprocessing), we achieved the alignment of images, the identification of spots and the final comparison of protein profiles. These workflow also allow analyze a large number of images automatically as well as enabling the parallelization in computational clusters to counteract the complexity of processing this type of data. Regarding the biological meaning, exposure to ciprofloxacin in *P aeruginosa* AG1 showed a similar pattern to control without treatment, and other groups were generated according to the antibiotic class. This information will be integrated with other molecular analyses using antibiotics in this multiresistant strain to gain insights regarding the mechanisms of tolerance to antibiotics and the modulation of biological processes in response to cellular stress.

Acknowledgments

We thank the student Daniel Solano Alvarado for his collaboration as an assistant in various activities of the project.

Author contributions

Conceptualization: Fernando García.
Formal analysis: Jose Arturo Molina-Mora, Diana Chinchilla-Montero, Carolina Castro-Peña.
Methodology: Jose Arturo Molina-Mora, Diana Chinchilla-Montero, Carolina Castro-Peña, Fernando García.
Software: Jose Arturo Molina-Mora.
Supervision: Fernando García.
Visualization: Jose Arturo Molina-Mora.
Writing – original draft: Jose Arturo Molina-Mora.
Writing – review & editing: Jose Arturo Molina-Mora, Diana Chinchilla-Montero, Carolina Castro-Peña, Fernando García.

References

[1] Goez MM, Torres-Madroñero MC, Röthlisberger S, et al. Preprocessing of 2-dimensional gel electrophoresis images applied to proteomic analysis: a review. Genomics Proteomics Bioinformatics 2018;16:63–72.
[2] O'Farrell PH. High resolution two-dimensional electrophoresis of proteins. J Biol Chem 1975;250:4007–21.
[3] Silva TS, Richard N, Dias JP, et al. Data visualization and feature selection methods in gel-based proteomics. Curr Protein Pept Sci 2014; 15:4–22.
[4] Natale M, Maresca B, Abrescia P, et al. Image analysis workflow for 2-D electrophoresis gels based on imageJ. Proteomics Insights 2011;4:37–49.
[5] Abdallah C, Dumas-Gaudot E, Renaut J, et al. Gel-based and gel-free quantitative proteomics approaches at a glance. Int J Plant Genomics 2012;2012.
[6] Dowsey AW, Morris JS, Gutstein HB, et al. Informatics and statistics for analyzing 2-D gel electrophoresis images. Methods Mol Biol 2010; 604:239–55.
[7] Abramoff MD, Magalhães PJ, Ram SJ. Image processing with Image. J Biophotonics Int 2004;11:36–42.
[8] Lamprecht M, Sabatini D, Carpenter A. CellProfilerTM: free, versatile software for automated biological image analysis. Biotechniques 2007; 42:71–5.

[9] Schindelin J, Rueden CT, Hiner MC, et al. The ImageJ ecosystem: an open platform for biomedical image analysis. Mol Reprod Dev 2015;82:518–29.

[10] Cirz RT, O'Neill BM, Hammond JA, et al. Defining the Pseudomonas aeruginosa SOS response and its role in the global response to the antibiotic ciprofloxacin. J Bacteriol 2006;188:7101–10.

[11] Toval F, Guzmán-Marte A, Madriz V, et al. Predominance of carbapenem-resistant Pseudomonas aeruginosa isolates carrying blaIMP and blaVIM metallo-β-lactamases in a major hospital in Costa Rica. J Med Microbiol 2015;64:37–43.

[12] Arganda-carreras I, Sorzano COS, Kybic J, *et al.* bUnwarp: Consistent and Elastic Registration in ImageJ. Methods and Applications. *Image (Rochester, NY)*; 2006.

[13] Collins T. ImageJ for microscopy. Biotechniques 2007;43(S1):S25–30.

[14] Kamentsky L, Jones TR, Fraser A, et al. Improved structure, function and compatibility for CellProfiler: modular high-throughput image analysis software. Bioinformatics 2011;27:1179–80.

[15] Ames GF, Prody C, Kustu S. Simple, rapid, and quantitative release of periplasmic proteins by chloroform. J Bacteriol 1984;160:1181–3.

[16] Virmani R, Sajid A, Singhal A, et al. The Ser/Thr protein kinase PrkC imprints phenotypic memory in Bacillus anthracis spores by phosphorylating the glycolytic enzyme enolase. J Biol Chem 2019;294:8930–41.

[17] Arora G, Sajid A, Virmani R, et al. Ser/Thr protein kinase PrkC-mediated regulation of GroEL is critical for biofilm formation in Bacillus anthracis. Npj Biofilms Microbiomes 2017;3:7.

[18] Schulte F, Hardt M, Niehaus K. A robust protocol for the isolation of cellular proteins from Xanthomonas campestris to analyze the methionine effect in 2D-gel experiments. Electrophoresis 2017;38:2603–9.

[19] Kindle LM, Kakadiaris IA, Ju T, et al. A semiautomated approach for artefact removal in serial tissue cryosections. J Microsc 2011;241:200–6.

[20] Seiler C, Reyes M. Displacement Vector Field Regularization for Modelling of Soft Tissue Deformations; 2008. Available at: https://christofseiler.github.io/MasterThesis.pdf. [Accessed June 4, 2019].

[21] Moy TI, Conery AL, Larkins-Ford J, et al. High throughput screen for novel antimicrobials using a whole animal infection model. ACS Chem Biol 2009;4:527.

[22] Appel R, Hochstrasser D, Roch C, et al. Automatic classification of two-dimensional gel electrophoresis pictures by heuristic clustering analysis: a step toward machine learning. Electrophoresis 1988;9:136–42.

[23] Supek F, Peharec P, Krsnik-Rasol M, et al. Enhanced analytical power of SDS-PAGE using machine learning algorithms. Proteomics 2008;8:28–31.

[24] Castillejo MÁ, Fernández-Aparicio M, Rubiales D. Proteomic analysis by two-dimensional differential in gel electrophoresis (2D DIGE) of the early response of Pisum sativum to Orobanche crenata. J Exp Bot 2012;63:107–19.

[25] Kush A, Raghava GPS. AC2DGel: analysis and comparison of 2D Gels. J Proteomics Bioinform 2008;01:043–6.

[26] Kelchtermans P, Bittremieux W, De Grave K, et al. Machine learning applications in proteomics research: how the past can boost the future. Proteomics 2014;14:353–66.

CHAPTER 4

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

Molina-Mora, J., Montero-Manso, P., Batán, R. G., Sánchez, R. C., Fernández, J. V., & García, F. (2020). A first Pseudomonas aeruginosa perturbome: Identification of core genes related to multiple perturbations by a machine learning approach. BioRxiv, 2020.05.05.078477. https://doi.org/10.1101/2020.05.05.078477

Summary

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.

A first *Pseudomonas aeruginosa* perturbome: 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 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 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 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 data partitioning, classification approaches and metrics. Altogether, this work offers a different and robust framework to select genes using a machine learning approach.

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

1. 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].

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

Several studies have used machine learning algorithms at the transcriptomic level to recognize gene expression patterns [2,14–16]. However, for many common biological contexts, applicability and utility of these machine learning approaches have not been fully explored and utilized [17], 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 approach to identify sets of genes that correlate with multiple culture media, but without other conditions [18].

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

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

Thus, the aims for 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.

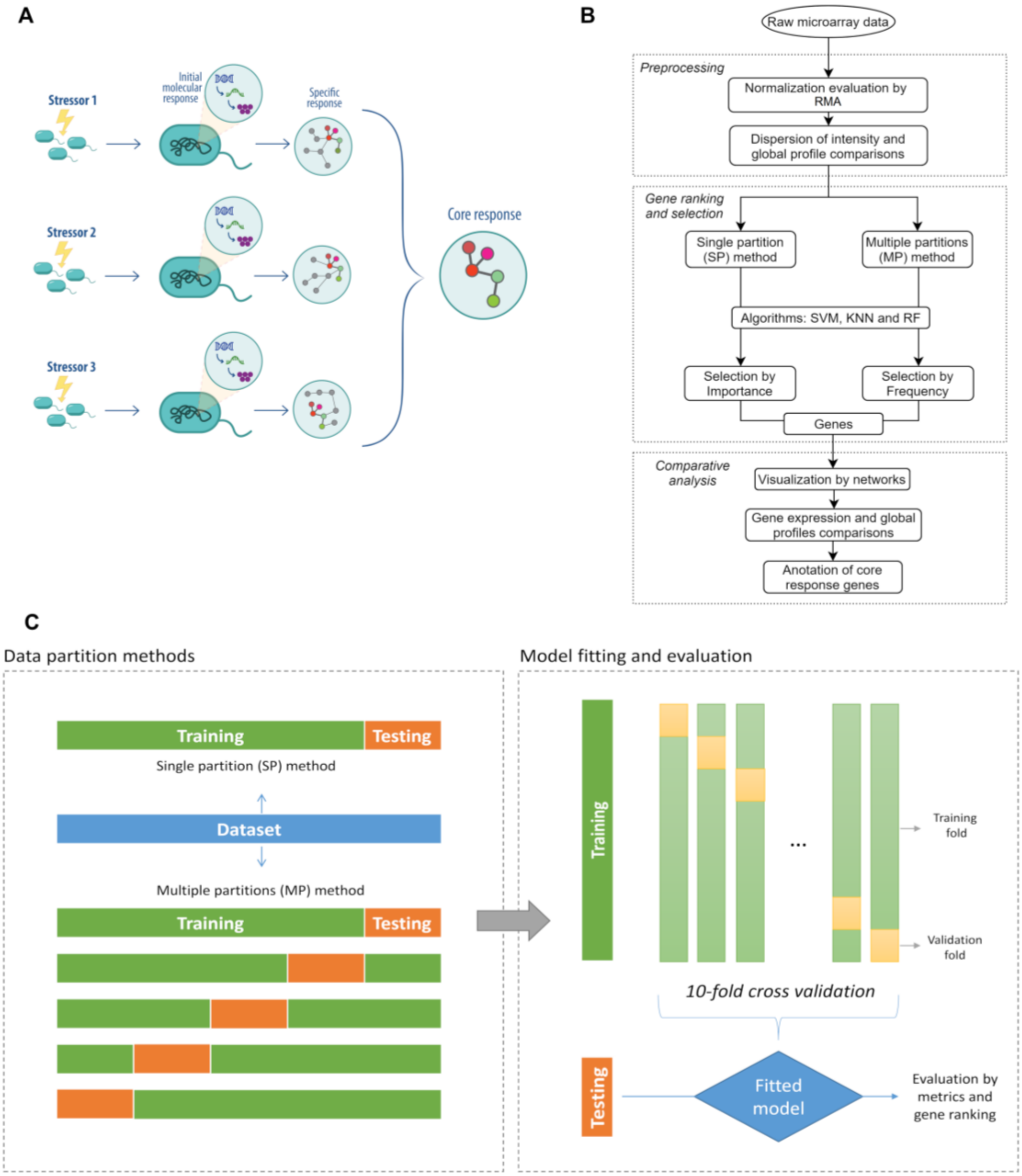

A
Stressor 1
Initial molecular response
Specific response
Stressor 2
Stressor 3
Core response
B
Raw microarray data
Preprocessing
Normalization evaluation by RMA
Dispersion of intensity and global profile comparisons
Gene ranking and selection
Single partition (SP) method
Multiple partitions (MP) method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

Figure 1. General pipeline to identify 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.

2. Materials and Methods

Overall methodology of the study is presented in Figure 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.

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

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

2.2 Pre-processing and comparison of global transcriptomic profiles

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

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

In order to robustly select a number of genes that could separate experimental conditions (controls and perturbations) and to identify the core response of *P. aeruginosa*, two approaches of feature selection protocols were implemented, as detailed below.

2.3 Gene ranking and selection by Single Partition SP method

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

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

As similarly reported [11], to evaluate the number of genes that should be selected in the top group (the first K ranked genes, as candidates for the core response by each algorithm), multiple classification models were systematically run, starting only with the highest-score ranked gene. In brief, after the training with one gene, model performance was evaluated by calculating metrics using the testing dataset. Then, 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 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, selection of top K genes between algorithms were compared by metrics and 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 1-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 training set and remaining for 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 top K of ranked genes (K value determined using the criteria described in the SP method), new set of training/testing sets were used for evaluating 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 *replicate* function, starting with a new random partition and finishing with the list of the K genes and the metrics associated to that partition. Finally, for each algorithm, full data of all the runs were analyzed for determining frequency of the appearance of genes (*table* function) and calculating average and dispersion of metrics across all the 100 runs. Definitive list of the K more frequent genes was established for each algorithm after this comparison by frequency.

2.5 Identification of core response genes

After selection of top K genes in each algorithm by SP and MP methods, comparison of genes was done using Venn diagrams in order to identify all the candidate genes using Venn-tool (http://bioinformatics.psb.ugent.be/webtools/Venn/). Genes identified by at least four algorithms were considered part of the perturbome (this guarantees that a gene was identified by the two methods and at least by two different classification algorithms).

Gene relationships were represented using molecular networks using a large scale model (using a top-down systems biology approach). The network was built using protein-protein interaction (PPI) graphs in PseudomonasNet database (www.inetbio.org/pseudomonasnet/). Network was downloaded and visualized using the Cytoscape software (https://cytoscape.org/).

2.6 Description and comparison of core response genes

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

Second, using the PseudomonasNet database, a small world network was built and then exported into Cytoscape software with the genes of the core response. The information of topological features (including connectivity) and expression levels of kmedoids were incorporated into different versions of the network.

Third, a classic analysis of differentially expressed genes (DEGs, $p<0.05$) was implemented in R with Limma package (https://www.rdocumentation.org/packages/limma/versions/3.28.14) using empirical Bayes moderated t-statistic (eBayes) with Benjamini and Hochberg method for *p* value correction [40]. This led us to compare our results with a classical approach for gene expression.

Finally, in order to give biological interpretation to the selected genes, to determine levels of expression reported in databases and to study metabolic pathways involved under each perturbation, an exhaustive annotation was made using the databases PseudomonasDB (gene ontology), GEO database and particular literature. This information was integrated with the results obtained by all the analysis and the DEGs in order to fully describe the genes that make up the core response or perturbome of *P. aeruginosa* PAO1.

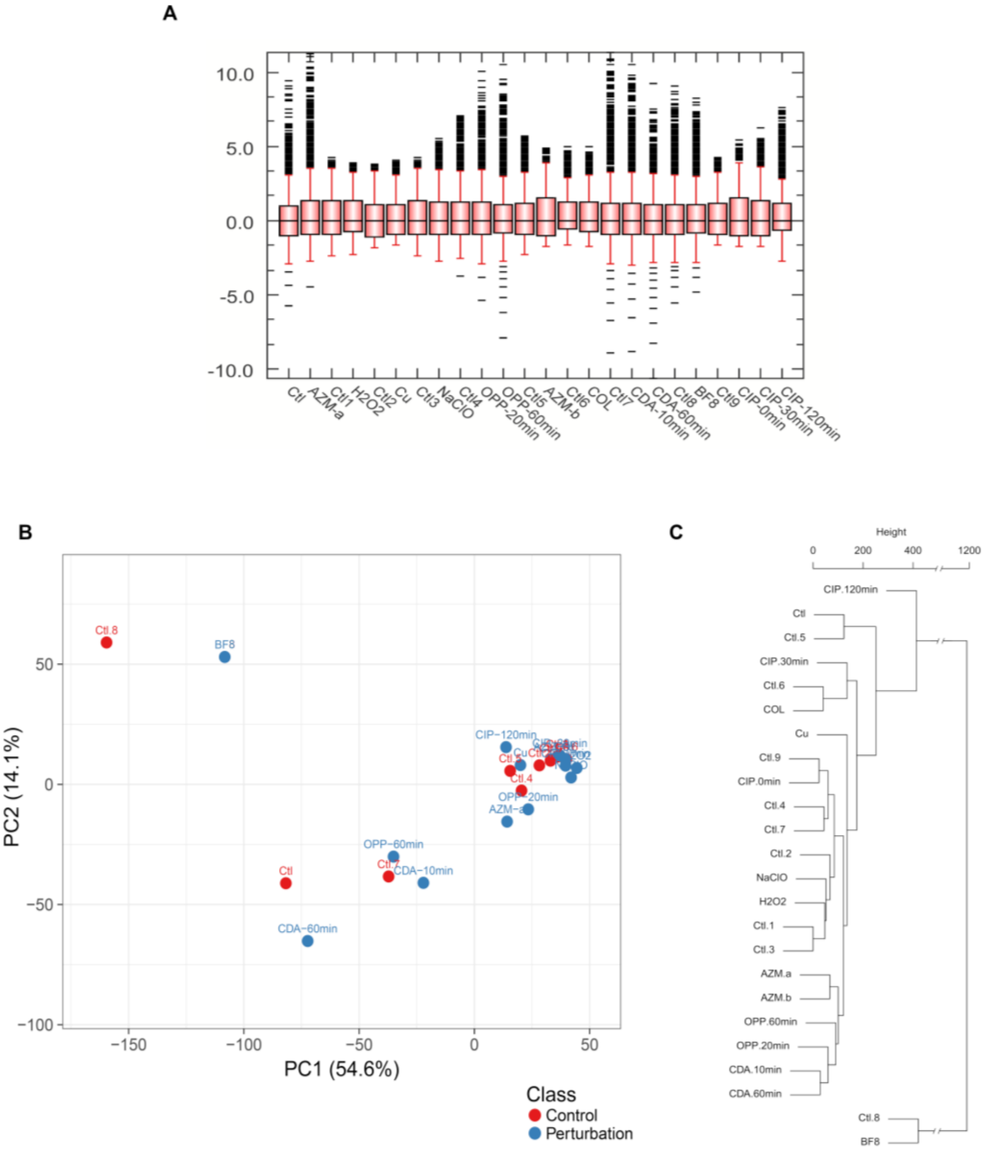


Figure 2. Normalization and comparison of samples by global profiles using all genes. (A) Dispersion of intensities of samples, showing similar distribution between samples. (B-C) Global profiles were compared by both PCA and HC clustering algorithms, showing mixed patterns between classes.

3. Results

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

A total of 71 samples of 10 controls and 14 perturbations were considered for the study, with comparable expression levels (Figure 2-A). Global transcriptomic profiles (all 5549 genes) were compared by both PCA and HC algorithms, revealing a mixed pattern (no separation) between perturbations and controls. Two samples (BF8 and Control 8) resulted with extreme global profiles (Figure 2 B-C). Two methods using machine learning (SP and MP) were implemented in order to 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 Figure 3, and supplementary Figure S1 for SVM and KNN.

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 (Figures 3-A and supplementary Figure S1-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. 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 S1). 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 (Figures 3-B and supplementary Figure S1-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 Figures 3-C for RF and Figure S1-E for SVM/KNN.

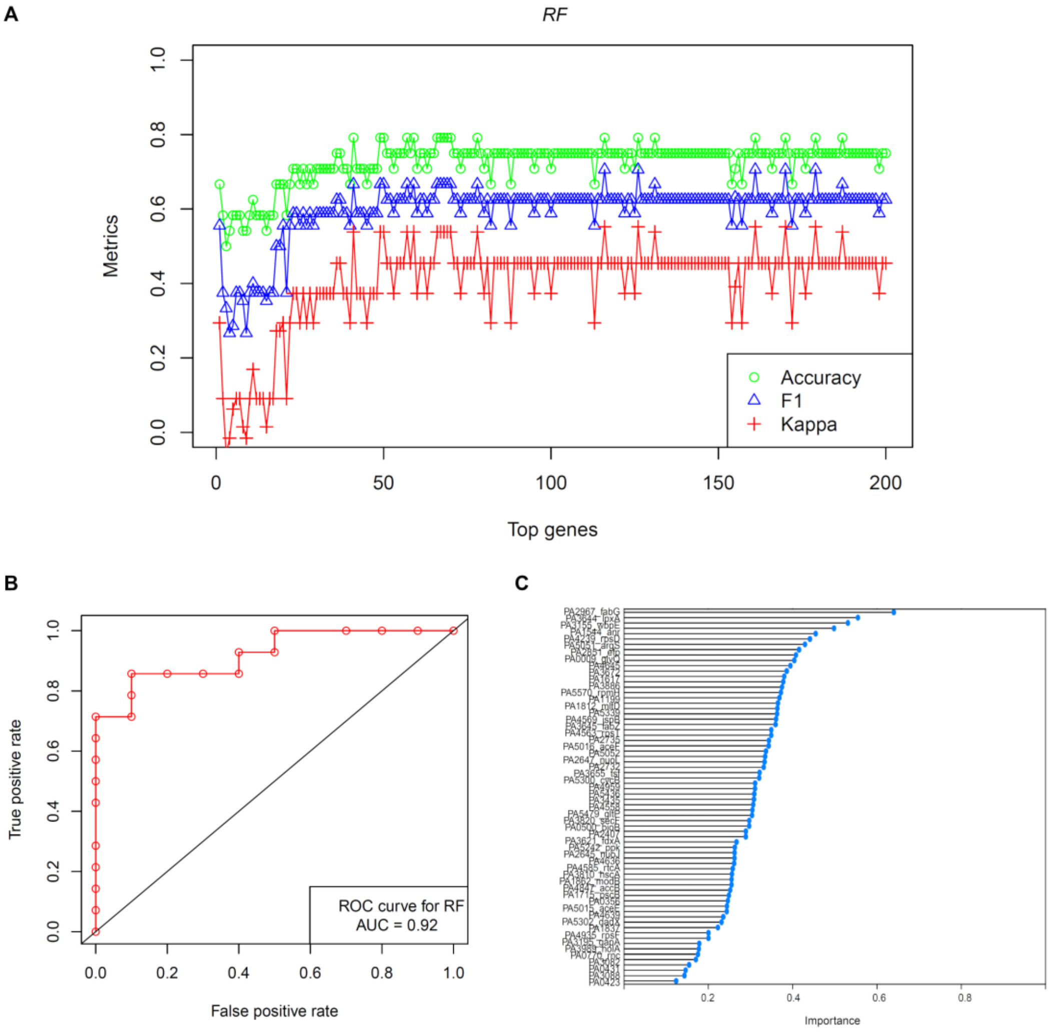


Figure 3. **Evaluation of SP method for gene ranking by importance, case for RF algorithm.** (A) Accuracy, F1 and kappa values after iterations of classification with the first top 200 genes (adding genes 1-to-1). (B) ROC plot using selected top 56 genes for evaluation of performance of the algorithm. (C) Ranking and importance value of top 56 genes. Similar results are shown for SVM and KNN algorithms in supplementary Figure S1.

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 Figure 4-A (SVM) and supplementary Figure S2 (KNN and RF) and metrics for all 100 iterations are presented in Table S1, Figure 4-B and supplementary Figure S2. 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.

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 (Figure 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 118 genes on large scale molecular networks is presented in Figure 5-A. Results show that selected genes are connected but they do not establish a defined cluster. These genes seem to be associated to different subgroups of highly connected genes.

Final version of the core perturbome components was established by selecting genes recognized for at least four algorithms. A total of 46 genes were finally associated to core response. In the representation as small world network, relationships between the 46 genes revealed different topological patterns of connectivity between molecules, being *nuoC* and *nuoF* the ones with higher connectivity (connection degree). In addition, only six genes had no connections between them (Figure 5-C).

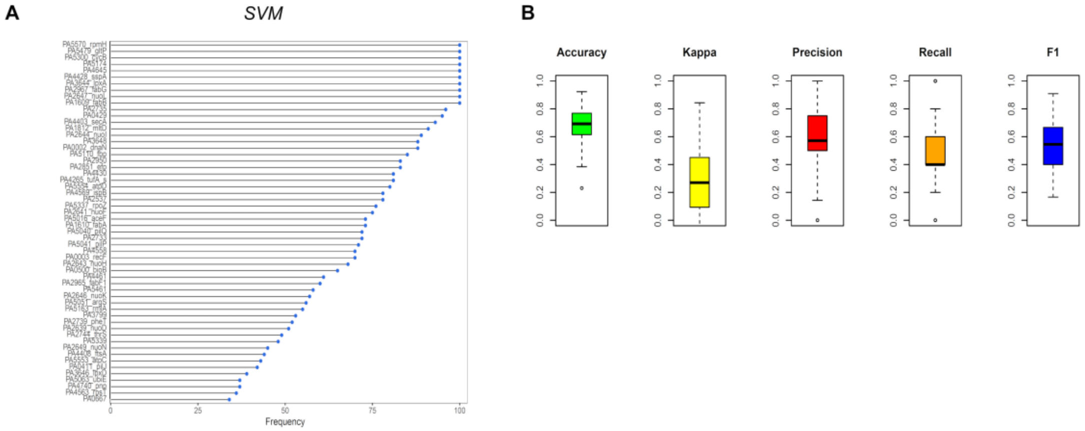


Figure 4. **Evaluation of MP method for gene ranking by frequency, case for SVM algorithm.** (A) Ranking of top 56 genes by frequency after iterations of 100 data partitions and classification model fitting. (B) Dispersion of metrics across 100 iterations. Similar results are shown for KNN and RF algorithms in supplementary Figure S2.

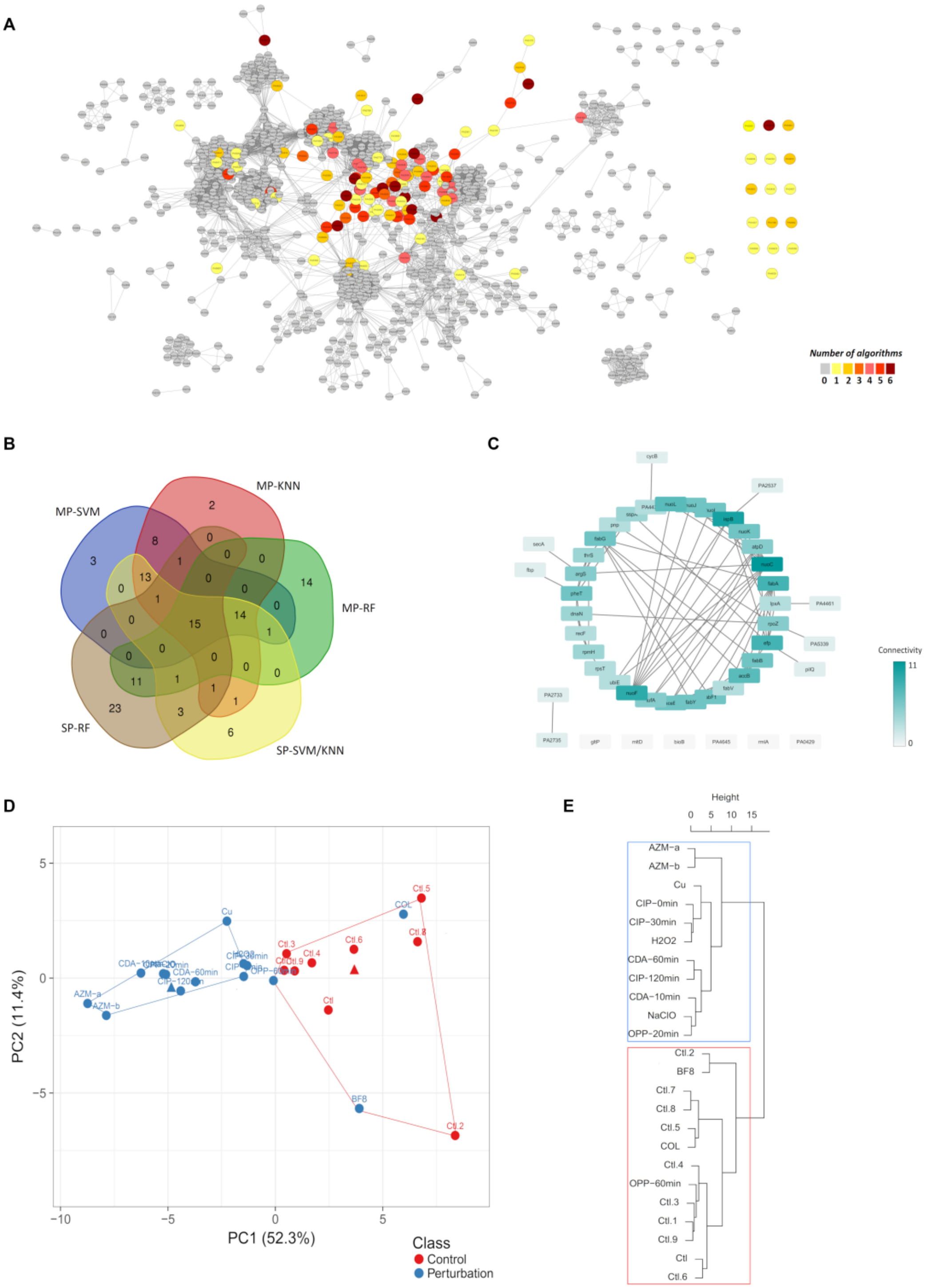


Figure 5. Identification, systems level description and global profiles comparison given by core response genes. (A) Distribution and number of algorithms of preliminary 118 selected genes on a basal large scale network of functional associations in *P. aeruginosa.* (B) Set comparison of genes given by classification algorithms of SP and MP methods, with a final selection of 46 genes which were identified by at least 4 algorithms. (C) Small world network showing relationships between the 46 core response genes and connectivity metric for each gene. (D-E) Global profile comparison of samples by both PCA and HC clustering algorithms, showing separation of conditions. For reference, centroids of each cluster were plotted as triangles in each cluster. More details are shown in supplementary Figure S3.

3.2 Comparisons of core response genes show separation of global transcriptomic profiles according to experimental conditions and biological functions related to tolerance to stress

After the selection of the core response genes, comparisons between controls and perturbations were done to characterize these genes by global profiles. First, clustering algorithms to compare global profiles were run using the 46 genes of the perturbome. In the case of PCA (Figure 5-D), samples distribution let to differentiate controls and perturbations. A Kmedoid algorithm (k = 2) was able to identify two clusters enriched by samples of each condition: one consisting entirely of perturbation samples (11 samples, blue color) and the other mostly by control samples (10 controls and 3 disturbances, red color). In the case of the blue cluster, the kmedoid was sample CIP-120min, meanwhile control 6 was selected for the another cluster.

For the case of HC (Figure 5-E), the same distribution of samples was obtained. Supplementary analysis of gene expression was included by comparing levels for all core response genes (Figure S3-A) and comparing expression levels of the kmedoids on the small world network (Figure S3-B-C).

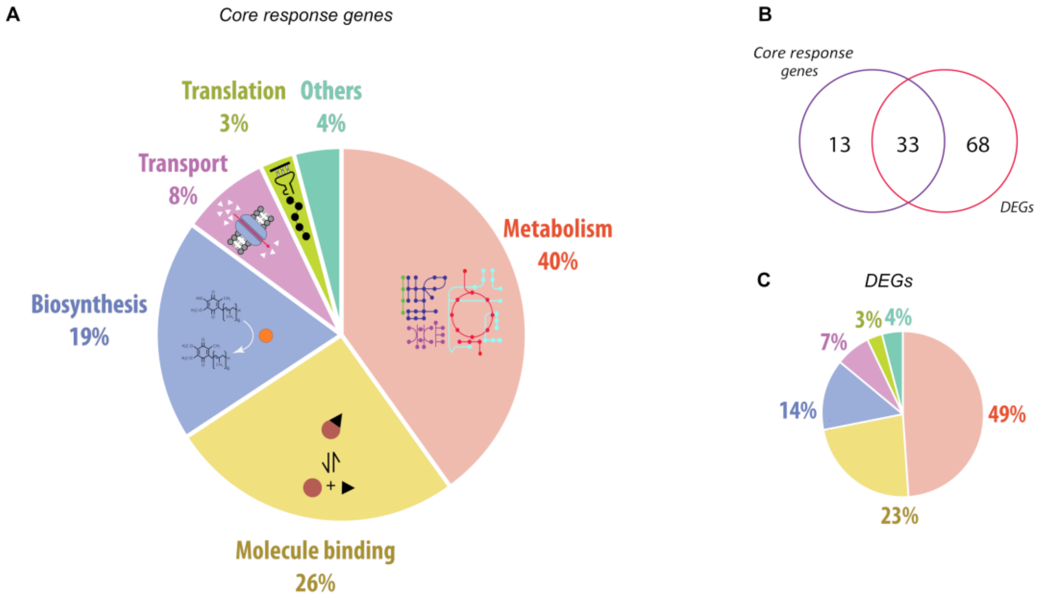


Figure 6. Annotation of core response genes and comparison with a differential expression analysis. (A) General annotation profile of identified genes by core response genes showing associated biological processes. (B) Comparison of identified genes by our machine learning approach and DEGs lists. (C) General annotation of DEGs showing similar profile than our approach. Specific annotation per gene is shown in supplementary Table S2.

Gene annotation revealed that biological processes related to most of the genes are metabolism, molecule binding and biosynthesis (Figure 6-A). Specific information about participation in processes of DNA damage response, DNA repairing and response to general/specific stimuli was also searched for each gene, showing that most of genes includes participation in such processes. Literature support shows variable patterns of expression, depending on the disturbance as shown in supplementary Table S2. Finally, in order to compare the results of the machine learning strategy with another approach, a differential expression analysis was run. A total of 101 DEGs were identified, which 33 were shared by the core response (Figure 6-B). This means that 72% of core response genes were also identified by another single and independent method. Annotation profile of DEGs (Figure 6-C) showed a similar pattern as our machine learning approach.

4. Discussion

Living organisms face external and internal conditions that compromise cellular functions at molecular, metabolic and structural levels, disrupting their homeostasis [41,42]. Cell stress response is crucial for organismal survival and complex networks are usually involved in the molecular mechanism related to tolerance [1]. However, few studies have identified central and possible universal regulation of the response to multiple disturbances, a concept termed as perturbome [4,5]. Common molecular response was previously reported as a network of common set of genes and pathways that can be generically associated to multiple perturbations in plants [7], pathogenic bacteria [3] or cell lines models [4], and others.

In our approach using *P. aeruginosa,* but applicable to other organisms, we hypothesized the existence of a set of core genes regulating the response to stressors in a generic sense of different pathways. *P. aeruginosa* has a high proportion (about 5%) of its genome dedicated to regulatory mechanisms, probably explaining its adaptability to such a broad range of growth conditions [25]. Since strain *P. aeruginosa* PAO1 is a clinical isolate with a profile of multiresistance to many antibiotics [43], characterization of molecular mechanisms involved in the tolerance to stressors in this strain could eventually help to modulate sensitivity and overcome resistance. Exhaustive integration of -omics data and network analysis are required in order to clarify the molecular mechanisms related to stress conditions and eventually use them for modulating cell response.

4.1 Insights of algorithms to identify core response genes or perturbome

In our study, initial global transcriptomic profiles showed mixed patterns between samples according to their experimental condition. Because its complexity, raw microarray data showed noise and redundant information that can explain the poor resolution of this clustering algorithms to identify classes [20]. Thus, a feature (gene) selection analysis was implemented for not only identifying genes associated to multiple perturbations in *P. aeruginosa,* but also to improve performance of predictive/descriptive models capable of separate control and perturbation categories.

In our case, these potential patterns were investigated using a robust machine learning approach by implementing six protocols, using SP and MP methods and SVM, KNN and RF algorithms in each case. This was a critical step because gene selection from microarray data is complex *per se*. Feature reduction remains as a challenging task in transcriptomic studies because thousands of genes to select from, and it introduces an additional layer of complexity in the modelling task [12,44]. To avoid bias and overfitting, implementations of diverse strategies of data partitioning such as bootstrapping, random partitions and cross validation are recommended. In general, these methodologies can robustly minorate the influence from noise, outliers, absence of ground truth sets, and to reduce variance [2,24,45].

The single partition SP method consisted of a particular and invariable data for training (with internal cross validation) and another to test, and it is probably the most common approach used in machine learning. In the case of the multiple partitions MP method, 100 random partitions of dataset were run. MP method had a dual consideration when splitting data (multiple partitions and the internal cross validation). This method can be considered as an *ensemble based on different data partitioning*, as it had been previously proposed [23]. Datasets were divided using multiple random partitioning procedures and then genes were ranked. After all runs, a final feature subset is determined by calculating the frequency of features in all the runs [23]. A equivalent approach was implemented by Pai and collaborators to classify gene expression data in a cancer model [46].

However, one possible problem with MP approach is that cross validation results may depend on similarity of testing and training sets. A classification prediction method is only expected to learn how to predict on unseen samples that are drawn from the same distribution as training samples [24,45], and MP method could violate this assumption. SP method guarantees this because was built always using a replicate for each perturbation and control. Thus, both methods SP and MP are required to robustly select features.

On the other hand, many algorithms for dimension reduction have been proposed [6,19,20] but no standard machine learning algorithm can be selected due multiple evaluations results on completely variable metrics associated to performance [26]. Many studies have shown that SVM, KNN, and RF generally outperform other traditional supervised classifiers [17,26,47]. In our case, a variable pattern was found for different metrics in all evaluated implementations. For example, based on accuracy SVM for SP method and RF for MP method resulted in higher scores, which agree with other studies using machine learning and biological data [26,46].

In our subsequent analysis, we were interested in identifying a consensus list of candidate genes for the core response in *P. aeruginosa*, resulting in 46 genes of the perturbome. When global profiles were compared using these genes (PCA and HC), control and perturbation classes (Figure 5-D) were clearly separated. This gene number seems to be a modest number of elements (less than 1% of all the available dataset with 5549 genes) but it agrees completely with other studies, including machine learning methods [11,17] or other approaches [11,22,48,49]. In addition, 72% of core response genes were also identified as DEGs with similar annotation profiles; differences can be explained by significant fluctuations in the differential expression results as previously reported, mainly because it is not a consensus strategy (only based on *p*-value) and it does not incorporate the estimates of the test performance (true positive/negative rates and other metrics) on the results [2].

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

Core response genes or perturbome can be related to a central regulation network, and as convergent point of signal transduction, transcriptional regulation and stress-related pathways, as it has been suggested [2,4,5,42]. Annotation of the 46 genes shows that most of them are functionally related to biosynthesis, molecule binding and metabolism (including an important number of hits for lipids), including additional functions for regulation of DNA damage repair, response to stimuli and aerobic respiration. Interestingly, these processes are represented by genes with high connectivity in the small world network. For example, main functions associated to *fabA* are lipid metabolism, fatty acid and lipid biosynthesis, meanwhile for *nuoC* are aerobic respiration, electron/proton transport and DNA damage. Finally, *nuoF* is associated to aerobic respiration, electron/proton transport and Krebs cycle.

In this sense, cells are equipped with systems and mechanisms to recover from the environmental stress and stimuli to maintain all necessary physiological functions [50]. Other common stimuli such as low ATP, slow growth, and ROS production can also occur before cells express stress specific factors, but mediating a common effect. For example, response to stress includes modulation of energy production and aerobic electron transfer chain components. As it has been reported in *E. coli*, aerobic electron transport chain components are down-regulated in response to growth arrest [51]. This corresponds with the global profiles of expression of 46 core response genes. Also, regulation of lipid metabolism is relevant for survival in the wide range of environmental conditions where bacteria thrive [52], even for biofilm-living forms [3] as *P. aeruginosa*. Core response genes *fabG*, *lpxA*, and PA5174 could be implied in this process.

In the case of DNA damage repair (including the case of *cycB* and *gltP* genes), responses mediated by SOS and *rpoS* help to maintain genome integrity, colonization, and virulence [39,53]. These responses are activated under multiple disturbances and modulating a low energy production and shutdown of the metabolism, promotes formation of antibiotic resistance and biofilms [3]. Other related pathways for some specific genes included regulation of the transcription during stress by RNA-binding proteins in order to reprogram or shut down translation and to rescue the ribosome stalled by a variety of mechanisms induced [54]. Three core response genes (*rpmH*, *tsf* and PA2735) were annotated with these functions.

Jointly, the relatively few diversity of metabolic functions and pathways makes sense in order to ensure redundancy and robustness in the response to stimuli. Similar results, regarding enriched pathways, have been obtained in other studies with eukaryotic models, including disturbed human cell lines [4,5,42], *Arabidopsis thaliana* under physical and genotoxic stresses [2] or a genome-wide association study of a generic response to stress conditions [6]. In the case of prokaryotic organisms, two studies have used *Escherichia coli* as model to identify differentially expressed genes after exposure to stress conditions [3] and to create networks associated with the response to fluctuating environments [7]. Differences with other organisms and disturbances can suggest that response cell stress can be organismal specific, although heterogeneity has also been suggested a reasonable explanation because differences in the response in a apparently homogeneous cell population [42,55].

As in our case, the response to stimuli and stressors is orchestrated by a pleotropic modulation [56] which can be associated to a central regulation. Alternative mechanisms such as cross stress protection (ability of a stress condition to provide protection against other stressors) [7], role of sigma factors and specific two component systems [3] can contribute to explain this phenomenon. The molecular response can lead to regulate multiple biological activities including metabolism, replication, transcription and translation, changes in membrane composition, motility, modification gene expression, expression of virulence factors, multi-drug resistant phenotypes and biofilm formation, and others [42].

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

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

5. Conclusions

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

Abbreviations

AUC: Area under the curve

AZM: Azithromycin

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

CDA: Chlorhexidine diacetate

CIP: Ciprofloxacin

COL: Colistin

Cu: Copper

DEGs: Differentially expressed genes

GEO: Gene Expression Omnibus

H2O2: Hydrogen peroxide

HC: Hierarchical Clustering

KNN: K-Nearest Neighbor

LB: Luria Bertani

mRNA: Messenger RNA

NaClO: Sodium hypoclorite

OPP: Ortho-phenylphenol

PCA: Principal Component Analysis

PPI: Protein-protein interaction

RF: Random Forest

ROC: Receiver-operating characteristic

SVM: Support Vector Machine

Declarations

Ethics approval and consent to participate

Not Applicable

Consent for Publication

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, 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. Zhao, W.; Chen, J.J.; Perkins, R.; Wang, Y.; Liu, Z.; Hong, H.; Tong, W.; Zou, W.; Metzker, M.; Didelot, X.; et al. A novel procedure on next generation sequencing data analysis using text mining algorithm. *BMC Bioinformatics* **2016**, *17*, 213, doi:10.1186/s12859-016-1075-9.

15. Cornforth, D.M.; Dees, J.L.; Ibberson, C.B.; Huse, H.K.; Mathiesen, I.H.; Kirketerp-Møller, K.; Wolcott, R.D.; Rumbaugh, K.P.; Bjarnsholt, T.; Whiteley, M. Pseudomonas aeruginosa transcriptome during human infection. *Proc. Natl. Acad. Sci. U. S. A.* **2018**, *115*, doi:10.1073/pnas.1717525115.

16. Glaab, E.; Bacardit, J.; Garibaldi, J.M.; Krasnogor, N. Using rule-based machine learning for candidate disease gene prioritization and sample classification of cancer gene expression data. *PLoS One* **2012**, *7*, doi:10.1371/journal.pone.0039932.

17. 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 multi-staged strategy based on machine learning and mathematical modeling to predict genotype-phenotype risk patterns in diabetic kidney disease: A prospective case-control cohort analysis. *BMC Nephrol.* **2013**, *14*, 1, doi:10.1186/1471-2369-14-162.

18. 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, C.S. Unsupervised Extraction of Stable Expression Signatures from Public Compendia with an Ensemble of Neural Networks. *Cell Syst.* **2017**, *5*, 63–71.e6, doi:10.1016/j.cels.2017.06.003.

19. Kar, S.C. Comparing Prediction Accuracy for Supervised Techniques in Gene Expression Data. *Math. Theory Model.* **2014**, *4*, 108–116.

20. Raza, K.; Hasan, A. A Comprehensive Evaluation of Machine Learning Techniques for Cancer Class Prediction Based on Microarray Data. *Int. J. Bioinform. Res. Appl.* **2015**, *11*, 397–416, doi:10.1504/IJBRA.2015.071940.

21. Savli, H.; Karadenizli, A.; Kolayli, F.; Gundes, S.; Ozbek, U.; Vahaboglu, H. Expression stability of six housekeeping genes: a proposal for resistance gene quantification studies of Pseudomonas aeruginosa by real-time quantitative RT-PCR. *J. Med. Microbiol.* **2003**, *52*, 403–408, doi:10.1099/jmm.0.05132-0.

22. Casares, F.M. A Simple Method for Optimization of Reference Gene Identification and Normalization in DNA Microarray Analysis. *Med. Sci. Monit. Basic Res.* **2016**, *22*, 45–52, doi:10.12659/MSMBR.897644.

23. Yang, P.; Zhou, B.B.; Yang, J.Y.-H.; Zomaya, A.Y. Stability of feature selection algorithms and ensemble feature selection methods in bioinformatics. *Biol. Knowl. Discov. Handb. Preprocessing, mining, postprocessing Biol. data* **2013**, 333–352, doi:10.1002/9781118617151.ch14.

24. Tabe-Bordbar, S.; Emad, A.; Zhao, S.D.; Sinha, S. A closer look at cross-validation for assessing the accuracy of gene regulatory networks and models. *Sci. Rep.* **2018**, *8*, 1–11, doi:10.1038/s41598-018-24937-4.

25. Alqarni, B.; Colley, B.; Klebensberger, J.; McDougald, D.; Rice, S.A. Expression stability of 13 housekeeping genes during carbon starvation of Pseudomonas aeruginosa. *J. Microbiol. Methods* **2016**, *127*, 182–187, doi:10.1016/j.mimet.2016.06.008.

26. Noi, P.T.; Kappas, M. Comparison of random forest, k-nearest neighbor, and support vector machine classifiers for land cover classification using sentinel-2 imagery. *Sensors (Switzerland)* **2018**, *18*, doi:10.3390/s18010018.

27. Li, L.; Weinberg, C.R.; Darden, T.A.; Pedersen, L.G. Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. *Bioinformatics* **2001**, *17*, 1131–42.

28. Vapnik, V.N.; Vladimir *Estimation of dependences based on empirical data*; Springer-Verlag, 1982; ISBN 0387907335.

29. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32, doi:10.1023/A:1010933404324.

30. Yusuf Nalca; Lothar Jänsch; Florian Bredenbruch; Robert Geffers, J.B.; Susanne Häussler Quorum-sensing antagonistic activities of azithromycin in Pseudomonas aeruginosa PAO1: a global approach. *Antimicrob Agents Chemother* **2008**, *50*, 1680–1688, doi:10.1128/AAC.50.5.1680.

31. Kai, T.; Tateda, K.; Kimura, S.; Ishii, Y.; Ito, H.; Yoshida, H.; Kimura, T.; Yamaguchi, K. A low concentration of azithromycin inhibits the mRNA expression of N-acyl homoserine lactone synthesis enzymes, upstream of lasI or rhlI, in Pseudomonas aeruginosa. *Pulm. Pharmacol. Ther.* **2009**, *22*, 483–486, doi:10.1016/j.pupt.2009.04.004.

32. Chang, W.; Small, D.A.; Toghrol, F.; Bentley, W.E. Microarray analysis of Pseudomonas aeruginosa reveals induction of pyocin genes in response to hydrogen peroxide. *BMC Genomics* **2005**, *6*, 1–14, doi:10.1186/1471-2164-6-115.

33. Teitzel, G.M.; Geddie, A.; De Long, S.K.; Kirisits, M.J.; Whiteley, M.; Parsek, M.R. Survival and Growth in the Presence of Elevated Copper: Transcriptional Profiling of Copper-Stressed Pseudomonas aeruginosa. *J. Bacteriol.* **2006**, *188*, 7242–7256, doi:10.1128/JB.00837-06.

34. Small, D.A.; Chang, W.; Toghrol, F.; Bentley, W.E. Comparative global transcription analysis of sodium hypochlorite, peracetic acid, and hydrogen peroxide on Pseudomonas aeruginosa. *Appl. Microbiol. Biotechnol.* **2007**, *76*, 1093–1105, doi:10.1007/s00253-007-1072-z.

35. Nde, C.W.; Jang, H.-J.; Toghrol, F.; Bentley, W.E. Toxicogenomic response of Pseudomonas aeruginosa to ortho-phenylphenol. *BMC Genomics* **2008**, *9*, 473, doi:10.1186/1471-2164-9-473.

36. Cummins, J.; Reen, F.J.; Baysse, C.; Mooij, M.J.; O'Gara, F. Subinhibitory concentrations of the cationic antimicrobial peptide colistin induce the pseudomonas quinolone signal in Pseudomonas aeruginosa. *Microbiology* **2009**, *155*, 2826–2837, doi:10.1099/mic.0.025643-0.

37. Nde, C.W.; Jang, H.J.; Toghrol, F.; Bentley, W.E. Global transcriptomic response of Pseudomonas aeruginosa to chlorhexidine diacetate. *Environ. Sci. Technol.* **2009**, *43*, 8406–8415, doi:10.1021/es9015475.

38. Pan, J.; Bahar, A.A.; Syed, H.; Ren, D. Reverting Antibiotic Tolerance of Pseudomonas aeruginosa PAO1 Persister Cells by (Z)-4-bromo-5-(bromomethylene)-3-methylfuran-2(5H)-one. *PLoS One* **2012**, *7*, doi:10.1371/journal.pone.0045778.

39. Cirz, R.T.; O'Neill, B.M.; Hammond, J.A.; Head, S.R.; Romesberg, F.E. Defining the Pseudomonas aeruginosa SOS response and its role in the global response to the antibiotic ciprofloxacin. *J. Bacteriol.* **2006**, *188*, 7101–7110, doi:10.1128/JB.00807-06.

40. Smyth, G.K. Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. *Stat. Appl. Genet. Mol. Biol.* **2004**, *3*, 1–25, doi:10.2202/1544-6115.1027.

41. Richter, K.; Haslbeck, M.; Buchner, J. The Heat Shock Response: Life on the Verge of Death. *Mol. Cell* **2010**, *40*, 253–266, doi:10.1016/j.molcel.2010.10.006.

42. Vihervaara, A.; Duarte, F.M.; Lis, J.T. Molecular mechanisms driving transcriptional stress responses. *Nat. Rev. Genet.* **2018**, *19*, 385–397, doi:10.1038/s41576-018-0001-6.

43. Holloway, B.W. Genetic Recombination in Pseudomonas aeruginosa. *Microbiology* **1955**, *13*, 572–581, doi:10.1099/00221287-13-3-572.

44. Piao, J.; Sun, J.; Yang, Y.; Jin, T.; Chen, L.; Lin, Z. Target gene screening and evaluation of prognostic values in non-small cell lung cancers by bioinformatics analysis. *Gene* **2018**, *647*, 306–311, doi:10.1016/j.gene.2018.01.003.

45. Touw, W.G.; Bayjanov, J.R.; Overmars, L.; Backus, L.; Boekhorst, J.; Wels, M.; van Hijum, S.A.F.T. Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle? *Brief. Bioinform.* **2013**, *14*, 315–326, doi:10.1093/bib/bbs034.

46. Pai, S.; Hui, S.; Isserlin, R.; Shah, M.A.; Kaka, H.; Bader, G.D. netDx: interpretable patient classification using integrated patient similarity networks. *Mol. Syst. Biol.* **2019**, *15*, e8497, doi:10.15252/msb.20188497.

47. Park, H.; Shimamura, T.; Imoto, S.; Miyano, S. Adaptive NetworkProfiler for Identifying Cancer Characteristic-Specific Gene Regulatory Networks. *J. Comput. Biol.* **2017**, *25*, cmb.2017.0120, doi:10.1089/cmb.2017.0120.

48. Falciani, F.; Diab, A.; Sabine, V.; Williams, T.; Ortega, F.; George, S.; Chipman, J. Hepatic transcriptomic profiles of European flounder (Platichthys flesus) from field sites and computational approaches to predict site from stress gene responses following exposure to model toxicants. *Aquat. Toxicol.* **2008**, *90*, 92–101, doi:10.1016/j.aquatox.2008.07.020.

49. Nota, B.; Verweij, R.A.; Molenaar, D.; Ylstra, B.; van Straalen, N.M.; Roelofs, D. Gene Expression Analysis Reveals a Gene Set Discriminatory to Different Metals in Soil. *Toxicol. Sci.* **2010**, *115*, 34–40, doi:10.1093/toxsci/kfq043.

50. Krämer, R. Bacterial stimulus perception and signal transduction: Response to osmotic stress. *Chem. Rec.* **2010**, *10*, 217–229, doi:10.1002/tcr.201000005.

51. Schurig-Briccio, L.A.; Farías, R.N.; Rodríguez-Montelongo, L.; Rintoul, M.R.; Rapisarda, V.A. Protection against oxidative stress in Escherichia coli stationary phase by a phosphate concentration-dependent genes expression. *Arch. Biochem. Biophys.* **2009**, *483*, 106–110, doi:10.1016/j.abb.2008.12.009.

52. Parsons, J.B.; Rock, C.O. Bacterial lipids: metabolism and membrane homeostasis. *Prog. Lipid Res.* **2013**, *52*, 249–76, doi:10.1016/j.plipres.2013.02.002.

53. Storvik, K.A.M.; Foster, P.L. RpoS, the stress response sigma factor, plays a dual role in the regulation of Escherichia coli's error-prone DNA polymerase IV. *J. Bacteriol.* **2010**, *192*, 3639–44, doi:10.1128/JB.00358-10.

54. Starosta, A.L.; Lassak, J.; Jung, K.; Wilson, D.N. The bacterial translation stress response. *FEMS Microbiol. Rev.* **2014**, *38*, 1172–1201, doi:10.1111/1574-6976.12083.

55. Adamson, B.; Norman, T.M.; Jost, M.; Cho, M.Y.; Nuñez, J.K.; Chen, Y.; Villalta, J.E.; Gilbert, L.A.; Horlbeck, M.A.; Hein, M.Y.; et al. A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. *Cell* **2016**, *167*, 1867–1882.e21, doi:10.1016/j.cell.2016.11.048.

56. Molina-Mora, J.A.; Campos-Sanchez, R.; Garcia, F. Gene Expression Dynamics Induced by Ciprofloxacin and Loss of Lexa Function in Pseudomonas aeruginosa PAO1 Using Data Mining and Network Analysis. In Proceedings of the 2018 IEEE International Work Conference on Bioinspired Intelligence (IWOBI); IEEE, 2018; pp. 1–7.

CHAPTER 5

Transcriptomic determinants of the response of ST-111 *Pseudomonas aeruginosa* AG1 to ciprofloxacin identified by a top-down systems biology approach

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

https://www.nature.com/articles/s41598-020-70581-2

Summary

Ciprofloxacin (CIP) is an antibiotic commonly used to treat *P. aeruginosa* infections, and it is known to produce DNA damage, triggering a complex molecular response. In order to evaluate the effects of a sub-inhibitory CIP concentration on the multi-resistant PaeAG1, growth curves using increasing CIP concentrations were compared. We then measured gene expression using RNA-Seq at three time points (0, 2.5 and 5 hours) after CIP exposure to identify the transcriptomic determinants of the response (i.e. hub genes, gene clusters and enriched pathways). Changes in expression were determined using differential expression analysis and network analysis using a top-down systems biology approach. A hybrid model using database-based and co-expression analysis approaches was implemented to predict gene-gene interactions.

We observed a reduction of the growth curve rate as the sub-inhibitory CIP concentrations were increased. In the transcriptomic analysis, we detected that over time CIP treatment resulted in the differential expression of 518 genes, showing a complex impact at the molecular level. The transcriptomic determinants were 14 hub genes, multiple gene clusters at different levels (associated to hub genes or as co-expression modules) and 15 enriched pathways. Down-regulation of genes implicated in several metabolism pathways, virulence elements and ribosomal activity was observed. In contrast, amino acid catabolism, RpoS factor, proteases, and phenazines genes were up-regulated. Remarkably, >80 resident-phage genes were up-regulated after CIP treatment, which was validated at phenomic level using a phage plaque assay. Thus, reduction of the growth curve rate and increasing phage induction was evidenced as the CIP concentrations were increased.

In summary, transcriptomic and network analyses, as well as the growth curves and phage plaque assays provide evidence that PaeAG1 presents a complex, concentration-dependent response to sub-inhibitory CIP exposure, showing pleiotropic effects at the systems level. Manipulation of these determinants, such as phage genes, could be used to gain more insights about

the regulation of responses in PaeAG1 as well as the identification of possible therapeutic targets. To our knowledge, this is the first report of the transcriptomic analysis of CIP response in a ST-111 high-risk *P. aeruginosa* strain, in particular using a top-down systems biology approach.

SCIENTIFIC REPORTS
natureresearch

OPEN

Transcriptomic determinants of the response of ST‑111 *Pseudomonas aeruginosa* AG1 to ciprofloxacin identified by a top‑down systems biology approach

José Arturo Molina‑Mora[1]✉, Diana Chinchilla‑Montero[1], Maribel Chavarría‑Azofeifa[1], Alejandro J. Ulloa‑Morales[2], Rebeca Campos‑Sánchez[3], Rodrigo Mora‑Rodríguez[1], Leming Shi[4] & Fernando García[1]

Pseudomonas aeruginosa is an opportunistic pathogen that thrives in diverse environments and causes a variety of human infections. *Pseudomonas aeruginosa* AG1 (PaeAG1) is a high-risk sequence type 111 (ST-111) strain isolated from a Costa Rican hospital in 2010. PaeAG1 has both blaVIM-2 and blaIMP-18 genes encoding for metallo-β-lactamases, and it is resistant to β-lactams (including carbapenems), aminoglycosides, and fluoroquinolones. Ciprofloxacin (CIP) is an antibiotic commonly used to treat *P. aeruginosa* infections, and it is known to produce DNA damage, triggering a complex molecular response. In order to evaluate the effects of a sub-inhibitory CIP concentration on PaeAG1, growth curves using increasing CIP concentrations were compared. We then measured gene expression using RNA-Seq at three time points (0, 2.5 and 5 h) after CIP exposure to identify the transcriptomic determinants of the response (i.e. hub genes, gene clusters and enriched pathways). Changes in expression were determined using differential expression analysis and network analysis using a top–down systems biology approach. A hybrid model using database-based and co-expression analysis approaches was implemented to predict gene–gene interactions. We observed a reduction of the growth curve rate as the sub-inhibitory CIP concentrations were increased. In the transcriptomic analysis, we detected that over time CIP treatment resulted in the differential expression of 518 genes, showing a complex impact at the molecular level. The transcriptomic determinants were 14 hub genes, multiple gene clusters at different levels (associated to hub genes or as co-expression modules) and 15 enriched pathways. Down-regulation of genes implicated in several metabolism pathways, virulence elements and ribosomal activity was observed. In contrast, amino acid catabolism, RpoS factor, proteases, and phenazines genes were up-regulated. Remarkably, > 80 resident-phage genes were up-regulated after CIP treatment, which was validated at phenomic level using a phage plaque assay. Thus, reduction of the growth curve rate and increasing phage induction was evidenced as the CIP concentrations were increased. In summary, transcriptomic and network analyses, as well as the growth curves and phage plaque assays provide evidence that PaeAG1 presents a complex, concentration-dependent response to sub-inhibitory CIP exposure, showing pleiotropic effects at the systems level. Manipulation of these determinants, such as phage genes, could be used to gain

[1]Centro de Investigación en Enfermedades Tropicales (CIET), Facultad de Microbiología, Universidad de Costa Rica, San José, Costa Rica. [2]Chemical Genomics Centre (CGC), Max-Planck-Institute for Molecular Physiology, Dortmund, Germany. [3]Centro de Investigación en Biología Celular Y Molecular (CIBCM), Facultad de Microbiología, Universidad de Costa Rica, San José, Costa Rica. [4]Human Phenome Institute (HuPI), Fudan University, Shanghai, China. ✉email: jose.molinamora@ucr.ac.cr

more insights about the regulation of responses in PaeAG1 as well as the identification of possible therapeutic targets. To our knowledge, this is the first report of the transcriptomic analysis of CIP response in a ST-111 high-risk *P. aeruginosa* strain, in particular using a top-down systems biology approach.

Abbreviations

CIP	Ciprofloxacin
DEGs	Differentially expressed genes
GGI	Gene–gene interaction
GSEA	Gene set enrichment analysis
KEGG	Kyoto encyclopedia of genes and genomes
MIC	Minimum inhibitory concentration
MLST	Multilocus sequence typing
OD_{600nm}	Optical density measured at 600 nm
PaeAG1	*Pseudomonas aeruginosa* Strain AG1
PCA	Principal components analysis
QC	Quality control
RNA-Seq	RNA sequencing
ST-111	Sequence type 111
WGCNA	Weighted gene co-expression network analysis
WHO	World Health Organization (WHO)

Pseudomonas aeruginosa is a ubiquitous Gram-negative organism which thrives in diverse environments and acts as an opportunistic pathogen[1]. The ability of this pathogen to cause a variety of human infections is facilitated by its nutritional versatility[2], resistance to a wide spectrum of antibiotics, and virulence factors[3,4]. *Pseudomonas aeruginosa* AG1 (PaeAG1) is a multiresistant high-risk sequence type 111 (ST-111) strain (GenBank CP045739)[5]. It was isolated from a Costa Rican hospital and it was the first report of an isolate of *P. aeruginosa* carrying both blaVIM-2 and blaIMP-18 genes encoding for metallo-β-lactamases enzymes (carbapenemases), located in two independent integrons[5,6]. PaeAG1 is resistant to β-lactams (including carbapenems), aminoglycosides, and fluoroquinolones, being only sensitive to colistin. In addition to this multidrug-resistant feature, as in other *P. aeruginosa* strains, the ability to colonize nosocomial environments makes this strain a high-risk clone[7]. Owed to this antibiotic resistance profile, including resistance to carbapenems, PaeAG1 is classified as a Priority 1 (critical) organism according to the World Health Organization (WHO)[8].

Antibiotic resistance is a major threat to public health because it compromises the administration of appropriate antibiotic therapy, and reduces the therapeutic options to treat infections, increasing patient morbidity and mortality[9,10]. This situation is aggravated by the emergence of strains resistant to multiple antibiotics[11], limitation of the knowledge of interactions with pathogens and mechanisms of action of antimicrobial agents, and development of new antibiotics[12]. Use of antibiotics below the minimum inhibitory concentration (MIC) or sub-inhibitory concentrations also contributes to antibiotic resistance as they allow strains to continue growing and can select for pre-existing resistant organisms[13]. Since sub-inhibitory antibiotic concentrations are found in many natural environments, bacteria can naturally trigger mechanisms of tolerance[14]. However, the fundamental mechanisms of bacterial tolerance to antibiotics have not been fully elucidated[15].

It has been shown that the perturbation induced by many antibiotics leads to stress conditions in prokaryotic cells[16], which can induce DNA damage[17]. Stressors activate the regulation of gene expression or the activity and stability of existing proteins to induce adaptation mechanisms[16]. Organisms have evolved numerous DNA repair pathways to eliminate DNA damage and restart DNA replication[18]. Regulatory networks of transcriptional responses to DNA damage involves not only DNA repair enzymes, but also diverse proteins with roles in cell division, metabolism modulation, genetic rearrangements and exchange, mutation, and virulence factor production[19].

Ciprofloxacin (CIP) is a fluoroquinolone antibiotic used to treat *P. aeruginosa* infections[20]. CIP is well-known to produce DNA damage by inhibiting DNA gyrase and topoisomerase IV, leading to DNA strand breaks[21]. Mutations in these genes are responsible for CIP resistance by losing drug affinity[22]. CIP has been used to study stress responses in this bacterial group[12,23], in particular with the induction of the SOS response as a mechanism of DNA damage repair[17,24,25]. In *P. aeruginosa,* the SOS response regulon is composed of 15 genes, including *recA* and *lexA* genes[26]. Upon DNA damage, RecA recognizes the single-stranded DNA (ssDNA) forming filaments and induces the autocleavage of the repressor LexA. This response leads to the expression of genes related to DNA damage repair[27]. Other LexA-like repressors are regulated during SOS activation, including elements of phages and pyocines[19]. SOS also mediates responses to resistance element transfer, generation of mutations and evolution of resistance[26], as well as appearance of persister cells[24].

However, modulation of stress responses after DNA damage is not limited to SOS response. RpoS is a general stress sigma factor (σS) known as a central element in a regulatory network that governs the expression of stationary-phase-induced genes[28] to maintain cell viability[29]. This regulator is strongly induced when cells are exposed to various stress conditions, including antibiotics, pH downshift, starvation, and hyperosmolarity[30]. RpoS regulates more than 50 genes in *Pseudomonas aeruginosa*[31], including virulence factors[32].

The SOS and RpoS regulons are complementary mechanisms in response to certain stresses and that protect bacteria from DNA damage[33]. Lon protease[11] and AmpR[34] can modulate both SOS and RpoS regulons. In addition, both responses can regulate key genes such as *polB*[18], *iraD*[19], and *dinB*[33]. The connection between RpoS and SOS responses seems to be associated with a mechanism to maximize survival and fitness of cells, and to maintain genome stability[18]. These responses can modulate virulence factors (including quorum sensing and biofilm formation), and increase homologous recombination and mutation frequencies[33,35]. However, other SOS and RpoS independent mechanisms are also known to be present in bacteria[36], including *P. aeruginosa* after CIP treatment[12,26] with variable results depending on strains and showing a mosaic response[12].

Although the full mechanisms of all these molecular responses are not well understood, it is known that cells respond to stress conditions by complex regulatory systems that control gene expression[37]. Since a key objective in biological research is to describe molecular interactions[38], the use of networks analysis is a common approach to describe complex biological systems and to mathematically model gene–gene interactions (GGI) with graphical representations (genes as nodes and interactions as edges)[39]. Molecules are thereby studied not only at a single level, but emergent properties are identified to describe and understand the complexity of the gene networking response and the emergent properties towards the stress condition. Functional status of genes by a top-down systems biology perspective, starting from "whole"-omics data to identify specific determinants or elements of biological importance, can be evaluated by construction of large scale networks[40]. For this purpose, data analysis from high-throughput technologies such as microarrays and RNA sequencing (RNA-Seq) can be used to describe molecular interactions at transcriptomic level[38,41]. Thus, to understand or to infer mechanisms associated with the transcriptional response, it is possible to build gene regulatory networks either using databases or based on co-expression data[39,42,43]. These networks allow to gain insight into response to stress conditions[44], leading to the identification of gene clusters or even hub genes as candidate biomarkers or modulators with potential to become key therapeutic targets[43,45].

In *P. aeruginosa*, rapid adaptation to stress conditions is partially explained by the modulation of the global gene expression, which represents around 8% of all coding genes[3]. This regulation induces pleiotropic effects on its genomic regulatory network[46], as previously shown using systems biology[47], and the transcriptomic profiling of the response to CIP[12,26,48].

In this work we first evaluated PaeAG1 growth at sub-inhibitory CIP concentrations, showing growth reduction as CIP was increased. We hypothesized that after exposing PaeAG1 to ciprofloxacin, even at sub-inhibitory concentrations, transcriptomic determinants will be triggered, including bacterial growth modulators. Thus, the aim was to identify transcriptomic determinants associated with the response to CIP in PaeAG1 using RNA-Seq profiling and network analysis by a top-down systems biology approach. Results showed that PaeAG1 generates a complex response to CIP exposure, evidencing pleiotropic effects involving the regulation of multiple hub genes, gene clusters and enriched pathways (transcriptomic determinants), many of them related to growth. As evidenced at the transcriptomic and the phenomic levels, phage induction was a particular trait modulated by CIP in a concentration-dependent manner with a correlation with bacterial growth reduction.

Methods

The general pipeline followed in this study to identify the transcriptomic determinants associated with the response to CIP in PaeAG1 is shown in Fig. 1.

Bacterial isolate.

The PaeAG1 strain is a Costa Rican multiresistant isolate from a sputum sample of a patient with pneumonia at the Intensive Care Unit of the San Juan de Dios Hospital (San José, Costa Rica)[6]. PaeAG1 exhibits resistance to β-lactams (including carbapenems, $\mathrm{MIC}_{\mathrm{Meropenem}}$ 32 μg/mL and $\mathrm{MIC}_{\mathrm{Imipenem}} > 32$ μg/mL), aminoglycosides ($\mathrm{MIC}_{\mathrm{Gentamycin}}$ 128 μg/mL and $\mathrm{MIC}_{\mathrm{Tobramycin}} > 192$ μg/mL) and fluoroquinolones ($\mathrm{MIC}_{\mathrm{Ciprofloxacin}}$ 32 μg/mL), and it is only sensitive to colistin ($\mathrm{MIC}_{\mathrm{Colistin}}$ 2 μg/mL). We recently assembled and annotated the PaeAG1 genome[5], and genome sequence and annotation are available in Genbank under accession CP045739 (Bioproject PRJNA587210).

Growth curves assay.

Three independent cultures of PaeAG1 cells were grown to exponential-phase overnight in Lysogenic Broth (LB) at 37 °C with shaking (pre-culture to reach mid-log phase). Then, five aliquots were added to 50 mL of fresh LB broth to an initial optical density measured at 600 nm ($\mathrm{OD}_{600\mathrm{nm}}$) of 0.01. Each sample was treated with a specific CIP concentration of 0.0 (control), 5.0, 12.5, 25.0 or 50.0 μg/mL (final concentrations). Growth of cultures was monitored by $\mathrm{OD}_{600\mathrm{nm}}$ at times 0, 2, 4, 6, 8, 12 and 16 h. Comparison of different CIP concentrations was done by assessing growth curve kinetics, including lag and exponential phases. As a complementary assay, evaluation of two other antibiotics was done in exactly the same growth conditions, but antibiotic concentrations depended on the MIC: imipenem (carbapenem) and tobramycin (aminoglycoside). See results and supplementary Figure S1 for details.

The growth curves were statistically compared to the control growth curve using a two-way ANOVA with Bonferroni post-tests (significance level of 95%), similar to[49], using the time and concentrations as factors. We also ran a unpaired t-test (95% significance) comparing area under curve (AUC) of each growth curve against the control, similar to[50]. Analyses were done using Prism (GraphPad Software, Inc., La Jolla, CA). To perform the transcriptomic assay, we used the results from growth curves to select a specific sub-inhibitory CIP concentration at which there were no major changes in the growth rate after treatment.

RNA isolation and RNA sequencing.

In order to evaluate the molecular response of PaeAG1 to a sub-inhibitory CIP concentration, a transcriptomic assay was designed using RNA-Seq technology, as described below.

Growth conditions. PaeAG1 cells were grown under the same conditions as detailed before but treatment was done using a single CIP concentration of 12.5 μg/mL (see "Results" for details of concentration selection). Immediately after adding treatment, an aliquot was taken as control (time 0 h), and cells were kept growing for 2.5 and

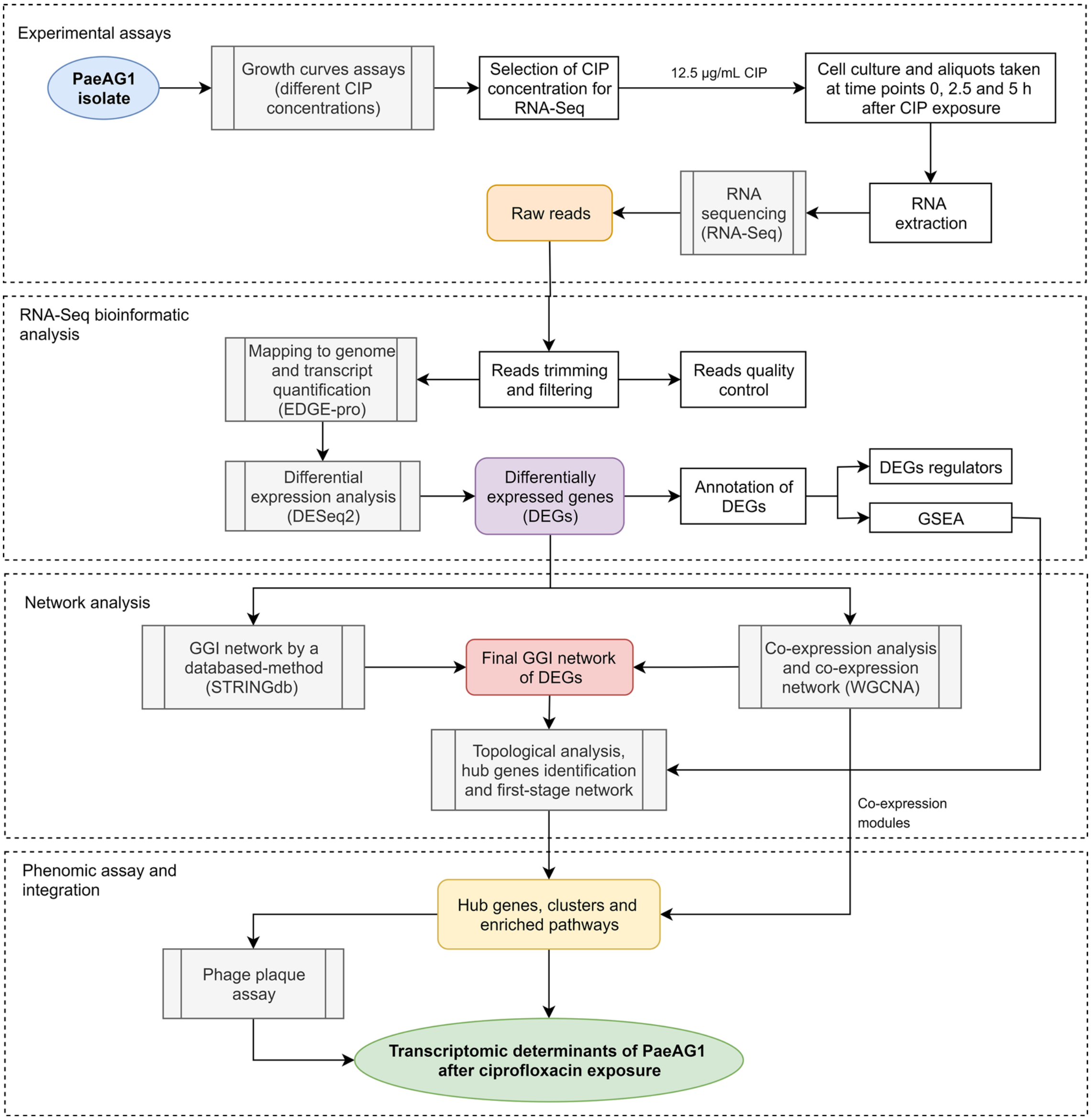

Experimental assays
PaeAG1 isolate
Growth curves assays (different CIP concentrations)
Selection of CIP concentration for RNA-Seq
12.5 μg/mL CIP
Cell culture and aliquots taken at time points 0, 2.5 and 5 h after CIP exposure
RNA extraction
RNA sequencing (RNA-Seq)
Raw reads
RNA-Seq bioinformatic analysis
Reads trimming and filtering
Reads quality control
Mapping to genome and transcript quantification (EDGE-pro)
Differential expression analysis (DESeq2)
Differentially expressed genes (DEGs)
Annotation of DEGs
DEGs regulators
GSEA
Network analysis
GGI network by a databased-method (STRINGdb)
Final GGI network of DEGs
Co-expression analysis and co-expression network (WGCNA)
Topological analysis, hub genes identification and first-stage network
Co-expression modules
Phenomic assay and integration
Hub genes, clusters and enriched pathways
Phage plaque assay
Transcriptomic determinants of PaeAG1 after ciprofloxacin exposure

Figure 1. General pipeline to identify the transcriptomic determinants of the response of *P. aeruginosa* AG1 to ciprofloxacin (CIP). After growth curves assessment, a specific CIP concentration was used to sequence RNA (RNA-Seq) at 0, 2.5 and 5 h after exposure. DEGs were identified and used to build GGI networks. Transcriptomic determinants were identified by network analysis. Findings were verified at phenomic level using a phage plaque assay.

5 h (times were selected according to preliminary results of phage induction, see "Methods" for Phage plaque assays). This was done with three independent cultures for a total of nine aliquots, three replicates per time.

RNA isolation. Aliquots from the cultures were preserved in two volumes of RNA protect reagent (QIAGEN) and cells were stored at 4 °C until RNA extraction. At the end of the sample collection period, total RNA was extracted using the RNeasy Mini kit (QIAGEN, UK) following the manufacturer's instructions. RiboZero Gold (Epicentre) was used to deplete bacterial rRNA from total RNA samples according to manufacturer's instructions. The quality and quantity of extracted RNA was determined using a Nanodrop (Nanodrop 2000, Thermo Scientific, UK). The RNA integrity was analyzed using Agilent 2,100 Bioanalyzer (Agilent Technologies, USA) to obtain the RNA integrity number (RIN) for all samples.

RNA sequencing. For RNA sequencing, TruSeq Stranded Total RNA library preparation kit (Illumina, USA) was used to generate cDNA (amplification with 13 PCR cycles) and libraries for 2 × 51 bp paired-end reads. Libraries were prepared and sequenced at the Genome Technology Center, New York University (New York, USA) on the Illumina HiSeq 2,500 platform. Sequencing generated more than 120 Gb of sequences (> 300 millions of reads in total) for all samples.

RNA-Seq data analysis. With the aim of quantifying transcripts and identifying DEGs in PaeAG1 after CIP treatment, RNA-Seq data was analyzed including a quality control step, reads mapping to genome for transcript quantification and differential expression analysis.

Quality control (QC). QC was done before and after trimming/filtering. Reads were trimmed using Trimmomatic v0.38[51] to discard sequences with per base phred sequence quality score < 30 and 35 minimum length. Reads were filtered using BBDuk (https://jgi.doe.gov/data-and-tools/bb-tools/) to remove adapters and reads mapping to rRNA. Sequence files were evaluated using FastQC v0.11.7[52] to obtain general quality control metrics. To evaluate the origin of reads sequences, FastQ-Screen[53] was used to quantify the proportion of reads that mapped to reference genomes (human, mouse, and adapters contaminants, included by default) and prokaryotic sequences specifically added for this work (PaeAG1 and *E. coli* genomes, and rRNA 16S and 23S databases). Reports were merged using MultiQC[54] to summarize all individual results. After selection, sequences for each of the nine samples had an average output of approximately 60 million reads.

Reads mapping and transcript quantification. We used EDGE-pro v1.3.1 software to: map RNA-Seq reads to the PaeAG1 genome (Genbank CP045739), filter out multialigned reads, and estimate expression levels of each gene by counts[55]. This program was run with the default parameters, using Bowtie2[56] as read alignment algorithm. The script "edgeToDeseq.perl", provided with the software, was used to convert raw counts (EDGE-pro output) to a count-table format for further differential expression analysis. Quality control of alignments per sample was done using: Qualimap RNA-Seq tool[57] to assess mapping quality, and RSeQC package[58] to estimate transcripts coverage uniformity (gene body coverage) and transcript integrity number (TIN). Required formats of genome annotation files for these analyses are available in https://github.com/josemolina6/PaeAG1_genome.

Differential expression analysis. We used raw counts of transcripts to estimate differential expression. For this purpose, DESeq2 package v1.26.0[59] in R program v3.5.1[60] was used based on the negative binomial generalized linear models, using default settings. DESeq2 based normalization, absolute expression comparisons by the regularized log transformation (rlog), Principal Component Analysis (PCA), counts dispersion plots and clustering analysis were run in the same program. Triplicates of each time after PaeAG1 exposure to CIP were considered as a factor level. Differential expression analysis was done comparing 2.5 h or 5 h data against the initial time point at 0 h. Hypothesis testing to select differentially expressed genes (DEGs) was done using Benjamini–Hochberg adjustment (to control false discovery rate, FDR) and $\log_2$[FoldChange] (logFC) of transformed and normalized mean counts. Genes were considered up-regulated if logFC > 1 or down-regulated if logFC < -1, considering an adjusted p-value < 0.05 for both cases. Gene list comparisons by Venn diagrams were performed using the Draw Venn Diagram Tool (https://bioinformatics.psb.ugent.be/webtools/Venn/).

Annotation of differentially expressed genes. DEGs annotation was retrieved from our previous work[5] for the assembly and annotation of PaeAG1 genome (Genbank CP045739). Particular features per gene (including molecular function, product, gene size and domains, and sub cellular location of proteins) were explored in more detail from *Pseudomonas* Genome Database (https://www.pseudomonas.com/)[61]. In addition, general regulators of the DEGs were investigated using PseudomonasNet tool (https://www.inetbio.org/pseudomonasnet/Network_regulon_form.php) with a p-value < 0.05 in a context-centric analysis. Using the same platform, it was possible to identify the DEGs and their regulators that corresponded to transcription factors genes.

Analysis of DNA–protein interactions. For selected genes, protein-DNA binding sites were investigated. The CollectTF database (https://www.collectf.org/) was primarily used to search for consensus DNA binding sequences of the protein of interest and to identify modulated genes. If no information was available, promoter consensus sequences were searched from particular studies and the identification of binding sites was done using the motif-based sequence analysis tool (MEME, using Find Individual Motif Occurrences FIMO, https://meme-suite.org/tools/fimo).

In order to identify DEGs as molecular determinants (hub genes, gene clusters and key pathways) of the response to CIP in PaeAG1, a large scale gene–gene interaction (GGI) network of DEGs was built using a top-down systems biology approach. Connections between genes were predicted using two independent methods, one using a database-based model and another from co-expression analysis, detailed as follows.

Database-based method for gene–gene interactions prediction and network construction. With the aim of obtaining a high confidence GGI between DEGs using a database-based method, the Search Tool for the Retrieval of Interacting Genes database (STRINGdb)[62] was used to construct a large scale GGI network for the DEGs using default parameters. All DEGs at any of the two times were used to build the main network. The resulting graph was exported and then visualized and topologically analyzed using Cytoscape software[63].

Co-expression analysis and co-expression network construction. To incorporate more interactions between DEGs, a data-driven systems biology approach was implemented using co-expression analysis with all the normalized counts of DEGs, as in recent studies[45,64–66].

Modules identification using co-expression analysis. Weighted gene co-expression network analysis (WGCNA) package[43] was run in R software. Briefly, a matrix of Pearson correlation between all pairs of genes was calculated. The adjacency matrix was then constructed using a power of $\beta = 9$ as a saturation level for a soft threshold of the correlation matrix based on the criterion of scale-free topology. The topological overlap matrix was calculated. Hierarchical clustering was used to generate a dendrogram to group highly co-expressed genes, creating gene clusters called modules (arbitrarily represented by colors) using the default dynamic tree cut algorithm. Default colors given to modules were kept.

Association of co-expression modules and traits. A t-test evaluated the association between the modules (using module eigengene ME, the first principal component gene of module expression matrix) and traits of PaeAG1 according to the experimental design. For this, the times (the experiment factors 0, 2.5 and 5 h) and data of phage induction at 2.5 and 5 h after 12.5 µg/mL CIP exposure were incorporated as traits (see "Phage plaque assay" section in "Methods").

Co-expression network. To visualize the whole network including the modules by colors, the WGCNA "exportNetworkToCytoscape" function was run, using a correlation threshold of 0.985 and weight = false to build an un-weighted graph of highly connected genes with very strict correlation. The data-driven graph was visualized using Cytoscape.

Integrated DEGs network construction. The final GGI network of DEGs was constructed joining the files of the well-known interactions predicted by STRING database and the strict data-driven interactions obtained from co-expression analysis (un-weighted graph). The definitive graph was visualized using Cytoscape software. Topological metrics of the graph were obtained using the defaults apps available in Cytoscape.

Enrichment analysis. For the gene set enrichment analysis (GSEA), STRINGdb was used to identify significantly enriched pathways according to KEGG database, using a cutoff of FDR < 0.05. This analysis was run for complete gene lists of DEGs at 2.5 h, DEGs at 5 h, and genes of each co-expression module. Results of enrichment were incorporated into the DEGs network using the Cytoscape app Omics Visualizer (https://apps.cytoscape.org/apps/omicsvisualizer).

Hub genes identification. In order to identify central or hub genes in the DEGs network of PaeAG1 after exposure to CIP, cytoHubba app[67] was run in Cytoscape. To address this, bottleneck and betweenness methods were implemented with default parameters. The top 10 nodes (genes) were selected for each method using calculated metrics. All selected genes in any of the methods were labeled as hub genes. In addition, cytoHubba was also used to build two subnetworks using the hub genes, one with the selected elements only, and another including the first-stage nodes (in direct connection with hub genes) to identify gene clusters. KEGG annotation information was kept from the DEGs network.

Expression profiles of hub genes were compared to expression levels obtained in other representative studies, including the following stressors: Cu (copper)[68], CIP (ciprofloxacin)[26], COL (colistin)[69], AZM (Azithromycin)[70] and H_2O_2 (hydrogen peroxide)[71]. Comparison was done using the general information of expression levels (down, up or variable regulation).

Phage plaques assay (validation assay at the phenomic level). To validate the transcriptomic results which showed an up-regulation of phage genes in PaeAG1 after exposure to CIP, we implemented a phage plaques assay and performed this assay in triplicate. To assess the CIP effect on phage induction, different CIP concentrations were evaluated. Evaluation was also done for imipenem and tobramycin as supplementary assays. Growth conditions were the same as described in the "Growth curve assays", until the addition of different antibiotic concentrations. At this point, cultures were kept growing for five hours and phages were isolated and quantified for each sample. During standardization, it was determined that five hours after CIP exposure was the minimum time for clear detection of phage plaques (see supplementary Figure S1-B for details). Phage plaque counts at 2.5 h and 5 h for 12.5 µg/mL CIP were used to associate the phage induction with co-expression modules (detailed in "Co-expression analysis" section).

Phages isolation. Protocols of [72] and [73] were adapted. Briefly, the culture was centrifuged for 20 min at 4,000 rpm, 40 mL of the supernatant was taken and 1 mL of chloroform was added to residual bacterial cells. After overnight incubation, cell debris was removed by centrifugation for 20 min at 3,000 rpm. The supernatant was filtered through a 0.45 µm filter to select phages. A volume of 30 mL of the filtered supernatant was mixed with 7.5 mL of polyethylene glycol (20%) and NaCl (2.5 M) to precipitate the phages. After overnight incubation, the sample was centrifuged for 30 min at 4,000 rpm, the supernatant was discarded and the pellet was resuspended in 250 µL of phage buffer (10 mM $MgSO_4$, 10 mM Tris–HCl and 150 mM NaCl).

Phages quantification. Phages were quantified by means of Plaque Forming Units (PFU) using *P. aeruginosa* PAO1 as host cells. The numbers of PFU was determined using the double-agar-layer method[74]. Briefly, medium was composed of two agar layers, a first layer 1.5% and another to 0.5% agar concentration. *P. aeruginosa* PAO1 and phages were added on the second layer and phage plaques were visualized after incubation for 24 h at 25 °C.

An exponential regression between the CIP concentrations and the PFU was run to associate the effect of CIP exposure on the phage induction.

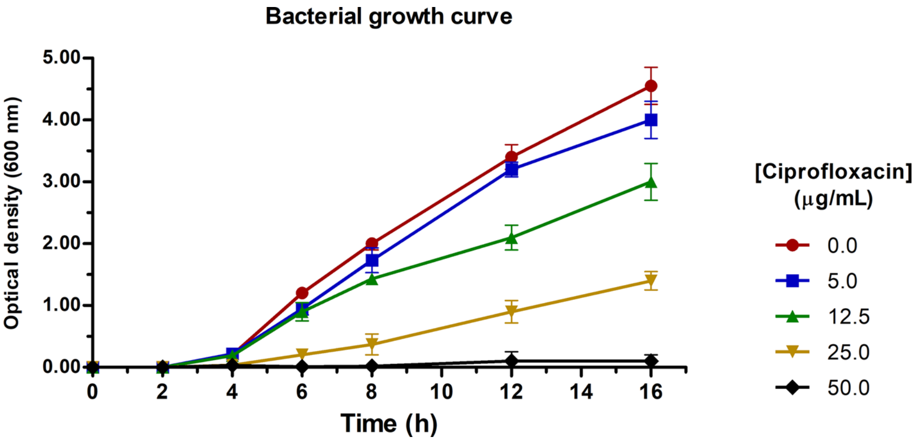


Figure 2. In vitro effects of ciprofloxacin on growth curve of PaeAG1. A growth rate reduction was observed as the CIP concentration was incremented. Area under curve (AUC) was compared using t-test ($p < 0.05$), showing a statistical difference between all curves when compared to control (0.0 mg/mL). In a similar manner, two-way ANOVA found differences in the OD_{600nm} and time for each case.

Ethical considerations. No animals or human participants were included in this study. Both the scientific committee of the Centro de Investigación en Enfermedades Tropicales (CIET) and Vicerrectoría de Investigación of Universidad de Costa Rica approved the study and the access to the PaeAG1 strain from the CIET collection of bacterial specimens.

Results

Concentration-dependent effect of CIP compromises the growth rate of PaeAG1.

To evaluate the effects of CIP in the growth rate of PaeAG1, increasing concentrations of the antibiotic were added to exponential-phase PaeAG1, and growth was monitored over time for 16 h. As shown in Fig. 2, OD_{600nm} values were highly consistent between replicates (error bars represent standard deviation). All CIP curves showed a statistical significant difference on OD_{600nm} compared to control ($p < 0.05$ for both AUC and two-way ANOVA). Lag phase for the control and two lower CIP concentrations (5 and 12.5 μg/mL) lasted approximately 4 h, while the higher CIP concentration of 25.0 μg/mL showed a lag phase of 8 h.

Kinetics at the exponential phase showed more variable results. There was a decrease in cell growth for 12.5 μg/mL CIP from 12 h onwards in comparison to 0 or 5.0 μg/mL, and more evident at same time for 25 μg/mL. For the case of 50.0 μg/mL (higher than MIC), the growth was drastically impaired and no exponential growth was observed. These results indicate that higher CIP concentrations have a stronger effect on the growth rate, even for sub-inhibitory concentrations ($MIC_{Ciprofloxacin}$ 32 μg/mL). Evaluation of the growth effects of other two antibiotics (imipenem and tobramycin) was also performed (supplementary Figure S1C–E, left). Unlike CIP, both cases showed no changes in the growth curves with different sub-inhibitory concentrations.

Due to the significant changes in growth curves with CIP (with respect to control) and considering a condition with enough cell mass for RNA-Seq analysis, 12.5 μg/mL CIP was used to evaluate the transcriptomic response of PaeAG1 to a sub-inhibitory concentration of the antibiotic.

RNA-Seq analysis identifies 518 DEGs in PaeAG1 over time after exposure to CIP.

A transcriptomic analysis was conducted to evaluate the molecular response to sub-inhibitory CIP concentration in PaeAG1. To this end, samples were taken at 0 (control), 2.5 and 5 h after CIP treatment. To ensure exponential growth at these times, the growth curve was monitored using OD_{600nm} measurements (successfully reproduced as Fig. 2), in addition to counting of Colony Forming Units (CFU), as shown in supplementary Figure S1A. After RNA was extracted, RNA integrity RIN > 9 was obtained for all samples and paired-end RNA sequencing was performed. For all samples, quality control of raw sequence data showed good results in terms of mean quality (> 30), no adapters, and no reads mapping to rRNA after filtering. Read mapping quality control showed that 98.6% were mapped to the PaeAG1 genome, with expected uniform coverage for gene body, and TIN > 90 for all samples. Details of assessment of transcriptomic data (counts per gene) is shown in supplementary Figure S2.

Identification of DEGs was conducted by comparing times 2.5 or 5 h against the initial 0 h time after CIP exposure (Fig. 3A,B). As shown in Table 1, 355 DEGs were identified at time 2.5 h, with 204 (57.5%) up-regulated and 151 (42.5%) down-regulated. At 5 h, 248 (56.6%) genes were up-regulated, meanwhile 190 (43.4%) were found to be down-regulated, for a total of 438 DEGs.

A total of 518 DEGs were found at any time points (union ∪), as shown in Fig. 3C and Table 1. These represent around 7% of the genes of PaeAG1. In addition, as presented in Fig. 3D, a total of 85 DEGs (at any time) belong to phages (27.6% of the 308 phage genes identified in the PaeAG1 genome), most of them up-regulated as shown in Table 2, Fig. 4 and supplementary Figure S3. The phages regulated include phiCTX, F10, JBD44 and JDO24 for which 3, 10, 65 and 7 DEGs were respectively observed at any time (Table 2).

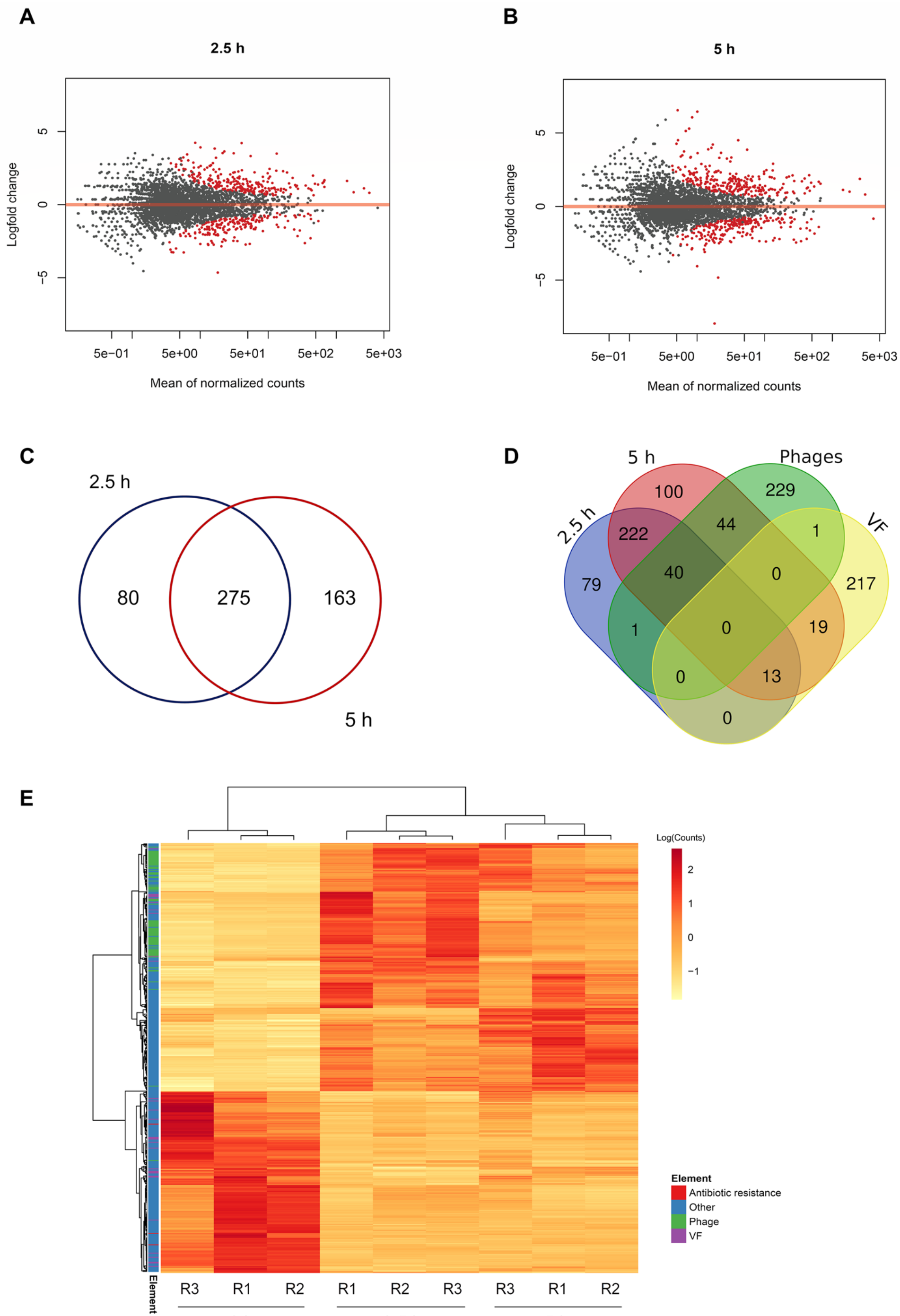


Figure 3. Differential expression analysis in PaeAG1 exposed to ciprofloxacin compared to initial time 0 h. Selection of DEGs according to adjusted *p*-value ($p < 0.05$) and logFC ($logFC < -1$ or $logFC > 1$) at 2.5 h (**A**) or 5 h (**B**) post-exposure to antibiotic. (**C**) Venn diagram showing the comparison of DEGs in the two evaluated times, with 275 shared genes (intersection) and total 518 genes at any time (union) with respect to time 0 h (control). More details in Table 1. (**D**) Venn diagram showing the comparison of DEGs and phage genes or virulence factors (more details in Table 2). (**E**) Heatmap of normalized counts and gene clustering of the total 518 DEGs at the three evaluated time points.

	Sets			
DEGs	2.5 h	5 h	2.5 h∩5 h	2.5 h∪5 h
Up regulated genes	204	248	153	299
Down regulated genes	151	190	118	223
Total DEGs	355	438	275	**518**

Table 1. Comparison of DEGs of PaeAG1 at 2.5 and 5 h after treatment with Ciprofloxacin, including counts of down or up regulated genes, shared genes (intersection) and total genes at both times (union).

Determinants			Sets of DEGs				
Type	Specific elements	Total genes (in PaeAG1 genome)	2.5 h	5 h	2.5 h∩5 h	2.5 h∪5 h	Regulation* and observations
Antibiotic resistance	Total	56	3	2	2	3	Down, lactamases
Phages	PPpW	12	0	0	0	0	No DEGs
	phiCTX	25	2	3	2	3	Up
	F10	62	1	9	0	10	Up
	JBD44	105	34	65	34	65	Up
	JDO24	59	4	7	4	7	Up
	phi3	45	0	0	0	0	No DEGs
	Total	308	41	84	40	85	–
Virulence factors	Adherence	96	11	19	11	19	Down
	Antimicrobial activity	17	1	6	1	6	Up, phenazines
	Antiphagocytosis	25	0	0	0	0	No DEGs
	Phospholipases	3	0	0	0	0	No DEGs
	Biosurfactant	3	0	0	0	0	No DEGs
	Iron uptake	28	0	1	0	1	Up, Pyochelin
	Protease	4	1	2	1	2	Up, elastases
	Quorum sensing	5	0	1	0	1	Up, RhlR
	Regulation GacS/GacA system	2	0	0	0	0	No DEGs
	Secretion system	63	0	2	0	2	Down, T3SS
	Toxins	4	0	1	0	1	Up, hydrogen cyanide
	Total	250	13	32	13	32	–

Table 2. Comparison of DEGs of PaeAG1 at 2.5 and 5 h after treatment with ciprofloxacin, and specific phages or categories of virulence factors, including shared genes (intersection) and total genes at both times (union), the regulation and the type of elements. *Based in logFC of genes for both times 2.5 and 5 h. Type of elements is also shown.

In the case of the 250 known virulence factors of PaeAG1, 32 (12.8%) were identified as DEGs at all of the assessed times (arrowheads of Fig. 4 and supplementary Figure S3). The virulence factors are mainly associated with adherence (19) and phenazines (6) genes (see Table 2). Regarding antibiotic resistance genes, only three out the 56 genes were found to be differentially expressed (Table 2).

A heatmap of normalized counts and gene clustering of the total 518 DEGs are shown in Fig. 3E. Well-defined clusters were found for genes and samples, showing similar expression patterns.

Out of all the DEGs at 2.5 h, seven genes corresponded to transcription factors, including *psrA, rpoH* and *prtN*. At 5 h, 14 DEGs including *psrA*, *rpoH*, *prtN*, *rpoS*, *rhlR* and *ptrB* were identified as transcription factors. All transcription factors activated at 2.5 h remained active at 5 h (Supplementary Table S2). Identification of regulators by a context-centric analysis revealed a total of 22 transcription factors modulating all the DEGs at 2.5 h, and most of them are part of the 28 transcription factors recognized as DEGs at 5 h (see Supplementary Table S2).

Genes of the SOS response were not identified as DEGs. The *rpoS* factor was up-regulated at 5 h. Due to the preponderant role of LexA (SOS response) and RpoS as essential genes in the response to CIP in *P. aeruginosa*, we further investigated the DNA binding sites for these elements. The CollectTF database provided the consensus binding sequence for LexA as CTG-TATAA-ATATA-CAG, described by[26]. Analysis revealed the role of LexA modulating all 15 genes in the SOS response in *P. aeruginosa,* as well as other sequences at promoter regions of *psrA* (coding for a transcription factor as described before), *grpE, hemO* and other genes. In PaeAG1, *psrA* and *grpE* genes were up-regulated at 2.5 and 5 h after CIP treatment. For RpoS, no sequence information was available in CollectTF, therefore we used the RpoS-dependent promoter consensus sequence CTATACT found by[75]. A total of 49 sites for RpoS were predicted to be associated with promoter regions of PaeAG1 genes, but none as DEGs in PaeAG1.

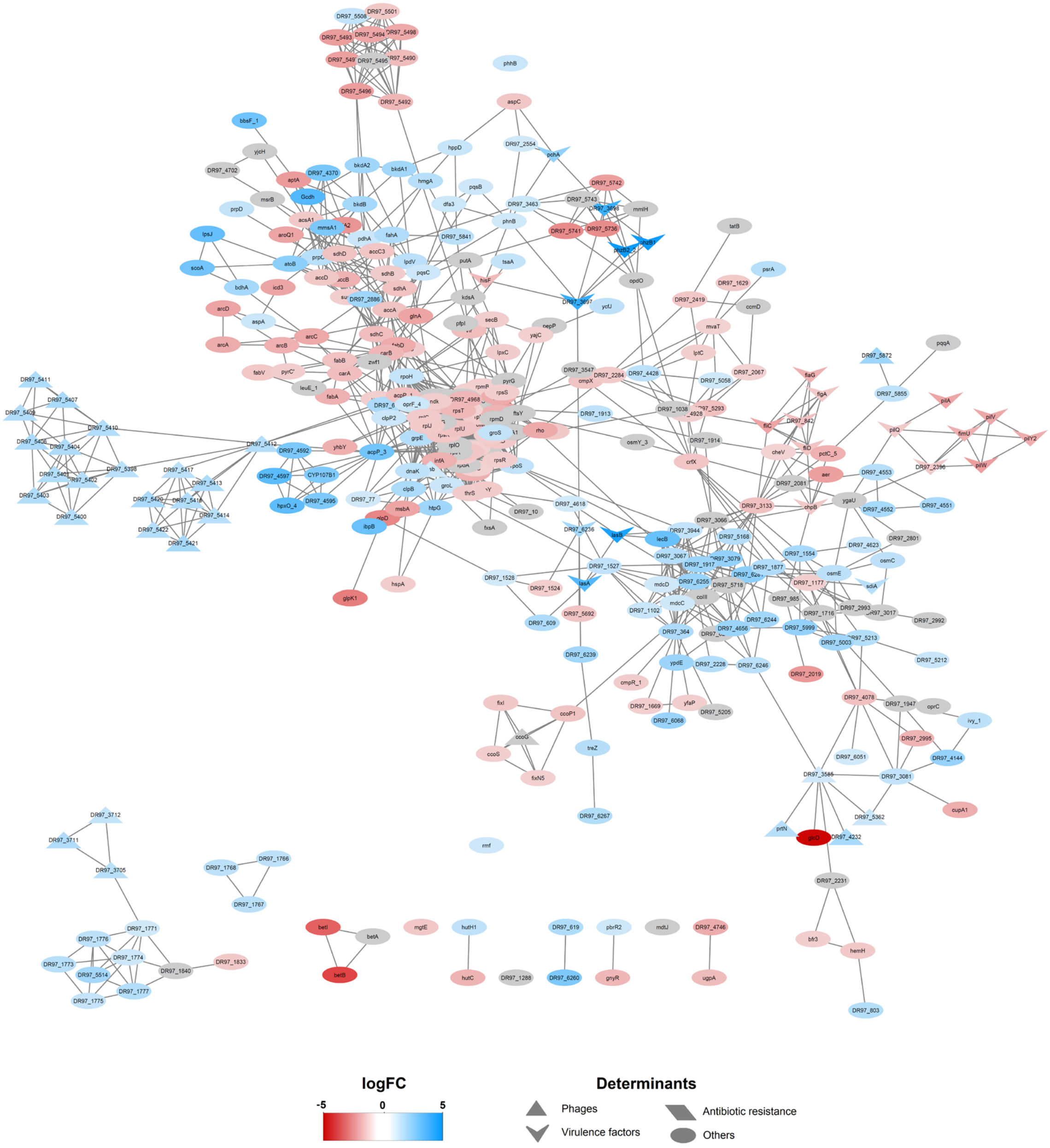


Figure 4. Gene–gene interaction (GGI) large scale network of differentially expressed genes in PaeAG1 after ciprofloxacin treatment, using a database-based method for prediction of interactions. Using STRINGdb, interactions between genes were predicted. To build the network all the DEGs in both times 2.5 and 5 h were included. A total of 342 genes resulted connected (66.0% of all DEGs) with 1685 edges in total (not connected nodes are not shown). The logFC is shown for 5 h. Gray nodes represent genes that were differentially expressed only at time 2.5 h (i.e. no logFC value is displayed at time 5 h). Details of the network by time is shown in supplementary Figure S3. Phages genes, virulence factors and antibiotic resistance genes are represented as triangles, arrowheads and rhomboids, respectively. Down-regulation (red tones) and up-regulation (blue tones).

Networks analysis shows pleiotropic effects of CIP exposure in PaeAG1. Using a top-down systems biology approach, a large scale GGI network of DEGs was built to identify molecular determinants associated with the response to CIP in PaeAG1.

GGI predictions by a database-based model: All of the 518 DEGs were incorporated as nodes and edges (high confidence connections or interactions). A total of 342 (66.0% of all DEGs) nodes were found to be connected with at least one other gene, as well as 1685 edges were established (Fig. 4). When selecting DEGs for each time, 248 nodes (69.9%) of the 355 DEGs at 2.5 h were connected with a total of 1,156 edges (supplementary Figure S2A). Out of all the 438 DEGs at 5 h, 284 (64.8%) were connected with 1,041 edges in total (supplementary Figure S2B).

As shown in Fig. 4, some determinants of virulence factors (adherence) and antibiotic resistance genes showed a down-regulation after CIP treatment, meanwhile, phage genes and other virulence factors (phenazines) were

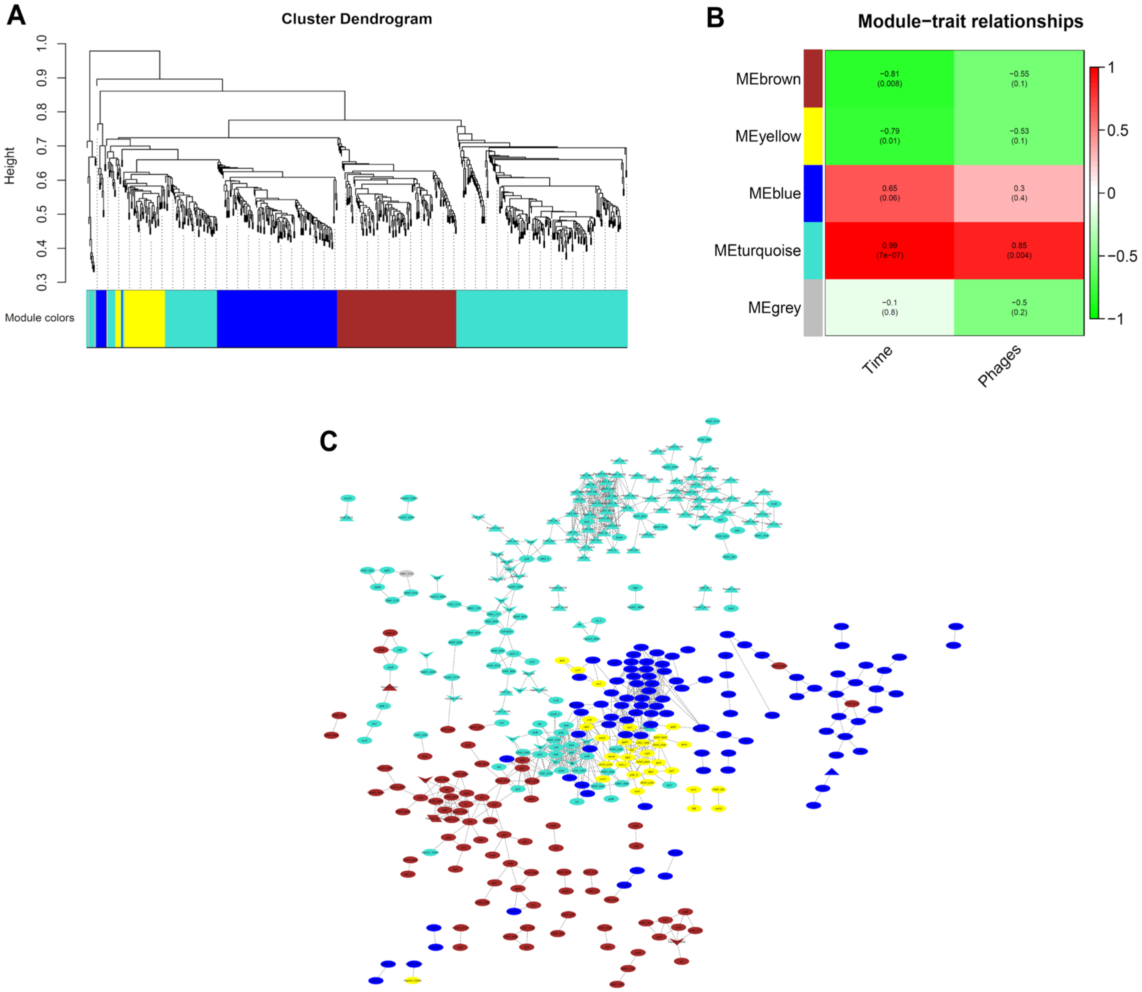


Figure 5. Co-expression analysis to identify modules of genes and the data-driven co-expression network in PaeAG1 after Ciprofloxacin treatment. (**A**) Modules identification (clusters by colors) using correlated expression genes (along times 0, 2.5 and 5 h) and clustering analysis after WGCNA was implemented. (**B**) Association of modules to traits, showing relations between turquoise and blue modules with exposure time to antibiotic and phages induction. (**C**) Data-driven co-expression network using correlation of gene expression by WGCNA analysis (correlation > 98.5%). A total of 388 DEGs were found to be connected, with a total of 1,073 edges. Only correlated genes are shown. More details in supplementary Figure S3A. Phages genes, virulence factors and antibiotic resistance genes are represented as triangles, arrowheads and rhomboids, respectively.

found to be up-regulated. In addition, gene clusters of highly connected DEGs showed the same expression pattern, suggesting a coordinated regulation.

The observed unconnected genes (107 DEGs for 2.5 h and 154 for 5 h) are inherent to limitations in the database (incomplete inclusion of phage genes) or the current state of the gene annotation (without information, hypothetical protein, etc.). To improve the associations between genes creating more connections, a data-driven co-expression analysis was run.

Co-expression analysis. Modules of highly connected genes (represented using color groups) were created using normalized counts for all the 518 DEGs. As shown in Fig. 5A, genes were clustered into four main modules, showing similar expression along samples. The number of genes belonging to the turquoise module was 239, 124 for blue, 114 brown and 39 for yellow module. In the co-expression network (Fig. 5C), a total of 388 DEGs (74.9% of the 518 DEGs) were found to be connected, with a total of 1,073 edges. Of these interactions, 385 were also found using the database-based model and 688 novel gene interactions were suggested by our co-expression analysis. The turquoise module includes most of the phage genes and virulence factors.

Integrated GGI network of DEGs. Integration of predicted connections between genes by both the database-based model and co-expression analysis was done to build a definitive large scale network, shown in Fig. 6. A total of 449 (86.7%) of DEGs were connected, in contrast with the 342 nodes from the preliminary network, an increment of ~ 20%. In addition, 2,373 edges were identified, 1685 from the database-based method (solid lines in the network) and the 688 new interactions suggested by the co-expression analysis (dashed lines). Further-

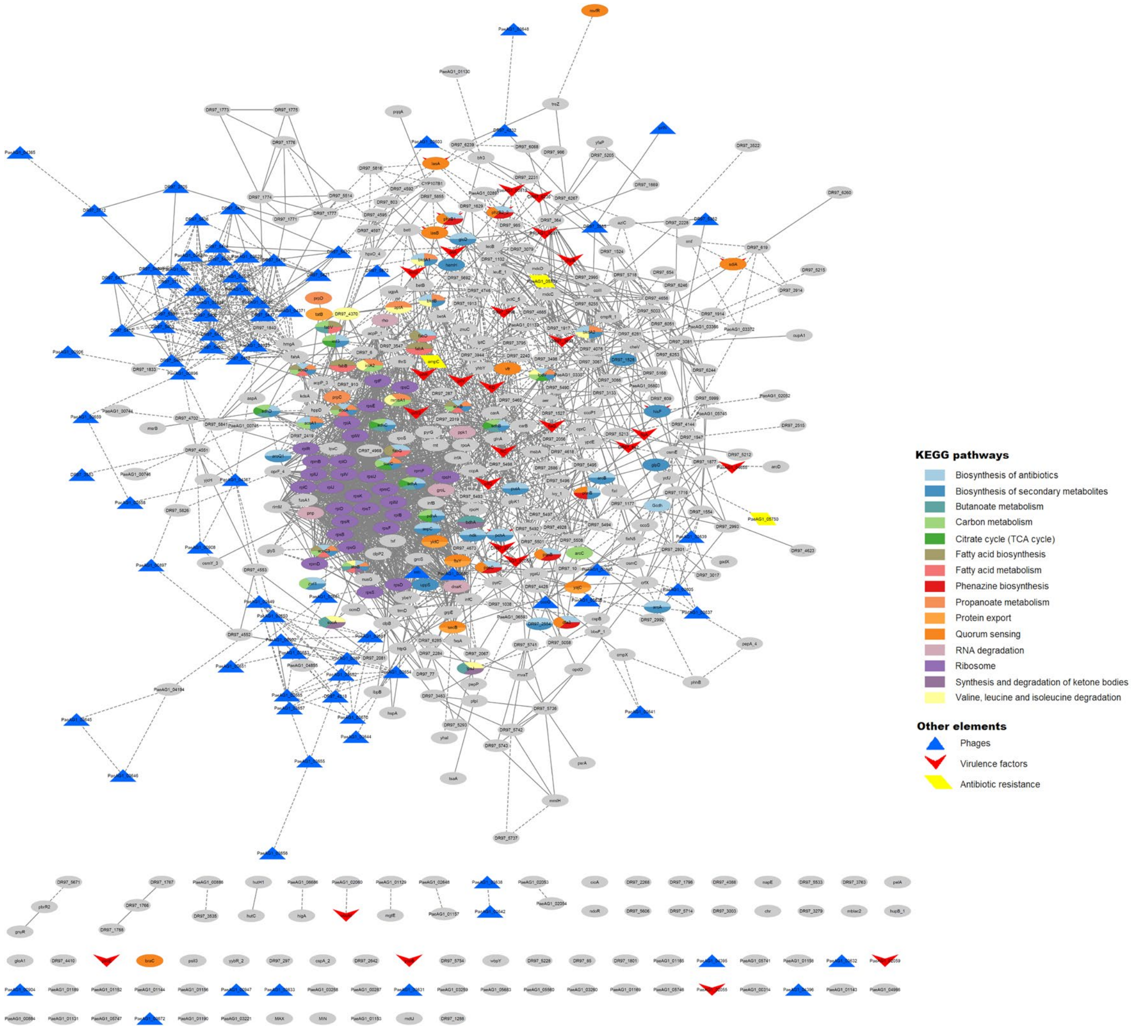


Figure 6. Definitive large scale network of DEGs, identification of hub genes and associated groups in PaeAG1 after treatment with ciprofloxacin. Network showing all 518 DEGs genes and their interactions (449 genes have at least one connection). Known interactions according to STRINGdb (database-based method) are shown as solid lines and data-driven interactions according to data-driven co-expression analysis as dashed lines. Enriched nodes associated to KEGG annotation are colored according to each pathway (more details in Table 3). Phages genes, virulence factors and antibiotic resistance genes are represented as triangles, arrowheads and rhomboids, respectively. Other genes are represented as ellipses.

more, a separated cluster was observed with high connectivity between phage genes (cluster of blue triangles, Fig. 6 left top). Remarkably, this cluster appears to have a critical bottleneck at the *fahA* gene, since many genes are connected to this node but, for the majority of the cluster nodes, this gene is the only connection to the rest of the network. Thus, the cluster becomes a clearly separated module. In addition, another smaller and less distinct cluster of phage genes was formed (Fig. 6 left down).

The same GGI network is presented in supplementary Figure S4A to show the distribution of genes by co-expression modules. A high functional interaction of genes across different clusters is observed. The logFC values at time 5 h are shown in the network in Figure S4B.

Enrichment analysis. In order to gain insight about the biological meaning of DEGs, gene set enrichment analysis (GSEA) was performed. The 518 DEGs were shown to be implemented in a total of 15 KEGG pathways (Figs. 6 and 7, and Table 3). The enriched pathways included ribosomal functions, RNA degradation, biosynthesis of antibiotics, fatty acids metabolism, propanoate metabolism, fatty acids biosynthesis, quorum sensing, amino acid degradation, carbon metabolism and citrate cycle, butanoate metabolism, phenazine biosynthesis, among others (see Fig. 7). Details of gene counts, FDR and regulation are shown in Table 3. Additionally, pathways by co-expression modules (Table 3) showed that some of them are enriched in specific pathways. For example, the blue module is down-regulated for ribosomal activity and RNA degradation (exclusive functions for this module), meanwhile the yellow module has multiple but tightly related pathways, most of them associated to interconnected metabolism pathways, down-regulated.

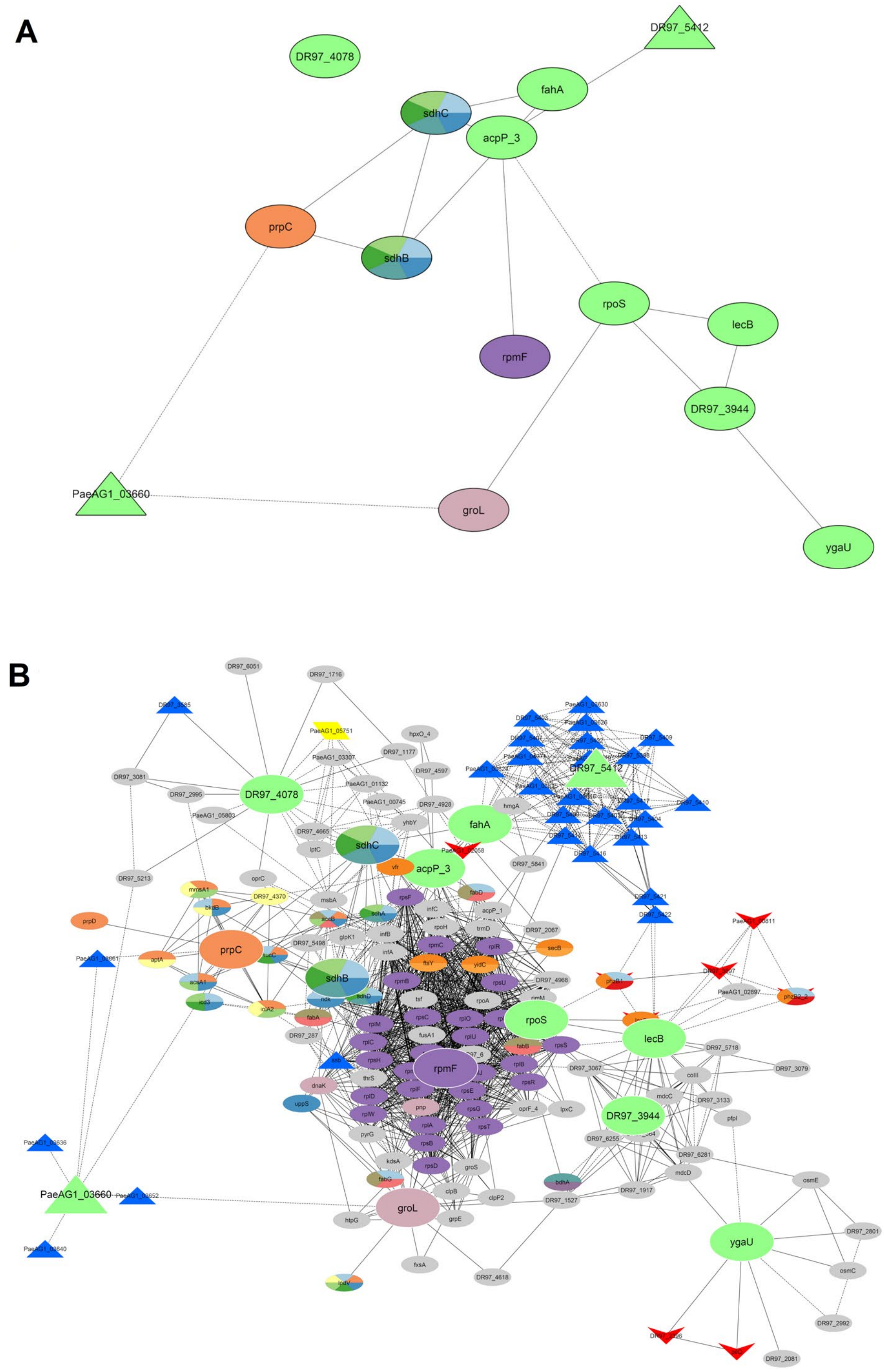


Figure 7. Identification of hub genes and first-stage subnetwork of their associated groups in PaeAG1 after treatment with ciprofloxacin. (**A**) Hub genes identification using cytoHubba (betweenness and bottleneck methods) in the network of DEGs (large nodes). Details in Table 4. (**B**) Subnetwork of nodes that directly interact with the 14 hub genes were used to build a first-stage elements network. Details of node shapes and colors are the same as described in Fig. 6.

Only 14 hub genes are able to represent the key pathways regulated by CIP in PaeAG1. With the aim of identifying an inter-modular key or central genes in the DEGs network of PaeAG1 after exposure to CIP, an analysis of hub gene identification was conducted. This approach revealed 14 connected hub genes (Fig. 7A and details in Table 4). Two genes, identified as PaeAG1_03660 and PaeAG1_03610, are part of the phage JBD44 and they were up regulated at 5 h. Topologically, they are part of the two identified phage gene clusters in the main network (Fig. 6). Two genes, *sdhB* and *sdhC,* (down-regulated) have functions related to

KEGG term ID	Term description	Total gene count	DEGs 2.5 h		DEGs 5 h		Modules	Regulation (% DEGs)*
			Observed gene count	FDR	Observed gene count	FDR		
paeb01130	Biosynthesis of antibiotics	266	30	0.0015	34	0.00047	Brown, Yellow	Down (61%)
paeb01110	Biosynthesis of secondary metabolites	320	30	0.0352	31	0.0205	Yellow	Down (70%)
paeb00650	Butanoate metabolism	37	8	0.0133	9	0.0068	Yellow	Down (55%)
paeb01200	Carbon metabolism	126	15	0.0258	18	0.0068	Yellow	Down (80%)
paeb00020	Citrate cycle (TCA cycle)	30	7	0.0158	8	0.0068	Yellow	Down (75%)
paeb00061	Fatty acid biosynthesis	27	7	0.0131	9	0.0014	Yellow	Down (100%)
paeb01212	Fatty acid metabolism	49	8	0.0309	10	0.0068	Yellow	Down (90%)
paeb00405	Phenazine biosynthesis	20	5	0.0309	6	0.0127	Brown	Up (100%)
paeb00640	Propanoate metabolism	47	12	0.00061	16	3.87e-06	Brown, Yellow	Variable (50/50)
paeb03060	Protein export	15	5	0.026	3	0.0014	Yellow	Down (100%)
paeb02024	Quorum sensing	86	11	0.0317	14	0.0068	Brown	Up (69%)
paeb03010	Ribosome	55	27	1.95e-14	27	2.63e-13	Blue	Down (100%)
paeb03018	RNA degradation	17	5	0.0258	5	0.0273	Blue	Down (60%)
paeb00072	Synthesis and degradation of ketone bodies	10	4	0.0258	4	0.0273	Brown, Turquoise	Up (100%)
paeb00280	Valine, leucine and isoleucine degradation	46	11	0.0015	11	0.0023	Brown, Turquoise	Up (82%)

Table 3. Pathways related to DEGs network of PaeAG1 exposed to ciprofloxacin, according to KEGG annotation. Annotation of modules of co-expressed genes and the general regulation are also included. *Based on logFC of DEGs at both times 2.5 and 5 h.

carbon and butanoate metabolism, and biosynthesis of secondary metabolites. Interestingly, the ribosomal protein L32 (*rpmF*, down-regulated), a chaperonin (*groL*, up-regulated) and the sigma factor (*rpoS*, up-regulated) were also identified as single molecular determinants of the network. Also, the *fahA* gene, which was previously recognized as a bottleneck for the phage genes cluster and coding for fumarylacetoacetase enzyme, was identified as a hub gene.

Analysis of gene clusters of first-stage connected genes (Fig. 7B) showed not only the same profile of enriched pathways for those hub genes (Fig. 7A), but also other pathways such as lipids metabolism, phenazine biosynthesis, quorum sensing and others. These groups include many elements of phages, virulence factors and multiple uncharacterized genes, as well as one antibiotic resistance gene (PaeAG1_05751). The logFC values at time 5 h are shown in Figure S4C.

Six hub genes were consistently identified by both bottleneck and betweenness approaches (Table 4). Together with *rpoS* and *groL*, eight hub genes (57%) are part of the turquoise module, and all of them are up-regulated by CIP. All other genes are part of the brown (4) and blue modules (2). Only four genes were found to be down regulated, three of them belonging to the brown module.

To compare the expression profiles of hub genes to other studies, we included information in Table 4 of the effect of perturbations or stressors of *P. aeruginosa* in the modulation of gene expression. Similar effects of CIP on hub genes were found when comparing our results to a previous report[26]. The effects of azithromycin seem to be opposite to CIP for these genes. More variable results were found for other perturbations (e.g. colistin, copper and H_2O_2); and *lecB* was the only hub gene that was up-regulated for all perturbations.

Thus, as expected, hub genes are strongly linked to elements of highly connected gene clusters and at the same time with the key pathways in response to CIP. Together, these three elements (hub genes, gene clusters and enriched pathways) represent the determinants of the response to CIP in PaeAG1, many of them related to the bacterial growth modulation, as initially hypothesized.

Concentration dependent effect of CIP in PaeAG1 phage induction. According to transcriptomic analysis, phage genes were up-regulated under 12.5 µg/mL CIP treatment in PaeAG1. To validate these results at phenomic level, evaluation of lytic plaque formation was done using a phage plaque assay. As shown in Fig. 8A, after treatment with 12.5 µg/mL CIP, phage induction was increased by tenfold (1,000 PFU/mL) with respect to control condition without antibiotics, in concordance with the molecular findings. More drastic changes were evidenced for higher concentrations, where more than 10 000 or 100 000 PFU/mL were quantified for PaeAG1 after treatment with 25.0 and 50.0 µg/mL CIP concentrations, respectively. Figure 8C shows phage plaques on culture plate during in vitro assays. Unlike CIP, when the same analysis was done for imipenem and tobramycin (supplementary assay), no induction was evidenced. Indeed, a slight reduction was observed for imipenem (Supplementary Figure S1C–E, right).

PaeAG1 Locus ID	Gene name	Betweenness score*	Bottleneck score*	logFC 2.5 h*	logFC 5 h*	Co-expression module	KEGG Annotation**	Annotation details	Other studies***
PaeAG1_01864	acpP (PA2966)	6,268.3	17	2.64	3.63	Turquoise	Metabolic pathways, biosynthesis of antibiotics	Acyl carrier protein; fatty acid biosynthesis	↑ AZM, ↕ CIP COL
PaeAG1_06246	ygaU	6,340.4	–	1.79	0.9	Blue	–	LysM domain/BON superfamily protein	–
PaeAG1_04068	sdhB (PA1584)	6,401.4	–	− 1.37	− 1.17	Blue	Biosynthesis of antibiotics, Carbon metabolism, Citrate cycle (TCA cycle), Butanoate metabolism, Biosynthesis of secondary metabolites	Succinate dehydrogenase and fumarate reductase iron-sulfur family protein	↑ COL AZM ↓ CIP Cu
PaeAG1_04991	prpC (PA0795)	6,485.7	14	1.58	1.7	Turquoise	Propanoate metabolism	Belongs to the citrate synthase family	↑ H_2O_2 CIP ↓ AZM ↕ COL
PaeAG1_03610	DR97_5412	7,285.4	15	0.9	1.84	Turquoise	–	Phage: JBD44; Tail tape measure protein	–
PaeAG1_05221	groL or groEL (PA4385)	8,440.2	–	1.16	1.21	Turquoise	RNA degradation	60 kDa chaperonin; Prevents misfolding and promotes the refolding and proper assembly of unfolded polypeptides generated under stress conditions	↑ CIP Cu ↓ AZM ↕ H_2O_2
PaeAG1_04071	sdhC (PA1581)	8,716.8	–	− 1.68	− 1.54	Brown	Biosynthesis of antibiotics, Carbon metabolism, Citrate cycle (TCA cycle), Butanoate metabolism, Biosynthesis of secondary metabolites	Succinate dehydrogenase, cytochrome b556 subunit	↑ AZM ↓CIP Cu
PaeAG1_03660	PaeAG1_03660	9,477.2	17	1.05	1.23	Turquoise	–	Phage: JBD44	–
PaeAG1_03555	fahA (PA2008)	11,245.9	16	1.19	1.93	Turquoise	Tyrosine metabolism	Fumarylacetoacetase	↑ CIP ↕ COL ↓ Cu AZM
PaeAG1_01837	lecB (PA3361)	13,150.8	17	1.71	3.88	Turquoise	Quorum sensing	fucose-binding lectin PA-IIL	↑ CIP COL AZM
PaeAG1_01229	DR97_3944	–	15	1.3	1.45	Brown	–	Uncharacterized protein	–
PaeAG1_01591	rpoS (PA3622)	–	15	1.03	1.49	Turquoise	Transcription machinery	RNA polymerase sigma factor RpoS	↑ COL CIP ↓ AZM ↕ Cu
PaeAG1_01361	DR97_4078	–	19	-1.22	-1.48	Brown	–	Uncharacterized protein	–
PaeAG1_02250	rpmF (PA2970)	–	22	-1.17	-1.39	Brown	Ribosome	Ribosomal protein L32; Belongs to the bacterial ribosomal protein bL32 family	↓ CIP H_2O_2 ↕ COL Cu

Table 4. Characterization of hub genes in the DEGs network of PaeAG1 after treatment with ciplofloxacin. *Cases with gray numbers refer to genes which were no selected as a DEG at that time (logFC and adjusted *p*-value). **Cases with "-" refer to no annotation information. ***Results from other studies: ↑ up-regulated, ↓down-regulated, ↕ variable regulation or "– " no information. All results from GEO-NCBI according to stress conditions: Cu (copper) from (Teitzel et al., 2006), CIP (ciprofloxacin) from (Cirz, O'Neill, Hammond, Head, & Romesberg, 2006), COL (colistin) from (Cummins, Reen, Baysse, Mooij, & O'Gara, 2009), AZM (Azithromycin) from (Kai et al., 2009) and H_2O_2 (hydrogen peroxide) from (Chang, Small, Toghrol, & Bentley, 2005).

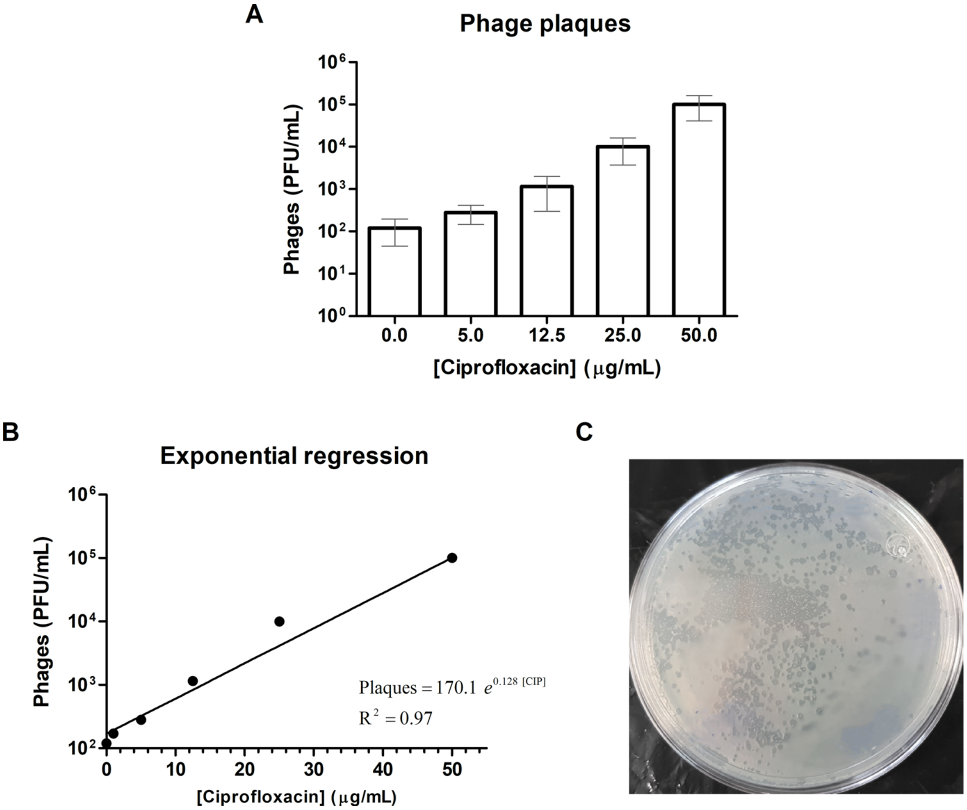


Figure 8. Phage plaques assay of PaeAG1 after exposure to ciprofloxacin. (**A**) Phages of PaeAG1 are induced under CIP exposure, with a pattern of higher induction of phage plaques at higher concentration of the drug, evidenced with an exponential regression as shown in (**B**). (**C**) Example of visualization of phage plaques on culture plate during in vitro assays.

Analysis of module genes to traits of PaeAG1 (phage production and time after CIP exposure) is presented in Fig. 5B. This analysis revealed a significant association of gene expression of the blue module, with changes at 2.5 h after CIP treatment and the low phage induction at this same time point. In a similar way, the turquoise module was significantly associated with changes of gene expression at 5 h and stronger phage induction. Other modules were not directly associated with these traits.

Altogether, these results indicate that phage induction in PaeAG1 is strongly dependent on CIP concentration, as shown with an exponential regression ($R^2 = 0.97$) in Fig. 8B.

Discussion

P. aeruginosa is a remarkable organism that can successfully resist, adapt, and survive in a wide variety of environments[29]. This versatility is conferred by the large proportion (>8%) of regulatory genes encoded in its large genome (6–7.5 Mb, 7.2 Mb in the case of PaeAG1)[5,22]. This particular case of PaeAG1 strain is a high-risk ST-111 strain isolated from an immune-compromised patient in a Costa Rican Hospital, with resistance to multiple antibiotics including CIP and carbapenems. Although many *P. aeruginosa* strains are resistant to CIP[6,10,12,48] and other antibiotics, the effects of sub-lethal concentrations on the development of antibiotic resistance had been ignored for decades due to the assumption that resistance emerges only with lethal concentrations (>MIC)[14].

Therefore, we evaluated the effect of different CIP concentrations on PaeAG1 growth rate (Fig. 2). We detected a concentration-dependent reduction of growth rate as the CIP concentration was increased, similar to another study with CIP in *P. aeruginosa*[12]. We then employed RNA-Seq analysis to investigate the influence of a sub-inhibitory CIP concentration on the gene expression of PaeAG1 and its relationship with the bacterial growth, similar to recent studies in *P. aeruginosa*[76,77] and other bacteria[16,44,78–81]. Differential expression analysis (Fig. 3) highlighted 518 DEGs at 2.5 and 5 h. Contrasting results have been previously reported in *P. aeruginosa* after CIP exposure, with some variations attributed mainly to differences in CIP concentration, time after exposure and/or the technical approach[12,26,48].

We used a top-down systems biology approach to build the interaction network across the 518 DEGs. Interactions were modeled using a database-based method and co-expression analysis. A total of 14 hub genes, gene clusters and 15 KEGG pathways were associated with the molecular response to CIP, many of them related to bacterial growth, in line with other studies[26,82,83]. Discovery and description of these strong relationships between genes provided not only biological insights of the molecular regulation under stress conditions[42], but also helped to reduce data complexity to only several central elements[40], as other studies in *P. aeruginosa* PAO1[47] and *E. coli*[40].

Sigma factor RpoS as a hub gene. Not surprisingly, one of the identified hub genes in PaeAG1 after CIP treatment was *rpoS*. This gene was only up-regulated at 5 h after exposure, suggesting a late regulation in comparison with other DEGs. RpoS is considered a master regulator of the general stress response[35] which is induced when bacterial growth decreases, or under starvation, antibiotics and osmotic or oxidative stress[18]. In addition, RpoS participates in the protection of cellular macromolecules[18], modulation of metabolism, virulence, and changes in cell envelope and morphology[11]. The overexpression of RpoS suggests that bacteria enter a stationary phase-like state upon stress conditions, as reported previously[44]. This is further supported by the observed significant lack of growth of bacteria under CIP treatment of various concentrations.

According to growth curves, PaeAG1 was in exponential phase at the time points used for the transcriptomic analysis (Fig. 2 and supplementary Figure S1A). This is a key point to ensure that RpoS induction (and all the response) is explained by the antibiotic and not due to stationary-phase entry (i.e. experimental design). The reliance of the observed changes on CIP treatment was further supported by the fact the curves at same conditions showed no changes for imipenem or tobramycin antibiotics (supplementary Figure S1C–E). Other fluoroquinolones were not tested for their effect on the production of phages in PaeAG1.

In addition, DNA binding site analysis using consensus sequence described in[75] revealed 49 sites for RpoS in PaeAG1 genes, however none of these were found to be DEGs. In the same work, RpoS was regulating 772 genes at the stationary phase, of which 41 genes (5%) were identified as DEGs in our study. Since our analysis was performed at the exponential phase, the small number of common genes could be attributed to growth phase differences in each study. In another study using a de novo approach to identify binding sites using ChIP-Seq, RpoS showed to have 199 binding motifs in *P. aeruginosa* PA14[37], including six transcription factors. In PaeAG1, 23 of these 199 genes corresponded to promoter regions of DEGs, including the RhlR and RpoS (itself) transcription factor genes. This suggests that 12% of the RpoS regulon was modulated by CIP in PaeAG1. Interestingly, context-centric analysis revealed that up to 28 transcription factors (including RpoS) are associated with the response to CIP, regulating gene expression with pleiotropic consequences and defining a crosstalk among factors in *P. aeruginosa*[37].

On the other hand, the RpoS response contributes to the robustness of bacterial cells facing stress conditions, acting synergistically with the SOS response[18]. Although SOS response is known to be induced by CIP in *P. aeruginosa* and other bacteria[20,26,27,84], in this study the SOS response was not significantly induced in response to CIP treatment at 2.5 and 5 h. The absence of SOS induction may be due to the timing and concentration of CIP treatment. In *E. coli,* dynamic models have shown that the time of response to cell stress is very fast, and stability of the SOS response can be achieved in minutes, around 30 min according to[85] or up to 90 min according to[86], until homeostasis is recovered or stronger stress responses are induced. Also, the SOS regulon of *P. aeruginosa* was established using a supra-inhibitory CIP concentration (8 × MIC) at times 30 and 120 min[26]. These differences in concentration and time (0.4 × MIC at 2.5 and 5 h for PaeAG1) could explain absence of SOS elements as DEGs. Our results are similar to another proteomic study using *P. aeruginosa*; profiles at 1.5, 5.5 and 14.5 h after CIP treatment were evaluated, and neither LexA nor other SOS proteins were differentially expressed, except for RecA, which was found to be up-regulated[87].

Phage induction as a response determinant. Regarding phage genes, two gene clusters with hub genes were defined in PaeAG1 after CIP treatment. Phage induction is known to be modulated upon stress conditions, including the SOS response[88]. As found recently for some antimicrobials, phage activity is product of pleiotropic regulation[89]. In the presence of sub-lethal concentrations of certain antibiotics, phages have been observed to be induced or to form larger phage plaques[88,90]. Under fluoroquinolones exposure, *P. aeruginosa* DNA is affected and the SOS response is triggered. In a similar manner to LexA, repressor cleavage reaction is stimulated by activated RecA, allowing virus assembly[91,92], and killing of the bacterium[93]. In some cases, alternative RecA-independent mechanisms have been described[91,94].

PaeAG1 has six prophages in the genome, including two complete elements[5]. After CIP exposure 85 phage genes were up-regulated, most of them from JBD44 (65 genes out of 105 JBD44 genes). In the co-expression analysis, when association between modules and traits was assessed, the turquoise module (Fig. 5) was significantly related to CIP exposure time and phage induction, indicating a coordinated gene expression activity belonging to this cluster/traits (Fig. 5B).

Although general information on PaeAG1 phages is scarce, there is evidence to suggest that JBD44 is one of the most prevalent in *P. aeruginosa*[95]. Effects of JBD44 induction on growth have been previously described in *P. aeruginosa* PAO1, showing that JBD44 expression significantly decreased the growth of PAO1, unlike other phages[96]. Similarly, SOS-mediated phage induction has been reported in *P. aeruginosa* PAO1[12,26] and LESB58[97]. In addition, effect evaluation of several antibiotics found that CIP and norfloxacin (another fluoroquinolone) caused a high level of phage induction, but variable results were found for other antibiotics[92]. As observed in our experiments, no induction was found for imipenem nor tobramycin (supplementary Figure S1C–E).

The underlying relationship between the up-regulation of multiple phage genes in PaeAG1 after CIP exposure and the effect on bacterial lysis was validated through the effect of CIP concentrations in the phage induction. A concentration-dependent effect of CIP on both growth curves (rate reduction, Fig. 2) and phage plaques formation (exponential increment, Fig. 8) was demonstrated. This validated the transcriptomic findings of up-regulation of phage genes in PaeAG1.

In congruence with this and the enriched pathways in PaeAG1, it has been reported that cells can adapt to stresses by disrupting their own metabolism in such a way that will impair the success of phage activity[98]. This implies that effects are observed not only on the host cell fate but also modulation of different responses, including RpoS regulation. These changes can be a product of tight modulation of functions reliant on molecular interactions from both phage and bacteria[99]. Similarly, as phages generally appear to consume amino acid metabolites[100], the bacterial up-regulation response of genes involved in amino acid catabolism has been suggested as a strategy for reducing the infection success[98] and disrupting phage propagation[100]. Blasdel et al. 2017 found that *maiA, fahA, hmgA* and *hpd* genes of tyrosine catabolism were up-regulated by *P. aeruginosa* during phage activity[98]. In our study, all four genes were up-regulated, including *fahA* as a hub gene and a bottleneck element for the main phage gene cluster, indicating a catabolic effect after exposure to CIP that may be related to phage induction. More details of the *fahA* gene are discussed later.

Although different possibilities of the regulation of phage genes have been suggested, in the case of PaeAG1 phages, most of the predicted phage genes cannot be associated with a putative function, as in other studies[26]. This complicates the interpretation of the results for particular genes[99]. Validation of phage induction at phenomic level in congruence with transcriptomic results suggests that modulation of phages by CIP (but not for imipenem or tobramycin as discussed before) in PaeAG1 is possible. This is particularly relevant since this strain is a ST-111 high-risk clone and a critical organism Priority 1 (resistant to carbapenems) according to WHO[8]. Modulation could be achieve targeting phage production as a therapeutic option, with the advantage that the induced phages are resident elements of the genome and not exogenous elements as in other studies. Thus, treatment of antibiotic-resistant bacterial infections can potentially be improved by using phage therapy and traditional antibiotics, regardless if cells are growing in biofilms or as planktonic bacteria[88]. In addition, phage therapy can be used as a bactericidal element against multiresistant strains[93]. However, this does not necessarily apply to all *P. aeruginosa* strains since phage induction in other cases (with different strains and antibiotics) have been shown to be variable[92].

Other transcriptomic determinants. Of the 15 pathways recognized as enriched in PaeAG1 after CIP treatment, ribosomal activity, RNA degradation and several metabolic routes were prominently enriched with respect to others. Reduction in the abundance of ribosomal proteins and protein implicated in cell division over time indicate a shift by tolerant cells away from growth[87], as it was evidenced by the changes in the growth curves under different CIP concentrations in PaeAG1. In the case of ribosomal activity, a cluster is clearly recognized in the whole network and the subnetwork of hub genes, where the *rpmF* gene is the up-regulated hub element. The *rpmF* gene encodes for the 50S ribosomal subunit protein L32, which is responsible for protein synthesis and membrane lipid synthesis[101]. It is also involved in multidrug tolerance by modulating biofilm formation and persister cell induction[102].

Regarding metabolism, several reports have shown a down-regulation of energy production and carbohydrates, amino acids and lipids metabolism[15,36,87,103, 104]. Five hub genes (*sdhB, sdhC, prpC, acpP* and *fahA*) are particularly associated with metabolism. For instance, *fahA* is key in the inhibition of amino acid metabolism[105], coding for a fumarylacetoacetase necessary for the tyrosine catabolism pathway. In addition, *fahA* is a topological bottleneck in the networks (Fig. 6A–C), separating the main phage genes cluster from the rest of the nodes. As detailed before, regulation of this gene could be used to restrict amino acids access to the phage and thus restraining the full phage activity[98].

In the case of RNA degradation pathways, we identified groL (or groEL) as a hub gene, a homolog of heat shock protein 60[106]. DnaK and GroL are major ubiquitous chaperones that play crucial roles in promoting protein folding during normal growth and under stress conditions[107] such as oxidative stress, antibiotics or heat[26,107,108]. In PaeAG1, both chaperones were up-regulated.

In relation to virulence factors, CIP modulated adherence and phenazines. A total of 19 DEGs implicated in adherence were identified with down-regulation observed for LPS O-antigen, flagella, and type IV pili biosynthesis elements. Similar results were found for *P. aeruginosa* after CIP treatment in another study[26]. Under other stress conditions, this down-regulation has been suggested to be a mechanism to avoid biofilm formation as a possible way to escape as planktonic cells[46] and, in general, to modulate mechanisms for colonization, survival and invasion within the host tissues[93].

Regarding phenazines, six genes were up-regulated. This profile is associated with tolerance to oxidative stress, iron availability, biofilms, virulence and killing microbial competitors[109]. Phenazine biosynthesis is regulated by the Rhl[76] and PQS[110] quorum sensing systems in *P. aeruginosa*. The *rhlR* gene was found to be up-regulated, suggesting a possible regulation of the phenazines.

More details of specific genes and their relationship with other virulence factors, antibiotic resistance and other responses (all with few number of DEGs) are discussed in the supplementary material "Extended discussion: Other transcriptomic determinants of PaeAG1 in response to CIP".

Altogether, the transcriptomic analysis in PaeAG1 allowed us to identify key molecular determinants of the response to CIP, many of them related to the bacterial grown, such as RpoS and phage induction. This agrees completely with our hypothesis in which transcriptomic response to CIP was related to bacterial growth modulation. After a DNA damage response is induced by sub-inhibitory CIP treatment, there is a subsequent pathway modulation and transcriptional changes that define changes in the bacterial growth. A conceptual representation of these results is shown in Fig. 9, aiming to integrate our results, literature reports and possible unknown connections.

All these features are particularly relevant for high-risk strains, such as PaeAG1. As it has been suggested, the biological markers of *P. aeruginosa* high-risk clones could be useful for the future design of specific treatments and infection control strategies[7]. Thus, more detailed analyses are needed to study the different levels of transcriptomic regulation in PaeAG1, including targeted expression analysis, other stress conditions, genetic and phenotypic variability, validation of the effect and power of hub genes, explorations of the relationship between presence of specific virulence traits and severity, and phage induction as a potential therapy.

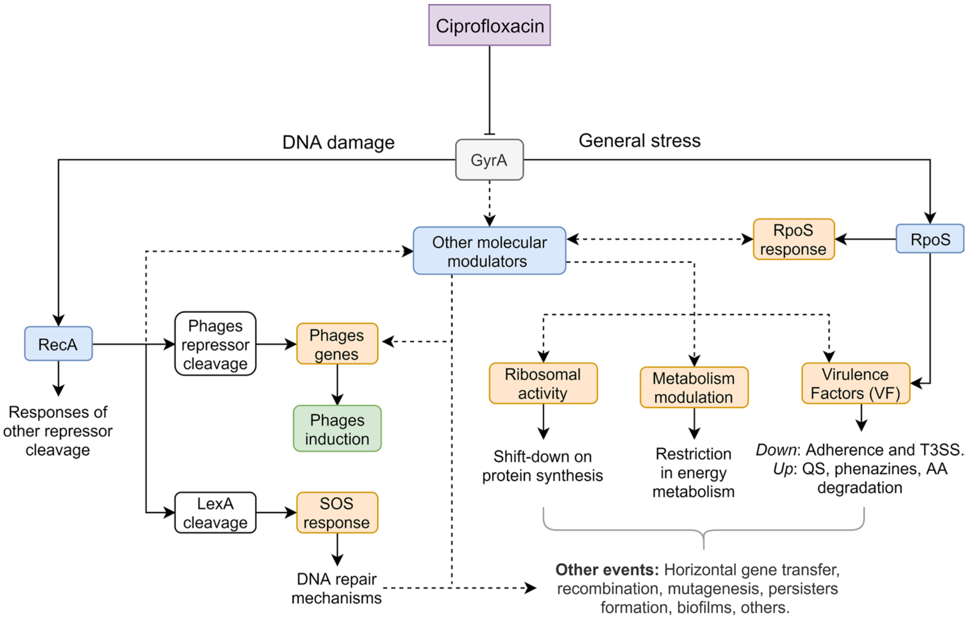

Ciprofloxacin
DNA damage
GyrA
General stress
Other molecular modulators
RpoS response
RpoS
RecA
Phages repressor cleavage
Phages genes
Phages induction
Responses of other repressor cleavage
LexA cleavage
SOS response
DNA repair mechanisms
Ribosomal activity
Metabolism modulation
Virulence Factors (VF)
Shift-down on protein synthesis
Restriction in energy metabolism
Down: Adherence and T3SS.
Up: QS, phenazines, AA degradation
Other events: Horizontal gene transfer, recombination, mutagenesis, persisters formation, biofilms, others.

Figure 9. Conceptualization of effects of ciprofloxacin treatment in PaeAG1 at the molecular level. Effects of DNA damage triggers RecA increment, which cleaves different repressors such as LexA, inducing the SOS response, but also phages induction repressors, and other elements. The general stress induces the RpoS response, modulating different responses and virulence factors. Other modulators induce changes in the metabolic state of cells, expression of virulence factors, as well as the down-shift in ribosomal activity. Together, all changes imply modulation of multiple responses with pleiotropic effects at a molecular level and regulation of phenotypes to face the stress given by the antibiotic.

Conclusions

In this work, we report a concentration-dependent reduction of PaeAG1 growth rate upon increasing sub-inhibitory CIP concentrations by comparing growth curves. The RNA-Seq analysis of PaeAG1 after treatment with a sub-inhibitory CIP concentration allowed us to identify 518 DEGs along time at 2.5 and 5 h. Using a top-down systems biology approach, we identified diverse transcriptomic determinants: 14 hub genes, multiple gene clusters and 15 enriched pathways. These included down-regulation of pathways related to metabolism, ribosomal activity and adherence factors, most of them related to bacterial growth reduction. Phages, phenazines and specific virulence factors were found to be up-regulated. In most cases, hub genes and complex relationships were identified, showing pleiotropic effects that are mainly illustrated by clusters of highly connected genes. Two particular clusters of phages genes were up-regulated by CIP. Validation of CIP effects on phage induction was done at phenomic level with a phage plaque assay, showing an exponential induction as CIP was increased. To our knowledge, this is the first report of the analysis of CIP response in a ST-111 high-risk *P. aeruginosa* strain, in particular by a combined strategy using a top-down systems biology approach. This led us to identify transcriptomic determinants in response to CIP, including resident phages induction as a potential therapeutic strategy to overcome antibiotic resistance.

Data availability

The RNA-seq raw data and processed files of transcripts quantification are available at the NCBI Gene Expression Omnibus (GEO) database under accession number GSE139866. Processed data and scripts for bioinformatics analyses (RNA-Seq data, differential expression using DESeq2 and co-expression analyses) are available at https://github.com/josemolina6/PaeAG1_CIP_RNA-Seq). Genome sequence and annotation files in all required formats for mapping and quality control of the RNA-Seq reads alignment are available from our previous work at https://github.com/josemolina6/PaeAG1_genome. More details of the genome assembly and annotation in[5].

Received: 11 March 2020; Accepted: 25 June 2020
Published online: 13 August 2020

References

1. Lyczak, J. B., Cannon, C. L. & Pier, G. B. Establishment of *Pseudomonas aeruginosa* infection: Lessons from a versatile opportunist1*Address for correspondence: Channing Laboratory, 181 Longwood Avenue, Boston, MA 02115, USA. *Microbes Infect.* **2**, 1051–1060 (2000).
2. Goldberg, J. B. 'Pseudomonas '99, The Seventh International Congress on Pseudomonas: biotechnology and pathogenesis', organized by the American Society for Microbiology, was held in Maui, HI, USA, 1–5 September 1999. *Trends Microbiol.* **8**, 55–57 (2000).
3. Wu, W. & Jin, S. PtrB of *Pseudomonas aeruginosa* suppresses the type III secretion system under the stress of DNA damage. *J. Bacteriol.* **187**, 6058–6068 (2005).

4. Silby, M. W., Winstanley, C., Godfrey, S. A. C., Levy, S. B. & Jackson, R. W. *Pseudomonas* genomes: Diverse and adaptable. *FEMS Microbiol. Rev.* **35**, 652–680 (2011).
5. 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.* **10**, 1392 (2020).
6. Toval, F. *et al.* Predominance of carbapenem-resistant *Pseudomonas aeruginosa* isolates carrying blaIMP and blaVIM metallo-β-lactamases in a major hospital in Costa Rica. *J. Med. Microbiol.* **64**, 37–43 (2015).
7. Mulet, X. *et al.* Biological markers of *Pseudomonas aeruginosa* epidemic high-risk clones. *Antimicrob. Agents Chemother.* **57**, 5527–5535 (2013).
8. World Health Organization. *Guidelines for the prevention and control of carbapenem-resistant Enterobacteriaceae, Acinetobacter baumannii and Pseudomonas aeruginosa in health care facilities*. (2017).
9. Woodford, N., Turton, J. F. & Livermore, D. M. Multiresistant Gram-negative bacteria: The role of high-risk clones in the dissemination of antibiotic resistance. *FEMS Microbiol. Rev.* **35**, 736–755 (2011).
10. Farajzadeh Sheikh, A. *et al.* Molecular epidemiology of colistin-resistant *Pseudomonas aeruginosa* producing NDM-1 from hospitalized patients in Iran. *Iran. J. Basic Med. Sci.* **22**, 38–42 (2019).
11. Firme, M., Kular, H., Lee, C. & Song, D. RpoS contributes to variations in the survival pattern of *Pseudomonas aeruginosa* in response to ciprofloxacin. *J. Exp. Microbiol. Immunol.* **14**, 21–27 (2010).
12. Brazas, M. D., Brazas, M. D., Hancock, R. E. W. & Hancock, R. E. W. Ciprofloxacin induction of a susceptibility determinant in *Pseudomonas aeruginosa*. *Antimicrob. Agents Chemother.* **49**, 3222–3227 (2005).
13. McVicker, G. *et al.* Clonal expansion during *Staphylococcus aureus* infection dynamics reveals the effect of antibiotic intervention. *PLoS Pathog.* **10**, 2 (2014).
14. Andersson, D. I. & Hughes, D. Microbiological effects of sublethal levels of antibiotics. *Nat. Rev. Microbiol.* **12**, 465–478 (2014).
15. Stewart, P. S. *et al.* Contribution of stress responses to antibiotic tolerance in *Pseudomonas aeruginosa* biofilms. *Antimicrob. Agents Chemother.* **59**, 3838–3847 (2015).
16. Matern, W. M., Rifat, D., Bader, J. S. & Karakousis, P. C. Gene enrichment analysis reveals major regulators of *Mycobacterium tuberculosis* gene expression in two models of antibiotic tolerance. *Front. Microbiol.* **9**, 1–10 (2018).
17. Hocquet, D. *et al.* Evidence for induction of integron-based antibiotic resistance by the SOS response in a clinical setting. *PLoS Pathog.* **8**, 2 (2012).
18. Dapa, T., Fleurier, S., Bredeche, M.-F. & Matic, I. The SOS and RpoS regulons contribute to bacterial cell robustness to genotoxic stress by synergistically regulating DNA polymerase Pol II. *Genetics* **206**, 1349–1360 (2017).
19. Kreuzer, K. N. DNA damage responses in prokaryotes: Regulating gene expression, modulating growth patterns, and manipulating replication forks. *Cold Spring Harbor Perspect. Biol.* https://doi.org/10.1101/cshperspect.a012674 (2013).
20. Valencia, E. Y., Esposito, F., Spira, B., Blázquez, J. & Galhardo, R. S. Ciprofloxacin-mediated mutagenesis is suppressed by subinhibitory concentrations of amikacin in *Pseudomonas aeruginosa*. *Antimicrob. Agents Chemother. AAC* https://doi.org/10.1128/AAC.02107-16 (2016).
21. Siqueira, V. L. D. *et al.* Structural changes and differentially expressed genes in *Pseudomonas aeruginosa* exposed to meropenem-ciprofloxacin combination. *Antimicrob. Agents Chemother.* **58**, 3957–3967 (2014).
22. Cabot, G. *et al.* Evolution of *Pseudomonas aeruginosa* antimicrobial resistance and fitness under low and high mutation rates. *Antimicrob. Agents Chemother.* **60**, 1767–1778 (2016).
23. Knezevic, P., Curcin, S., Aleksic, V., Petrusic, M. & Vlaski, L. Phage-antibiotic synergism: A possible approach to combatting *Pseudomonas aeruginosa*. *Res. Microbiol.* **164**, 55–60 (2013).
24. Dörr, T., Lewis, K. & Vulić, M. SOS Response induces persistence to fluoroquinolones in *Escherichia coli*. *PLoS Genet.* **5**, e1000760 (2009).
25. Recacha, E. *et al.* Quinolone resistance reversion by targeting the SOS response. *MBio* **8**, 2 (2017).
26. Cirz, R. T., O'Neill, B. M., Hammond, J. A., Head, S. R. & Romesberg, F. E. Defining the *Pseudomonas aeruginosa* SOS response and its role in the global response to the antibiotic ciprofloxacin. *J. Bacteriol.* **188**, 7101–7110 (2006).
27. Breidenstein, E. B. M., Bains, M. & Hancock, R. E. W. Involvement of the lon protease in the SOS response triggered by ciprofloxacin in *Peudomonas aeruginosa* PAO1. *Antimicrob. Agents Chemother.* **56**, 2879–2887 (2012).
28. Shiba, T., Tsutsumi, K., Ishige, K. & Noguchi, T. Inorganic polyphosphate and polyphosphate kinase: Their novel biological functions and applications. *Biochem.* **65**, 315–323 (2000).
29. Suh, S. J. *et al.* Effect of rpoS mutation on the stress response and expression of virulence factors in *Pseudomonas aeruginosa*. *J. Bacteriol.* **181**, 3890–3897 (1999).
30. Weber, H. *et al.* Genome-wide analysis of the general stress response network in *Escherichia coli*: sigmaS-dependent genes, promoters, and sigma factor selectivity. *Society* **187**, 1591–1603 (2005).
31. Kayama, S. *et al.* The role of *rpoS* gene and quorum-sensing system in ofloxacin tolerance in *Pseudomonas aeruginosa*. *FEMS Microbiol. Lett.* **298**, 184–192 (2009).
32. Hong, S. H., Wang, X., O'Connor, H. F., Benedik, M. J. & Wood, T. K. Bacterial persistence increases as environmental fitness decreases. *Microb. Biotechnol.* **5**, 509–522 (2012).
33. Baharoglu, Z. & Mazel, D. SOS the formidable strategy of bacteria against aggressions. *FEMS Microbiol. Rev.* **38**, 2 (2014).
34. Balasubramanian, D. *et al.* The regulatory repertoire of pseudomonas aeruginosa AmpC ß-lactamase regulator AmpR includes virulence genes. *PLoS ONE* **7**, 2 (2012).
35. Nguyen, H. *et al.* Negative control of RpoS synthesis by the sRNA ReaL in *Pseudomonas aeruginosa*. *Front. Microbiol.* **9**, 1–10 (2018).
36. Müller, A. U., Imkamp, F. & Weber-Ban, E. The mycobacterial LexA/RecA-independent DNA damage response is controlled by PafBC and the pup-proteasome system. *Cell Rep.* **23**, 3551–3564 (2018).
37. Schulz, S. *et al.* Elucidation of sigma factor-associated networks in *Pseudomonas aeruginosa* reveals a modular architecture with limited and function-specific crosstalk. *PLoS Pathog.* **11**, 1–21 (2015).
38. van Dam, S., Võsa, U., van der Graaf, A., Franke, L. & de Magalhães, J. P. Gene co-expression analysis for functional classification and gene-disease predictions. *Brief. Bioinform.* **19**, 575–592 (2018).
39. Linde, J., Schulze, S., Henkel, S. G. & Guthke, R. Data- and knowledge-based modeling of gene regulatory networks: An update. *EXCLI J.* **14**, 346–378 (2015).
40. Liu, W. *et al.* Construction and analysis of gene co-expression networks in *Escherichia coli*. *Cells* **7**, 19 (2018).
41. Khaledi, A. *et al.* Transcriptome profiling of antimicrobial resistance in *Pseudomonas aeruginosa*. *Antimicrob. Agents Chemother.* **60**, 4722–4733 (2016).
42. Fang, G. *et al.* Transcriptomic and phylogenetic analysis of a bacterial cell cycle reveals strong associations between gene co-expression and evolution. *BMC Genom.* **14**, 2 (2013).
43. Langfelder, P. & Horvath, S. WGCNA: An R package for weighted correlation network analysis. *BMC Bioinform.* **9**, 2 (2008).
44. Lovelace, A. H., Smith, A. & Kvitko, B. H. Pattern-triggered immunity alters the transcriptional regulation of virulence-associated genes and induces the sulfur starvation response in pseudomonas syringae pv. tomato DC3000. *Mol. Plant-Microbe Interact.* **31**, 750–765 (2018).

45. Dai, H., Zhou, J. & Zhu, B. Gene co-expression network analysis identifies the hub genes associated with immune functions for nocturnal hemodialysis in patients with end-stage renal disease. *Med. (United States)* **97**, 1–8 (2018).
46. Chan, K.-G. *et al.* Transcriptome analysis of *Pseudomonas aeruginosa* PAO1 grown at both body and elevated temperatures. *PeerJ* **4**, e2223 (2016).
47. Anupama, R., Sajitha Lulu, S., Mukherjee, A. & Babu, S. Cross-regulatory network in *Pseudomonas aeruginosa* biofilm genes and TiO2 anatase induced molecular perturbations in key proteins unraveled by a systems biology approach. *Gene* **647**, 289–296 (2018).
48. Molina-Mora, J. A., Campos-Sanchez, R. & Garcia, F. Gene Expression Dynamics Induced by Ciprofloxacin and Loss of Lexa Function in Pseudomonas aeruginosa PAO1 Using Data Mining and Network Analysis. in *2018 IEEE International Work Conference on Bioinspired Intelligence (IWOBI)* 1–7 (IEEE, 2018). doi: 10.1109/IWOBI.2018.8464130
49. Stojakovic, A., Mastronardi, C. A., Licinio, J. & Wong, M.-L. Long-term consumption of high-fat diet impairs motor coordination without affecting the general motor activity. *J. Transl. Sci.* **5**, 1–10 (2018).
50. Bjursell, M. *et al.* Ageing Fxr deficient mice develop increased energy expenditure, improved glucose control and liver damage resembling NASH. *PLoS ONE* **8**, 2 (2013).
51. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
52. Andrews, S. FastQC a quality control tool for high throughput sequence data. (2010). Available at: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/. (Accessed: 10th April 2018)
53. Wingett, S. W. & Andrews, S. FastQ Screen: A tool for multi-genome mapping and quality control. *F1000Research* **7**, 1338 (2018).
54. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).
55. Magoc, T., Wood, D. & Salzberg, S. L. EDGE-pro: estimated degree of gene expression in prokaryotic genomes. *Evol. Bioinform. Online* **9**, 127–136 (2013).
56. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
57. Okonechnikov, K., Conesa, A. & García-Alcalde, F. Qualimap 2: Advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* **32**, 292–294 (2016).
58. Wang, L., Wang, S. & Li, W. RSeQC: Quality control of RNA-seq experiments. *Bioinformatics* **28**, 2184–2185 (2012).
59. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
60. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. (2020).
61. Winsor, G. L. *et al.* Enhanced annotations and features for comparing thousands of *Pseudomonas* genomes in the *Pseudomonas* genome database. *Nucleic Acids Res.* **44**, D646–D653 (2016).
62. Szklarczyk, D. *et al.* STRING v11: Protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).
63. Shannon, P. *et al.* Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
64. Mine, A. *et al.* The defense phytohormone signaling network enables rapid, high-amplitude transcriptional reprogramming during effector-triggered immunity[OPEN]. *Plant Cell* **30**, 1199–1219 (2018).
65. Wang, X. *et al.* Weighted gene co-expression network analysis for identifying hub genes in association with prognosis in Wilms tumor. *Mol. Med. Rep.* **19**, 2041–2050 (2019).
66. Cao, L. *et al.* Identification of hub genes and potential molecular mechanisms in gastric cancer by integrated bioinformatics analysis. *PeerJ* **6**, e5180 (2018).
67. Chin, C.-H. *et al.* cytoHubba: identifying hub objects and sub-networks from complex interactome. *BMC Syst. Biol.* **8**, S11 (2014).
68. Teitzel, G. M. M. *et al.* Survival and growth in the presence of elevated copper: Transcriptional profiling of copper-stressed *Pseudomonas aeruginosa*. *J. Bacteriol.* **188**, 7242–7256 (2006).
69. Cummins, J., Reen, F. J., Baysse, C., Mooij, M. J. & O'Gara, F. Subinhibitory concentrations of the cationic antimicrobial peptide colistin induce the pseudomonas quinolone signal in *Pseudomonas aeruginosa*. *Microbiology* **155**, 2826–2837 (2009).
70. Kai, T. *et al.* A low concentration of azithromycin inhibits the mRNA expression of N-acyl homoserine lactone synthesis enzymes, upstream of lasI or rhlI, *Pseudomonas aeruginosa*. *Pulm. Pharmacol. Ther.* **22**, 483–486 (2009).
71. Chang, W., Small, D. A., Toghrol, F. & Bentley, W. E. Microarray analysis of *Pseudomonas aeruginosa* reveals induction of pyocin genes in response to hydrogen peroxide. *BMC Genom.* **6**, 1–14 (2005).
72. Ceyssens, P.-J. *Isolation and characterization of lytic bacteriophages infecting Pseudomonas aeruginosa* (Katholieke Universiteit Leuven, Flanders, 2009).
73. Schwab, K. J., De Leon, R. & Sobsey, M. D. Concentration and purification of beef extract mock eluates from water samples for the detection of enteroviruses, hepatitis A virus, and Norwalk virus by reverse transcription-PCR. *Appl. Environ. Microbiol.* **61**, 531–537 (1995).
74. Paterson, W. D., Douglas, R. J., Grinyer, I. & McDermott, L. A. Isolation and preliminary characterization of some *Aeromonas salmonicida* bacteriophages. *J. Fish. Res. Board Canada* **26**, 629–632 (1969).
75. Schuster, M., Hawkins, A. C., Harwood, C. S. & Greenberg, E. P. The *Pseudomonas aeruginosa* RpoS regulon and its relationship to quorum sensing. *Mol. Microbiol.* **51**, 973–985 (2004).
76. 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.* **10**, 1–14 (2019).
77. Fernández, M., Corral-Lugo, A. & Krell, T. The plant compound rosmarinic acid induces a broad quorum sensing response in *Pseudomonas aeruginosa* PAO1. *Environ. Microbiol.* **20**, 4230–4244 (2018).
78. Salmon-Divon, M., Zahavi, T. & Kornspan, D. Transcriptomic analysis of the brucella melitensisrev.1 vaccine strain in an acidic environment: Insights into virulence attenuation. *Front. Microbiol.* **10**, 1–12 (2019).
79. Thode, S. K. *et al.* Construction of a fur null mutant and RNA-sequencing provide deeper global understanding of the *Aliivibrio salmonicida* Fur regulon. *PeerJ* **2017**, 2 (2017).
80. Mets, T. *et al.* Fragmentation of *Escherichia coli* mRNA by MazF and MqsR. *Biochimie* **156**, 79–91 (2019).
81. Cabezas, C. E. *et al.* The transcription factor SlyA from Salmonella Typhimurium regulates genes in response to hydrogen peroxide and sodium hypochlorite. *Res. Microbiol.* **169**, 263–278 (2018).
82. Fornelos, N., Browning, D. F. & Butala, M. The use and abuse of LexA by mobile genetic elements. *Trends Microbiol.* **24**, 391–401 (2016).
83. Stockwell, V. O. & Loper, J. E. The sigma factor RpoS is required for stress tolerance and environmental fitness of *Pseudomonas fluorescens* Pf-5. *Microbiology* **151**, 3001–3009 (2005).
84. Goerke, C., Koller, J. & Wolz, C. Ciprofloxacin and trimethoprim cause phage induction and virulence modulation in *Staphylococcus aureus*. *Antimicrob. Agents Chemother.* **50**, 171–177 (2006).
85. Friedman, N., Vardi, S., Ronen, M., Alon, U. & Stavans, J. Precise temporal modulation in the response of the SOS DNA repair network in individual bacteria. *PLoS Biol.* **3**, e238 (2005).

86. Ronen, M., Rosenberg, R., Shraiman, B. I. & Alon, U. Assigning numbers to the arrows: Parameterizing a gene regulation network by using accurate expression kinetics. *Proc. Natl. Acad. Sci. USA*. **99**, 10555–10560 (2002).
87. Babin, B. M. *et al.* Selective proteomic analysis of antibiotic-tolerant cellular subpopulations in pseudomonas aeruginosa biofilms. *MBio* **8**, 2 (2017).
88. Kamal, F. & Dennis, J. J. Burkholderia cepacia complex phage-antibiotic synergy (PAS): Antibiotics stimulate lytic phage activity. *Appl. Environ. Microbiol.* **81**, 1132–1138 (2015).
89. Burmeister, A. R. *et al.* Pleiotropy complicates a trade-off between phage resistance and antibiotic resistance. *Proc. Natl. Acad. Sci. USA*. https://doi.org/10.1073/pnas.1919888117 (2020).
90. Ryan, E. M., Alkawareek, M. Y., Donnelly, R. F. & Gilmore, B. F. Synergistic phage-antibiotic combinations for the control of *Escherichia coli* biofilms *in vitro*. *FEMS Immunol. Med. Microbiol.* **65**, 395–398 (2012).
91. Golais, F., Hollý, J. & Vítkovská, J. Coevolution of bacteria and their viruses. *Folia Microbiol. (Praha)* **58**, 177–186 (2013).
92. Fothergill, J. L. *et al.* Effect of antibiotic treatment on bacteriophage production by a cystic fibrosis epidemic strain of *Pseudomonas aeruginosa*. *Antimicrob. Agents Chemother.* **55**, 426–428 (2011).
93. Chatterjee, M. *et al.* Antibiotic resistance in *Pseudomonas aeruginosa* and alternative therapeutic options. *Int. J. Med. Microbiol.* **306**, 48–58 (2016).
94. Rozanov, D. V., D'Ari, R. & Sineoky, S. P. RecA-independent pathways of lambdoid prophage induction in *Escherichia coli*. *J. Bacteriol.* **180**, 6306–6315 (1998).
95. Xie, X. T. *Characterization of the fecal virome and fecal virus shedding patterns of commercial mink (Neovison vison)* (University of Guelph, Guelph, 2017).
96. Tsao, Y. F. *et al.* Phage morons play an important role in *Pseudomonas aeruginosa* phenotypes. *J. Bacteriol.* **200**, 1–15 (2018).
97. Winstanley, C. *et al.* Newly introduced genomic prophage islands are critical determinants of in vivo competitiveness in the Liverpool epidemic strain of *Pseudomonas aeruginosa*. *Genome Res.* **19**, 12–23 (2008).
98. Blasdel, B. G., Chevallereau, A., Monot, M., Lavigne, R. & Debarbieux, L. Comparative transcriptomics analyses reveal the conservation of an ancestral infectious strategy in two bacteriophage genera. *ISME J.* **11**, 1988–1996 (2017).
99. Chevallereau, A. *et al.* Next-generation-omics approaches reveal a massive alteration of host RNA metabolism during bacteriophage infection of *Pseudomonas aeruginosa*. *PLoS Genet.* **12**, 1–20 (2016).
100. De Smet, J. *et al.* High coverage metabolomics analysis reveals phage-specific alterations to *Pseudomonas aeruginosa* physiology during infection. *ISME J.* **10**, 1823–1835 (2016).
101. Podkovyrov, S. & Larson, T. J. Lipid biosynthetic genes and a ribosomal protein gene are cotranscribed. *FEBS Lett.* **368**, 429–431 (1995).
102. Liu, S. *et al.* Identification of novel genes including rpmF and yjjQ critical for Type II 1 persister formation in *Escherichia coli*. *bioRxiv* https://doi.org/10.1101/310961 (2018).
103. Cornforth, D. M. *et al. Pseudomonas aeruginosa* transcriptome during human infection. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 2 (2018).
104. Quintana, J., Novoa-Aponte, L. & Argüello, J. M. Copper homeostasis networks in the bacterium *Pseudomonas aeruginosa*. *J. Biol. Chem.* **292**, 15691–15704 (2017).
105. Zheng, X., Su, Y., Chen, Y., Huang, H. & Shen, Q. Global transcriptional responses of denitrifying bacteria to functionalized single-walled carbon nanotubes revealed by weighted gene-coexpression network analysis. *Sci. Total Environ.* **613–614**, 1240–1249 (2018).
106. Shin, H., Jeon, J., Lee, J.-H., Jin, S. & Ha, U.-H. *Pseudomonas aeruginosa* GroEL stimulates production of PTX3 by activating the NF-κB pathway and simultaneously downregulating MicroRNA-9. *Infect. Immun.* **85**, 2 (2017).
107. Ito, F., Tamiya, T., Ohtsu, I., Fujimura, M. & Fukumori, F. Genetic and phenotypic characterization of the heat shock response in *Pseudomonas putida*. *Microbiologyopen* **3**, 922–936 (2014).
108. Michta, E. *et al.* Proteomic approach to reveal the regulatory function of aconitase AcnA in oxidative stress response in the antibiotic producer Streptomyces viridochromogenes Tü494. *PLoS ONE* **9**, 1 (2014).
109. Wang, Y., Kern, S. E. & Newman, D. K. Endogenous phenazine antibiotics promote anaerobic survival of *Pseudomonas aeruginosa* via extracellular electron transfer. *J. Bacteriol.* **192**, 365–369 (2010).
110. Higgins, S. *et al.* Differential regulation of the phenazine biosynthetic operons by quorum sensing in *Pseudomonas aeruginosa* PAO1-N. *Front. Cell. Infect. Microbiol.* **8**, 2 (2018).

Acknowledgements

We thank the students John Rodríguez Fernández and Daniel Ulate Rodríguez for their collaboration in multiple activities of the project. We also thank the Genome Technology Center of New York University for generating all sequencing data used in this work, and to all members of Centro de Investigación en Enfermedades Tropicales (Universidad de Costa, Costa Rica) for their support in the experimental assays, and PGx group of The Human Phenome Institute (Fudan University, Shanghai-China) for their support in the bioinformatic analysis. Finally we thank Rachel O'Dea for the critical reading of this work. This work was funded by project "B8114 Definición de la red transcriptómica y de las alteraciones genómicas inducidas por ciprofloxacina en *Pseudomonas aeruginosa* AG1", Vicerrectoría de Investigación, Universidad de Costa Rica (period 2017-2019).

Author contributions

J.M.M. and F.G. participated in the conception, design of the study and data selection. D.C.M., M.C.A. and A.U.M. run the experimental assays. J.M.M. implemented all bioinformatic analysis. J.M.M., R.C.S., R.M.R. and L.S. were involved in bioinformatic analysis interpretation. J.M.M. and F.G. participated in the interpretation of the data in the biological context. J.M.M. drafted the manuscript and all authors were involved in its revision. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-70581-2.

Correspondence and requests for materials should be addressed to J.A.M.-M.

Reprints and permissions information is available at www.nature.com/reprints.

GENERAL DISCUSSION AND CONCLUSIONS

Antibiotic resistance is a major threat to public health because of its continuous emergence, worldwide spread, and increasing prevalence (Hong et al., 2015). Unlike highly host-adapted pathogens and symbionts undergoing genome reduction, as a versatile environmental organism, *P. aeruginosa* continually expands its genomic repertory (Mathee et al., 2008). With a high-risk ST-111 profile, PaeAG1 is a critical organism given its resistance to multiple antibiotics, including carbapenems (World Health Organization, 2017). In this context, a comprehensive multi-omics approach was implemented to study the molecular determinants of antibiotic tolerance in this strain.

The case of PaeAG1 genome assembly was a first and important step to understand the genomic architecture of an ST-111 high-risk strain. A *de novo* approach was preferred since PaeAG1 has around 1.0 Mb of an additional DNA sequence when compared to the reference genome. These exclusive regions are composed of 57 genomic islands harboring two MBL-carrying integrons, pro-phages, and many other genes. The annotation also revealed all the genomic content and molecular determinants related to phenotypes, which for PaeAG1 are related to multi-resistance and virulence mainly.

As it was shown here, those advances in sequencing technology play an outstanding and determinant role in infection investigation and tracking evolution of international lineage of high-risk bacterial clones in clinical context over long times and in great detail (Dößelmann et al., 2017). However, genome assembly is not obvious and it is challenged by sequencing technology, genomic features, and all bioinformatics algorithms, making it a real and open problem. An exhaustive comparison of different strategies to assembly the genome and their assessment give a better way to get close to the real genome sequence. Benchmarking using the 3C criterion is a consensus approach that includes different levels and aims of comparison for the robust selection of a final assembly. A hybrid assembly was the best approach to achieve a single circular sequence with high-quality 3C for the case of the genome of a high-risk *P. aeruginosa* strain. Thus, the best features of short and long-read sequencing technologies are included and their drawbacks are compensated.

Second, since PaeAG1 is a high-risk and critical organism due to its resistance to carbapenems, we performed a comparative genomic analysis to describe the genomic context associated with the MBL-carrying integrons. We analyzed 211 complete genome sequences using a pan-genome analysis, separating strains by MLST profile. Then, the analysis of the 57 PaeAG1 genomic islands showed a varying pattern of the presence/absence among all the strains, in particular for the closest genomes to PaeAG1. Two selected genomic island clusters, $\text{GIC}_{\text{VIM-2}}$ and $\text{GIC}_{\text{IMP-18}}$, were studied in-depth. $\text{GIC}_{\text{VIM-2}}$ sequence was completely found in other two known ST-111 strains, which contained the VIM-2-carrying integron as an old-acquaintance In59-like element. $\text{GIC}_{\text{IMP-18}}$ was partially found in another genome, but the IMP-18-carrying integron has an architecture never reported before, being considered as a novel In1666 integron. We provided new insights about the genomic determinants associated with this high-risk *P. aeruginosa* clone and its resistance to carbapenems using comparative genomics.

Third, proteomic profiles of PaeAG1 after exposure to antibiotics demonstrated that ciprofloxacin effects are similar to the control without antibiotics, contrasting with the results for other antibiotics and the growth curves. In a subsequent analysis, to study the central response to multiple perturbations in the *P. aeruginosa* group, the core perturbome, and to identify gene expression patterns, we used a machine learning approach. Using public microarray data, two independent partition strategies (single and multiple with SP and MP methods respectively) and three classification algorithms, we were able to identify 46 perturbome elements. Both, network analysis and functional annotation of these genes showed coordinated modulation of biological processes in response to multiple perturbations (including metabolism, biosynthesis and molecule binding, associated with DNA damage repairing, and aerobic respiration), all related to tolerance to stressors, growth arrest, and molecular regulation.

In the last step, the particular gene expression response to CIP in PaeAG1 was studied using RNA-Seq. A concentration-dependent reduction of the PaeAG1 growth rate upon increasing sub-inhibitory CIP concentrations was reported when comparing growth curves. The RNA-Seq analysis of PaeAG1 after treatment with a sub-inhibitory CIP concentration allowed us to identify 518 DEGs along time at 2.5 and 5 h. Using a top-down systems biology approach, we identified diverse transcriptomic determinants: 14 hub genes, multiple gene clusters, and 15 enriched pathways. These included down-regulations of pathways related to metabolism, ribosomal activity, and adherence factors, most of them related to bacterial growth reduction. Phages, phenazines, and specific virulence factors were found to be up-regulated. In most cases, hub genes and complex relationships were identified showing pleiotropic effects that are mainly illustrated by clusters of highly connected genes. Two particular clusters of phage genes were up-regulated by CIP. The validation of CIP effects on phage induction was done at a phenomic level with a phage plaque assay, showing an exponential induction as CIP was increased. To our knowledge, this is the first report of the analysis of CIP response in an ST-111 high-risk *P. aeruginosa* strain, in particular by a combined strategy using a top-down systems biology approach. This led us to identify transcriptomic determinants in response to CIP, including resident phage induction as a potential therapeutic strategy to overcome antibiotic resistance.

Together, these genomic and transcriptomic elements are molecular determinants of antibiotic tolerance and resistance in PaeAG1. This is particularly relevant for critical clones with the ability to conquer nosocomial environments and to develop a multi-resistance profile. As has been suggested, the biological markers of high-risk clones could be useful for future design of specific treatments and infection control strategies (Mulet et al., 2013). Thus, in order to study the implications of these genomic and transcriptomic determinants in PaeAG1, more detailed analyses are needed, which include: different levels of molecular regulation, other expression analyses (including proteomic level), other stress conditions to define the perturbome, genetic and phenotypic variability, validation of the effect and power of hub genes, modeling molecular circuits, explorations of the relationship between the presence of specific virulence traits and severity, and phage induction as a potential therapy to overcome resistance.

Finally, as shown here, the study of the molecular determinants in PaeAG1 was possible thanks to the integration of sequencing data, phenotypes, and bioinformatics pipelines. In view of the data complexity and results depending on algorithms, benchmarking strategies were required to analyze the data and to select the best protocols according to different criteria. Although we studied a bacterial genome (small in comparison to eukaryotic models), high-performance computational infrastructure was necessary mainly for comparative genomic and transcriptomic analyses. In addition, isolation and antibiotic resistance profiling, genome and RNA sequencing, as well as proteomic and other phenomic assays have been implemented for the last 10 years to study this bacterial model, implying a cost that can be estimated at more than $30 000, only considering sequencing and experimental assays. All these considerations remind us that these types of projects demand high-performance computational infrastructure, best bioinformatics practices, and investment in scientific research in general.

SUPPLEMENTARY MATERIAL

Two-dimensional gel electrophoresis image analysis of two *Pseudomonas aeruginosa* clones

Jose Arturo Molina-Mora[1,2], Diana Chinchilla-Montero[3], Carolina Castro-Peña[1,2], Fernando García[1,2]

[1] Centro de Investigación en Enfermedades Tropicales, Universidad de Costa Rica, San José, Costa Rica

[2] Facultad de Microbiología, Universidad de Costa Rica, San José, Costa Rica

[3] Instituto Costarricense de Investigación y Enseñanza en Nutrición y Salud (INCIENSA), Tres Ríos, Costa Rica

jose.molinamora@ucr.ac.cr

Abstract. A classical strategy to analyze the protein content of a biological sample is the two-dimensional gel electrophoresis (2D-GE). This technique separates proteins by both isoelectric point and molecular weight, and images are taken for subsequent analyses. However, analyses of 2D-GE images require standardized image analysis due to susceptibility of gels to get deformed, presence of overlapping spots and stripes, fuzzy and unstained spots, and others. This represent a difficulty for final users (researchers), which demand for free and user-friendly solutions. We have previously reported the standardization of a protocol to analyze 2D-GE images, and in the current study we applied it to two new bacterial isolates *Pseudomonas aeruginosa* C25 and C50. We first extracted periplasmic proteins after exposure to antibiotics, and we then run a 2D-GE analysis. Images were analyzed using our standardized protocol, achieving the identification of protein spots using CellProfiler after pre-processing step. Comparison between strains was done using differential spot analysis, revealing a specific pattern in the protein expression between bacteria. These results will help to study the biological meaning of these strains using proteomic profiling under different conditions.

Keywords: 2D-GE, Image analysis, CellProfiler, *P. aeruginosa* C25, *P. aeruginosa* C50.

1 Introduction

The study of the protein content in biological systems is the main study subject of proteomics. This included not only to identify the particular proteins that are expressed that can explain a biological context, but also the comparison between conditions to recognize differential proteomic patterns [1].

A classical strategy to analyze the proteomic profile of a sample is the two-dimensional gel electrophoresis (2D-GE) [2]. This technique separates proteins in a layer of polyacrylamide gel by both isoelectric point (pI, pH at which a molecule is electrically neutral) and molecular weight [3], creating spots that are then stained.

Analyses of 2D-GE images require standardized image analysis [3], due to susceptibility of gels to get deformed, presence of overlapping spots and stripes, fuzzy and unstained spots, and others. [1], [4]. However, the 2D-GE image analysis is not straightforward. This represent a difficulty for final users (such as microbiologist, biologist and researchers in general), which demand for user-friendly solutions. However, these user-friendly software are expensive commercial packages. Free options regularly requires command-line work, making it a drawback for researchers.

In this scenario, we have previously reported the standardization of a protocol to analyze 2D-GE images using the Costa Rican bacteria *Pseudomonas aeruginosa* AG1 as model [5]. Now, in this work we applied our protocol to two new isolates, *P. aeruginosa* C25 and C50, which are two clones obtained from the former strain when exposed to high ciprofloxacin (antibiotic) concentrations. *P. aeruginosa* is an opportunistic bacteria able to infect immunocompromised hosts, which is frequently associated with antibiotic multiresistance [6]. The three Costa Rican isolates have a multiresistance profile. They are categorized as a high risk clones because are coming from a strain causing infections in hospitals. Thus, the goal of this study was to implement and assess an image analysis protocol using our previously reported protocol to identify protein spots in 2D-GE gels images from two *P. aeruginosa* strains C25 and C50.

To achieve this, we first extracted periplasmic proteins of *P. aeruginosa* C25 and C50 after exposure to antibiotics, and we then run a 2D-GE analysis. Images were analyzed using our standardized protocol, by identifying spots using CellProfiler. Then, comparison between conditions was done using differential spot analysis.

2 Methods

For the extraction of periplasmic proteins of *P. aeruginosa* C25 and C50, we followed the protocol by [5], [7]. Briefly, cells were cultured until the exponential phase in LB medium. The 2D-GE was performed using strips for separation by isoelectric point (GE HealthCare Immobiline Dry Strip GelsTM), and a SDS-GE gradient was done for the molecular weight separations. Images were taken using ChemiDoc™ photo viewer (BioRad®).

The processing step included an image alignment using bUnwarpJ package in the ImageJ program [8]. In this program, five spots were used as reference for the deformation of images and to achieve the alignment. Identification of spots was done using our previously reported protocol [5]. Briefly, CellProfiler (https://CellProfiler.org/) was used to analyze images following the next steps: images inversion, primary object recognition and segmentation, manual editing, intensity measuring and visualization of objects.

To compare 2D-GE images, a differential spot analysis was implemented. Pairs of images were compared to identify shared spots using an analysis of primary objects (segmentation) of overlapping spots, identification of exclusive spots in each image using the no-overlapping regions, and the subsequent representation spot borders separating shared (red circles) or exclusive dots (green or blue circles).

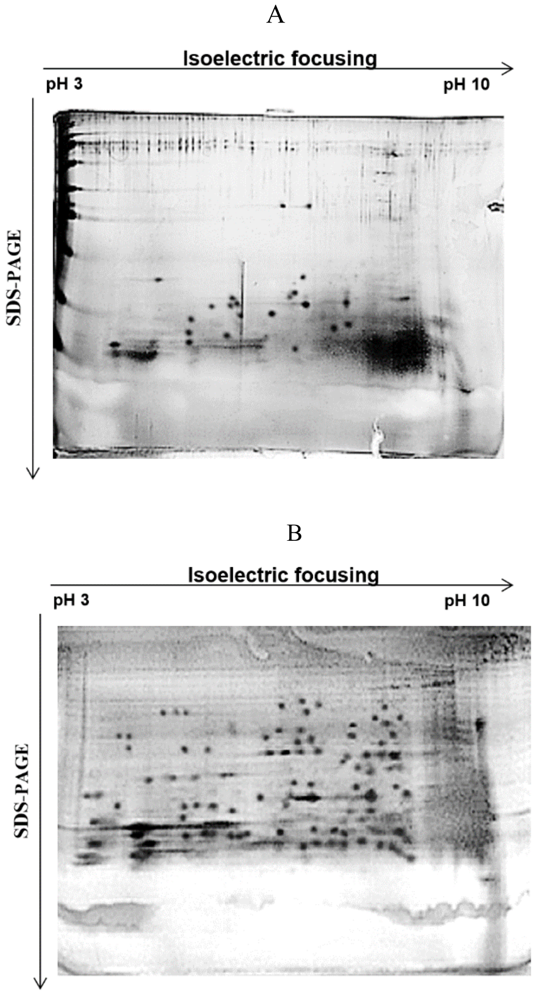


Fig. 1. Example of two-dimensional gel electrophoresis (2D-GE) of *P. aeruginosa* C25 (A) and C50 (B) after growing in LB medium. Assays was performed after cells were growth in LB medium.

3 Results and discussion

Proteomics is considered an essential field for the systematic analysis of biological systems, an assessment of changes in the abundance of proteins that occur in living organisms and that can be studied at various levels [4].

The two-dimensional gel electrophoresis 2D-GE is a classical technique used to analyze the protein content in biological samples [1]. Here we first performed a 2D-GE assay for the bacterial clones P. aeruginosa C25 and C50, as shown in Fig. 1-A-B.

However, 2D-GE image analysis requires specific protocols due to image complexity [3]. In this way, we previously established a standardized protocol to identify protein spots using CellPro-filer and other image analysis tools [5].

For the pre-processing step, bUnwarpJ package in the ImageJ program was used to align images. According to this pipelines, five points between the target image (to be modified) and a reference image are selected as common denominator to make the alignment, creating a deformation field and grid (Fig. 2-A-B).

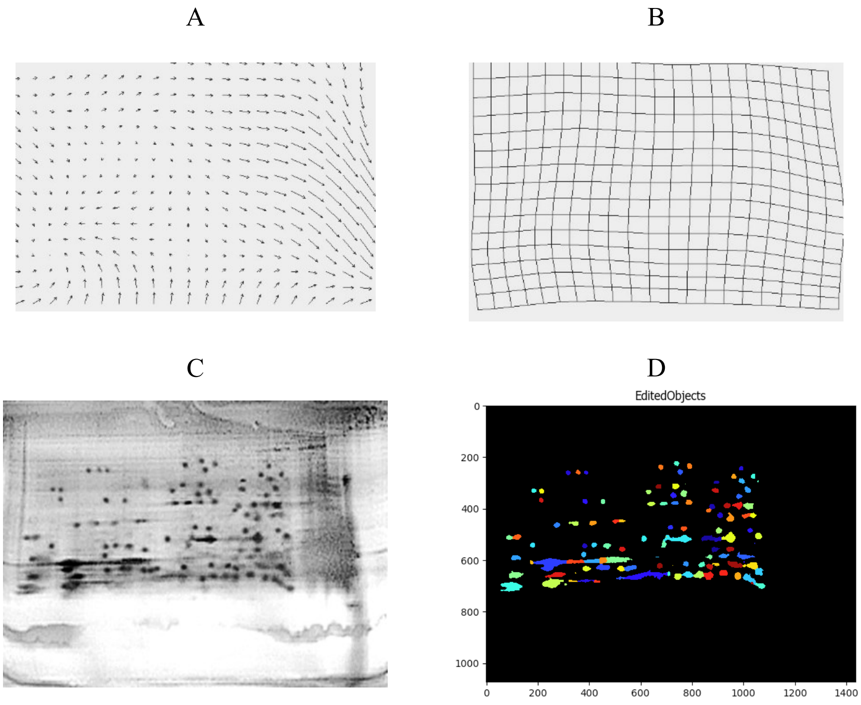


Fig. 2. Analysis of 2D-GE images. Examples of deformation field (A) and deformation grid (B) to align images against a reference in the pre-processing step. (C) Example of a raw image used for the identification of spots using CellProfiler pipeline, as resulted in (D).

As shown in Fig. 2-C-D, identification of spots was achieved using CellProfiler software. Different metrics were used to optimize the segmentation algorithm, as previously described [5]. Although automatic spot recognition is sensitive to complex regions, manual edition helped to solve these drawbacks. Commercial solutions have similar tools to deal with this particular features that are common un 2D-GE image analyses [3].

With a modified protocol, the pipeline was also able to recognize common and shared spots when comparison of proteomic profiles of the two strains was done.

For this, a new consensus image was built using image operations (pixel operations), making possible the identification of common spots, which were identified in the same way as before but using the new image. After subtraction of shared dots, exclusive spots were marked and a final visualization was done in the initial images, as shown in Fig. 3.

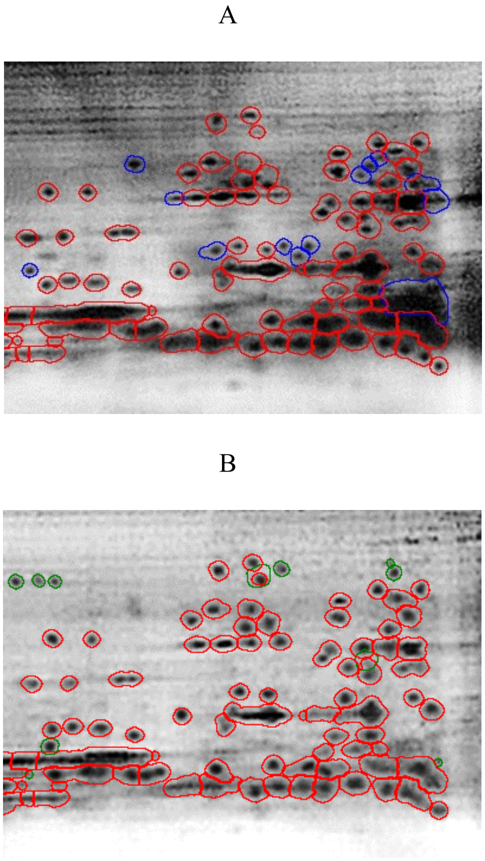


Fig. 3. Example of the differential spot identification with 2D-GE images from two *P. aeruginosa* strains C25 (A) and C50 (B). Shared spots were identified using red circles, and exclusive spots were marked as blue or green spots.

Regarding the CellProfiler program, this is a known tool used for cell imaging, for example for microscopy images. However, as we have demonstrated before [5] and here, it is possible to use the algorithms to recognize spots in 2D-GE images. See our previous work for details of the implementations, more details of the pipeline and comparison of samples [5].

In summary, in this work we presented a new analysis of 2D-GE images using a standardized protocol to identify spots and compare conditions by proteomic profile. This was done using two *P. aeruginosa* clones, in which was possible to identify both shared and exclusive dots. Although this work is focused on the image analysis, these results will help us to apply this protocol to study *P. aeruginosa* strains under different experimental conditions, including antibiotics or other stressors and their effect on the proteomic profile of the bacteria.

References

[1] M. M. Goez, M. C. Torres-Madroñero, S. Röthlisberger, and E. Delgado-Trejos, "Preprocessing of 2-Dimensional Gel Electrophoresis Images Applied to Proteomic Analysis: A Review.," *Genomics. Proteomics Bioinformatics*, vol. 16, no. 1, pp. 63–72, 2018.

[2] P. H. O'Farrell, "High resolution two-dimensional electrophoresis of proteins.," *J. Biol. Chem.*, vol. 250, no. 10, pp. 4007–21, May 1975.

[3] M. Natale, B. Maresca, P. Abrescia, and E. M. Bucci, "Image analysis workflow for 2-D electrophoresis gels based on imageJ," *Proteomics Insights*, vol. 4, pp. 37–49, 2011.

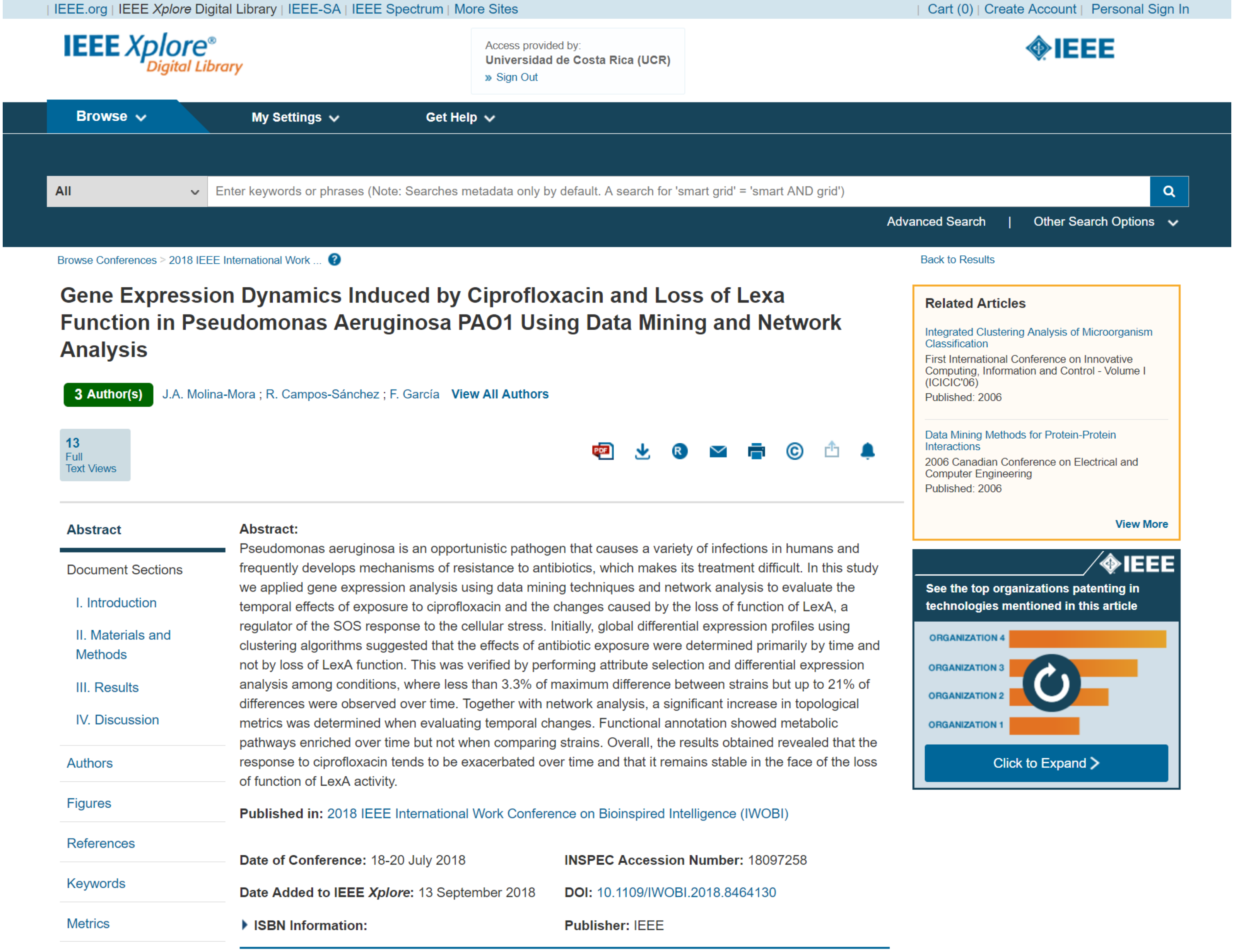
[4] T. S. Silva, N. Richard, J. P. Dias, and P. M. Rodrigues, "Data visualization and feature selection methods in gel-based proteomics.," *Curr. Protein Pept. Sci.*, vol. 15, no. 1, pp. 4–22, Feb. 2014.

[5] J. A. Molina-Mora, D. Chinchilla-Montero, C. Castro-Peña, and F. Garcia, "Two-dimensional gel electrophoresis (2D-GE) image analysis based on CellProfiler," *Medicine.*, vol. IN-PRESS, 2020.

[6] R. T. Cirz, B. M. O'Neill, J. A. Hammond, S. R. Head, and F. E. Romesberg, "Defining the Pseudomonas aeruginosa SOS response and its role in the global response to the antibiotic ciprofloxacin," *J. Bacteriol.*, vol. 188, no. 20, pp. 7101–7110, Oct. 2006.

[7] G. F. Ames, C. Prody, and S. Kustu, "Simple, rapid, and quantitative release of periplasmic proteins by chloroform.," *J. Bacteriol.*, vol. 160, no. 3, pp. 1181–3, Dec. 1984.

[8] I. Arganda-Carreras, C. O. S. Sorzano, J. Kybic, and C. Ortiz-de-solorzano, "bUnwarpJ : Consistent and Elastic Registration in ImageJ . Methods and Applications .," *Image (Rochester, N.Y.)*, 2006.



| IEEE.org | IEEE *Xplore* Digital Library | IEEE-SA | IEEE Spectrum | More Sites | Cart (0) | Create Account | Personal Sign In

IEEE *Xplore®* Digital Library

Access provided by:
Universidad de Costa Rica (UCR)
» Sign Out

IEEE

Browse My Settings Get Help

All | Enter keywords or phrases (Note: Searches metadata only by default. A search for 'smart grid' = 'smart AND grid')

Advanced Search | Other Search Options

Browse Conferences > 2018 IEEE International Work ... ?

Back to Results

Gene Expression Dynamics Induced by Ciprofloxacin and Loss of Lexa Function in Pseudomonas Aeruginosa PAO1 Using Data Mining and Network Analysis

3 Author(s) J.A. Molina-Mora ; R. Campos-Sánchez ; F. García View All Authors

13 Full Text Views

Abstract

Document Sections

I. Introduction

II. Materials and Methods

III. Results

IV. Discussion

Authors

Figures

References

Keywords

Metrics

Abstract:
Pseudomonas aeruginosa is an opportunistic pathogen that causes a variety of infections in humans and frequently develops mechanisms of resistance to antibiotics, which makes its treatment difficult. In this study we applied gene expression analysis using data mining techniques and network analysis to evaluate the temporal effects of exposure to ciprofloxacin and the changes caused by the loss of function of LexA, a regulator of the SOS response to the cellular stress. Initially, global differential expression profiles using clustering algorithms suggested that the effects of antibiotic exposure were determined primarily by time and not by loss of LexA function. This was verified by performing attribute selection and differential expression analysis among conditions, where less than 3.3% of maximum difference between strains but up to 21% of differences were observed over time. Together with network analysis, a significant increase in topological metrics was determined when evaluating temporal changes. Functional annotation showed metabolic pathways enriched over time but not when comparing strains. Overall, the results obtained revealed that the response to ciprofloxacin tends to be exacerbated over time and that it remains stable in the face of the loss of function of LexA activity.

Published in: 2018 IEEE International Work Conference on Bioinspired Intelligence (IWOBI)

Date of Conference: 18-20 July 2018

INSPEC Accession Number: 18097258

Date Added to IEEE *Xplore*: 13 September 2018

DOI: 10.1109/IWOBI.2018.8464130

ISBN Information:

Publisher: IEEE

Related Articles

Integrated Clustering Analysis of Microorganism Classification
First International Conference on Innovative Computing, Information and Control - Volume I (ICICIC'06)
Published: 2006

Data Mining Methods for Protein-Protein Interactions
2006 Canadian Conference on Electrical and Computer Engineering
Published: 2006

View More

Gene expression dynamics induced by ciprofloxacin and loss of LexA function in *Pseudomonas aeruginosa* PAO1 using data mining and network analysis

Molina-Mora J.A.
Research Center on Tropical Diseases (CIET)
Faculty of Microbiology, University of Costa Rica (UCR)
San José, Costa Rica
jose.molinamora@ucr.ac.cr

Campos-Sánchez R.
Research Center in Cellular and Molecular Biology (CIBCM)
Faculty of Microbiology, University of Costa Rica (UCR)
San José, Costa Rica

García F.
Research Center on Tropical Diseases (CIET)
Faculty of Microbiology, University of Costa Rica (UCR)
San José, Costa Rica

***Abstract*— *Pseudomonas aeruginosa* is an opportunistic pathogen that causes a variety of infections in humans and frequently develops mechanisms of resistance to antibiotics, which makes its treatment difficult. In this study we applied gene expression analysis using data mining techniques and network analysis to evaluate the temporal effects of exposure to ciprofloxacin and the changes caused by the loss of function of LexA, a regulator of the SOS response to the cellular stress. Initially, global differential expression profiles using clustering algorithms suggested that the effects of antibiotic exposure were determined primarily by time and not by loss of LexA function. This was verified by performing attribute selection and differential expression analysis among conditions, where less than 3.3% of maximum difference between strains but up to 21% of differences were observed over time. Together with network analysis, a significant increase in topological metrics was determined when evaluating temporal changes. Functional annotation showed metabolic pathways enriched over time but not when comparing strains. Overall, the results obtained revealed that the response to ciprofloxacin tends to be exacerbated over time and that it remains stable in the face of the loss of function of LexA activity.**

Keywords*— *P. aeruginosa; Data mining; Network analysis; Differential expression; Ciprofloxacin.

I. Introduction

Pseudomonas aeruginosa is a Gram-negative bacterium, metabolically and genomically versatile, found in natural environments, but it also causes infections in animals and plants [1]. In humans, it is an important opportunistic pathogen, being the third most common cause of nosocomial infections [2]. Many *P. aeruginosa* infections can be controlled with antibiotics, but are difficult to eradicate [3] due in part to the ability of this pathogen to carry out progressive modifications that facilitate infection and persistence, between those that emphasize their ability to adapt to environmental stress [4], [5].

As in other bacterial groups, cellular stress induces changes in DNA architecture, either by direct damage to DNA or indirectly in the replication process as a result of stress, which culminates with the exposure of single stranded DNA (ssDNA) and which represents the start signal of the SOS response [6]. In this process, protein RecA binds to ssDNA mediating recombinational repair but it also joins to LexA, a SOS repressor gene, and induces its autocleavage. The loss of repression by LexA results in the induction of proteins that mediate the SOS response for DNA repair and regulate damage tolerance mechanisms [7].

Ciprofloxacin, an antibiotic of the fluoroquinolone family and classically used for the treatment of *P. aeruginosa* infections, is an inducer of the SOS response in this bacterium. The antibiotic alters the activity of the bacterial enzymes DNA gyrase and topoisomerase IV, so it affects the correct replication of DNA, its recombination, repair and transcription [8], [9]. This condition causes activation of the SOS response; however it has been characterized that in *P. aeruginosa* the SOS response is mediated by 15 genes, which is much lower than that reported for other bacterial groups [5]. In addition to SOS response, *P. aeruginosa* generates a LexA-independent response after exposition to ciprofloxacin [2], [5].

The biological aim of the present study is to describe the dynamics of the global differential expression response to perturbation with ciprofloxacin and the effects of loss of LexA function in *P. aeruginosa*. For this, curated data of the bacterial strain *P. aeruginosa* PAO1, a reference strain, were used. In addition, data of a PAO1 mutant, produced by mutagenesis with loss of function of LexA, were available. Both strains were exposed to ciprofloxacin and data from the global differential expression profiles were obtained at 0, 30 and 120 minutes post exposure and using microarray technology.

Due to the type of high-throughput technology used, amount of data (5900 genes per replicate, for 12 samples) and the complexity and diversity of data available, an analysis mediated by data mining was required for classification, clustering and selection of genes. Additionally, an analysis for the creation, interpretation and evaluation of networks with a large-scale systems biology approach was implemented.

II. Materials and Methods

A. Data source

Data from 12 gene expression microarrays (GPL84 Affymetrix *P. aeruginosa* Array, with 5900 genes per sample) was available in NCBI database, accession number GSE5443 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE5443). The data included duplicates of each sample at 3 time points (0, 30 and 120 minutes) for two strains, PAO1 (wildtype, called WT strain) and an isogenic mutant strain with affectation of LexA activity (S125A in *lexA* gene, called mlexA strain), which were exposed to ciprofloxacin.

B. Normalization and evaluation of quality

The temporal data from microarray readouts were analyzed with two quality protocols. First, an RMA algorithm (Robust Multi-array Average, molmine.com/magma/loading/rma.htm) was applied to create an expression matrix from the row data. In short, the values of unprocessed intensity were normalized and corrected compared to the background noise, with a subsequent logarithmic transformation (log2) and quantification by quartiles. Next, a linear model was fitted to the normalized data to obtain an adjusted measure of expression for each set of probes.

Second, in order to carry out a global study of the differential expression profiles, a hierarchical clustering (HC) algorithm and a PCA (Principal Components Analysis) were implemented. In both cases, data were loaded into the program MATLAB and defined functions were directly called (*pca()* and *clustering()*). For visualization, PCA components were exported as a table and the three components with the greatest variation were plotted in 3 dimensions using the *Plot3D* function of the *Wolfram-Mathematica* software. In the case of HC, a 95% confidence and Euclidean metric were used.

C. Expression level comparisons with subsets of genes

In order to evaluate expression profile variations among strains and times, we proceeded to select defined sets of genes using three different methodologies.

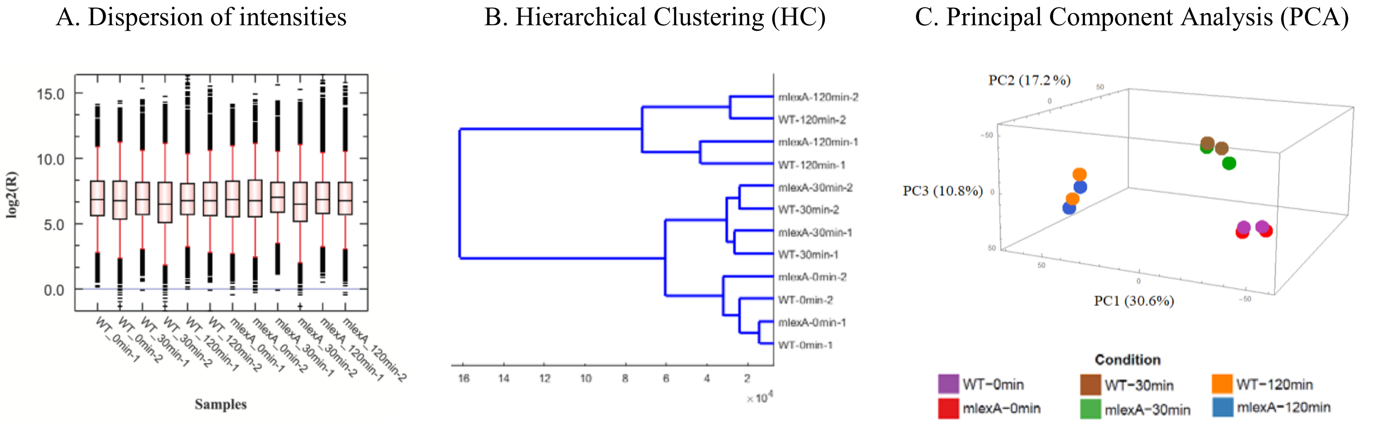
First, for a strict quantitative comparison based on housekeeping genes, as normally done in classic studies of gene expression, we proceeded to compare the relative values of expression of the housekeeping genes *proC* and *rpoD.* Both genes have been previously reported as the most stable for *P. aeruginosa* faced with disturbances [10]. This was done to verify that there was no significant difference between genes under the effects of ciprofloxacin, mutation and time. Differential expression protocol is detailed in the next section.

Second, an algorithm for ranking and selection of attributes was applied, using the classes (experimental conditions) to verify that it was possible to separate the conditions by selecting groups of genes. For this, we applied a classification by SVM Suppport Vector Machine algorithm [11], creating a ranked list of the total genes (tolerance=1.0e^{-10}, complexity=1). Using quartiles as the parameter for selecting attributes, we eliminated the last 25% of the ranked list of genes and evaluated with HC the top 75% remaining. The elimination was repeated leaving 50 and 25% of the top genes, and in each step repeating the clustering by HC until obtaining the separation of classes (by strains and times).

Third, in order to validate the loss of LexA function in the mutated strain, we proceeded to compare the expression levels of the 15 genes associated with the SOS response at all times and for both strains (with respect to the initial time for WT). In addition, for the time 120 minutes a representation of the connections was made with a non-directed graph and comparing the values of expression between strains. To do this, Cytoscape (http://www.cytoscape.org) was used to create the subnetwork of 15 genes, and coloring according to the level of expression.

D. Differential expression analysis

In order to identify differentially expressed genes among the evaluated conditions, the algorithm of Benjamini & Hochberg [12] was implemented, using as a criterion a two-fold change with respect to the control (all with values p>0.001). The relative comparison was made in two different ways. First, the gene expression profile of the mlexA mutant strain was compared with the corresponding WT strain at the same time, allowing to evaluate the differences between the strains at each time when exposed to ciprofloxacin.


A. Dispersion of intensities

B. Hierarchical Clustering (HC)

C. Principal Component Analysis (PCA)

Figure 1. Evaluation of quality, distribution and global differential expression profiles of samples. Quality was evaluated using dispersion of intensities for all samples using Robust Multi-array Average (RMA) algorithm, showing similar distribution (A). Global differential expression profiles were compared by two clustering techniques. HC shows separation of samples by time but not by strain (B). Similar results were found when a PCA was applied (C).

TABLE 1. Relative comparison of expression of housekeeping genes *proC* and *rpoD*

Condition	logFC (*p*-value)	
	proC	*rpoD*
mlexA_0min/WT_0min	0.2619 (0.4483)	0.1043 (0.7342)
mlexA_30min/WT_30min	0.1150 (0.8076)	0.0439 (0.9015)
mlexA_120min/WT_120min	-0.4476 (0.1859)	0.3062 (0.3237)
mlexA_30min/WT_0min	0.0337 (0.9373)	0.4769 (0.2138)
mlexA_120min/WT_0min	-0.3480 (0.2838)	0.4040 (0.2112)

In order to study the dynamics over time in the mlexA strain, a second analysis was performed comparing the differential expression of the mlexA mutant strain at different times with the WT strain (time 0 minutes).

E. *Annotation and construction of gene networks*

In order to annotate and characterize the differentially expressed genes, an ontology analysis was carried out by biological process and metabolic pathways using the PANTHER database (http://pantherdb.org/). The genes, both up and down regulated, were directly incorporated into the functional modules of the resource and with the specifications for *P. aeruginosa*. In addition, using the PseudomonasNET database (www.inetbio.org/pseudomonasnet/), we performed a screening of biological functions and their relationships at the network level. To do this, the genes were incorporated and prioritized by candidate functions for *P. aeruginosa* using the Gene-centric search module. The resultant network was exported in graph format and was incorporated into the Cytoscape program. The network analysis was established to visualize expression levels. The expression data matrix was adjusted to the PseudomonasNET identifiers, and specifications were selected to differentiate down and up regulated genes by color.

III. Results

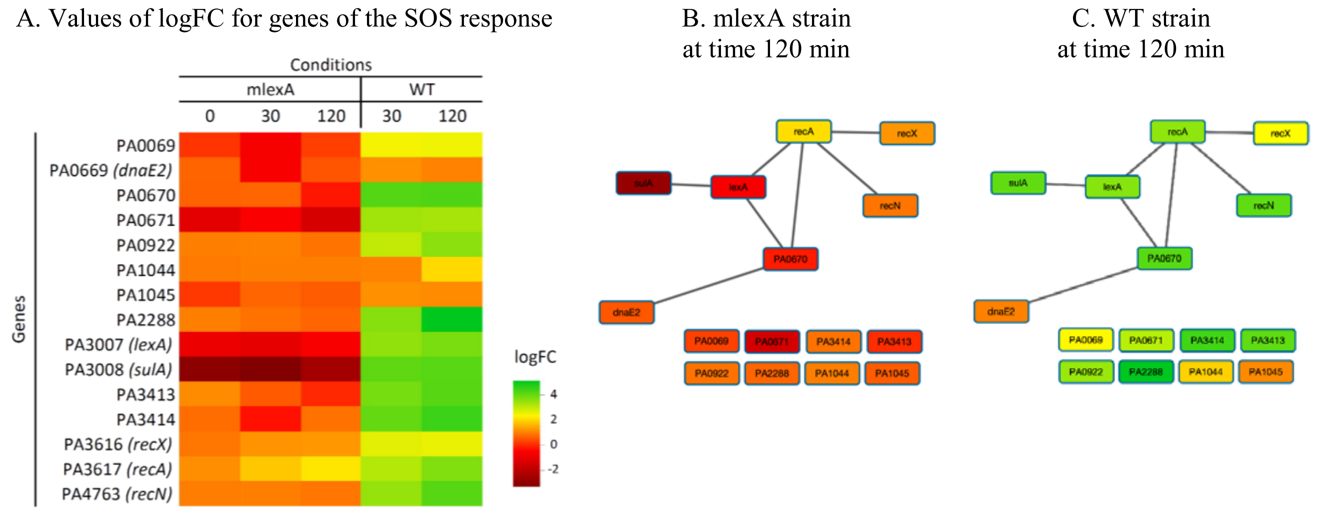
A. *Clustering algorithms suggest that global differential expression due to exposure to ciprofloxacin in* P. aeruginosa *tend to differ by time than by loss of LexA function*

With the aim to study the time effect due to exposure to ciprofloxacin and the loss of LexA function in *P. aeruginosa*, we analyzed microarray data. For this, we implemented different analyses to evaluate dispersion, quality and comparability of total data. First, an RMA algorithm was applied to evaluate dispersion of total data to obtain normalized, logarithmically transformed and adjusted values (Figure 1-A). The general dispersion of all the tests shows equivalence in the intensity signals for each of the samples and their replicates, which is consistent with the criteria proposed by Bolstad et al. regarding the correction of variation [13].

Second, a hierarchical clustering (HC) algorithm based on Euclidean distance was applied to the data set with the aim of conducting a study of the global differential expression profiles and the relationship between conditions and replicates. According to the result of the HC, it is observed that the separation between the experimental conditions is achieved by time but not clearly by strain (Figure 1-B). A second evaluation criterion with PCA algorithm provided congruent results with HC (Figure 1-C). Moreover, differences obtained by both algorithms showed a cluster between samples at time 0 and 30 minutes.

Third, in order to evaluate the variation of differential expression profiles from the perspective of housekeeping genes, we compared gene expression from *proC* and *rpo*. We proved that there are no statistical significant differences among expression values (presented as a ratio between conditions) for both genes (p>0.001) under the perturbations by ciprofloxacin, the mutation and the different times (Table 1).

Finally, given the biological importance of LexA in the SOS response, we verified its inactivity in mlexA strain by comparing the expression level of the 15 characterized genes in this pathway (relative to the initial profile of WT). As shown in


A. Values of logFC for genes of the SOS response

B. mlexA strain at time 120 min

C. WT strain at time 120 min

Figure 2. Comparison of logFC of strains mlexA and WT for genes of the SOS response. All logFC values were obtained by relative comparison with WT-0min. In strain mlexA, which has no LexA function, SOS response is completely inactive as expected, in contrast to WT which has high levels of expression (A). Representation of values using relations between genes was done by building a subnetwork of the SOS response (B-C).

Figure 2-A, the loss of LexA function affects gene expression of SOS genes, however all them are active for the WT at 30 and 120 minutes. The representation as a non-directed graph alternatively shows the same observations for 120 minutes, although some connections are unknown (Figure 2-B-C).

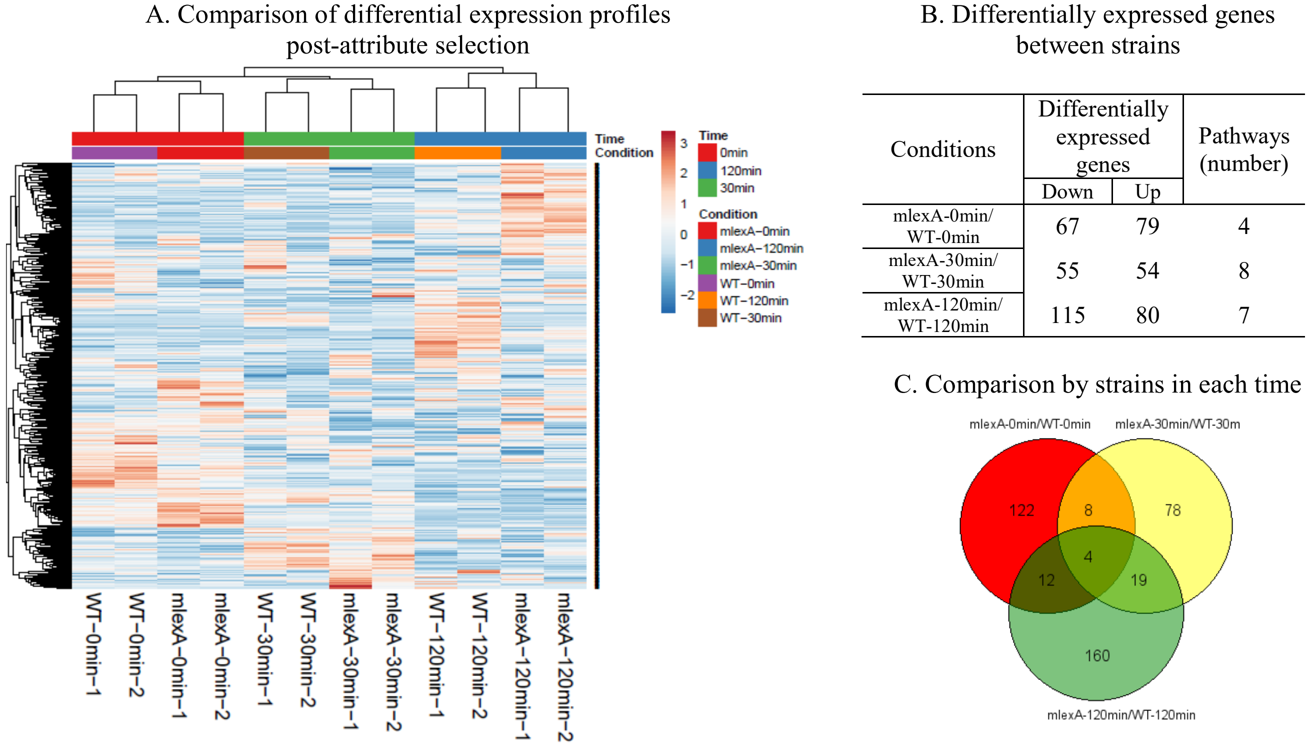
Altogether, these results suggest that differential expression profiles of the samples exposed to ciprofloxacin differentiate better in time than among strains. To study genes differentially expressed post exposition to ciprofloxacin, we conducted two comparisons: (i) strain-to-strain in the same time, and (ii) dynamical (time course) analysis of mlexA strain relative to the initial profile of WT, as detailed in the next sections.

B. SVM and differential expression analysis show that exposure to ciprofloxacin has a similar effect independently of LexA activity in P. aeruginosa

In order to contrast the differential expression profiles of the mlexA and WT strains at each specific time and to verify that there are few differences among them, we performed a screening analysis based on data mining to select attributes. For this, variation of differential expression profiles was evaluated with defined sets of genes by first ranking all genes using SVM algorithm, then last genes were eliminated from the rank, and finally we evaluated the clustering with HC algorithm. Because successive elimination of genes was done by quartiles, it was necessary to remove the last three quartiles of ranked genes (leaving the top 25% genes or the top quartile) to generate a separation of classes. This means that differences in expression level between strains are defined by less than 25% of genes. As shown in Figure 3-A, separation of the experimental classes and the generation of groups was achieved first by time and then by strain. This is consistent with previous results of global profiles, however now we clearly observed the separation of strains.

In addition, an analysis of differential expression between the mlexA strain and WT was carried out at the same time in order to estimate accurately differences and variation with statistical meaning at single gene resolution and not by global profiles (Figure 3-B-C). The statistically significant differences (at least 2-fold change with p>0.001) provided evidence that a discrete number of genes would be affected by the loss of function of LexA. This in turn corresponded to a minimum number of pathways affected. For example, at 30 minutes, a total of 109 genes were differentially expressed beloging to 8 metabolic pathways; while at 120 minutes 195 genes were identified corresponding to only 7 routes. Moreover, only 4 genes were identified for the 3 times and no metabolic pathway was common for all 3 times. With these results, we concluded that a low number of transcripts, at most 195 genes (about 3.3%), differentiate the strains when exposed to ciprofloxacin.

Notably, when screening was done for expression networks using the PseudomonasNET database (networks not shown), we observed that differentially expressed genes were not significantly associated with any particular metabolic pathway at any time (Table 2, last row). This is consistent with previous results and suggests that expression differences between the two strains (given by LexA activity) at any time have no greater biological effects (they have similar response). Topological metrics of the networks are included and compared in Table 2 to contrast with networks obtained by time (next section).



A. Comparison of differential expression profiles post-attribute selection

B. Differentially expressed genes between strains

Conditions	Differentially expressed genes		Pathways (number)
	Down	Up	
mlexA-0min/ WT-0min	67	79	4
mlexA-30min/ WT-30min	55	54	8
mlexA-120min/ WT-120min	115	80	7

C. Comparison by strains in each time

Figure 3. Differentially expressed genes by strain, using WT strain at same time for comparisons. Because initial analysis showed less differences betweens strain than by time, a features selection was done by SVM algorithm. An analysis by quartiles show that no more than 25% of top ranked genes can separate conditions, however, in order to incorporate variation at single gene level, an differential expression analysis was done, showing relatively few changes (differences were no higher than 3.3% when comparing mlexA and WT) (B-C).

A. Differentially expressed genes by time

Conditions	Differentially expressed genes	
	Down	Up
mlexA-0min/ WT-0min	67	79
mlexA-30min/ WT-0min	288	161
mlexA-120min/ WT-0min	614	609

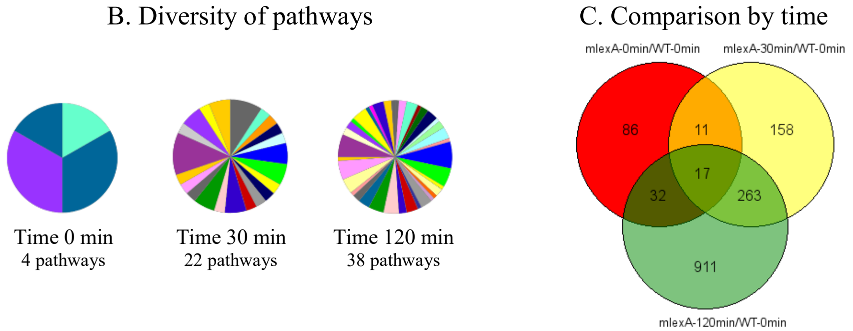


Figure 4. Differentially expressed genes of strain mlexA by time, using WT-0min for comparison. Differences shows an increment by time (A) with enrichment of pathways (B). 280 differentially expressed genes were found to be shared between 30 and 120 minutes (C).

C. Large-scale network approach and differential expression analysis reveals time-intensified effects in P. aeruginosa *mlexA after exposure to ciprofloxacin*

In order to analyze the temporal dynamics of differential gene expression in *P. aeruginosa* mlexA strain after exposure to ciprofloxacin, we compared the expression of this strain at all times with respect to the WT strain at 0 minutes using the same criteria applied to previous analysis of gene expression. As shown in the Figure 4-A, the number of differentiated genes triples from time 0 to 30 minutes, and increases 10 times between time 0 and 120 minutes. Between time 30 and 120 minutes, 280 differentially expressed genes were shared, representing more than 62% of the genes at 30 minutes (Figure 4-C). At 120 minutes, 1223 genes were differentially expressed, representing 20.7%, which contrasts with the analyses among strains, where the differences did not exceed 3.3%.

In addition, ontologies with characterized genes showed a significant increase in the metabolic pathways involved, with 22 pathways at 30 minutes and 38 pathways at 120 minutes, as detailed in Figure 4-B.

When annotation and screening of expression networks was done with PANTHER and PseudomonasNET databases, at 0 minutes the differentially expressed genes are not significantly associated with any metabolic pathway, however, this changes at 30 and 120 minutes (Figure 5 and Table 2, last row). When performing a general analysis of the networks obtained for each of the comparisons, as shown in Table 2, the topological metrics revealed relevant changes by time of exposure to the antibiotic in mlexA strain, but not significant when considering the difference between strains at the same time. For example, as shown in Table 2, when comparing the networks obtained between strains in each time (networks not shown), the number of nodes and edges was oscillating but with relatively stable variations compared to the other cases, the same for the degree of the nodes (1.92 at the beginning and then it passes to 1.49 at 30 minutes and 2.41 at 120 minutes). Despite this observation, all networks generated presented significant relationships (based on the p-value of the PPI) but none were significantly associated with any metabolic pathway (based on functional enrichment).

When comparing the time for mlexA strain, the changes were significant with a drastic increase in various metrics. For example, between 0 and 120 minutes, the number of nodes changed from 146 to 1223 and the average number of connections per node increased from 1.49 to 16.1.

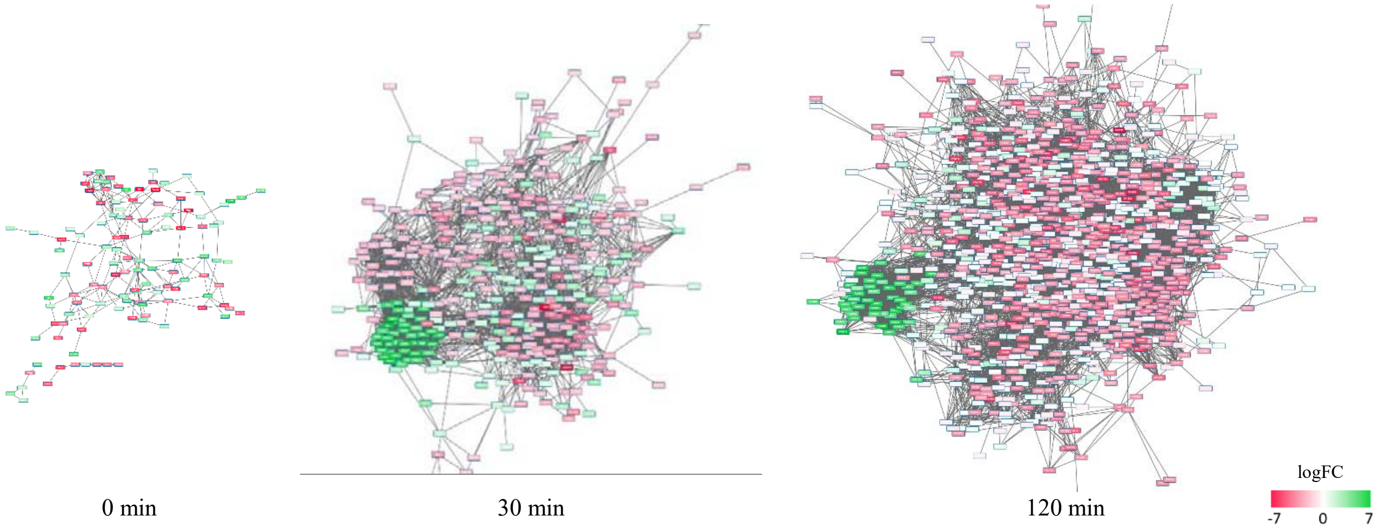


Figure 5. Transcriptional networks of differentially expressed genes of strain mlexA by time, using WT-0min for comparison. Identification of one cluster was possible at times 30 and 120 minutes (logFC of up-regulated genes are shown in green and down-regulated genes in red).

TABLE 2. Comparison of topological metrics of networks created using the differentially expressed genes

Topological metrics	Comparison of conditions				
	mlexA-0min/ WT-0min	mlexA-30min/ WT-30min	mlexA-120min/ WT-120min	mlexA-30min/ WT-0min	mlexA-120min/ WT-0min
Number of nodes	146	109	193	448	1223
Number of edges	140	81	233	2080	9841
Average node degree	1.92	1.49	2.41	9.29	16.1
Average local clustering coefficient	0.506	0.39	0.431	0.476	0.384
PPI enrichment p-value	$5.99e^{-08}$	$1.42e^{-08}$	$2.58e^{-12}$	$< 1.0e^{-16}$	$< 1.0e^{-16}$
Functional enrichments detected	No	No	No	Yes	Yes

At 30 and 120 minutes, using the annotation with PANTHER (Figure 4-B), the diversity of metabolic pathways was significantly linked to the differential expression profile by functional enrichment.

On the other hand, in order to compare the values of expression of the down and up regulated genes in the network, the preliminary graph was imported into Cytoscape and was edited to incorporate expression data. When performing the representation of the relative expression values (Figure 5, up-regulated genes are shown in green and down-regulated genes in red), we observed a random distribution of genes, except a group of up-regulated genes that formed a cluster (at 30 and 120 minutes). These clusters are also formed when networks for WT are created at 30 and 120 minutes (with differential expression relative to WT-0min), so they are independent of LexA activity (networks not shown). When carrying out the characterization of the genes that conformed these clusters, the functional annotation revealed that the majority corresponded to: hypothetical proteins of *P.aeruginosa* (not characterized), phage-associated proteins (mostly hypothetical), pyocin metabolism, transcription, SOS response and other metabolic processes (Table 3).

IV. Discussion

The complexity of biological systems and the amount of data obtained with high-performance technologies continue to represent a limitation when extracting relevant information [14]. This is also true for prokaryotic biological systems like *P. aeruginosa,* whose physiology and regulatory mechanisms at the global level are barely understood. At the transcriptomic level, the need to associate RNA molecules to decipher their complex interactions can be solved with pattern recognition within the data sets. This can be complemented with the knowledge stored in databases to characterize interactions and analyze them as a complex network.

P. aeruginosa is a bacterium of high relevance for human health due to the infections it causes and the common loss of susceptibility to antibiotics leading to multiple drug resistance [2], [15], [16]. This is worsened by the absence of new antimicrobials and the inability to control the development of antibiotic resistance. Moreover, lack of knowledge of the antibiotic-pathogen interactions and the mechanisms of action of antimicrobial agents at the complete system level delay the formulation of new strategies to control infections [15].

The model for studies, the strain PAO1 with loss of LexA function, it is of particular relevance because LexA is a regulator of the SOS response, which constitutes a mechanism of tolerance to DNA damage. For this strain, the SOS response is induced with the exposure to ciprofloxacin.

To perform an initial evaluation, the data was normalized with RMA, an algorithm regularly used to evaluate the quality and comparability of the data [13]. This process guarantees that the differences in expression are biologically significant, referred to as *interesting variation*. In contrast with variation introduced during sample preparation, array manufacture and array processing (labeling, hybridization, and scanning), referred to as *obscuring variation* [17].

Notably, clustering algorithms such as PCA and HC (Figure 1) showed that the effect of the mutation is reduced compared to the effect of time when the bacteria is exposed to ciprofloxacin. In the PCA, the trajectory of global profiles is the same for the two strains, to the point that it is not possible to clearly differentiate the strains. Additionally, the global profiles were validated with housekeeping genes (Table 1) showing no significant differences between conditions as expected.

When comparing the two strains at each time, with or without LexA function, transcriptomic responses showed less than 3.3% differences between strains when a single gene analysis was performed. The initial screening was done by gene ranking and elimination of quartiles; indicating that no more than 25% of the genes could differentiate the classes using the SVM algorithm.

TABLE 3. Cluster genes annotation of mlexA strain (30min and 120min)

Functional Annotation	Number of genes	
	30 min	120 min
Bacteriophage protein	10	10
Hypothetical protein	39	40
Pyocin metabolism	5	5
SOS response regulator	4	-
Transcriptional regulator	3*	3*
Others (with only 1 gene)	7	2
Total	66	58

* Two genes were also counted in pyocin metabolism

This combination of algorithms and analyses has not been previously reported for expression analyses of *P. aeruginosa*, since they are regularly performed separately. The aim was to verify with two different techniques the variations among conditions, where global profiles are used (using SVM by quartiles) and then individual gene level is applied for separation of classes (with analysis of differential expression).

These results are consistent with what was expected for *P. aeruginosa*, because LexA seems to regulate a discrete number of genes, including 15 of the SOS response, as well as others in various metabolic pathways [5]. Therefore, global profiles in response to ciprofloxacin do not allow a clear separation of the strains conditions, i.e. both are affected in a similar way. The functional effects of LexA loss were verified by comparing the strains, as shown in Figure 2. In other organisms such as *Escherichia coli* and *Bacillus subtilis* [5], the SOS response also involves a relatively small number of genes, although higher than for *P. aeruginosa*, so the loss of LexA function could also have discrete effects on the global differential expression profiles. Additionally, no significantly enriched pathways for diferentially expressed genes were found and the topological metrics of correlation networks remained oscillating in relatively narrow ranges (Table 2).

However, when making the temporal comparison of the mlexA strain with the initial profile of WT, a significant increase of differentially expressed genes is evidenced, both for 30 minutes and much larger for 120 minutes. For the latter case reaching almost 21% of differences compared to the initial control. For instance, the gene networks showed increasing exponential changes in the number of genes and interactions among them (according to the topological metrics), and that corresponded to an increase in metabolically enriched routes. Inclusively, a cluster was identified at 30 and 120 minutes, whose characterization revealed many hypothetical proteins involved. This gap in knowledge regarding the function and importance of these proteins is a limitation of the current state of knowledge in *P. aeruginosa*, because although there is experimental evidence that they are related, it is not possible to characterize them completely from the curated databases.

As previously reported, the knowledge of the transcriptional dynamics during exposure to ciprofloxacin contributes to the understanding of its pathogenicity by the biological processes it regulates (phages, mobility, toxin production, others) and can potentially offer alternatives to modulate the stress response [5], particularly the SOS response. This has an impact not only in the susceptibility to antibiotics, but could also regulate other biological processes such as the induction of errors in the DNA polymerases and emergence of spontaneous mutations [18]. Such mechanisms are of particular importance in *P. aeruginosa* since resistance to multiple drugs originate mainly from point mutations [19]. Moreover, SOS response affects other determinants of production of pyocins and transference and expression of exogenous genes of resistance [18]. Potentially, all these biological processes could be targets for modulation of the SOS response based on the knowledge obtained from this and other studies.

References

[1] M. V. Olson *et al.*, “Complete genome sequence of Pseudomonas aeruginosa PAO1, an opportunisticpathogen,” *Nature*, vol. 406, no. 6799, pp. 959–964, Aug. 2000.

[2] E. B. M. Breidenstein, M. Bains, and R. E. W. Hancock, “Involvement of the lon protease in the SOS response triggered by ciprofloxacin in Peudomonas aeruginosa PAO1,” *Antimicrob. Agents Chemother.*, vol. 56, no. 6, pp. 2879–2887, 2012.

[3] E. Drenkard and F. M. Ausubel, “Pseudomonas biofilm formation and antibiotic resistance are linked to phenotypic variation,” *Nature*, vol. 416, no. 6882, pp. 740–743, Apr. 2002.

[4] J. R. Govan and V. Deretic, “Microbial pathogenesis in cystic fibrosis: mucoid Pseudomonas aeruginosa and Burkholderia cepacia.,” *Microbiol. Rev.*, vol. 60, no. 3, pp. 539–74, Sep. 1996.

[5] R. T. Cirz, B. M. O’Neill, J. A. Hammond, S. R. Head, and F. E. Romesberg, “Defining the Pseudomonas aeruginosa SOS response and its role in the global response to the antibiotic ciprofloxacin,” *J. Bacteriol.*, vol. 188, no. 20, pp. 7101–7110, 2006.

[6] E. Recacha *et al.*, “Quinolone Resistance Reversion by Targeting the SOS Response.,” *MBio*, vol. 8, no. 5, 2017.

[7] C. Y. Mo, L. D. Birdwell, and R. M. Kohli, “Specificity Determinants for Autoproteolysis of LexA, a Key Regulator of Bacterial SOS Mutagenesis,” *Biochemistry*, vol. 53, no. 19, pp. 3158–3168, May 2014.

[8] R. L. Gibson, J. L. Burns, and B. W. Ramsey, “Pathophysiology and Management of Pulmonary Infections in Cystic Fibrosis,” *Am. J. Respir. Crit. Care Med.*, vol. 168, no. 8, pp. 918–951, Oct. 2003.

[9] K. Drlica and X. Zhao, “DNA gyrase, topoisomerase IV, and the 4-quinolones.,” *Microbiol. Mol. Biol. Rev.*, vol. 61, no. 3, pp. 377–92, Sep. 1997.

[10] H. Savli, A. Karadenizli, F. Kolayli, S. Gundes, U. Ozbek, and H. Vahaboglu, “Expression stability of six housekeeping genes: a proposal for resistance gene quantification studies of Pseudomonas aeruginosa by real-time quantitative RT-PCR,” *J. Med. Microbiol.*, vol. 52, no. 5, pp. 403–408, May 2003.

[11] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, “Gene Selection for Cancer Classification using Support Vector Machines,” *Mach. Learn.*, vol. 46, no. 1/3, pp. 389–422, 2002.

[12] Y. Benjamini and Y. Hochberg, “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing,” *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 57. WileyRoyal Statistical Society, pp. 289–300, 1995.

[13] B. M. Bolstad, T. P. Speed, R. A. Irizarry, and M. Astrand, “A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.,” *Bioinformatics*, vol. 19, no. 2, pp. 185–193, 2003.

[14] J. A. Molina-Mora, M. Kop-Montero, I. Quirós-Fernández, S. Quiros, J. L. Crespo-Mariño, and R. A. Mora-Rodríguez, “A hybrid mathematical modeling approach of the metabolic fate of a fluorescent sphingolipid analogue to predict cancer chemosensitivity,” *Comput. Biol. Med.*, vol. 97, pp. 8–20, Jun. 2018.

[15] M. D. Brazas, M. D. Brazas, R. E. W. Hancock, and R. E. W. Hancock, “Ciprofloxacin Induction of a Susceptibility Determinant in Pseudomonas aeruginosa,” *Antimicrob. Agents Chemother.*, vol. 49, no. 8, pp. 3222–3227, 2005.

[16] F. Toval, A. Guzman-Marte, V. Madriz, T. Somogyi, C. Rodriguez, and F. Garcia, “Predominance of carbapenem-resistant Pseudomonas aeruginosa isolates carrying blaIMP and blaVIM metallo- -lactamases in a major hospital in Costa Rica,” *J. Med. Microbiol.*, vol. 64, no. Pt_1, pp. 37–43, Jan. 2015.

[17] A. J. Hartemink, D. K. Gifford, T. S. Jaakkola, and R. A. Young, “Combining location and expression data for principled discovery of genetic regulatory network models.,” *Pac. Symp. Biocomput.*, pp. 437–49, 2002.

[18] E. Y. Valencia, F. Esposito, B. Spira, J. Blázquez, and R. S. Galhardo, “Ciprofloxacin-mediated mutagenesis is suppressed by subinhibitory concentrations of amikacin in *Pseudomonas aeruginosa*,” *Antimicrob. Agents Chemother.*, no. December, p. AAC.02107-16, 2016.

[19] E. B. M. Breidenstein, C. de la Fuente-Núñez, and R. E. W. Hancock, “Pseudomonas aeruginosa: all roads lead to resistance,” *Trends Microbiol.*, vol. 19, no. 8, pp. 419–426, Aug. 2011.

REFERENCES

Andersson, D. I., & Hughes, D. (2014). Microbiological effects of sublethal levels of antibiotics. *Nature Reviews Microbiology*, *12*(7), 465–478. https://doi.org/10.1038/nrmicro3270

Bermingham, M. L., Pong-Wong, R., Spiliopoulou, A., Hayward, C., Rudan, I., Campbell, H., … Haley, C. S. (2015). Application of high-dimensional feature selection: Evaluation for genomic prediction in man. *Scientific Reports*, *5*, 1–12. https://doi.org/10.1038/srep10312

Berti, A. D., & Hirsch, E. B. (2020, January 10). Tolerance to antibiotics affects response. *Science*. American Association for the Advancement of Science. https://doi.org/10.1126/science.aba0150

Brauner, A., Fridman, O., Gefen, O., & Balaban, N. Q. (2016, May 1). Distinguishing between resistance, tolerance and persistence to antibiotic treatment. *Nature Reviews Microbiology*. Nature Publishing Group. https://doi.org/10.1038/nrmicro.2016.34

Brazas, M. D., Brazas, M. D., Hancock, R. E. W., & Hancock, R. E. W. (2005). Ciprofloxacin Induction of a Susceptibility Determinant in Pseudomonas aeruginosa. *Antimicrobial Agents and Chemotherapy*, *49*(8), 3222–3227. https://doi.org/10.1128/AAC.49.8.3222

Cabot, G., Zamorano, L., Moyà, B., Juan, C., Navas, A., Blázquez, J., & Oliver, A. (2016). Evolution of Pseudomonas aeruginosa antimicrobial resistance and fitness under low and high mutation rates. *Antimicrobial Agents and Chemotherapy*, *60*(3), 1767–1778. https://doi.org/10.1128/AAC.02676-15.Address

Caldera, M., Müller, F., Kaltenbrunner, I., Licciardello, M. P., Lardeau, C. H., Kubicek, S., & Menche, J. (2019). Mapping the perturbome network of cellular perturbations. *Nature Communications*, *10*(1). https://doi.org/10.1038/s41467-019-13058-9

Chinchilla, D. (2018). *Patrones de expresión de los genes de las metalo-b-lactamasas blaIMP-18 y* *blaVIM-2 e IMP-18 en la cepa Pseudomonas aeruginosa AG1 resistente a carbapenems. Tesis del Posgrado en Microbiología con énfasis en Bacteriología.* Universidad de Costa Rica, San José, Costa Rica.

Ciofu, O., & Tolker-Nielsen, T. (2019, May 3). Tolerance and resistance of pseudomonas aeruginosabiofilms to antimicrobial agents-how P. aeruginosaCan escape antibiotics. *Frontiers in Microbiology*. Frontiers Media S.A. https://doi.org/10.3389/fmicb.2019.00913

Civelek, M., & Lusis, A. J. (2014). Systems genetics approaches to understand complex traits. *Nature Reviews. Genetics*, *15*(1), 34–48. https://doi.org/10.1038/nrg3575

Cornforth, D. M., Dees, J. L., Ibberson, C. B., Huse, H. K., Mathiesen, I. H., Kirketerp-Møller, K., … Whiteley, M. (2018). Pseudomonas aeruginosa transcriptome during human infection. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(22). https://doi.org/10.1073/pnas.1717525115

DeLong, E. F. (2012). *Prokaryotes : prokaryotic physiology and biochemistry*. Springer.

Dößelmann, B., Willmann, M., Steglich, M., Bunk, B., Nübel, U., Peter, S., & Neher, R. A. (2017). Rapid and Consistent Evolution of Colistin Resistance in Extensively Drug-Resistant Pseudomonas aeruginosa during Morbidostat Culture. *Antimicrobial Agents and Chemotherapy*, *61*(9), e00043-17. https://doi.org/10.1128/AAC.00043-17

Farajzadeh Sheikh, A., Shahin, M., Shokoohizadeh, L., Halaji, M., Shahcheraghi, F., & Ghanbari, F. (2019). Molecular epidemiology of colistin-resistant Pseudomonas aeruginosa producing NDM-1 from hospitalized patients in Iran. *Iranian Journal of Basic Medical Sciences*, *22*(1), 38–42. https://doi.org/10.22038/ijbms.2018.29264.7096

Fernández, M., Corral-Lugo, A., & Krell, T. (2018). The plant compound rosmarinic acid induces a broad quorum sensing response in Pseudomonas aeruginosa PAO1. *Environmental Microbiology*, *20*(12), 4230–4244. https://doi.org/10.1111/1462-2920.14301

Firme, M., Kular, H., Lee, C., & Song, D. (2010). RpoS Contributes to Variations in the Survival Pattern of Pseudomonas aeruginosa in Response to Ciprofloxacin. *Journal of Experimentall Microbiology and Immunology (JEMI)*, *14*(April), 21–27. Retrieved from https://microbiology.ubc.ca/sites/default/files/roles/drupal_ungrad/JEMI/14/JEMI14_21-27.pdf

Fothergill, J. L., Mowat, E., Walshaw, M. J., Ledson, M. J., James, C. E., & Winstanley, C. (2011). Effect of antibiotic treatment on bacteriophage production by a cystic fibrosis epidemic strain of Pseudomonas aeruginosa. *Antimicrobial Agents and Chemotherapy*, *55*(1), 426–428. https://doi.org/10.1128/AAC.01257-10

Glaab, E., Bacardit, J., Garibaldi, J. M., & Krasnogor, N. (2012). Using rule-based machine learning for candidate disease gene prioritization and sample classification of cancer gene expression data. *PLoS ONE*, *7*(7). https://doi.org/10.1371/journal.pone.0039932

Hasin, Y., Seldin, M., & Lusis, A. (2017). Multi-omics approaches to disease. *Genome Biology*, *18*(1), 83. https://doi.org/10.1186/s13059-017-1215-1

Hong, D. J., Bae, I. K., Jang, I. H., Jeong, S. H., Kang, H. K., & Lee, K. (2015). Epidemiology and characteristics of metallo-ß-lactamase-producing Pseudomonas aeruginosa. *Infection and Chemotherapy*, *47*(2), 81–97. https://doi.org/10.3947/ic.2015.47.2.81

Kamal, F., & Dennis, J. J. (2015). Burkholderia cepacia complex phage-antibiotic synergy (PAS): Antibiotics stimulate lytic phage activity. *Applied and Environmental Microbiology*, *81*(3), 1132–1138. https://doi.org/10.1128/AEM.02850-14

Klockgether, J., Munder, A., Neugebauer, J., Davenport, C. F., Stanke, F., Larbig, K. D., … Tümmler, B. (2010). Genome diversity of Pseudomonas aeruginosa PAO1 laboratory strains. *Journal of Bacteriology*, *192*(4), 1113–1121. https://doi.org/10.1128/JB.01515-09

Lu, P., Wang, Y., Zhang, Y., Hu, Y., Thompson, K. M., & Chen, S. (2016). RpoS-dependent sRNA RgsA regulates Fis and AcpP in Pseudomonas aeruginosa. *Molecular Microbiology*, *102*(2), 244–259. https://doi.org/10.1111/mmi.13458

Ma, C., Xin, M., Feldmann, K. A., & Wang, X. (2014). Machine Learning-Based Differential Network Analysis: A Study of Stress-Responsive Transcriptomes in Arabidopsis. *The Plant Cell*, *26*(2), 520–537. https://doi.org/10.1105/tpc.113.121913

Mathee, K., Narasimhan, G., Valdes, C., Qiu, X., Matewish, J. M., Koehrsen, M., … Lory, S. (2008). Dynamics of Pseudomonas aeruginosa genome evolution. *Proceedings of the National Academy of Sciences*, *105*(8), 3100–3105. https://doi.org/10.1073/PNAS.0711982105

McVicker, G., Prajsnar, T. K., Williams, A., Wagner, N. L., Boots, M., Renshaw, S. A., & Foster, S. J. (2014). Clonal Expansion during Staphylococcus aureus Infection Dynamics Reveals the Effect of Antibiotic Intervention. *PLoS Pathogens*, *10*(2). https://doi.org/10.1371/journal.ppat.1003959

Molina-Mora, J.-A., Campos-Sánchez, R., Rodríguez, C., Shi, L., & García, F. (2020). High quality 3C de novo assembly and annotation of a multidrug resistant ST-111 Pseudomonas aeruginosa genome: Benchmark of hybrid and non-hybrid assemblers. *Scientific Reports*, *10*(1), 1392. https://doi.org/10.1038/s41598-020-58319-6

Molina-Mora, J.-A., Garcia-Batan, R., & Garcia, F. (2020). From pan-genome to the genomic context of the two integrons of ST-111 Pseudomonas aeruginosa AG1: A VIM-2-carrying old-acquaintance and a novel IMP-18-carrying integron. *Research Square (Pre-Print)*. https://doi.org/10.21203/RS.3.RS-41474/V1

Molina-Mora, J., Montero-Manso, P., Batán, R. G., Sánchez, R. C., Fernández, J. V., & García, F. (2020). A first Pseudomonas aeruginosa perturbome: Identification of core genes related to multiple perturbations by a machine learning approach. *BioRxiv*, 2020.05.05.078477. https://doi.org/10.1101/2020.05.05.078477

Molina-Mora, J.A., Campos-Sanchez, R., & Garcia, F. (2018). Gene Expression Dynamics Induced by Ciprofloxacin and Loss of Lexa Function in Pseudomonas aeruginosa PAO1 Using Data Mining and Network Analysis. In *2018 IEEE International Work Conference on Bioinspired Intelligence (IWOBI)* (pp. 1–7). IEEE. https://doi.org/10.1109/IWOBI.2018.8464130

Molina-Mora, Jose Arturo, Chinchilla-Montero, D., Castro-Peña, C., & Garcia, F. (2020). Two-dimensional gel electrophoresis (2D-GE) image analysis based on CellProfiler: *Pseudomonas aeruginosa* AG1 as model. *Medicine*, *IN-PRESS*.

Molina-Mora, Jose Arturo, Chinchilla-Montero, D., Castro-Peña, C., & García, F. (2020). Two-dimensional gel electrophoresis image analysis of two Pseudomonas aeruginosa clones. *2020 IEEE International Work Conference on Bioinspired Intelligence (IWOBI)*, 1–6.

Molina-Mora, Jose Arturo, Chinchilla, D., Chavarría, M., Ulloa, A., Campos-Sanchez, R., Mora-Rodríguez, R. A., … García, F. (2020). Transcriptomic determinants of the response of ST-111 Pseudomonas aeruginosa AG1 to ciprofloxacin identified by a top-down systems biology approach. *Scientific Reports*, *10*, 1–23. https://doi.org/10.1038/s41598-020-70581-2

Morales-Berrocal, M. (2016). *Descripción de un modelo infeccioso murino de Pseudomonas aeruginosa AG1. Tesis para optar por el grado de Licenciatura en Microbiología y Química Clínica.* Universidad de Costa Rica, San José, Costa Rica.

Mulet, X., Cabot, G., Ocampo-Sosa, A. A., Dominguez, M. A., Zamorano, L., Juan, C., … Spanish Network for Research in Infectious Diseases (REIPI). (2013). Biological Markers of Pseudomonas aeruginosa Epidemic High-Risk Clones. *Antimicrobial Agents and Chemotherapy*, *57*(11), 5527–5535. https://doi.org/10.1128/AAC.01481-13

O'Donnell, S. T., Ross, R. P., & Stanton, C. (2020). The Progress of Multi-Omics Technologies: Determining Function in Lactic Acid Bacteria Using a Systems Level Approach. *Frontiers in Microbiology*, *10*, 3084. https://doi.org/10.3389/fmicb.2019.03084

Oliver, A., Mulet, X., López-Causapé, C., & Juan, C. (2015). The increasing threat of Pseudomonas aeruginosa high-risk clones. *Drug Resistance Updates*, *21–22*, 41–59. https://doi.org/10.1016/j.drup.2015.08.002

Petitjean, M., Martak, D., Silvant, A., Bertrand, X., Valot, B., & Hocquet, D. (2017). Genomic characterization of a local epidemic Pseudomonas aeruginosa reveals specific features of the widespread clone ST395. *Microbial Genomics*, *3*(10), e000129. https://doi.org/10.1099/mgen.0.000129

Stewart, P. S., Franklin, M. J., Williamson, K. S., Folsom, J. P., Boegli, L., & James, G. A. (2015). Contribution of stress responses to antibiotic tolerance in Pseudomonas aeruginosa biofilms. *Antimicrobial Agents and Chemotherapy*, *59*(7), 3838–3847. https://doi.org/10.1128/AAC.00433-15

Subramanian, I., Verma, S., Kumar, S., Jere, A., & Anamika, K. (2020). Multi-omics Data Integration, Interpretation, and Its Application. *Bioinformatics and Biology Insights*, *14*, 1177932219899051. https://doi.org/10.1177/1177932219899051

Toval, F., Guzmán-Marte, A., Madriz, V., Somogyi, T., Rodríguez, C., & García, F. (2015). Predominance of carbapenem-resistant Pseudomonas aeruginosa isolates carrying blaIMP and blaVIM metallo-β-lactamases in a major hospital in Costa Rica. *Journal of Medical Microbiology*, *64*(1), 37–43. https://doi.org/10.1099/jmm.0.081802-0

Turton, J. F., Wright, L., Underwood, A., Witney, A. A., Chan, Y. T., Al-Shahib, A., … Woodford, N. (2015). High-resolution analysis by whole-genome sequencing of an international lineage (Sequence Type 111) of pseudomonas aeruginosa associated with metallo-carbapenemases in the United Kingdom. *Journal of Clinical Microbiology*, *53*(8), 2622–2631. https://doi.org/10.1128/JCM.00505-15

Woodford, N., Turton, J. F., & Livermore, D. M. (2011). Multiresistant Gram-negative bacteria: the role of high-risk clones in the dissemination of antibiotic resistance. *FEMS Microbiology Reviews*, *35*(5), 736–755. https://doi.org/10.1111/j.1574-6976.2011.00268.x

World Health Organization. (2017). *Guidelines for the prevention and control of carbapenem-resistant Enterobacteriaceae, Acinetobacter baumannii and Pseudomonas aeruginosa in health care facilities*. Geneva. Retrieved from https://apps.who.int/iris/bitstream/handle/10665/259462/9789241550178-eng.pdf?sequence=1&ua=1

Zhao, W., Chen, J. J., Perkins, R., Wang, Y., Liu, Z., Hong, H., … Strain, E. (2016). A novel procedure on next generation sequencing data analysis using text mining algorithm. *BMC Bioinformatics*, *17*(1), 213. https://doi.org/10.1186/s12859-016-1075-9