UNIVERSIDAD DE CHILE -FACULTAD DE CIENCIAS
ESCUELA DE CIENCIAS AMBIENTALES Y BIOTECNOLOGÍA

Desarrollo de una herramienta bioinformática para la anotación detallada de plásmidos del patógeno bacteriano *Klebsiella pneumoniae*

Seminario de Título entregado a la Universidad de Chile en cumplimiento parcial de los requisitos para optar al Título de

Ingeniero en Biotecnología Molecular

Joaquín Elías Acosta Aguilera

Director: Dr. Andrés Marcoleta Caldera

Codirector: Dr.(c) Camilo Berríos Pastén

Mayo 2023

Santiago - Chile

.

UNIVERSIDAD DE CHILE -FACULTAD DE CIENCIAS
ESCUELA DE CIENCIAS AMBIENTALES Y BIOTECNOLOGÍA

INFORME DE APROBACIÓN SEMINARIO DE TÍTULO

Se informa a la Escuela de Pregrado de la Facultad de Ciencias, de la Universidad de Chile que el Seminario de Título, presentado por

Joaquín Elías Acosta Aguilera

"Desarrollo de una herramienta bioinformática para la anotación detallada de plásmidos del patógeno bacteriano *Klebsiella pneumoniae*"

Ha sido aprobado por la Comisión de Evaluación, en cumplimiento parcial de los requisitos para optar al Título de

Ingeniero en Biotecnología Molecular

Dr. Andrés Marcoleta Caldera
Director Seminario de Título _____

Dr.(c) Camilo Berríos Pastén
Codirector Seminario de Título _____

Comisión Revisora y Evaluadora

Dra. Rosalba Lagos Mónaco
Presidenta _____

Dr. Gastón Higuera Guajardo
Evaluador _____

Santiago de Chile, 2023

BIOGRAFÍA

Nací en la ciudad de Concepción el 27 de noviembre de 1996. Desde ese momento mi vida siempre estuvo en movimiento, dado que gracias al esfuerzo y cariño de mis padres, nos íbamos trasladando de ciudad en ciudad para aprovechar mejores oportunidades. Recorriendo Chile de Sur a Centro tuve que aprender a adaptarme y saber priorizar responsabilidades desde muy chico, no me aburría tanto de este estilo de vida ya que siempre podía encontrar algo nuevo que hacer.

Y no solo cosas nuevas sino que personas interesantes, mi educación fue muy enriquecedora especialmente en ciencias y, donde creo yo, nació mi pasión por ser científico. Si me preguntan creo que todo comenzó en el colegio Osorno College donde tuve mi primera clase de Taxidermia y tuvimos que reconstruir un esqueleto de mamífero. Recuerdo que mi profesora en ese entonces, Renate Yunginger, fomentaba atreverse y equivocarse y eso es una lección que llevo muy dentro de mi corazón ya que esa es la esencia de practicar ciencias, ensayo y error.

Terminando mi enseñanza media llegó la hora de tomar la decisión de mi vida, sobre qué estudiar. Si bien no estaba seguro, sabía que lo que eligiera no debía limitarme a mí y a mi curiosidad por aprender, por lo tanto, afortunadamente elegí la carrera de Ing. en Biotecnología Molecular. Y digo afortunadamente porque efectivamente esta carrera me otorgó la oportunidad de mezclar también otra pasión que tengo que es la electrónica y la instrumentación, lo que me dio acceso a otros mundos y perspectivas por conocer. Cada vez que aprendo algo nuevo me siento aún más ignorante, pero no importa, ya que disfruto el viaje.

A mi tío Horacio.
Le dedico el fruto de todos estos años de trabajo y
espero algún día podamos reunirnos con una cervecita en mano.

A mis padres.
Por confiar en mí y nunca rendirse, el camino fue empedrado pero
ya estamos cosechando lo sembrado.

La mente no es una vasija que deba ser llenada
sino una llama que debe ser avivada (...)
–Plutarch

AGRADECIMIENTOS

Primero que todo me gustaría agradecer el trabajo y confianza de mis directores de seminario de título, Dr. Andrés Marcoleta y Dr (c). Camilo Berríos. Llegué con la esperanza de aprender sobre bioinformática y me recibieron con los brazos abiertos, su dirección y guía me permitió nutrir esa pasión y desarrollar mis habilidades.

A Rodrigo Pulgar, Inmaculada Vaca, Juan Carlos Letelier y a sus respectivos equipos, por recibir a un novato y enseñarme desde el trabajo en mesón hasta el análisis de datos.

Al equipo UChile_Biotec 2017 y 2018, por darme la oportunidad de participar en su proyecto, que fue una experiencia muy enriquecedora en cuanto a lo científico y personal.

A Marcelo Trejo y Rosita Barchaj, por su labor docente, cariño y apoyo al desarrollo de mi vocación científica durante la enseñanza media.

A Francisco Tapia por ser el mejor, tu apoyo y amor incondicional durante el colegio, la universidad y la vida son impagables. Como ya sabes, siempre tendrás un techo en mi hogar.

A Juan Carlos, por ser un tremendo primo y enseñarme de computación y hardware.

A Javier Bustamante, Catalina Ávila, Carlos Vidal, Karla Villalobos, Camila Arancibia, Amelia Cox, Camilo Berríos, Valentina Carrasco, Macarena Collao, Paulina Aguilera, Patricio Arros, Matías Gálvez, David Arancibia, Alam Núñez, Daniel Acuña, Francisca Vera, Carlos Serrano, Pablo Lorca, Roberto Rojas y tantos otros. Por su apoyo, cariño, risas y paciencia.

A mi familia en casa, en Concepción y Curicó, hemos hecho muchos sacrificios, pero su apoyo, oraciones y buenas vibras me han ayudado mucho y les estoy eternamente agradecido.

Finalmente, me gustaría agradecer a mi familia laboral, los integrantes del laboratorio BEM, quienes infundieron mis jornadas de trabajo de alegría y momentos especiales. Es un privilegio participar en su equipo.

INDICE

RESUMEN

La cada vez más frecuente detección de patógenos bacterianos resistentes a múltiples antibióticos y con mayor virulencia constituye actualmente una de las amenazas más críticas para la salud global. El desarrollo de estas cepas depende, en gran medida, de la adquisición de genes de virulencia y de resistencia mediada por elementos genéticos móviles, donde los plásmidos juegan un rol preponderante. La creciente incorporación de técnicas de secuenciación masiva de ADN y de herramientas genómicas para la vigilancia de cepas patogénicas ha traído un aumento explosivo en la cantidad de genomas que deben ser analizados en un tiempo acotado. Por ello, se requiere de estrategias que permitan un análisis exhaustivo, rápido, automatizado, y amigable para el usuario de esta información. En particular, se requiere de herramientas que permitan la anotación integrada de plásmidos, para identificar genes y elementos de secuencia de relevancia clínica, o que permitan predecir su potencial de diseminación.

En este trabajo se creó una herramienta bioinformática llamada "*GMI-PAN*", bajo el código "*Contannotate*", para la evaluación y anotación de perfiles de resistencia a antibióticos y metales, factores de virulencia y elementos genéticos móviles en secuencias nucleotídicas ensambladas de naturaleza procarionte, optimizado para la familia *Enterobacteriaceae*. La herramienta permite automatizar un *pipeline* de anotación que integra varias herramientas previamente desarrolladas, lo que reduce significativamente el tiempo de trabajo.

Adicionalmente, *GMI-PAN* fue utilizada exitosamente para el análisis detallado del genoma de diez cepas de *Klebsiella pneumoniae* resistentes a carbapenémicos aisladas, en territorio nacional, por el Instituto de Salud Pública de Chile (ISP). En dichas cepas, se identificaron más de 40 plásmidos, los cuales fueron anotados con el programa desarrollado. Los resultados destacan la construcción exitosa de un *pipeline* optimizado para la anotación de plásmidos, donde se reporta que de los 43 *contigs* analizados por *GMI-PAN*, 24 están relacionados

a genes asociados a betalactamasas y 9 a genes asociados a carbapenemasas, específicamente *OXA*, *NDM* y *KPC*, siendo consecuente con los resultados reportados de los ensayos fenotípicos del ISP.

Al comparar los resultados automatizados y manuales, se reportó una mejora cuantitativa en el rendimiento para la anotación de determinantes tales como resistencia a antibióticos, sistemas de secreción, transposones y resistencia a metales. Por otro lado, se detectó la presencia de grupos de plásmidos al hacer clusterización en base a resistoma, donde dentro de estos se encontró un grupo asociado a resistencia a arsénico, cobre y betalactamasas de espectro extendido *CTX-M*, otro asociado a *OXA* y bombas de flujo de tipo *MFS* que confieren resistencias a sustancias tóxicas, otro asociado a *KPC* y una batería de resistencias a antibióticos y, por último, otro asociado a *NDM* y resistencia a mercurio.

En conclusión, *GMI-PAN* es una herramienta que disminuye notablemente el tiempo de trabajo en la anotación de secuencias genómicas bacterianas de semanas a solo minutos, sin sacrificar la cantidad o la calidad de la información que se obtiene.

ABSTRACT

The increasingly frequent detection of multidrug-resistant bacterial pathogens with enhanced virulence is currently one of the most critical threats to global health. The development of these strains depends largely on the acquisition of virulence and resistance genes mediated by mobile genetic elements, where plasmids play a prominent role. The growing adoption of high-throughput DNA sequencing techniques and genomic tools for pathogenic strain surveillance has resulted in an explosive increase in the number of genomes to be analyzed within a limited timeframe. Thus, there is a need for comprehensive, rapid, automated, and user-friendly strategies of analyzing this information. Specifically, tools that enable integrated plasmid annotation are required to identify clinically relevant genes and sequence elements, or to predict their potential dissemination.

In this study, we developed a bioinformatics tool called *"GMI-PAN"*, under the code *"Contannotate"*, for the evaluation and annotation of antibiotic and metal resistance profiles, virulence factors, and mobile genetic elements in assembled nucleotide sequences of prokaryotic nature, optimized for the *Enterobacteriaceae* family. This tool allows the automation of an annotation pipeline that integrates several previously developed tools, significantly reducing the required work time.

Furthermore, *GMI-PAN* was successfully utilized for the detailed analysis of the genomes of ten carbapenem-resistant *Klebsiella pneumoniae* strains isolated in Chile by the National Institute of Public Health (ISP). In these strains, over 40 plasmids were identified and annotated using the developed program. The results highlight the successful construction of an optimized pipeline for plasmid annotation, reporting that out of the 43 contigs analyzed by *GMI-PAN*, 24 are related to genes associated with betalactamases and 9 to genes associated with carbapenemases, specifically *OXA*, *NDM*, and *KPC*, which is consistent with the results reported

by ISP's phenotypic assays.

When comparing automated and manual results, a quantitative improvement in performance is reported for annotating determinants such as antibiotic resistance, secretion systems, transposons, and metal resistance. Additionally, the presence of plasmid groups was detected through cluster analysis based on the resistome, including a group associated with resistance to arsenic, copper, and extended-spectrum betalactamases *CTX-M*, another associated with *OXA* and multidrug efflux pumps of the *MFS* type conferring resistance to toxic substances, another associated with *KPC* and a range of antibiotic resistances, and finally, another associated with *NDM* and mercury resistance.

In conclusion, *GMI-PAN* reduces time required for bacterial genomic sequence annotation from weeks to minutes, without sacrificing quantity or quality of the obtained information.

1. INTRODUCCIÓN

1.1. Multiresistencia en bacterias y abuso de antibióticos

El surgimiento de las resistencias a antimicrobianos (RAM) es un fenómeno natural en microorganismos, sin embargo, se ha acelerado su aparición en cepas con relevancia clínica dado el uso abusivo de agentes antimicrobianos en humanos y en animales de granja (Elshamy & Aboshanab, 2020). Además, el uso indiscriminado de antibióticos ha sido declarado el catalizador clave para la generación de multirresistencia, debido a la presión selectiva que se genera desde una perspectiva ecológica (WHO, Geneva, 2001).

El rol de los antibióticos no sólo se limita a los tratamientos de enfermedades y a la resistencia, sino que también pueden funcionar como agentes de señalización y, dependiendo del contexto y dosificación, pueden actuar promoviendo la expresión de genes de virulencia como por ejemplo la producción de toxinas, adherencia, formación de biofilms, y sistema de secreción, entre otros (Goneau et al., 2020).

En este sentido, la presencia de bacterias patógenas multirresistentes representa una amenaza, pues es altamente probable que en algún momento estas bacterias sean resistentes a todos los antibióticos disponibles, situación principalmente grave en países subdesarrollados con acceso restringido a los antimicrobianos y peores condiciones de higiene. El año 2017 la OMS publicó una lista de patógenos contra los cuales el desarrollo de nuevos antibióticos es urgente. Dentro de esta lista se encuentran *Enterococcus faecium, Staphylococcus aureus, Klebsiella pneumoniae, Acinetobacter baumannii, Pseudomonas aeruginosa y Enterobacter cloacae (ESKAPE)* fueron patógenos designados como un grupo prioritario (De Oliveira et al., 2020). Estas cepas tienen un origen tanto hospitalario como comunitario (Walsh & Amyes, 2004). Dentro de las bacterias gram-negativas tenemos casos como *Klebsiella pneumoniae*, de la familia *Enterobacteriaceae*, en la que se han identificado cepas portadoras de betalactamasas de espectro

extendido que son capaces de resistir incluso la última generación de antibióticos desarrollados (Levy & Marshal, 2004).

1.2. Elementos genéticos móviles y transferencia horizontal

¿Cómo se transfieren los determinantes genéticos de la multirresistencia y virulencia? Se sabe que la transferencia horizontal de genes (THG) es uno de los mecanismos relevantes en la evolución de procariontes, y sus principales vectores corresponden a elementos genéticos móviles (EGM) tales como plásmidos, transposones, bacteriófagos y elementos integrativos/conjugativos como secuencias de inserción e integrones (Che et al., 2021; Botelho & Schulenburg, 2020).

Los plásmidos son probablemente los elementos más estudiados. Estos han sido divididos en tres tipos: Conjugativos, no-conjugativos movilizables y no-conjugativos no-movilizables. Adicionalmente, los mecanismos en que se basan para su diseminación consisten en la conjugación, movilización conjugativa, transducción y transformación (Che et al., 2021) (Beltrán et al., 2021).

Los plásmidos se asocian con frecuencia a otros EGM tales como elementos transponibles e integrones. Una vez incorporado en el nuevo hospedero, dichos elementos pueden ser transferidos a la región cromosomal del genoma bacteriano, y potencialmente promover una ventaja adaptativa. Por ejemplo, se ha descrito la transferencia de transposones codificantes para resistencia a mercurio, entre distintas especies de *Pseudomonas* usando plásmidos conjugativos como vehículos (capaces de moverse desde el cromosoma hasta el plásmido o viceversa), de esta forma pueden colonizar ambientes contaminados con mercurio (Beltrán et al., 2021).

En este seminario de título, el enfoque estuvo centrado principalmente en los plásmidos, como EGM capaces de adquirir y diseminar RAM. Las razones detrás de esto están relacionadas con que se ha descrito que los plásmidos son un punto de convergencia en muchos fenómenos de THG. De todos modos, no está limitado únicamente a este tipo de elementos movilizables, pero

son de gran importancia ya que disponen de una arquitectura genómica lo suficientemente flexible y funcional para la inclusión y activación de mecanismos de otros EGM (Beltrán et al., 2021).

1.3. Factores relevantes asociados a RAM y Virulencia

Las bacterias pueden utilizar varios mecanismos de resistencia, unos intrínsecos, donde la célula puede utilizar genes que ya posee para sobrevivir exposición a un antibiótico, y otros adquiridos, en los que la ganancia de nuevo material genético puede proveer la capacidad de mediar la supervivencia (Darby et al., 2022). Algunas de las clases de antibióticos más importantes y de especial interés en este trabajo son los carbapenémicos, ya que estos agentes antimicrobianos son reservados como la última línea de defensa en el combate contra infecciones causadas por bacterias multirresistentes (Elshamy & Aboshanab, 2020).

Entre los mecanismos de resistencia a antibióticos betalactámicos se encuentran mutaciones en genes codificantes para ciertas porinas, la sobreexpresión de bombas de flujo y, por último, la síntesis de betalactamasas capaces de hidrolizar algunos tipos de betalactámicos, algunas activas inclusive sobre carbapenémicos, que en consecuencia estas pasan a llamarse carbapenemasas (Sawa et al., 2020). Las betalactamasas se dividen en cuatro clases. Las clases A, C y D contienen enzimas con un sitio activo compuesto por serina, mientras que la clase B agrupa enzimas que contienen iones Zinc en su sitio activo para la hidrólisis de betalactámicos.

Una de las formas de adquirir nuevas capacidades de resistencia y virulencia, es a través de transposones, elementos cuyo tamaño varía entre 2,5 y 60 kb y que poseen un repetido inverso terminal largo más uno o varios genes accesorios que confieren un fenotipo ventajoso para el hospedero bacteriano, tales como resistencia a antibióticos (RA) o resistencia a metales (RM) (Wozniak & Waldor, 2010; Darmon & Leach, 2014). Otros elementos que codifican resistencias a antimicrobianos son los integrones, los cuales capturan marcos abiertos de lectura (ORFs) sin

promotor y mediante recombinación de sitios específicos (attI y attC) los insertan en un operón, facilitando su expresión. Estos casetes de genes pueden ser variables ya que dependiendo de su contenido estos pueden codificar diversos determinantes, como por ejemplo, proteínas relacionadas con RAM (Domingues et al., 2012; Darmon & Leach, 2014). Por otro lado, se puede mediar la transferencia de genes asociados a RAM a través de bacteriófagos, los cuales pueden transponer su ADN dentro del cromosoma bacteriano, un plásmido o un profago (Darmon & Leach, 2014).

También se ha descrito la presencia de genes RAM en secuencias de inserción (IS), los cuales son segmentos de ADN de entre 0,7 y 2,5 kb que contienen uno o dos ORFs. Éstos codifican proteínas responsables para su movilidad como la transposasa, la cual va ligada a una secuencia terminal corta *IR*. La inserción de un IS siempre va a cambiar el genoma hospedero, mientras que la escisión puede producir una restauración o una mutación, por lo que no es de extrañar que estos elementos tengan un rol en la creación de ensambladores modulares de genes (Siguier et al., 2006; Darmon & Leach, 2014; Larsson et al., 2020).

Finalmente, existen diversos mecanismos que tienen una incidencia importante en la virulencia bacteriana, entre estos están los sistemas de secreción bacteriana. Específicamente, el sistema de secreción tipo IV que es un complejo proteico asociado a eventos de conjugación bacteriana y es capaz de transportar proteínas o ADN a través de la membrana celular (Che et al., 2021). También existen otros sistemas de secreción como el tipo III y VI, sin embargo, se ha descrito que estos conjuntos adoptan un rol comunicacional entre bacterias, mediado por el transporte de proteínas efectoras tanto beneficiosas como perjudiciales (Green & Mecsas, 2016).

1.4. Iniciativas de seguimiento genómico en base a secuenciación

Dada la gran escala, impacto y dinámica de la multirresistencia e hipervirulencia, surge entonces la necesidad de generar iniciativas de seguimiento genómico para dichos organismos. De esta forma, debemos contar con herramientas que permitan monitorear, explorar y recolectar evidencia de las características epidemiológicas de poblaciones bacterianas. Los principales objetivos de un estudio de seguimiento genómico son la determinación de la distribución geográfica, las dinámicas poblacionales de clones multirresistentes (MDR) y elementos transmisibles de resistencia para después, en base a esto, construir políticas de prevención de riesgo, prevención y control ante una potencial amenaza epidemiológica.

Por ejemplo, una revisión de la genómica poblacional de *Klebsiella pneumoniae* (Wyres et al., 2020) describe muy detalladamente el rol y complejidad que tiene *K. pneumoniae* en infecciones oportunistas intrahospitalarias basadas en RAM. Sin embargo, no toda su población es homogénea, también se revela un paralelismo en las dinámicas poblacionales de *K. pneumoniae*, donde una comunidad se caracteriza por la existencia de cepas exclusivamente MDR (B-lactamasas; carbapenemasas) y otras exclusivamente hipervirulentas. Por lo tanto, resulta clave entender y conocer el potencial de la población a analizar, ya que la manera en que se desee abordar algún evento epidemiológico también depende del tipo de interacciones que puedan tener esta con otras comunidades diferentes entre sí.

Dada esta situación, los autores se enfocaron en analizar cómo estas aproximaciones de seguimiento genómico pueden mejorar nuestro entendimiento de la taxonomía, ecología, evolución, diversidad y distribución de determinantes de patogenicidad y RAM relevantes en el ámbito clínico, revelando la importancia de herramientas como el *whole genome sequencing* (WGS) para la detección de especies, linajes, loci de síntesis de cápsula, determinantes basales y accesorios (RAM y virulencia), y el grado de relación entre cepas para la evaluación de

transmisibilidad. Esto permite, finalmente, hacer un rastreo de los distintos grupos clonales en base a la convergencia o divergencia de estos determinantes y de esta forma profundizar en la exploración de esta relación entre las infecciones clínicas y potenciales reservorios ecológicos.

Otro ejemplo de seguimiento genómico es el análisis del resistoma de la microbiota de suelo Antártico (Marcoleta et al., 2022), donde se encontró una variedad extensa de resistencias tanto a metales como antibióticos, y se destaca que una proporción significativa del resistoma se encuentra principalmente en plásmidos y otros EGM. Adicionalmente, se describe su potencial para ser transferido a bacterias patogénicas, similar al caso previamente presentado donde se hablaba de un paralelismo en estas comunidades, lo que entrega una oportunidad para que este microbioma antártico actúe como una fuente de genes de resistencia relevante para el ámbito clínico (Marcoleta et al., 2022). Dentro de las herramientas utilizadas para llevar a cabo el ejercicio de seguimiento genómico, se encuentran WGS mediado por secuenciación *Illumina*^{TM} y *Nanopore*^{TM} más ensamblaje híbrido genómico, caracterización/anotación de genomas mediante diversos *pipelines* con bases de datos asociados a RAM, etc.

1.5. Herramientas de anotación, contexto y automatización

La anotación genómica es el proceso de identificación y rotulación de todos los determinantes relevantes en una secuencia genómica, incluyendo como mínimo las coordenadas de las regiones codificantes predichas y sus productos proteicos (Richardson & Watson, 2012).

El avance en las tecnologías de secuenciación de ADN ha ido en crecimiento, mejorando el rendimiento de estos instrumentos cada vez más por cada generación tecnológica y a tal nivel, que la secuenciación y ensamblaje de genomas bacterianos completos se ha vuelto un ejercicio rutinario. El paso siguiente, que consiste en la anotación de determinantes genómicos relevantes en los ensamblados, se ha abordado lentamente mediante la utilización de herramientas web. Si bien una interfaz web resulta atractiva y, generalmente, más simple de aprender y ejecutar, la

utilización de estas plataformas no es una estrategia aplicable para información sensible e integración óptima de protocolos computacionales, además del análisis masivo de secuencias (Seeman, 2014).

Mientras más genomas son secuenciados, más datos son generados y menos es el tiempo que se puede destinar al ejercicio de anotación manual. Además, dado que el seguimiento genómico suele contemplar cantidades masivas de datos, como en los ejemplos anteriormente descritos, es necesario contar con herramientas que sean capaces de automatizar ciertas etapas dentro de este, como lo es la anotación de genomas ensamblados. De no ser así, estas etapas pueden dilatarse entre semanas a meses.

Actualmente, se han desarrollado herramientas que permiten la automatización de la anotación genómica y posterior análisis. En particular, para el trabajo de genómica poblacional de *K. pneumoniae* se ha descrito la automatización de la etapa de anotación a través de programas como *Kleborate* (Lam et al., 2021). Este programa está escrito en lenguaje de programación *Python* y tiene como objetivo identificar y anotar genomas ensamblados de *K. pneumoniae* para encontrar elementos como: tipo de secuencia según *MLST*, subespecie, *loci* asociados a virulencia *ICEKp*, loci asociados a plásmidos de virulencia, RAM y predicción de serotipo capsular *K/O*. El resultado de este esfuerzo permitió la genotipificación de aproximadamente 10.000 genomas en bases de datos públicas.

Una de las herramientas de anotación genómica más utilizadas es *PROKKA* (Seeman, 2014), superando más de 10 mil citas según Google Scholar. Este programa está escrito en base al lenguaje *Pearl* y su objetivo es la anotación de características tales como secuencias codificantes de proteínas (*CDSs*), genes de ARN ribosomal (ARNr), genes de ARN de transferencia (ARNt), péptido señal y ARN no codificante. El autor de *PROKKA* especifica “La forma tradicional de predecir qué codifica un gen es comparándolo con una base de datos grande

de secuencias conocidas, usualmente a nivel proteico, y transferir la anotación del mejor emparejamiento significativo". Si bien, *PROKKA* logró anotar para un genoma cromosomal aproximadamente 6.000 *CDSs*, hay que tener en cuenta que las bases de datos usadas como referencia no son algo final. Estas varían en tamaño y calidad conforme se actualizan los datos, y pueden tener cierto sesgo en su creación, ya que depende mucho del interés puesto sobre algunos géneros. Es por esto que *PROKKA* trata de solventar este problema estableciendo un método de anotación jerárquico, donde la prioridad se sitúa sobre el usuario ofreciendo la posibilidad de introducir su base de datos curada de proteínas. Sin embargo, ¿Qué ocurre en situaciones donde el genoma a anotar es desconocido? ¿O si el genoma es tan raro que la comparación con las bases de datos existentes no arroja muchos resultados? Y si se desea anotar EGM de bacterias multirresistentes, ¿se podría identificar el origen de transferencia o grupo de incompatibilidad de un plásmido?

Estas limitantes de *PROKKA* pueden resultar en la anotación de proteínas hipotéticas (HP), o simplemente en un espacio de anotación vacío para otros *pipelines*. Debido a esto, los autores del trabajo decidieron acoplar otras herramientas de anotación para afinar y completar dichos resultados, por ejemplo: *Phaster* (Arndt et al., 2016) para la anotación de profagos o *RGI* (Alcock et al., 2020) para la anotación de bombas de flujo u otros determinantes asociados a RA.

1.6. *GMI-PAN* como herramienta de anotación automatizada

En nuestro laboratorio el objetivo es estudiar el resistoma y los mecanismos de virulencia presentes en cepas aisladas de diversos organismos procariontes, con especial atención en los EGM que puedan contener debido a su potencial diseminación y riesgo epidemiológico. En específico, resulta interesante conocer: origen de transferencia (oriT), grupo de incompatibilidad (IncG), integrones (INT), transposones (Tn), profagos (PHAGE), sistema de secreción tipo IV (T4SS), resistencia a carbapenémicos y otros antibióticos, RM y naturaleza cromosomal o

plasmidial del *contig* analizado. Si bien la comunidad *open-source/libre-software* ofrece un gran abanico de posibilidades en cuanto a *pipelines* de anotación, la parcelación de la información y el enfoque específico de este trabajo limita esas opciones. En consecuencia, resulta necesario diseñar un *pipeline* a la medida con el objetivo de este trabajo.

Si bien, anteriormente ya se mencionan herramientas capaces de rescatar resultados con cierta precisión para determinantes específicos, estas no necesariamente se ajustan adecuadamente al enfoque del investigador o del proyecto que se lleva a cabo, dado que quien diseñó estos programas fue un tercero con otro objetivo en mente. Si el enfoque de la investigación requiere de una anotación rápida y lo más completa posible, que otorgue libertad al usuario para manipular variables dentro del análisis y resguarde la privacidad de los datos, es necesario diseñar software basado en líneas de comando dentro del mismo marco en que se da origen al trabajo, para así aprovechar al máximo los recursos utilizados y obtener datos de alta calidad.

Dada esta necesidad, en este seminario de título se crea *GMI-PAN*, un conjunto de módulos independientes de anotación integrados en una línea de trabajo escrita en lenguaje *Python* para el análisis completo de genomas bacterianos. Esta herramienta permite automatizar el ejercicio de evaluación y anotación de perfiles de RA y metales, factores de virulencia y elementos genéticos móviles, la cual está optimizada con bases de datos para la familia *Enterobacteriaceae,* ya que comprende un número significativo de patógenos con relevancia clínica.

El Grupo de Microbiología Integrativa trabajó en un estudio que consistió en el análisis de distintas cepas de *K. pneumoniae* resistentes a carbapenémicos aislados por el Instituto de Salud Pública de Chile (ISP), las cuales presentan un conjunto altamente diverso de EGMs que codifican factores de virulencia y resistencia. Para demostrar la funcionalidad de la herramienta

desarrollada, se ejecutó un análisis con *GMI-PAN* de las diez cepas de *K. pneumoniae*, aisladas por el ISP, las cuales se secuenciaron y ensamblaron de manera híbrida para obtener los genomas completos con sus replicones cerrados (tanto cromosoma como plásmidos) para analizar con este software.

2. OBJETIVOS

2.1. Objetivo General

Desarrollar un programa bioinformático para la anotación detallada de plásmidos de *Klebsiella pneumoniae* y otros patógenos bacterianos de la familia *Enterobacteriaceae*.

2.2. Objetivos Específicos:

1. Construir una herramienta que permita caracterizar y anotar plásmidos respecto a genes relevantes en patogénesis, resistencia a biocidas y otras funciones relevantes en plásmidos.

2. Evaluar la herramienta desarrollada en base al perfil de plásmidos anotados de *Klebsiella pneumoniae,* provenientes de diez cepas aisladas por el Instituto de Salud Pública.

3. Describir la diversidad de plásmidos de la colección de cepas analizadas. Evaluar la correlación del fenotipo de resistencia reportado para las cepas con los elementos codificados en plásmidos resueltos por el programa.

3. MATERIALES Y MÉTODOS

3.1. Aislados y test de resistencia

Las cepas analizadas de *K. pneumoniae* resistentes a carbapenémicos (CR-*Kp*) fueron obtenidas en el contexto de la vigilancia nacional de enterobacterias resistentes a carbapenémicos, llevada a cabo por el Instituto de Salud Pública (ISP), que depende en parte del Ministerio de Salud de Chile. Los ensayos de susceptibilidad a antimicrobianos para la validación del fenotipo de estas estas cepas se encuentran en Tablas Suplementarias 1 y 2.

3.2. Secuenciación y ensamblaje de *contigs*

El protocolo de extracción del ADN genómico y los pasos de ensamblaje híbrido de las 10 cepas de *K. pneumoniae* aisladas por el ISP fueron descritos en el *preprint* publicado de este estudio (Veloso et al., 2022). Dicho trabajo buscó estudiar resistencia microbiana y análisis genómico comparativo de diez aislados *CR-Kp* provenientes de la vigilancia nacional de *Enterobacteriaceae* resistentes a carbapenémicos. Dentro de sus resultados se describe resistencia a la mayoría de los antibióticos probados entre los aislados, 5 *ST25*, 3 *ST11*, 1 *ST45* y 1 *ST505*, los cuales contenían un total de 44 plásmidos, muchos de estos predichos como conjugativos y que albergan una batería de genes que confieren RA y RM. Finalmente se destaca la caracterización de diez plásmidos asociados a carbapenemasas que codifican tanto para *KPC-2*, *NDM-1* como *NDM-7*. En base a estos precedentes, se consideraron 43 *contigs* que cumplieron la condición de estar ensamblados.

3.3. Módulos de anotación independientes

En cuanto a los módulos utilizados por *GMI-PAN*, estos ejecutan sus protocolos en el marco de *Annotation & Filtering*, dentro de la etapa *Annotation Module,* descrito en la Figura 1B específicamente. A continuación, se presenta un resumen de las funciones de cada uno dentro del *pipeline*.

PROKKA (v1.14.6) es un *pipeline* de anotación que fue incorporado principalmente por su anotación basada en *CDSs* y también porque da la posibilidad al usuario de ingresar su propia base de datos de proteínas anotadas. Este realiza un alineamiento con proteínas o con evidencia de transcrito, contenidas en la base de datos *The Univesal Protein Resource* (*UniProt*), la cual alberga información sobre secuencias proteicas y su respectiva anotación (Apweiler et al., 2004). Posteriormente, ejecuta una búsqueda en la base de datos *The Reference Sequence* (*RefSeq*), del *National Center for Biotechnology Information* (*NCBI*), que contiene una colección no redundante e integrada sobre secuencias asociadas a ADN genómico y proteínas (O'Leary et al., 2016). Finalmente, *PROKKA* termina con un último nivel de búsqueda asociado a la base de datos *The protein families database* (*Pfam*), la cual es una base de datos que se construye a partir de perfiles ocultos de modelo de Márkov (Punta et al., 2012). Si en ninguno de los pasos anteriormente descritos se encuentra emparejamiento, entonces este se anota como *"hypothetical protein"* (Seeman, 2014).

PlasmidVerify (v04/20) es una herramienta que permite determinar la naturaleza plasmidial o cromosómica del *contig* a anotar. Primero predice los genes que pueden estar contenidos en el *contig* con *Prodigal* (Hyatt et al., 2010), luego ejecuta un alineamiento de las proteínas predichas con *hmmersearch* (http://hmmer.org/), donde principalmente se buscan genes plásmido-específico conocidos en bases de datos que fueron curadas e incluidas por los mismos autores, tales como *PlasmidDatabase* y *nonPlasmidDatabase*. Finalmente, *PlasmidVerify* identifica si el *contig* se representa mejor como un plásmido o cromosoma a través de un clasificador bayesiano (Antipov et al., 2019).

PlasmidFinder (v09/21) es un programa con una base de datos curada de replicones plasmidiales diseñada para la identificación de plásmidos en *WGS* originados desde especies de *Enterobacteriaceae*, usando *contigs* ensamblados (Carattoli et al., 2014). En cuanto a la

identificación de los IncG, su anotación se lleva a cabo con otra base de datos llamada *Plasmid Multi-Locus Sequence Typing* (*pMLST*), la cual también usa *blastn* tanto en secuencias crudas como ensambladas. El criterio de anotación considera un 85% de identidad como mínimo con los emparejamientos contenidos en la base de datos y este tiene que cubrir como mínimo un 66% de la longitud de la secuencia del replicón contenido en la base de datos. Después de identificar los alelos para todos los genes, el secuencio-tipo del plásmido se determina según una combinación específica de los alelos (Jolley et al, 2018).

RGI (v5.2.1) es una herramienta que se desarrolló en conjunto con la base de datos llamada *Comprehensive Antibiotic Resistance Database* (*CARD*). *CARD*, surgió con el objetivo de armonizar y estandarizar, a través de un proceso humano de curado de datos, el conocimiento sobre las secuencias de ADN y proteínas vinculadas con RAM, para así producir una base de datos confiable que permita su clasificación. Dentro, la información se encuentra organizada según una ontología de resistencia a antibióticos (*ARO*, por sus siglas en inglés), donde se describe en detalle la base molecular que confiere resistencia, clase de antibiótico, droga blanco y el mecanismo molecular de resistencia. La idea es entregar una fuente de información práctica y funcional para investigadores, industrias y agencias de salud públicas (Alcock et al, 2020).

PhageBoost (v0.1.7) es un *pipeline* de anotación basado en *machine learning* (ML) que permite la detección y descubrimiento de regiones relacionadas a fagos. Su aproximación calcula e identifica determinantes biológicos tanto en secuencias nucleotídicas como aminoacídicas para cada gen y luego usa ML para predecir qué gen o genes pertenecen a bacterias o profagos. La anotación se basa en características biológicas tales como el contenido GC, composición aminoacídica, largo del gen, sentido del gen, distancia intergénica, entre otros (Kimmo et al., 2020).

NHMMER (v3.3.2) es un programa enfocado en la comparación de una o más secuencias nucleotídicas contra una base de datos de secuencias de la misma naturaleza. Para cada búsqueda, el *pipeline* da como resultado una lista con los emparejamientos más significativos. Cada resultado en la lista consiste en una región con una similitud local entre una porción del *input* y una sub-secuencia de la base de datos, donde además contiene un alineamiento entre estas (Wheeler & Eddy, 2013).

IVIP (v01/22) es un *pipeline* que permite clasificar, anotar y visualizar INT de clase I. Este está separado en un módulo A de identificación/clasificación y un módulo B de extracción, anotación y visualización. Para identificar el integrón se consideran 3 elementos: el gen *intI1* (integrasa) como parte del segmento 5' conservado, el sitio *attC* asociado a los *cassettes* de genes, y la resistencia a *sul* en el segmento 3' conservado. Para cada elemento se hace un emparejamiento en la base de datos respectiva que se construyó (Zhang et al., 2018).

3.4. Anotación automatizada de diez genomas del ISP

Para el análisis de los 43 *contigs* rescatados de los genomas del ISP, se ejecutó de manera independiente el *pipeline* de *GMI-PAN* para los respectivos archivos FASTA, en donde participan los módulos previamente descritos.

3.5. Visualización de resultados

En cuanto a la visualización gráfica de los plásmidos, el archivo en formato .gbk obtenido de cada output de *GMI-PAN* puede ser analizado con *SnapGeneViewer* 6.1.1 (www.snapgene.com) para generar archivos que contienen un diagrama de cada determinante anotado del *contig*, los cuales además cuentan con una nomenclatura colorimétrica para una presentación opcional de los datos. Este programa tiene la ventaja de poder exportar la visualización del genoma anotado hacia archivos de imagen vectorizados para una etapa posterior de análisis y tratamiento.

3.6. Comparación funcional de plásmidos y genes

Usando el ejercicio de clusterización proteica de mmseq (v14.7e284)(Steinegger et al., 2017), tanto datos de *CDSs* como resistencias fueron clusterizados al 95% de identidad y 95% de cobertura bajo el argumento *easy-cluster*, donde se obtuvo como resultado un archivo .tsv respectivamente. Para cada archivo se ejecutó un programa escrito en *Python* que permite iterar y agrupar estos grupos de cluster representativos, asignando un 0 o 1, según corresponda, para reflejar la presencia o ausencia del cluster en cada *contig* de la muestra. Una vez obtenida la matriz representativa de cada archivo, este se hizo pasar por el *pipeline* del paquete *ComplexHeatmap* (Schlesner et al., 2016) para la visualización de los respectivos grupos de datos. La clusterización de plásmidos de *ComplexHeatmap* se contrastó directamente con la clusterización previamente descrita por el grupo (Veloso et al., 2022), basada en *CD-HIT*.

3.7. Anotación manual de diez genomas del ISP

Para validar los resultados obtenidos por *GMI-PAN* es necesario hacer un ejercicio comparativo entre los datos generados por automatización de código y datos generados por iteración humana. En este caso se construyó manualmente un perfil de anotación basado en las herramientas ya validadas: *PATRIC* (Wattam et al., 2017; Legacy version) para la anotación de *CDSs*, *PlasmidVerify* (Antipov et al., 2019; 04/20 version) para la naturaleza del *contig*, *oriTfinder* (Li et al., 2018; Version 1.1) para oriT, *PlasmidFinder* (Carattoli et al., 2014; 10/20 version) para IncG, *RGI-CARD* (Alcock et al., 2020; Version 5.1.1) para RAM, *ISfinder* (Siguier et al., 2006; 10/20 version) para IS y Tn, *IntegronFinder* (Nerón et al., 2022; Version 1.5.1) para INT y *PHASTER* (Arndt et al., 2016; 10/20 version) para profagos. Último acceso y actualización de datos 20/10/21.

3.8. Dependencias necesarias para el funcionamiento de *GMI-PAN*

GMI-PAN requiere de cierto ambiente y dependencias específicas para el funcionamiento adecuado del *pipeline* de anotación. Para este caso se utilizó *conda* 4.12.0 (https://docs.conda.io/projects/conda/en/stable/index.html) y el siguiente listado de dependencias, instalados respectivamente para su compatibilidad: *PROKKA* 1.14.6, *RGI* 5.2.1, *PlasmidVerify* 04/20, *PlasmidFinder* 09/21, *PhageBoost* 0.1.7 (*xgboost* 1.3.3), *I-VIP* 01/22, *NHMMER* 3.3.2, *bcbio-gff* 0.6.9. Variaciones de *conda* podrían seguir siendo funcionales con los respectivos paquetes, sin embargo, no se puede asegurar inter-compatibilidad en diferentes versiones de los distintos paquetes.

4. RESULTADOS

4.1. Esquema operativo y funcionamiento lógico de *GMI-PAN*

Para dar inicio al *pipeline* de anotación de *GMI-PAN* (Figura 1A), el usuario debe disponer de secuencias pre-ensambladas de ADN genómico en formato *FAST-ALL* (FASTA). Este es un formato de texto plano o ASCII que contiene un *header* con un identificador y, opcionalmente, un comentario. Luego del *header*, se encuentra una secuencia biológica nucleotídica o proteica (Pearson & Lipman, 1988). Este es el único archivo necesario y obligatorio para ejecutar *GMI-PAN*.

Para la anotación, *GMI-PAN* depende de herramientas de predicción externas (Figura 1B) para predecir oriT, IncG, INT, Tn, PHAGE, T4SS, resistencia a antibióticos, RM y naturaleza cromosomal o plasmidial del *contig* analizado. Además, se integró un conjunto de bases de datos adicionales que permiten agregar y refinar anotaciones en el *pipeline*.

Antes de analizar los resultados obtenidos, es necesario ordenar la información que se extrae de cada módulo y los futuros archivos que se vayan a generar, por lo que previo al tratamiento de datos, se construye un árbol de directorios que respete la jerarquía del *pipeline*. Esto es importante, dado que el manejo ordenado y cohesivo de datos o metadatos tiene un efecto en la reproducibilidad y validez de las conclusiones que se rescatan de los análisis que se realizan sobre estos.

Una vez que la anotación se haya completado, el árbol se haya establecido y todos los módulos hayan sido ejecutados, un archivo de texto único es extraído de los respectivos datos crudos generados por cada módulo. Para esta extracción, se diseñó un código de *Python* que permite reconocer patrones de datos en la fuente original y construir un archivo de texto plano con un formato estandarizado, desde donde se deposita la información rescatada de manera precisa para el output diseñado en *GMI-PAN*.

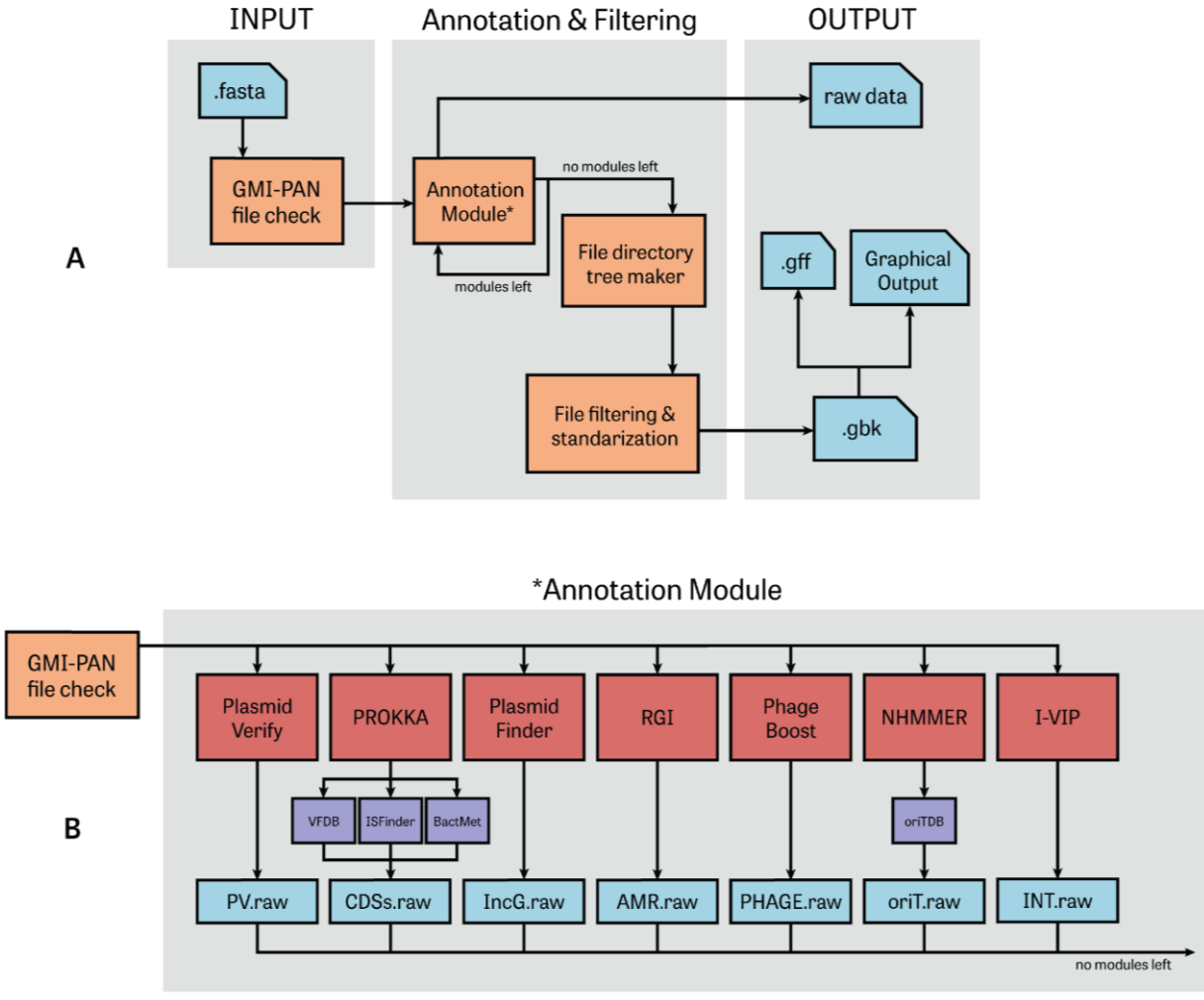

Figura 1.- *Pipeline* de anotación general de *GMI-PAN.* A) Esquema de flujo de las etapas y procesos inscritos dentro del código del software. El archivo presentado por el usuario es incorporado al *pipeline*, solo si cumple ciertas condiciones para, luego de ser tratado, expulsar archivos de salida en los respectivos formatos. B) Detalle del proceso *Annotation Module*. Cada módulo genera sus datos crudos para luego ser registrados para la construcción del árbol de directorios y, posteriormente, aislar y analizar los archivos de interés.

El código que logra sincronizar, ordenar, pulir, rescatar y generar datos estandarizados de este *pipeline* de análisis fue desarrollado en *Python v3.6* bajo el nombre clave *Contannotate*. Los módulos de predicción acoplados y las bases de datos adicionales utilizadas están listadas en la Tabla 1 y Tabla 2 respectivamente.

Tabla 1.- Herramientas de predicción de determinantes bacterianos utilizados por *GMI-PAN*.

Herramienta (referencia)	Determinante(s)
PROKKA (Seeman, T. 2014)	Secuencias codificantes (CDS), Resistencias a Metales, Transponsones (Tn) y Sistemas de Secreción (TXSS)
PlasmidVerify (Antipov, D. et al. 2019)	Genoma bacteriano de tipo: Cromosomal / Plasmidial
PlasmidFinder (Carattoli, A. et al. 2014)	Grupo de Incompatibilidad (IncG)
Resistance Genome Identifier (RGI) (Alcock, B. et al. 2020)	Resistencia a Antibióticos
PhageBoost (Kimmo, S. et al. 2020)	Profagos (PHAGE)
NHMMER (Wheeler, T. et Eddy, S. 2013)	Origen de Transferencia (oriT)
I-VIP (Zhang, A. et al. 2018)	Integrones (INT)

Tabla 2.- Bases de datos adicionales incorporadas al *pipeline* de anotación de *GMI-PAN*.

Base de Datos (referencia)	Determinante(s)	Fecha de descarga
VFDB (Chen, L. et al. 2004)	Factores de Virulencia	10/11/2021
BactMet (Pal, C. et al. 2013)	Resistencia a Metales	26/01/2022
ISFinder (Siguier, P. et al. 2006)	Secuencias de Inserción	18/11/2021
oriTDB (Li, X. et al. 2018)	Origen de Transferencia, Relaxasa, proteína de acoplamiento de tipo IV (T4CP) y proteínas auxiliares	07/12/2021

GMI-PAN comienza la anotación con *PROKKA* para así establecer un primer listado de características basado principalmente en *CDSs* y luego agrega o refina anotaciones gracias a la incorporación de otros módulos con bases de datos más especializadas.

Como resultado, *GMI-PAN* genera 2 directorios con un prefijo en común: Uno llamado *"c_<CONTIGNAME>_raw_results"* en donde están almacenados los datos brutos que salieron de cada módulo incluido en este *pipeline*, esto con el propósito de mantener un control positivo que contenga datos de anotación originales de *pipelines* ya validados, para así dar la posibilidad de someterlo a un ejercicio de comparación con los datos que obtengamos después de ser procesados. El otro se llama *"c_<CONTIGNAME>_final_results"*, este directorio contiene archivos de texto separado por tabuladores (`.tsv`), que contienen una lista estandarizada de los resultados de cada módulo de anotación.

El estándar se define con 5 columnas que aluden al: Identificador (*ID*), `Comienzo` (*START*), `Final` (*END*), `Hebra` (*STRAND*) y `Notas` (*NOTES*) respectivamente. Esta información cumple con el mínimo necesario para poder construir los formatos, *Genbank* (`.gbk`) y *General Feature Format* (`.gff`), de cada *contig* analizado con la biblioteca para *Python*, *Biopython*, y paquetes asociados como *Bio.Seq* (https://github.com/biopython/biopython).

Gracias a este último, es posible automatizar un proceso de iteración sobre cada determinante bacteriano anotado en el archivo *"c_<CONTIGNAME>_final_results"*, los cuales quedan registrados con el protocolo *Seq.Feature* (https://biopython.org/docs/1.76/api/Bio.SeqFeature.html). Posteriormente, cada *Seq.Feature* que contiene cada anotación se guarda en un *Seq.Record* (https://biopython.org/docs/1.76/api/Bio.SeqRecord.html), el cual adicionalmente posee información sobre el *contig* analizado como el tipo de molécula, topología, código de acceso, organismo, etc. Con *Seq.Record*, finalmente se puede transformar esta variable en un archivo de

tipo. gbk o `.gff` gracias a los argumentos adicionales que provee *Biopython*. En la Tabla 3 y Figura A1 (anexo) se describe el output y el árbol de directorios respectivamente.

Tabla 3.- Output generado por *GMI-PAN*.

OUTPUT	Descripción
c_<CONTIGNAME>_raw_results	Directorio con resultados brutos del pipeline de anotación
c_< CONTIGNAME>_final_results	Directorio con resultados estandarizados del pipeline de anotación
Contannotate_<CONTIGNAME>.gbk	Archivo GenBank, además de una lista de determinantes anotados posee una ficha estándar con datos de identificación y referencia sobre el contig analizado
Contannotate_<CONTIGNAME>.gff	Archivo *General Feature Format* con un listado estándar de todos los determinantes identificados en los diversos módulos

4.2. Uso y evaluación de GMI-PAN sobre un set de genomas de *Klebsiella pneumoniae* resistentes a carbapenémicos aislados en Chile.

GMI-PAN fue diseñado para ser preciso y rápido. Gracias a esto se comprimió un ejercicio de anotación de plásmidos bacterianos que normalmente tomaba semanas, a minutos por plásmido. Para validar el correcto funcionamiento y utilidad de esta herramienta, esta fue empleada para anotar detalladamente y comparar los plásmidos presentes en 10 cepas de *K. pneumoniae* resistentes a carbapenémicos aisladas por el ISP en Chile, cuyo genoma fue secuenciado previamente por nuestro grupo de investigación. Para ello, se ejecutó el *pipeline*, de forma completa, sobre 43 plásmidos previamente ensamblados, provenientes de dichas cepas. Para cada *contig* se obtuvieron dos directorios, uno que almacena la información cruda obtenida de cada módulo y otro que guarda los datos filtrados y tabulados. El procesamiento de cada archivo crudo consiste en la identificación de patrones, manipulación y organización de texto plano gracias a los comandos y argumentos incluidos en *Python.* Adicionalmente, se obtuvieron

dos archivos para cada *contig*, los cuales contienen el listado completo de todos los determinantes que se lograron identificar respectivamente, tanto en formato `.gbk` como `.gff`.

4.3. Datos estadísticos sobre los resultados de *GMI-PAN*

Para llevar un registro cuantitativo de los determinantes anotados para cada *contig*, se diseñó un *pipeline* que permite iterar para cada línea de texto del respectivo archivo `.gff` y buscar palabras claves que permitan reconocer y diferenciar un determinante de otro. En la Tabla A1 adjunta en el anexo se detalla la cantidad de determinantes específicos que se logró anotar para cada *contig* analizado. De un total de 4.744 determinantes anotados para todos los *contigs* analizados, 2.083 correspondieron a HP (44%), en consecuencia, 56% de los genes predichos lograron ser anotados. Adicionalmente, un 1% de los determinantes está asociado a RM como el cobre y cerca de un 7% corresponden a genes que tienen un rol en RAM, dentro del cual cerca de un 2% fueron asociados a betalactamasas. Cabe destacar que alrededor de un 12% de los determinantes anotados en total corresponden a genes relacionados a EGM, estos incluyen integrones, transposones, secuencias de inserción y posibles profagos.

En la Figura 2 se muestra una aproximación gráfica del *pipeline* de anotación de *GMI-PAN*, mostrando como ejemplo los resultados obtenidos para el *contig* pVA126-68. Primero, *GMI-PAN Contannotate* identifica los genes que discriminan al *contig* como un plásmido, en este caso el grupo de incompatibilidad *IncX3* y el origen de transferencia *R6K* mediado por el módulo de anotación de *PlasmidFinder* y la base de datos de *oriTDB*, respectivamente. Luego, la presencia de genes asociados a resistencia a antibióticos mediado por el módulo de *RGI-CARD* y *PROKKA*. Cabe destacar que el código de *Contannotate* utiliza *PROKKA* para realizar una anotación basal mientras que otros módulos como *RGI* tienen un enfoque más inclinado a pulir anotaciones. Como resultado de este análisis destacan las betalactamasas, específicamente *NDM* y *OXA*, más una batería adicional de genes de RA como

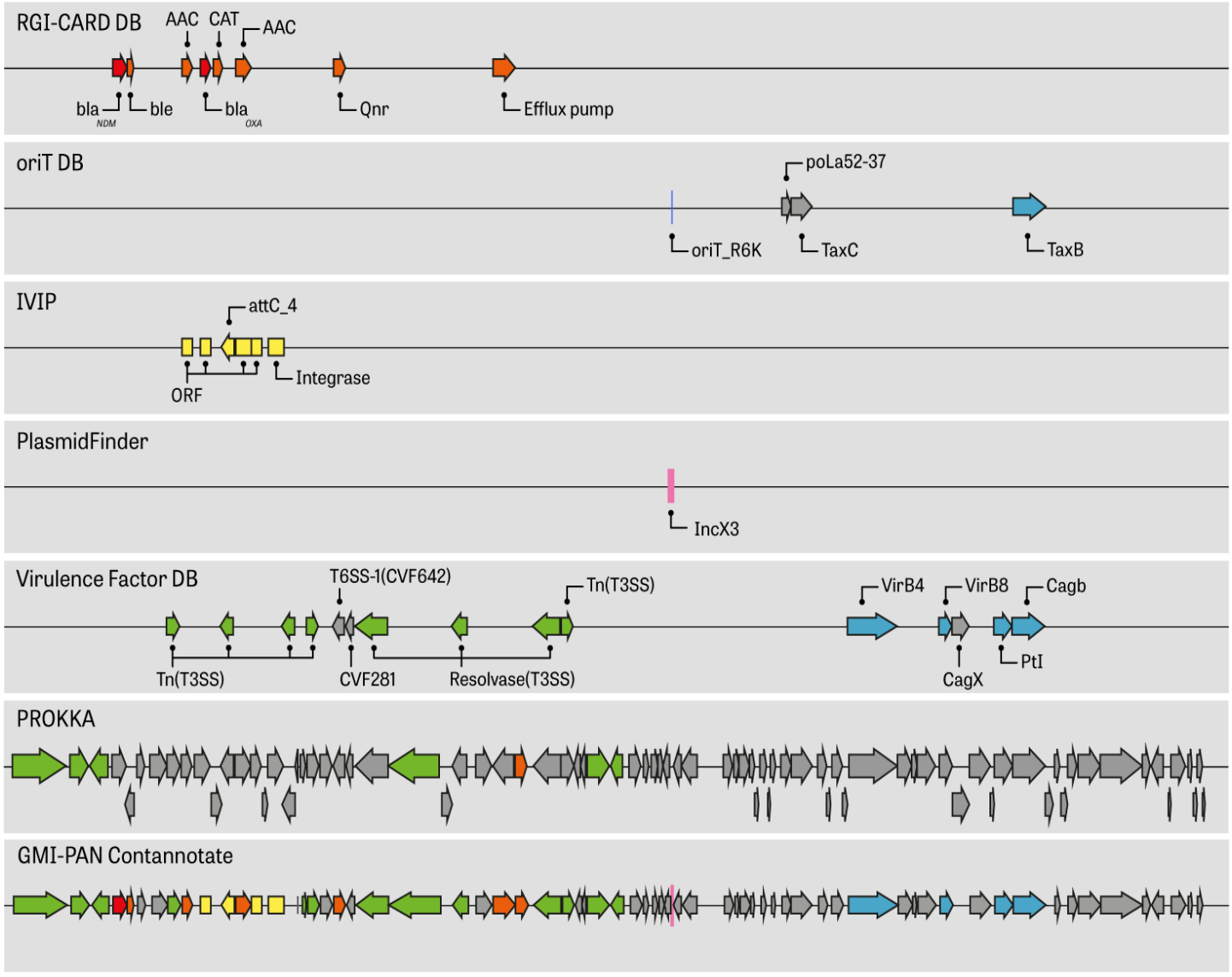

Figura 2.- Diagrama modular del *pipeline* de anotación de *GMI-PAN*. Cada determinante rescatado de los respectivos módulos o bases de datos, según corresponda, es condensado en el output gráfico final, llamado *GMI-PAN Contannotate*. Azul: origen de transferencia, Celeste: sistema de secreción tipo IV, Rojo: betalactamasas, Naranjo: resistencia a antibióticos, Verde: secuencias de inserción y transposones, Amarillo: Integrones, Rosa: grupo de incompatibilidad y Gris: Otros *CDSs*.

bleomicina, cloranfenicol, aminoglucósidos y quinolonas. También se encontraron genes que codifican proteínas con funciones asociadas a la maquinaria del T4SS tales como *VirB*, que tienen relación con proteínas de acople del aparato conjugativo. Por último, el software también logró predecir elementos genéticos móviles tales como INT, donde se identifican los ORFs, integrasa y sitios *attC*, además de IS, Tn y resolvasas, que permiten ejecutar la catálisis de un replicón cointegrado para obtener los productos finales de la transposición. Finalmente, *GMI-PAN* logra concatenar los resultados de cada módulo y entregar una sola predicción que incluya un panorama robusto del resistoma y mobiloma del *contig*.

4.4. Plásmidos asociados a carbapenemasas y otras resistencias

Del grupo de 43 *contigs* anotados, 23 fueron predichos como conjugativos (oriT + T4CP), 17 movilizables (oriT) y 3 carecían de determinantes conjugativos. Adicionalmente, se identificaron sobre 10 IncG, donde predominaban *IncR, IncX3, IncFIB(K)* y *IncFII(K)*.

Según lo expuesto en la Tabla A1, fueron anotados 81 genes de betalactamasas (*bla*) entre los *contigs* analizados. De estos, se lograron encontrar 24 plásmidos que codifican betalactamasas, de los cuales 9 son plásmidos codificantes para carbapenemasas tales como *KPC* y *NDM*. Un alineamiento de nucleótidos con *blastn* reveló que pVA04-46, pVA126-68, pVA591-46 y pVA684-49 se encuentran asociados a carbapenemasas de tipo *NDM-7* con un 100% de identidad, pVA172-90, pVA32-58, pVA681-58 y pVA833-92 están asociados a *KPC-2* con un 100% de identidad y pVA833-165 asociado a *NDM-1* con un 100% de identidad.

También hay que tener en cuenta la existencia de betalactamasas *OXA* en los plásmidos pVA681-191, pVA564-179, pVA32-200, pVA126-68 y pVA04-196 para *OXA-1*, pVA04-58, pVA126-42, pVA564-37, pVA569-35 y pVA833-176 para *OXA-10* y pVA04-58 para *OXA-11*. Sin embargo, estas no entran a la categoría de carbapenemasas como *KPC* y *NDM* dado que estas variantes no tienen capacidad de hidrolizar carbapenémicos, con excepción de las variantes *OXA*-

23 y *OXA-48* que no se encontraron en este set de plásmidos (Sawa et al., 2020).

También se encontraron otros plásmidos relevantes, tales como pVA04-196 y pVA172-90 que destacan por contener tres betalactamasas distintas, en el primero dos de espectro extendido (*ESBLs*) *CTX-M-15*, *TEM-1* y 1 de tipo *OXA-1* además de resistencias a metales como el cobre (*CopX*) y el arsénico (a*rsX*), mientras que el segundo posee dos *ESBLs*, *CTX-M-2* y *TEM-12*, y una carbapenemasa *KPC-2*. Además, una batería de genes asociados a proteínas del T4SS que podrían, potencialmente, fomentar la transferencia horizontal de múltiples genes de resistencias. Otro ejemplo es el caso de pVA833-92, en donde se presentan betalactamasas *KPC-2*, *TEM-1* y *CTX-M-3* más una batería de genes de resistencia a trimetoprima, aminoglucósidos, eritromicina y macrólidos. O también el caso de pVA591-160, que posee una betalactamasa de espectro extendido *SHV-1* y adicionalmente factores de virulencia tales como *heme uptake system e iron dicitrate transporter*, los cuales pueden ser transferibles horizontalmente y actuar como quelantes de hierro, compuesto esencial en funciones asociadas a virulencia en bacterias (Palmer & Skaar, 2016).

Cabe destacar que existen plásmidos que también codifican para resistencias a antibióticos y metales adicionales, entre estos están las resistencias a tetraciclina, sulfonamida, fosfomicina, rifamicina y fluoroquinolonas. Por otro lado, entre los metales también se encuentran las resistencias a mercurio, cobalto, magnesio, plata y telurio, donde este último es un metal raro en el planeta y que se usa principalmente en aleaciones de cobre y acero, siendo un elemento crítico para el avance en la industria tecnológica (Schuyler, 2022).

En la Figura 3 se presentan las distintas arquitecturas que pueden adoptar algunos de los plásmidos resueltos por GMI-PAN.

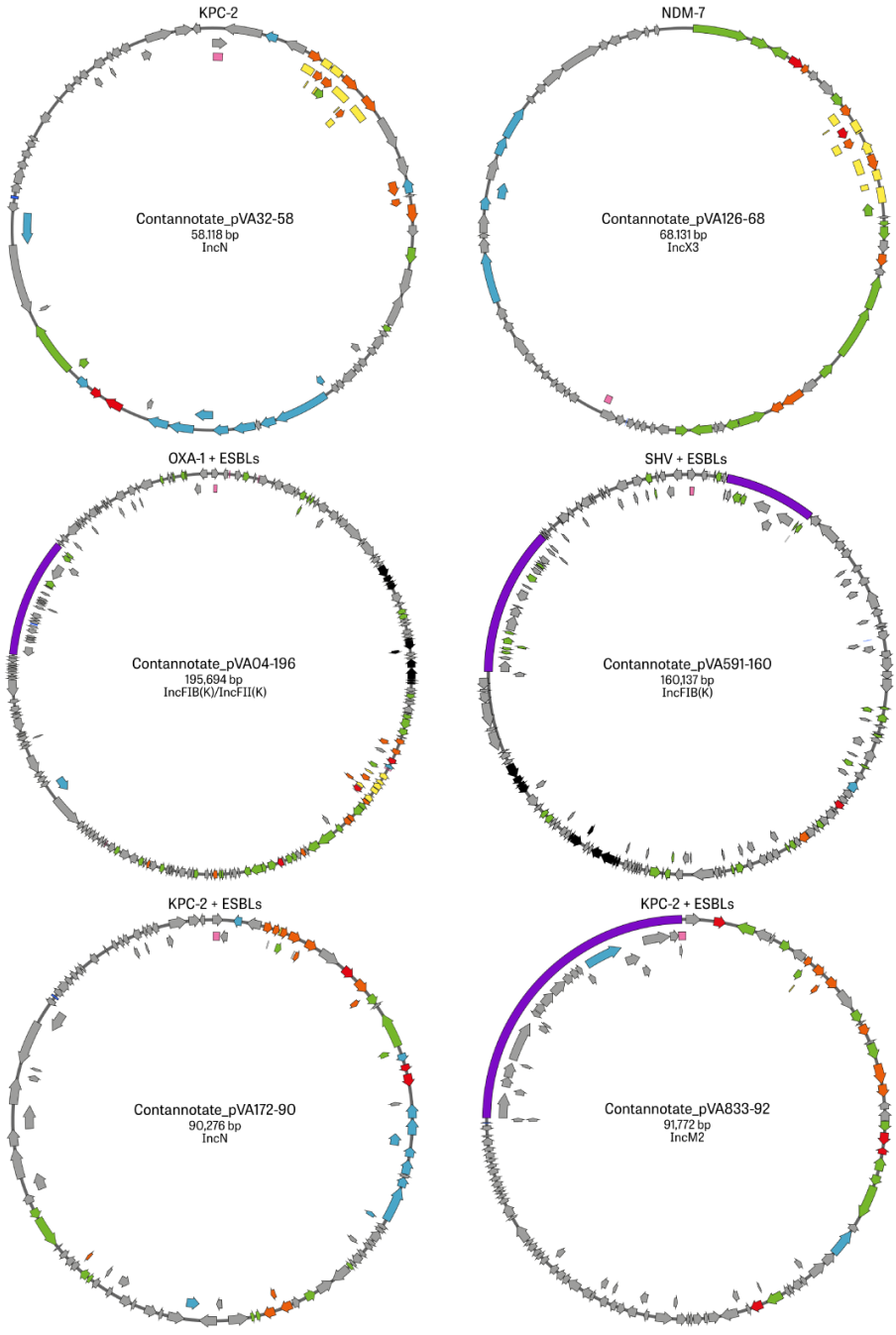

Figura 3.- Plásmidos asociados a genes de carbapenemasas y betalactamasas *ESBLs*.

Plásmidos asociados a betalactamasas tales como *NDM*, *KPC*, *OXA* y *ESBLs*, entre otras funciones. Rojo: Betalactamasas, Naranjo: Resistencia a Antibióticos, Amarillo: INT, Verde: Secuencias de IS y Tn, Celeste: T4SS o variedad, Azul: oriT, Rosa: IncG, Negro: Resistencia a Metales, Morado: PHAGE y Gris: *CDSs* u otros.

4.5. Comparación entre anotación automatizada y manual

Respecto a los resultados de anotación obtenidos por iteración manual, que se encuentran adjuntos en la Tabla A2, de un total de 5.316 determinantes anotados para todos los *contigs* analizados, 2.049 corresponden a HP lo que equivale aproximadamente a un 39%, en consecuencia, 61% de los genes predichos lograron ser anotados. Adicionalmente, un 2% de los determinantes está asociado a RM como el cobre y cerca de un 6% corresponden a genes que tienen un rol en RA, dentro del cual aproximadamente un 1.5% tienen relación con betalactamasas.

¿Cómo se comparan este set de resultados del *pipeline* de anotación manual respecto a los resultados del *pipeline* automatizado? En base a la cuantificación de *CDSs* respecto a cada plásmido (Tabla A1 y Tabla A2), *GMI-PAN* logró una mejora significativa en la anotación. Esta mejora fue calculada de acuerdo a la tasa de crecimiento porcentual entre el número de determinantes anotados de manera automatizada y manual, mediante la siguiente fórmula:

$$\nabla\ \% = \frac{X_{(Automatizado)} - Y_{(Manual)}}{Y_{(Manual)}} \times 100$$

Dado que el enfoque de este trabajo reside principalmente en betalactamasas, la cuantificación de *CDSs* asociados a resistencias a antibióticos se dividió en dos, donde RA representa las resistencias a antibióticos en general, excluyendo betalactamasas, y BL a los determinantes asociados a betalactamasas específicamente. Un criterio similar se aplicó a variables como sistemas de secreción y resistencia a metales.

Dentro de las mejoras obtenidas, se puede observar RA con un 5.96%, BLs en un 2.53%, SS (sin incluir el IV) con un 44.23%, Tn y IS un 124.40% y RM (sin incluir cobre) en un 70.45% adicional. Por otro lado, el *pipeline* de anotación manual destacó en la anotación de INT con un

43.85%, T4SS en un 30.77%, resistencia a cobre un 46.67%, 1.66% en HP y un 10.76% adicional de determinantes anotados. Si bien la anotación de profagos del *pipeline* automatizado resultó en una pérdida de rendimiento del 63.24% en comparación al manual, al momento de comparar los determinantes obtenidos. Hay que tener en cuenta que en el módulo *PhageBoost* contenido en el *pipeline* de *GMI-PAN,* cada determinante anotado consiste en la anotación del profago como un solo elemento inserto en el genoma hospedero sin distinguir los elementos que lo conforman. A diferencia de la herramienta *PHASTER,* contenida en el *pipeline* manual, la cual anota de acuerdo a los elementos que conforman el profago en su completitud.

En el preprint asociado a este seminario de título (Veloso et al., 2022) se detalla la utilización de una base de 44 *contigs* anotados manualmente que, a diferencia de la utilizada en este trabajo, ejecutó un ejercicio de reensamblaje en donde se refinaron algunos *contigs*. Se descartaron los contigs pVA126-339 y pVA591-7, y se sumaron los contigs pVA684-37, pVA833-45 y pVA833-56 a la lista de plásmidos encontrados en las 10 cepas chilenas. Esta mejora en la resolución del ensamblaje permitió la identificación adicional de *KPC* en la cepa pVA684, algo que no era posible en la base que se utilizó para este trabajo dado que fue analizada de manera previa al refinamiento del ensamblaje.

No obstante, en la parte de post-refinamiento de dicho preprint también se usó paralelamente *GMI-PAN* para los *contigs* adicionales, donde igualmente se logró identificar *KPC* en pVA684, por lo que, a pesar de haber anotado 572 determinantes totales menos (entre todos los plásmidos) en comparación al análisis manual, este no perdió precisión en cuanto a la identificación de genes putativos asociados a carbapenemasas y betalactamasas de espectro extendido. Por otro lado, la anotación automatizada de INT, proteínas del T4SS y proteínas de resistencia a cobre se vio mermada, a diferencia de la anotación manual, disminuyendo la resolución de algunos casetes o sistemas, principalmente por una limitación en el diseño del

programa dado que fue instalado en un solo ambiente con un conjunto de dependencias específicos, lo que en consecuencia fuerza a prescindir de módulos más fiables que presenten algún grado de incompatibilidad. Por ejemplo, en el caso específico de pVA591-160 donde *GMI-PAN* no logra identificar genes asociados a INT a diferencia del análisis manual del grupo.

4.6. Análisis comparativo de las proteínas codificadas en los plásmidos de las diez cepas de *K. pneumoniae* aisladas por el ISP

Para realizar un análisis funcional comparativo entre los plásmidos de las cepas del ISP se optó por una estrategia de clusterización de *CDSs* utilizando *mmseq*, lo que se resumió en una matriz binaria que indica la presencia o ausencia del clúster de *CDSs* para cada plásmido (Figura 4). Lo primero que resalta es la existencia de grupos de plásmidos codificantes de betalactamasas definidos. Un primer grupo incluye a pVA126-175, pVA564-179, pVA04-196, pVA681-191 y pVA32-200, los cuales están asociados a betalactamasas *CTX-M*. Además, estos plásmidos comparten otros conjuntos de *CDSs* asociados a resistencias a otros antibióticos y a metales. Se encontraron *CDSs* asociados a resistencia a arsénico como *arsA* y *arsB* (Rosen, 1990), que codifican proteínas de transmembrana, y a*rsD* (Li et al., 2002) y a*rsR* (Xu et al., 1997) que actúan a través de represión mediada por regulación trans del operón *ars*. Otro conjunto son los *CDSs* asociados a resistencia a cobre, por ejemplo, los genes del operón *copABCD*, que incluye proteínas periplasmáticas y de membrana que participan en el secuestro y compartimentalización del cobre (Cooksey, 1994). Adicionalmente, también se observa variabilidad en el conjunto de resistencias a antibióticos donde se aprecia *CDSs* asociados a resistencia a quinolonas, sulfonamidas, estreptomicina, aminoglucósidos, cloranfenicol, trimetoprima y macrólidos.

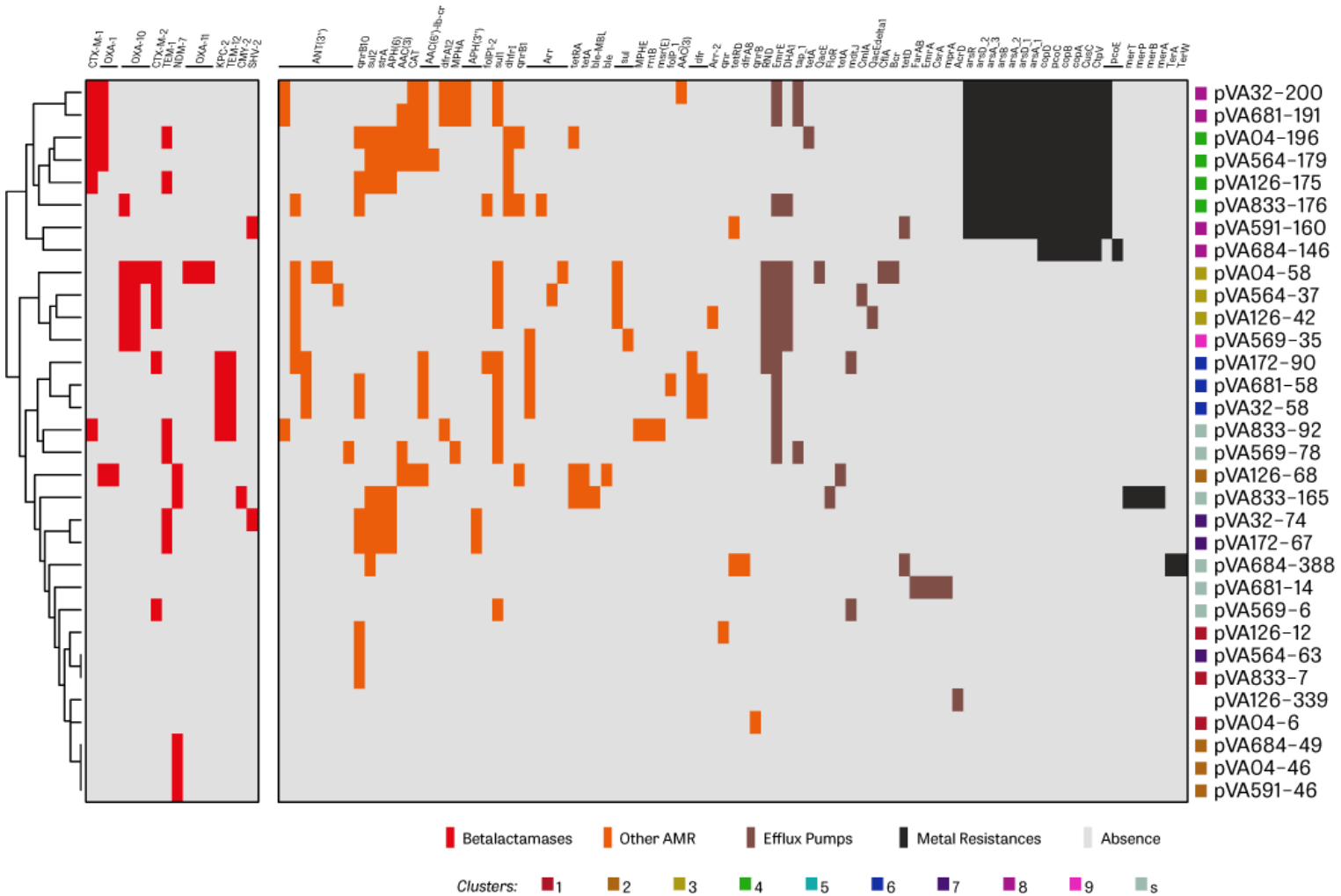

Figura 4.- Análisis comparativo de las proteínas codificadas en plásmidos de 10 cepas de *K. pneumoniae* resistentes a carbapenémicos aisladas por el ISP. Predicción bioinformática de *CDSs* que confieren resistencia a diferentes biocidas. La escala de colores del gráfico de presencia y ausencia representa el tipo de biocida y el cluster al cual pertenece el *contig* según corresponda.

Otro grupo son los plásmidos pVA569-35, pVA126-42, pVA564-37 y pVA04-58, los cuales están asociados a betalactamasas de tipo *OXA* y que comparten conjuntos de *CDSs* asociados a bombas de flujo como *EmrE* y *DHA1* que median el flujo e intercambio de un rango amplio de sustratos citotóxicos mediante cambios de gradiente quimiosmóticos, estos corresponden a una superfamilia de proteínas de transporte de membrana (*MFS*) (Yelin et al., 1999; Kim et al., 2021). Adicionalmente, también se observan bombas de tipo *RND*, las cuales son complejos de proteínas tripartitas que conectan (en el caso de gramnegativas) membrana interna con membrana externa para la expulsión de biocidas al espacio extracelular (Kim et al., 2021). Por último, también se aprecia otro conjunto asociado a RA como sulfonamidas, aminoglucósidos y rifamicina.

Por otro lado, el grupo de plásmidos pVA833-92, pVA32-58, pVA681-58 y pVA172-90 se encuentra asociado a betalactamasas *KPC* y comparte conjuntos de *CDSs* asociadas a antibióticos como aminoglucósidos, sulfonamidas, rifamicina, trimetoprima y macrólidos, además de incluir un conjunto asociado a bomba de flujo mediado por *EmrE*.

Finalmente, el grupo de plásmidos pVA126-68 y pVA833-165 se encuentran asociados a betalactamasas de tipo *NDM*, que contiene resistencias a antibióticos asociadas a tetraciclina, bleomicina, aminoglucósidos, cloranfenicol, sulfonamidas, quinolonas y streptomicina. Cabe destacar que en este grupo existe la presencia de un conjunto de genes asociados a resistencia a mercurio mediado por genes del operón *mer* (Osborn et al., 1997). No obstante, también existen plásmidos que comparten sólo un conjunto de *CDSs*, tales como pVA684-49, pVA04-46 y pVA591-49 que codifican sólo para la carbapenemasa *NDM-7*. O también pVA833-176, pVA591-160 y pVA684-146 que codifican para el operón resistencia a cobre *cop*.

En cuanto a la clusterización de plásmidos según sus genes de resistencia que está representada por el dendograma en la Figura 4, se puede apreciar que se condice con la clusterización colorimétrica realizada previamente por el grupo, en el trabajo que derivó de este seminario de título (Veloso et al., 2022), donde usaron el submódulo *Mob-cluster* del pipeline *Mob-suite*, el cual agrupa mediante distancias de *mash* y otras técnicas para clasificar plásmidos según sus distancias genómicas (Ondov et al., 2016).

5. DISCUSIÓN

5.1. *GMI-PAN* como herramienta de anotación bioinformática

Avances en tecnologías de secuenciación nos han permitido evaluar bacterias y microbiomas a una profundidad sin precedentes gracias a la genómica y metagenómica, lo cual permite a investigadores analizar grandes cantidades de datos a través de la bioprospección mediada por herramientas bioinformáticas (Roumpeka et al., 2017). Este tipo de aproximaciones también es muy útil para el monitoreo y exploración de patógenos (Gardy et al., 2017), como es el caso de *CR-Kp*, donde es necesario tener un entendimiento amplio de la diversidad y distribución de determinantes patogénicos o antimicrobianos relevantes en el ámbito clínico (Wyres et al., 2020).

Nuestro grupo de laboratorio ha provisto de un conocimiento base para el entendimiento de los mecanismos de resistencia y potencial patogénico de *Kp* en Chile a través del análisis de las primeras diez secuencias genómicas completas de *CR-Kp* chilenas (Veloso et al., 2022). Sin embargo, para poder seguir de cerca estos fenómenos epidemiológicos, es necesario conseguir resultados en el menor tiempo posible. Este trabajo presenta un software *open-source* para la anotación rápida y detallada de determinantes asociados a RAM y EGM en plásmidos.

GMI-PAN es un *pipeline* de anotación simple y completo, dado que es capaz de reducir un ejercicio de anotación del mobiloma y resistoma bacteriano desde semanas a solo minutos para cada plásmido, requiriendo únicamente un archivo. fasta como *input*. Por otro lado, también es una herramienta flexible, ya que utiliza líneas de código abierto que son editables para la consideración de otras bases de datos que sean de interés del usuario, lo que se traduce en un mayor repertorio de géneros, especies y tipos de determinantes bacterianos a poder anotar. Este *pipeline* puede ser instalado y ejecutado tanto en un computador de escritorio como en uno portátil.

5.2. Herramientas bioinformáticas y su espacio en la ciencia

Diseñar una herramienta bioinformática no es algo extraño, ya que el fenómeno de *open-source software* es algo que se remonta ya varios años atrás. El hecho de contar con una historia y comunidad alrededor de repositorios *open-source* en plataformas como *Github* (Wilson et al., 2017), nos permite contar con "bloques de construcción" para insertarlos dentro de *pipelines* o reutilizarlos para otro tipo de aplicaciones (Stajich et al., 2006).

Para el caso de *GMI-PAN* (y en general), no basta sólo con ejecutar cada bloque. Para desarrollar este *pipeline* se debe generar comunicación entre módulos para la sincronización y estandarización del *input/output* de cada bloque integrado al *pipeline*, lo que en consecuencia conlleva a la obtención de información cohesiva y válida. Si no se establece comunicación, los datos obtenidos pueden ser comprometidos y, en consecuencia, no son confiables para el objetivo que uno como investigador establece. Es por este motivo, que resulta importante entender el lenguaje y nomenclatura únicos utilizado por cada bloque dentro del *pipeline.* Resolver estos patrones habilita al desarrollador a diseñar subprocesos que adoptan la función de "traductor" y así generar comunicación entre bloques.

De acuerdo con lo anterior, *GMI-PAN* actúa como un Adaptador, o mejor conocido como *Wrapper* (Gamma et al., 1994), en donde se toman interfaces como *PROKKA*, *PlasmidVerify*, *PlasmidFinder*, *RGI*, *PhageBoost*, *NHMMER* y *IVIP* para que trabajen en conjunto, lo que no podría realizarse de otra manera, ya que son incompatibles entre sí. Gracias a esto, se logró ejecutar y completar la anotación detallada de 43 plásmidos de cepas de *K. pneumoniae* resistentes a carbapenémicos aisladas en nuestro país, permitiendo realizar un análisis comparativo de CDSs. Para ello, se ejecutaron los distintos módulos y bases de anotación incluidos, obteniendo un perfil estadístico de la naturaleza del *contig* sumado al IncG, y oriT. Además, una medida de la probabilidad de ser un plásmido, y la búsqueda y clasificación de

genes de interés tanto de tipo resistencia a biocidas como de movilización y conjugación. Si bien, un usuario puede proveer al pipeline de GMI-PAN de un archivo fasta asociado a un cromosoma bacteriano, no es recomendable hacerlo, aun cuando la herramienta permita anotarlo, dado que la calidad y variedad de los determinantes anotados no va a ser la misma, considerando que parte de las bases de datos y módulos están optimizados para anotar plásmidos.

Uno de los objetivos de este *wrapper* es ejecutar los módulos en un orden jerárquico, por lo tanto, para este caso *PROKKA* provee un piso de anotación basal sobre el cual otros módulos pueden agregar o refinar. En la Figura 2 se puede observar la concatenación de los distintos módulos insertos en *GMI-PAN*. En algunos ejemplos esto se puede dar de forma más "limpia", donde los determinantes anotados por bases más específicas rellenan espacios dentro del mapa que no fueron predichos o se definieron como hipotéticos por *PROKKA* en su inicio. Sin embargo, la realidad muestra que podemos tener múltiples anotaciones para un mismo segmento genómico o variaciones de éste tanto en el *START* como el *END*. Definir la prioridad de un determinante tampoco es trivial dado que, por razones que se discutirán más adelante, la fidelidad de estas predicciones no es 100% certera porque posee parámetros técnicos, objetivos y subjetivos que dependen del desarrollador de la respectiva herramienta. Es por este motivo que esta primera versión de *GMI-PAN* se inclina más por un código laxo en donde se prefiere que sobren anotaciones y no a que falten, para así darle la posibilidad al investigador de aplicar su propio criterio.

Es necesario destacar el hecho de que el diseño de programas no es un proceso estático, por ejemplo, un computador con sistema operativo *Windows* requiere de actualizaciones periódicas. Este es un fenómeno llamado "*Versioning*" (Speck et al., 2001) y que alude a la implementación de "versiones" como un modelo para expresar el deseo de la reutilización de piezas de *software* y el diseño de razonamiento, donde además se pueden ordenar para describir

conjuntos de características que se necesitan dentro del sistema.

En ese contexto, *GMI-PAN* también acoge este modelo, dado que se pueden diseñar versiones que sumen elementos/características al *pipeline* como versiones que corrigen procesos o resultados no deseados en éste. Por ejemplo, específicamente, los archivos `.gff` que se obtienen como output del *pipeline* no tienen definido la implementación de texto que haga referencia a la fecha de creación del archivo, esto se debe a que no es una característica esencial para el objetivo de este seminario de título y por ende se puede abordar en el futuro, entre otros. Si bien, *GMI-PAN* también se puede optimizar concatenando el input completo en un solo multifasta, el método resulta desventajoso ya que este programa está en una fase Beta de desarrollo y se requiere destinar tiempo adicional en el pulido e identificación adecuada de la diversidad total del potencial output de anotación.

¿Cómo se comparan los resultados de anotación mediados por *GMI-PAN* con los datos obtenidos de la prueba fenotípica del ISP para la producción de carbapenemasas? En la tabla suplementaria 2 se puede observar que los fenotipos reportados por el ISP se condicen con los resultados obtenidos por *GMI-PAN*, con excepción de la cepa VA684 donde se reporta la presencia de los genes de resistencia *KPC + NDM*. Esto podría deberse a que, tanto en el trabajo de Veloso, M et al. 2022 como en el análisis realizado en este seminario, se refinaron los ensamblajes de los genomas, logrando recuperar un contig que codifica la carbapenemasa KPC faltante de dicha cepa. De todos modos, eventualmente la no identificación de ciertos genes también se puede dar por una limitación en el programa diseñado, lo que podría tener que ver con la forma en cómo conversan los módulos dentro del *wrapper*. Por ejemplo, un módulo realiza sus tareas siempre y cuando haya una versión específica de *python* instalada, si se quiere insertar otro módulo que usa otra versión de *python* este puede generar conflictos con el primero y en consecuencia el *pipeline* no se ejecuta. Esto impide integrar módulos más complejos y por ende

puede disminuir la calidad de los análisis, lo que se traduce en una pérdida de información o, en este caso, genes. Por este motivo se requieren de más estudios a futuro para solucionar estas diferencias en versiones posteriores.

Los altos niveles de resistencia encontrados en estos aislados se correlacionaron con un amplio espectro de genes de resistencia adquiridos, algunos presentes en frecuencias bajas en otros aislados de *K. pneumoniae*, incluyendo bla_{NDM-7}, bla_{KPC-2}, bla_{SHV-31} y bla_{CMY-2}. Adicionalmente, la mayoría de los aislados también contaban con genes que confieren RM y desinfectantes, en concordancia con reportes previos que muestran su co-selección con genes de RA, ampliando las opciones para resistir antimicrobianos y persistir en ambientes clínicos, urbanos y contaminados (Li et al., 2017; Baker et al., 2006).

5.3. Desafíos en el desarrollo de *pipelines* de anotación

Si bien ya se ha mencionado que la anotación automatizada de determinantes en *contigs* por *GMI-PAN* tiene varios beneficios, hay que tener en cuenta algunos puntos de precaución. A continuación, se discuten ciertos aspectos expuestos por análisis posteriores de los datos obtenidos por *GMI-PAN*.

El objetivo de clusterizar proteínas consiste en una agrupación basada en similitudes estructurales y funcionales, de esta forma se puede identificar grupos de proteínas que tienen propiedades, funciones o relación evolutiva similares (Steinegger et al., 2017). Esto se puede abordar, por ejemplo, mediante métodos de alineamiento de secuencias, los cuales dependen de la comparación de secuencias aminoacídicas. En la Figura 4, se hizo una clusterización de genes con una identidad y cobertura del 95%, sin embargo, se pueden observar algunas columnas como las betalactamasas *OXA* u otros genes de resistencia que comparten una misma etiqueta.

¿Cómo se puede explicar, por ejemplo, la existencia de tres columnas OXA-10? En la Figura 5A se puede apreciar un ejemplo para el *contig* pVA04-58, en este caso, las predicciones

establecidas por *PROKKA* y *RGI-CARD* identifican dos genes putativos asociados a la betalactamasa *OXA*, no obstante, el tamaño de estos fragmentos bordea el límite de los 100 aminoácidos. Al comparar con literatura se puede observar que las betalactamasas *OXA* (completas) caracterizadas superan los 200 aminoácidos (Queenan et al., 2007). Por lo tanto, las distintas columnas identificadas como *OXA-10* corresponden a la proteína completa y a variantes truncas. Si bien, hay evidencia de betalactamasas truncas (Shi et al., 2022) y otros genes de resistencia (Bi et al., 2015; Wu et al., 2019), también existen otras posibilidades para explicar este fenómeno, por ejemplo, el caso del gen "*CsrA*" de la cepa pVA681-14. En este caso, una búsqueda por *NCBI* (*WP_000906486.1*) y *PROKKA* identifican el gen como un regulador de almacenamiento de carbono *CsrA*, mientras que *RGI-CARD* lo identifica como una bomba de flujo de tipo *RND*. Esto indica que también existe la posibilidad de que las bases de datos utilizadas por los módulos de *GMI-PAN* tengan errores, sea por precisión o una anotación errónea. En consecuencia, dado lo anterior y sabiendo que los ejercicios de ensamblaje y predicción de *ORFs* no son perfectos, resulta difícil ejercer un criterio de prioridad de anotación, por lo que siempre es recomendable verificar los resultados, además de realizar trabajo experimental en laboratorio para corroborar la actividad de los genes y sistemas putativos estudiados. Aún así, *GMI-PAN* es capaz de entregar gran cantidad de datos detallados en sólo minutos, y que, si bien se recomienda verificar los resultados y hacer estudios posteriores para desarrollar nuevas versiones que corrijan estos obstáculos, al final resulta ser una herramienta práctica para el seguimiento temprano de resistoma y mobiloma de cepas patogénicas y multirresistentes como *Klebsiella pneumoniae*.

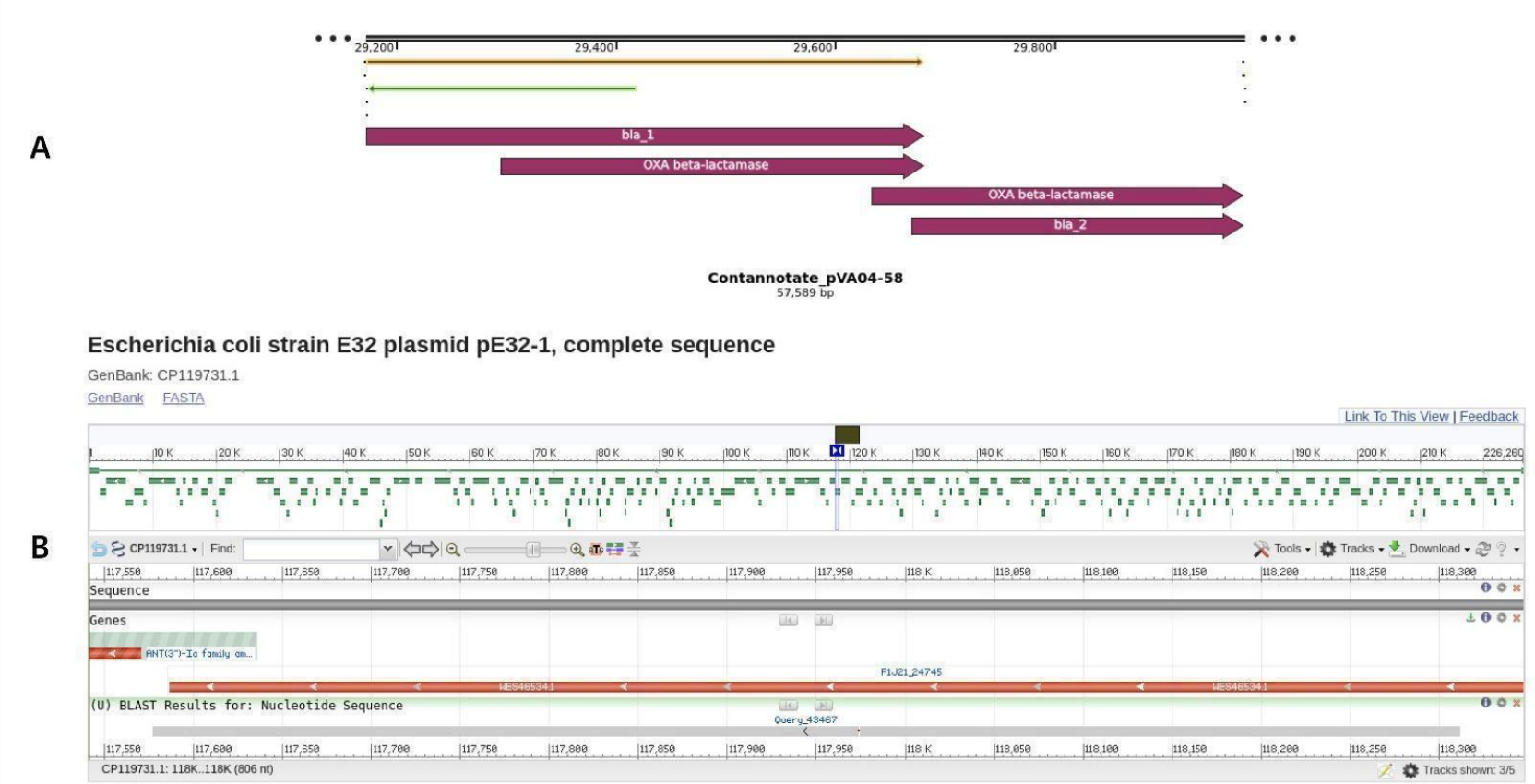

Figura 5. El problema de la clusterización de OXAs. A) Visualización de genes OXA-10 anotados por *GMI-PAN* a través de la herramienta SnapGene Viewer. B) Resultado de un alineamiento por blastn en NCBI sobre la secuencia que comprende la región completa de los CDSs truncos reportados (parámetros estándar).

6. CONCLUSIONES

- Se desarrolló el programa *"GMI-PAN"* para la anotación detallada de plásmidos en enterobacterias, incluyendo la determinación de grupo de incompatibilidad, origen de transferencia, genes de conjugación, anotación y clasificación de resistencia a antibióticos, factores de virulencia y elementos genéticos móviles.

- *GMI-PAN* automatiza un *pipeline* de anotación, basado en *Python* 3.6, integrando módulos previamente validados como *PROKKA*, *RGI*, *PlasmidVerify*, *PlasmidFinder*, *I-VIP*, *PhageBoost* y *NHMMER*.

- La herramienta *GMI-PAN* logró implementar el análisis de un conjunto de plásmidos presentes en cepas de *K. pneumoniae* resistentes a carbapenémicos aisladas en Chile. Los resultados respaldan la creación de un flujo de trabajo optimizado para la anotación de plásmidos, en el que las betalactamasas anotadas concuerdan con el análisis fenotípico realizado por el ISP y con la anotación manual llevada a cabo y revisada por nuestro grupo de investigación.

- El análisis de clusterización basado en la presencia o ausencia de *CDSs* reveló grupos de plásmidos que comparten genes de resistencia a antibióticos y biocidas, como las betalactamasas *KPC, NDM* y *OXA*. Estos grupos se condicen con el análisis de clusterización por similitud de secuencia realizado por nuestro grupo (Veloso, M. et al. 2022). Ambos tipos de clusterización brindan información complementaria: basándose en la estructura del plásmido, IncG, *CDSs* y sus regiones no codificantes como el oriT.

- *GMI-PAN* es capaz de reducir el tiempo de anotación de secuencias plasmidiales, sin sacrificar la calidad o la cantidad de la información obtenida. Sin embargo, dada la naturaleza de la herramienta, el desempeño dependerá de la fidelidad y actualización de las bases de datos.

ANEXOS

Population genomics, resistance, virulence, and mobile genetic elements in carbapenem-resistant *Klebsiella pneumoniae* causing infections in Chile

Marcelo Veloso[1], Joaquín Acosta[1], Patricio Arros[1], Camilo Berríos-Pastén[1], Roberto Rojas[1], Macarena Varas[1], Miguel L. Allende[3], Francisco P. Chávez[2], Pamela Araya[4], Juan Carlos Hormazábal[4], Rosalba Lagos[1], Andrés E. Marcoleta[1,*]

1Grupo de Microbiología Integrativa, Laboratorio de Biología Estructural y Molecular BEM, Departamento de Biología, Facultad de Ciencias, Universidad de Chile. Santiago, Chile.
2Laboratorio de Microbiología de Sistemas, Departamento de Biología, Facultad de Ciencias, Universidad de Chile. Santiago, Chile.
3Millenium Institute Center for Genome Regulation, Facultad de Ciencias, Universidad de Chile. Santiago, Chile.
4Instituto de Salud Pública. Santiago, Chile.

Supplementary Table 1. *K. pneumoniae* isolates studied in this work.

Isolate	Sex	Age	Collection date	Source	Region of origin
VA4	M	36	05-12-2018	Cerebrospinal fluid	Metropolitan
VA32	F	81	04-01-2019	Bronchoalveolar lavage	Metropolitan
VA126	M	33	16-01-2019	Blood	Metropolitan
VA172	M	71	25-01-2019	Bone tissue	VII Del Maule
VA564	M	79	02-04-2019	Blood	X De los Lagos
VA569	M	53	06-04-2019	Peritoneal fluid	Metropolitan
VA591	M	29	11-04-2019	Blood	Metropolitan
VA681	F	32	24-04-2019	Abscess	Metropolitan
VA684	F	15	25-04-2019	Catheter blood	Metropolitan
VA833	F	54	23-05-2019	Blood	Metropolitan

Supplementary Table 2. Phenotypic tests for carbapenemase production and beta-lactamase gene detection in the *K. pneumoniae* isolates studied in this work.

Isolate	Blue Carba	Boronic Acid	Triton Hodge	carbapenemase gene	ESBL
VA4	+	-	+	NDM+	+
VA32	+	+	+	KPC+	+
VA126	+	-	+	NDM+	+
VA172	+	+	+	KPC+	+
VA564	-	-	-	-	+
VA569	-	-	-	-	+
VA591	+	-	+	NDM+	+
VA681	+	+	+	KPC+	+
VA684	+	+	+	KPC+/NDM+	+
VA833	+	+	+	KPC+/NDM+	+

Figura A1.- Árbol de directorios y archivos generado por *GMI-PAN*.

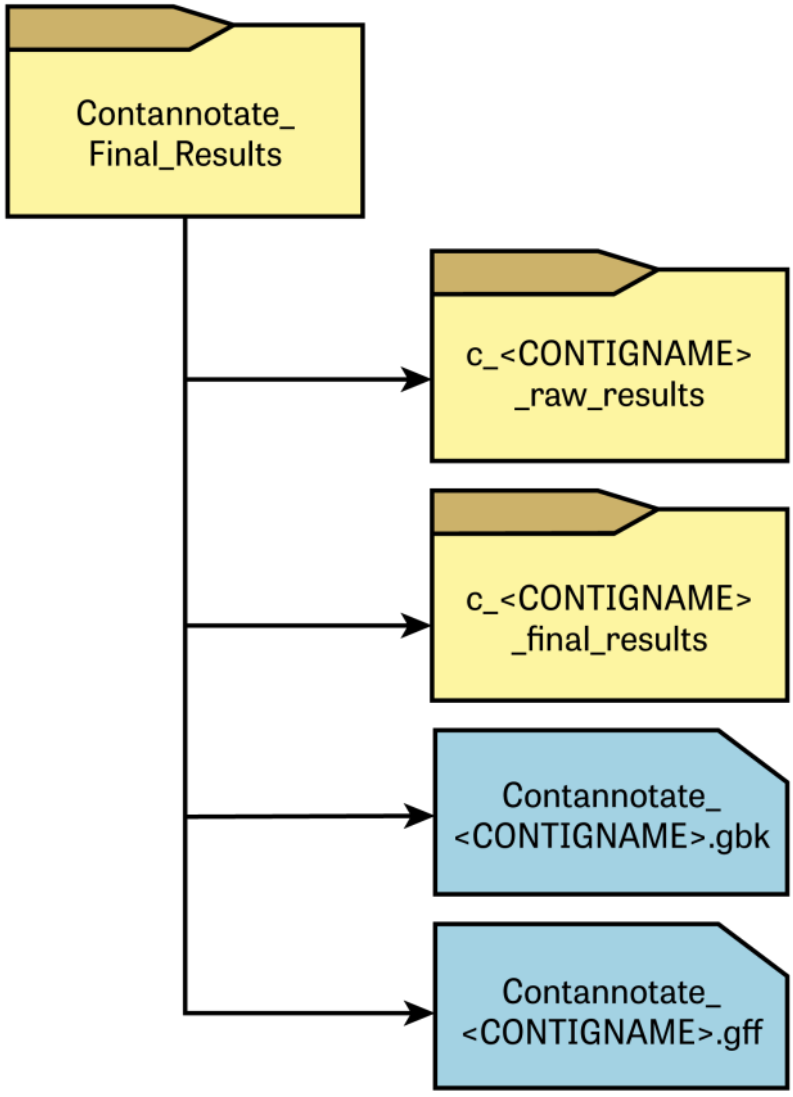

Tabla A1.- Número de determinantes anotados por *GMI-PAN* para cada plásmido, la leyenda para cada determinante incluido corresponde *Antibiotic Resistance* (AR), *B-Lactamase* (BL), *Integron* (INT), *Secretion System* (SS), *Type IV Secretion System* (T4SS), *Transposon* (Tn), *Metal Resistance* (MR), *Copper Resistance* (CR), *Phage* (Ph), *Hypothetical Protein* (HP) y *Features* (Fs).

ID (*auto*)	AR	BL	INT	SS	T4SS	Tn	MR	CR	Ph	HP	Fs
pVA04-6	1	0	0	0	0	0	0	0	0	3	7
pVA04-46	0	1	0	5	3	6	1	0	0	34	65
pVA04-58	21	8	4	1	0	14	0	0	0	35	104
pVA04-196	17	5	9	0	0	36	8	7	1	102	258
pVA126-12	2	0	0	0	0	0	0	0	0	8	18
pVA126-42	12	4	2	1	0	10	0	0	0	18	63
pVA126-68	8	4	10	5	3	13	1	0	0	36	98
pVA126-175	11	4	0	0	1	32	8	7	1	102	231
pVA126-339	1	0	0	0	3	1	2	0	4	175	369
pVA172-67	5	2	0	4	3	7	0	0	1	52	102
pVA172-90	16	5	3	10	1	12	0	0	0	51	128
pVA32-7	0	0	0	0	0	0	0	0	0	5	28
pVA32-58	11	3	9	8	1	5	0	0	0	33	90
pVA32-74	6	4	0	4	3	8	0	0	0	53	108
pVA32-104	0	0	0	0	2	14	0	0	0	69	136
pVA32-200	19	3	7	0	0	41	8	7	1	93	260
pVA564-33	0	0	0	4	3	1	2	0	0	33	49
pVA564-37	11	4	2	1	0	10	0	0	0	14	52
pVA564-63	1	0	0	3	2	5	0	0	0	51	90
pVA564-179	13	4	2	0	0	33	8	7	2	97	234
pVA569-6	3	2	0	0	0	0	0	0	0	1	7
pVA569-7	0	0	0	0	0	0	0	0	0	4	27
pVA569-35	8	2	2	1	0	9	0	0	0	16	52
pVA569-78	7	2	2	0	0	19	0	0	2	45	107
pVA591-7	0	0	0	0	0	1	0	0	0	3	18
pVA591-9	0	0	0	0	0	1	0	0	0	2	32
pVA591-46	0	1	0	5	3	6	1	0	0	34	65
pVA591-160	6	2	0	0	0	30	8	7	2	67	187
pVA681-5	0	0	0	0	0	1	0	0	0	4	25
pVA681-14	5	2	0	0	0	2	0	0	0	2	24
pVA681-36	0	0	0	5	2	2	1	0	0	38	59
pVA681-58	10	3	3	8	1	5	0	0	1	34	89
pVA681-191	17	3	13	0	0	38	8	7	1	86	241
pVA684-2	0	0	0	0	0	0	0	0	0	2	5
pVA684-6	0	0	0	0	0	2	0	0	0	5	25
pVA684-8	0	0	0	0	0	2	0	0	0	6	44
pVA684-49	0	1	0	5	3	5	1	0	0	35	65
pVA684-146	0	0	0	1	0	3	1	7	1	74	166
pVA684-388	3	0	0	2	0	44	3	0	4	254	426
pVA833-7	1	0	0	0	0	0	0	0	0	3	8
pVA833-92	11	7	3	2	2	10	0	0	1	69	141
pVA833-165	8	3	0	0	0	14	6	0	2	139	212
pVA833-176	15	2	2	0	0	27	8	7	1	96	229
Total	249	81	73	75	36	469	75	56	25	2083	4744

Tabla A2.- Número de determinantes anotados por análisis manual de cada plásmido, la leyenda para cada determinante incluido corresponde *Antibiotic Resistance* (AR), *B-Lactamase* (BL), *Integron* (INT), *Secretion System* (SS), *Type IV Secretion System* (T4SS), *Transposon* (Tn), *Metal Resistance* (MR), *Copper Resistance* (CR), *Phage* (Ph), *Hypothetical Protein* (HP) y *Features* (Fs).

ID (*manual*)	AR	BL	INT	SS	T4SS	Tn	MR	CR	Ph	HP	Fs
pVA04-6	0	0	0	0	0	0	0	0	0	4	9
pVA04-46	0	1	0	6	6	4	0	0	0	34	70
pVA04-58	13	5	14	0	0	4	0	0	1	36	128
pVA04-196	18	6	9	0	0	16	4	13	2	98	310
pVA126-12	1	0	0	0	0	0	0	0	0	9	20
pVA126-42	10	3	6	0	0	5	0	0	2	18	75
pVA126-68	8	3	5	6	6	9	0	0	2	39	118
pVA126-175	12	4	4	0	0	11	4	13	1	96	275
pVA126-339	2	0	0	0	0	2	0	1	36	107	369
pVA172-67	4	2	0	0	0	3	0	0	1	44	101
pVA172-90	16	6	9	6	5	6	0	0	2	71	165
pVA32-7	0	0	0	0	0	0	0	0	0	9	13
pVA32-58	12	3	9	5	5	2	0	0	0	45	108
pVA32-74	5	4	0	0	0	4	0	0	1	44	110
pVA32-104	0	0	0	0	0	9	0	0	2	60	151
pVA32-200	17	4	12	0	0	16	4	13	0	93	304
pVA564-33	0	0	0	6	6	0	0	0	0	37	55
pVA564-37	8	3	6	0	0	4	0	0	1	17	72
pVA564-63	1	0	0	0	0	4	0	0	0	44	88
pVA564-179	13	4	7	0	0	13	4	13	1	98	278
pVA569-6	4	2	0	0	0	0	0	0	0	3	10
pVA569-7	0	0	0	0	0	0	0	0	0	7	10
pVA569-35	5	1	6	0	0	5	0	0	2	14	65
pVA569-78	5	2	6	0	0	7	0	0	0	50	145
pVA591-7	0	0	0	0	0	1	0	0	0	6	12
pVA591-9	0	0	0	0	0	1	0	0	0	7	13
pVA591-46	0	1	0	6	6	4	0	0	0	33	73
pVA591-160	8	2	0	0	0	12	4	13	1	76	227
pVA681-5	0	0	0	0	0	1	0	0	0	5	8
pVA681-14	7	2	0	0	0	1	0	0	0	0	21
pVA681-36	0	0	0	6	6	1	0	0	0	43	65
pVA681-58	9	3	8	5	5	1	0	0	0	48	106
pVA681-191	16	4	12	0	0	15	4	13	0	85	267
pVA684-2	0	0	0	0	0	0	0	0	0	1	1
pVA684-6	0	0	0	0	0	1	0	0	0	8	15
pVA684-8	0	0	0	0	0	0	0	0	0	12	17
pVA684-49	0	1	0	6	6	4	0	0	0	34	72
pVA684-146	1	0	0	0	0	2	1	13	0	72	185
pVA684-388	4	0	0	0	0	22	7	0	5	254	520
pVA833-7	1	0	0	0	0	0	0	0	0	3	9
pVA833-92	11	7	7	0	0	2	0	0	3	75	152
pVA833-165	9	4	0	0	1	4	8	0	2	121	228
pVA833-176	15	2	10	0	0	13	4	13	3	91	276
Total	235	79	130	52	52	209	44	105	68	2049	5316

Notas del Autor

Disponibilidad e implementación

GMI-PAN está implementado en *Python* y está disponible como versión beta en la plataforma *GitHub* dentro de un repositorio privado (dada la sensibilidad de la información) creado por el autor *https://github.com/the-alquemist/* , acceso es otorgado previo a una solicitud formal que en este caso se limita al comité evaluador de este seminario de título. Dentro de los contenidos se encuentra tanto el código fuente como el *output* del programa.

Colaboración externa

El desarrollo gráfico de este seminario de título no hubiera sido posible sin la colaboración de Leonardo Acosta, estudiante de Diseño Gráfico de la Facultad de Arquitectura y Urbanismo, Universidad de Chile. Gracias a su talento y paciencia ha sido posible materializar de una forma más amena el contenido de este trabajo en las figuras y tablas adjuntas en este documento.

Bibliografía

1.- Alcock, B. P., Raphenya, A. R., Lau, T. T., Tsang, K. K., Bouchard, M., Edalatmand, A., ... & McArthur, A. G. (2020). CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. Nucleic acids research, 48(D1), D517-D525.

2.- Antipov, D., Raiko, M., Lapidus, A., & Pevzner, P. A. (2019). Plasmid detection and assembly in genomic and metagenomic data sets. Genome research, 29(6), 961–968.

3.- Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., ... & Yeh, L. S. L. (2004). UniProt: the universal protein knowledgebase. Nucleic acids research, 32(suppl_1), D115-D119.

4.- Arndt, D., Grant, J. R., Marcu, A., Sajed, T., Pon, A., Liang, Y., & Wishart, D. S. (2016). PHASTER: a better, faster version of the PHAST phage search tool. Nucleic acids research, 44(W1), W16-W21.

5.- Baker-Austin, C., Wright, M. S., Stepanauskas, R., & McArthur, J. V. (2006). Co-selection of antibiotic and metal resistance. Trends in microbiology, 14(4), 176-182.

6.- Bi, D., Jiang, X., Sheng, Z. K., Ngmenterebo, D., Tai, C., Wang, M., ... & Ou, H. Y. (2015). Mapping the resistance-associated mobilome of a carbapenem-resistant *Klebsiella pneumoniae* strain reveals insights into factors shaping these regions and facilitates generation of a ‘resistance-disarmed’model organism. Journal of Antimicrobial Chemotherapy, 70(10), 2770-2774.

7.- Botelho, J., & Schulenburg, H. (2021). The role of integrative and conjugative elements in antibiotic resistance evolution. Trends in microbiology, 29(1), 8-18.

8.- Carattoli, A., Zankari, E., García-Fernández, A., Voldby Larsen, M., Lund, O., Villa, L., Møller Aarestrup, F., & Hasman, H. (2014). In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. Antimicrobial agents and chemotherapy, 58(7), 3895–3903.

9.- Che, Y., Yang, Y., Xu, X., Břinda, K., Polz, M. F., Hanage, W. P., & Zhang, T. (2021). Conjugative plasmids interact with insertion sequences to shape the horizontal transfer of antimicrobial resistance genes. Proceedings of the National Academy of Sciences, 118(6), e2008731118.

10.- Cooksey D. A. (1994). Molecular mechanisms of copper resistance and accumulation in bacteria. FEMS microbiology reviews, 14(4), 381–386.

11.- Darby, E. M., Trampari, E., Siasat, P., Gaya, M. S., Alav, I., Webber, M. A., & Blair, J. M. (2022). Molecular mechanisms of antibiotic resistance revisited. Nature Reviews Microbiology, 1-16.

12.- Darmon, E., & Leach, D. R. (2014). Bacterial genome instability. Microbiology and Molecular Biology Reviews, 78(1), 1-39.

13.- De Oliveira, D. M., Forde, B. M., Kidd, T. J., Harris, P. N., Schembri, M. A., Beatson, S. A., ... & Walker, M. J. (2020). Antimicrobial resistance in ESKAPE pathogens. Clinical microbiology reviews, 33(3), 10-1128.

14.- Domingues, S., da Silva, G. J., & Nielsen, K. M. (2012). Integrons: vehicles and pathways for horizontal dissemination in bacteria. Mobile genetic elements, 2(5), 211-223.

15.- Elshamy, A. A., & Aboshanab, K. M. (2020). A review on bacterial resistance to carbapenems: epidemiology, detection and treatment options. Future science OA, 6(3), FSO438.

16.- Gamma, E., Helm, R., Johnson, R., Johnson, R. E., & Vlissides, J. (1995). Design patterns: elements of reusable object-oriented software. Pearson Deutschland GmbH.

17.- Gardy, J. L., & Loman, N. J. (2018). Towards a genomics-informed, real-time, global pathogen surveillance system. Nature Reviews Genetics, 19(1), 9-20.

18.- Goneau, L. W., Delport, J., Langlois, L., Poutanen, S. M., Razvi, H., Reid, G., & Burton, J. P. (2020). Issues beyond resistance: inadequate antibiotic therapy and bacterial hypervirulence. FEMS Microbes, 1(1), xtaa004.

19.- Green, E. R., & Mecsas, J. (2016). Bacterial secretion systems: an overview. Virulence mechanisms of bacterial pathogens, 213-239.

20.- Gu, Z., Eils, R., & Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics, 32(18), 2847-2849.

21.- Hyatt, D., Chen, G. L., LoCascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC bioinformatics, 11(1), 1-11.

22.- Jolley KA, Bray JE and Maiden MCJ. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications [version 1; peer review: 2 approved]. Wellcome Open Res 2018, 3:124

23.- Kim, J., Cater, R. J., Choy, B. C., & Mancia, F. (2021). Structural Insights into Transporter-Mediated Drug Resistance in Infectious Diseases. Journal of molecular biology, 433(16), 167005.

24.- Lam, M. M., Wick, R. R., Watts, S. C., Cerdeira, L. T., Wyres, K. L., & Holt, K. E. (2021). A genomic surveillance framework and genotyping tool for *Klebsiella pneumoniae* and its related species complex. Nature communications, 12(1), 4188.

25.- Levy, S. B., & Marshall, B. (2004). Antibacterial resistance worldwide: causes, challenges and responses. Nature medicine, 10(12 Suppl), S122–S129.

26.- Li, L. G., Xia, Y., & Zhang, T. (2017). Co-occurrence of antibiotic and metal resistance genes revealed in complete genome collection. The ISME journal, 11(3), 651-662.

27.- Li, S., Rosen, B. P., Borges-Walmsley, M. I., & Walmsley, A. R. (2002). Evidence for cooperativity between the four binding sites of dimeric ArsD, an As (III)-responsive transcriptional regulator. Journal of Biological Chemistry, 277(29), 25992-26002.

28.- Li, X., Xie, Y., Liu, M., Tai, C., Sun, J., Deng, Z., & Ou, H. Y. (2018). oriTfinder: a web-based tool for the identification of origin of transfers in DNA sequences of bacterial mobile genetic elements. Nucleic acids research, 46(W1), W229-W234.

29.- Marcoleta, A. E., Arros, P., Varas, M. A., Costa, J., Rojas-Salgado, J., Berríos-Pastén, C., ... & Lagos, R. (2022). The highly diverse Antarctic Peninsula soil microbiota as a source of novel resistance genes. Science of the Total Environment, 810, 152003.

30.- Néron, B., Littner, E., Haudiquet, M., Perrin, A., Cury, J., & Rocha, E. P. (2022). IntegronFinder 2.0: identification and analysis of integrons across bacteria, with a focus on antibiotic resistance in *Klebsiella*. Microorganisms, 10(4), 700.

31.- O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., ... & Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic acids research, 44(D1), D733-D745.

32.- Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., & Phillippy, A. M. (2016). Mash: fast genome and metagenome distance estimation using MinHash. Genome biology, 17(1), 1-14.

33.- ORE, I. (2021). 2018 Minerals Yearbook. US Geological Survey.

34.- Osborn, A. M., Bruce, K. D., Strike, P., & Ritchie, D. A. (1997). Distribution, diversity and evolution of the bacterial mercury resistance (mer) operon. FEMS microbiology reviews, 19(4), 239–262.

35.- Palmer, L. D., & Skaar, E. P. (2016). Transition Metals and Virulence in Bacteria. Annual review of genetics, 50, 67–91.

36.- Pearson, W. R., & Lipman, D. J. (1988). Improved tools for biological sequence comparison. Proceedings of the National Academy of Sciences, 85(8), 2444-2448.

37.- Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., ... & Finn, R. D. (2012). The Pfam protein families database. Nucleic acids research, 40(D1), D290-D301.

38.- Queenan, A. M., & Bush, K. (2007). Carbapenemases: the versatile beta-lactamases. Clinical microbiology reviews, 20(3), 440–458.

39.- Razavi, M., Kristiansson, E., Flach, C. F., & Larsson, D. J. (2020). The association between insertion sequences and antibiotic resistance genes. Msphere, 5(5), e00418-20.

40.- Richardson, E. J., & Watson, M. (2013). The automatic annotation of bacterial genomes. Briefings in bioinformatics, 14(1), 1-12.

41.- Rodríguez-Beltrán, J., DelaFuente, J., Leon-Sampedro, R., MacLean, R. C., & San Millan, A. (2021). Beyond horizontal gene transfer: the role of plasmids in bacterial evolution. Nature Reviews Microbiology, 19(6), 347-359.

42.- Rosen B. P. (1990). The plasmid-encoded arsenical resistance pump: an anion-translocating ATPase. Research in microbiology, 141(3), 336–341.

43.- Roumpeka, D. D., Wallace, R. J., Escalettes, F., Fotheringham, I., & Watson, M. (2017). A review of bioinformatics tools for bio-prospecting from metagenomic sequence data. Frontiers in genetics, 23.

44.- Sawa, T., Kooguchi, K., & Moriyama, K. (2020). Molecular diversity of extended-spectrum β-lactamases and carbapenemases, and antimicrobial resistance. Journal of intensive care, 8, 1-13.

45.- Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. Bioinformatics, 30(14), 2068-2069.

46.- Shi, Q., Han, X., Huang, Q., Meng, Y., Zhang, P., Wang, Z., ... & Yu, Y. (2022). The Genetic Characteristics and Carbapenem Resistance Mechanism of ST307 Klebsiella pneumoniae Coharbouring bla CMY-6, bla OXA-48, and a Truncated bla NDM-1. Antibiotics, 11(11), 1616.

47.- Siguier, P., Filée, J., & Chandler, M. (2006). Insertion sequences in prokaryotic genomes. Current opinion in microbiology, 9(5), 526-531.

48.- Siguier, P., Pérochon, J., Lestrade, L., Mahillon, J., & Chandler, M. (2006). ISfinder: the reference centre for bacterial insertion sequences. Nucleic acids research, 34(suppl_1), D32-D36.

49.- Sirén, K., Millard, A., Petersen, B., Gilbert, M. T. P., Clokie, M. R., & Sicheritz-Pontén, T. (2021). Rapid discovery of novel prophages using biological feature engineering and machine learning. NAR genomics and bioinformatics, 3(1), lqaa109.

50.- Speck, A., & Pulvermuller, E. (2001, December). Versioning in software engineering. In IECON'01. 27th Annual Conference of the IEEE Industrial Electronics Society (Cat. No. 37243) (Vol. 3, pp. 1856-1861). IEEE.

51.- Stajich, J. E., & Lapp, H. (2006). Open source tools and toolkits for bioinformatics: significance, and where are we?. Briefings in bioinformatics, 7(3), 287–296.

52.- Steinegger, M., & Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nature biotechnology, 35(11), 1026-1028.

53.- Veloso, M., Arros, P., Acosta, J., Berrios-Pasten, C., Rojas, R., Varas, M., ... & Marcoleta, A. E. (2022). Population genomics, resistance, pathogenic potential, and mobile genetic elements of carbapenem-resistant Klebsiella pneumoniae causing infections in Chile. bioRxiv, 2022-11.

54.- Walsh, F. M., & Amyes, S. G. (2004). Microbiology and drug resistance mechanisms of fully resistant pathogens. Current opinion in microbiology, 7(5), 439–444.

55.- Wattam, A. R., Davis, J. J., Assaf, R., Boisvert, S., Brettin, T., Bun, C., ... & Stevens, R. L. (2017). Improvements to PATRIC, the all-bacterial bioinformatics database and analysis resource center. Nucleic acids research, 45(D1), D535-D542.

56.- Wheeler, T. J., & Eddy, S. R. (2013). nhmmer: DNA homology search with profile HMMs. Bioinformatics, 29(19), 2487-2489.

57.- Wilson, G., Bryan, J., Cranston, K., Kitzes, J., Nederbragt, L., & Teal, T. K. (2017). Good enough practices in scientific computing. PLoS computational biology, 13(6), e1005510.

58.- World Health Organization. (2001). WHO global strategy for containment of antimicrobial resistance (No. WHO/CDS/CSR/DRS/2001.2). World Health Organization.

59.- Wozniak, R. A., & Waldor, M. K. (2010). Integrative and conjugative elements: mosaic mobile genetic elements enabling dynamic lateral gene flow. Nature Reviews Microbiology, 8(8), 552-563.

60.- Wu, F., Ying, Y., Yin, M., Jiang, Y., Wu, C., Qian, C., ... & Lu, J. (2019). Molecular characterization of a multidrug-resistant Klebsiella pneumoniae strain R46 isolated from a rabbit. International journal of genomics, 2019.

61.- Wyres, K. L., Lam, M. M., & Holt, K. E. (2020). Population genomics of Klebsiella pneumoniae. Nature Reviews Microbiology, 18(6), 344-359.

62.- Xu, C., & Rosen, B. P. (1997). Dimerization is essential for DNA binding and repression by the ArsR metalloregulatory protein of Escherichia coli. Journal of Biological Chemistry, 272(25), 15734-15738.

63.- Yelin, R., Rotem, D., & Schuldiner, S. (1999). EmrE, a small Escherichia coli multidrug transporter, protects Saccharomyces cerevisiae from toxins by sequestration in the vacuole. Journal of bacteriology, 181(3), 949-956.

64.- Zhang, A. N., Li, L. G., Ma, L., Gillings, M. R., Tiedje, J. M., & Zhang, T. (2018). Conserved phylogenetic distribution and limited antibiotic resistance of class 1 integrons revealed by assessing the bacterial genome and plasmid collection. Microbiome, 6(1), 1-14.