UNIVERSIDAD DE CHILE – FACULTAD DE CIENCIAS – ESCUELA DE PREGRADO

Desarrollo de una herramienta informática para la clasificación y anotación funcional de islas genómicas asociadas a genes de tRNAs y tmRNAs en el género *Klebsiella.*

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

Roberto Ignacio Rojas Contreras

Directores del Seminario de Título:

Dr. Andrés Marcoleta Caldera.

Dr. (c) Camilo Berríos Pastén

Noviembre 2022

Santiago – Chile

UNIVERSIDAD DE CHILE – FACULTAD DE CIENCIAS – ESCUELA DE PREGRADO

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 el Sr. Roberto Ignacio Rojas Contreras.

Desarrollo de una herramienta informática para la clasificación y anotación funcional de islas genómicas asociadas a genes de tRNAs y tmRNAs en el género *Klebsiella*.

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

Directores Seminario de Título:

Dr. Andrés Marcoleta C.

Dr. (c) Camilo Berríos P.

Comisión Revisora y Evaluadora:

Presidenta Comisión

Dra. Rosalba Lagos M.

Evaluador

Dr. Francisco Chávez E.

Santiago de Chile, ___ de _____ de 2023

Dedicado con mucho cariño

a mi mamá Ana,

a mi papá Robert y

a mi hermano Marcelo.

AGRADECIMIENTOS

Quiero agradecer al Dr. Andrés Marcoleta, director de mi seminario de título, por su valiosa ayuda y porque me dio la oportunidad de incorporarme al Laboratorio de Biología Molecular y Estructural (BEM) siendo parte de un equipo con gran espíritu de compañerismo y, que a la vez, trabaja mucho.

También quiero agradecer especialmente a Camilo Berríos, por su disposición a ayudar permanentemente, por su compensión y liderazgo dentro del grupo de bioinformática. Sin sus sugerencias y consejos, este trabajo no hubiera sido posible.

A todos los miembros del BEM por el buen ambiente dentro del laboratorio que permitió no solo dedicarse a lo académico sino a vivir el día a día durante el período del desarrollo de este trabajo. Especialmente a Carlos, Pablo y Joaquín quienes fueron un soporte muy importante durante todo el período que estuve en el laboratorio.

A mis compañeras y compañeros de carrera quienes han sido excelentes compañeros de ruta dentro de estos agitados años, permitiendo que hayamos podido ir llegando al final de este viaje, que al inicio parecía tan lejano. Especialmente a Moira, quien venida desde la lejana y fría Patagonia ha sido un soporte fundamental durante estos años con sus consejos, amabilidad y permanente alegría; estos años de amistad los agradeceré por siempre.

Por último, a mi familia, que sin su apoyo permanente no podría haber llegado hasta aquí, en lo académico, pero más importante, en la vida. Especialmente, a mi mamá Ana, mi papá Robert y mi hermano Marcelo, quienes siempre me han dado lo mejor. Muchísimas gracias.

ÍNDICE DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

ABREVIATURAS

BGC	Clúster génico biosintético
CDS	Secuencia codificante
CG	Grupo clonal
COG	Grupo de genes ortólogos
DNA	Ácido desoxirribonucleico
MGE	Elemento genético móvil
GI	Isla genómica
KpSC	Complejo de especies de *K. pneumoniae*
hvKp	*Klebsiella pneumoniae* hipervirulenta
ICE	Elemento integrativo y conjugativo
mRNA	RNA mensajero
ORF	Marco de lectura abierto
RNA	Ácido ribonucleico
tDNA	Gen que codifica un tRNA
tRNA	RNA de transferencia
tmDNA	Gen que codifica un tmRNA
tmRNA	RNA mensajero y de transferencia
t(m)DNA	Gen que codifica tRNA o tmRNA
t(m)RNA	tRNA o tmRNA

RESUMEN

La aparición de nuevas cepas microbianas resistentes a múltiples antibióticos junto a cepas hipervirulentas harán progresivamente más difícil el tratamiento de las infecciones, lo que constituye una de las principales amenazas a la salud mundial. Además, se conoce que la aparición de resistencia está estrechamente asociada al uso de antibióticos y esto ha ido en aumento durante las últimas décadas debido al uso en alta cantidad y en indicaciones inadecuadas en humanos y en otros animales. Unos de los microorganismos principales en esta amenaza son las bacterias integrantes del género *Klebsiella*, las cuales destacan por acceder a un grupo movilizable de genes de virulencia y por ser un reservorio de genes de resistencia antibiótica, los cuales podrían diseminarse a otras especies gramnegativas, posibilitando la aparición de nuevas cepas que combinen multirresistencia e hipervirulencia. La transferencia genética horizontal ha tenido un rol central en este proceso porque permite intercambiar una amplia variedad de genes. Los elementos genéticos móviles (MGE) son quienes permiten esta transferencia y, dentro de ellos, están las islas genómicas (GIs) que son un grupo importante, pero poco estudiado. Existe evidencia de que las GIs tienen un papel importante en la evolución y patogénesis de *K. pneumoniae*, por esto, se ha hecho necesario sistematizar y automatizar su estudio. El objetivo de este trabajo fue desarrollar una herramienta bioinformática para comparar, clasificar y anotar GIs asociadas a genes de tRNAs y tmRNAs de cepas del género *Klebsiella*. Para evaluar los resultados de esta herramienta se utilizaron GIs que previamente habían sido analizadas en detalle y publicadas por este y otros grupos de investigación. En un primer paso, esta herramienta es capaz de agrupar las GIs y generar un conjunto no redundante de islas que permite realizar el análisis en un menor tiempo y con un menor procesamiento computacional. Posteriormente, la herramienta utiliza diversas bases de datos para anotar los genes predichos en cada una de las GIs e identificar regiones asociadas a clústeres biosintéticos y fagos integrados gracias al uso de varios programas bioinformáticos. Con la herramienta funcionando correctamente, se procesó un conjunto ampliado de 1.004 cepas que, hasta ahora, no ha sido analizado. En este conjunto se identificaron 7.648 GIs, generándose un conjunto no redundante de 2.411 islas donde se identificaron 79.293 CDS. La anotación identificó 666 genes de integrasas, 822 de toxinas y antitoxinas, 8.221 factores de virulencia, 1.295 genes relacionados con resistencia a antibióticos y 2.954 genes relacionados con biocidas antibacterianos y resistencias a metales. En la identificación de regiones, se reconocieron 358 profagos y 107 clústeres biosintéticos. En un paso posterior se anotaron funcionalmente otros 29.898 CDS, quedando sin identificar 16.464 proteínas hipotéticas. Un análisis bioinformático como este es un buen punto de partida para analizar GIs desconocidas porque, en poco tiempo y con pocos recursos computacionales, anota los genes codificados en ellas y permite guiar el análisis para centrarse en estudios posteriores. En el futuro, con información más completa en las bases de datos, esta misma herramienta podrá tener mejores resultados.

ABSTRACT

The emergence of new microbial strains with resistance to multiple antibiotics combined to hypervirulent strains will make progressively more difficult to treat infections, which is one of the main threats to global health. Furthermore, the emergence of resistance is known to be closely associated with the use of antibiotics and this has been increasing over the last decades due to high usage and inappropriate indications in humans and other animals. One of the main microorganisms in this threat are bacteria belonging to the genus *Klebsiella*, which stand out for accessing a mobilizable pool of virulence genes and for being a reservoir of antibiotic resistance genes, which could spread to other gram-negative species, enabling the emergence of new strains combining multi-resistance and hypervirulence. Horizontal gene transfer has played a central role in this process because it allows a wide variety of genes to be exchanged. Mobile genetic elements (MGEs) are those that allow this transfer and, among them, there are genomic islands (GIs), which are an important but poorly studied group. There is evidence that GIs play an important role in the evolution and pathogenesis of *K. pneumoniae*, therefore, it has become necessary to systematize and automate their study. The aim of this work was to develop a bioinformatics tool to compare, classify and annotate GIs associated with tRNAs and tmRNAs genes of *Klebsiella* strains. GIs that had previously been analysed in detail and published by this and other research groups were used to evaluate the results of this tool. In a first step, this tool clusters the GIs and generates a non-redundant set of islands that allows the analysis to be performed in a shorter time and requires less computational processing. Subsequently, the tool uses various databases to annotate the predicted genes in each of the GIs and identifies regions associated with biosynthetic clusters and integrated phages using bioinformatics software. With the tool working properly, an expanded set of 1,004 strains was processed, which has not been analysed so far. In this set, 7,648 GIs were identified, generating a non-redundant set of 2,411 islands where 79,293 CDS were identified. The annotation identified 666 integrase genes, 822 toxin and antitoxin genes, 8,221 virulence factors, 1,295 antibiotic resistance-related genes and 2,954 genes related to antibacterial biocides and metal resistances. In region identification, 358 prophages and 107 biosynthetic clusters were recognized. In a subsequent step, a further 29,898 CDS were functionally annotated, leaving 16,464 hypothetical proteins unidentified. A bioinformatics analysis like this is a useful starting point to analyse unknown GIs because, in a shorter time and with fewer computational resources, it annotates the encoded genes and allows guiding the analysis to focus on further studies. In the future, with more comprehensive information in the databases, this same tool could have better results.

1. INTRODUCCIÓN

Una de las principales amenazas a la salud mundial, junto a la actual pandemia de COVID-19, es la aparición de nuevas cepas microbianas con resistencia a múltiples antibióticos simultáneamente y, por otro lado, las cepas hipervirulentas que podrían hacer progresivamente más difícil el tratamiento de estas infecciones. La Microbiología ha enfrentado esta amenaza en varios aspectos, pero principalmente a través de lograr un conocimiento más profundo de los mecanismos que permiten la adquisición de resistencias o de hipervirulencia, es decir, investigación de ciencia básica.

Diferentes grupos de investigación se han dedicado a identificar los mecanismos bacterianos que permiten la transferencia de estos fenotipos con relevancia clínica desde cepas donantes hacia otras cepas que, previo a este evento, no eran hipervirulentas o multirresistentes. Unos de los mecanismos que median este tipo de transferencia son los elementos genéticos móviles (MGE, por su sigla en inglés), correspondientes a un grupo diverso de fragmentos de DNA que son capaces de escindirse del cromosoma de una cepa donante y, a su vez, de integrarse al cromosoma de otra cepa receptora. Dentro de este grupo destacan las islas genómicas (GIs, por su sigla en inglés), que hasta ahora han sido relativamente poco estudiadas (como sí lo han sido los plásmidos); los elementos integrativos y conjugativos (ICEs, por sus siglas en inglés) y; los profagos, solo por mencionar algunos. Uno de los microorganismos más relevantes dentro de esta amenaza a la salud mundial son las bacterias integrantes del género *Klebsiella*, que son tratadas con mayor detalle en este trabajo. Dichas bacterias destacan por tener acceso a un grupo movilizable de genes

de virulencia y ser un reservorio conocido de genes de resistencia antibiótica que, también podrían dispersarse a otras especies gramnegativas, posibilitando la aparición de cepas que combinen la multirresistencia y la hipervirulencia (Bengoechea & Sa Pessoa, 2019). También es importante mencionar que estas bacterias aún conservan muchos aspectos desconocidos de su genoma y su funcionamiento, lo que representa un campo de estudio aún abierto, de hecho, existen estimaciones de que aún queda por conocer las anotaciones funcionales de aproximadamente un tercio de los genes que codifican proteínas dentro de sus genomas (Hoskisson & Seipke, 2020).

1.1 La amenaza de los microorganismos multirresistentes a la salud mundial

La resistencia a antimicrobianos ocurre cuando bacterias, virus, hongos y parásitos no responden efectivamente al tratamiento farmacológico para su eliminación. Actualmente, es una de las mayores amenazas mundiales, y así lo ha manifestado durante varios años la Organización Mundial de la Salud, destacándose como posibles consecuencias de esto millones de muertes, discapacidades de larga duración y mayores costos asociados a la atención sanitaria. También ha sido descrita como una pandemia silenciosa que tendría repercusiones en otros ámbitos diferentes a la salud humana como pérdida de vidas animales, severos efectos sobre los estilos de vida humanos y seguridad alimentaria (World Health Organization et al., 2021). Fue estimado por el Banco Mundial que, en caso de no ser enfrentado este problema, la economía mundial podría perder casi el 4% del producto interno bruto, siendo afectados principalmente países de ingreso bajo y medio, empujando a la pobreza hasta 28 millones de personas en 2050 por efectos sobre la productividad económica, producción ganadera y costos de atención sanitaria (World Bank Group, 2017).

La aparición de resistencia antibiótica es intrínseca al uso de antibióticos y ha ido en aumento durante las últimas décadas debido al uso en gran cantidad y en indicaciones inadecuadas tanto en humanos como en otros animales. Específicamente, se han descrito altas tasas de resistencia a antibióticos en algunos de los patógenos más frecuentemente involucrados en infecciones humanas, tales como *Staphylococcus aureus*, *Pseudomonas aeruginosa* o el complejo *Mycobacterium tuberculosis* que provocan enfermedades como infecciones cutáneas, abscesos, neumonía, sepsis o tuberculosis, entre otras (Seoane & Bou, 2021).

Se han descrito varios factores que permiten la propagación de la resistencia antibiótica entre los microorganismos, incluyendo el uso inapropiado de antibióticos en los ámbitos sanitario y agrícola y la ausencia de nuevas terapias antibióticas (Martin & Bachman, 2018). La movilización de genes de resistencia dentro y entre especies es uno de los factores evolutivos más importantes que dirigen el fenómeno, de hecho, varios estudios han mencionado la importancia del papel de los animales y de los reservorios medioambientales (Bush et al., 2011).

1.1.1 Formas de enfrentar la multirresistencia a los antimicrobianos

Hasta ahora han existido varias iniciativas para enfrentar la amenaza de la resistencia a antibióticos, originadas en instituciones de salud a nivel mundial y nacional en cada país. Por simplicidad, se mencionarán las más importantes categorizadas en 3 grupos. La primera forma es a través de la investigación científica y la difusión del conocimiento generado; es relevante que todas las personas conozcan la amenaza a la que se enfrenta la humanidad, porque solo así podrán ser parte de la solución y no del problema debido al desconocimiento. La investigación básica debe centrarse en conocer cómo funcionan y cómo se transfieren los mecanismos de

resistencia antibiótica. Con esta manera de enfrentar la amenaza se pueden ir actualizando las otras dos formas (O'Neill, 2016) porque hasta el día de hoy hay muchos mecanismos de acción desconocidos en los antibióticos utilizados (Goneau et al., 2020). También es importante conocer la verdadera magnitud de los efectos de la resistencia en la población, pero no solo en los países desarrollados donde abundan sus datos, sino también en el resto del mundo no desarrollado donde se encuentra la mayoría de la humanidad (Murray et al., 2022).

La segunda forma es eliminar de forma eficiente las bacterias patógenas, es decir, utilizar los fármacos en indicación y dosis correcta, junto al descubrimiento de nuevos antibióticos con mecanismos para los que las bacterias no hayan desarrollado resistencia todavía. En las últimas décadas, casi no se han descubierto nuevos antibióticos y de los que se han liberado al mercado, en realidad, se basan en compuestos similares a los ya existentes, favoreciendo que la resistencia aparezca en poco tiempo. Para esto se debe incentivar el desarrollo de nuevos compuestos, ya sea, a través de fondos provistos por los gobiernos nacionales o de un trabajo mancomunado con las empresas farmacéuticas (Duval et al., 2019; O'Neill, 2016).

La tercera forma es a través de no incentivar la aparición y transmisión de multirresistencia, aunque sea una obviedad, hay que evitar que las personas se infecten, así se usarán menos antibióticos, lo que conlleva a una menor aparición de resistencia. También hay que combatir las infecciones dentro de los hospitales, puesto que parte importante de los patógenos amenazantes están en esos recintos. La reducción del uso innecesario de antibióticos en agricultura también evitará la diseminación en el medio ambiente. Por último, el desarrollo de vacunas y tratamientos alternativos también ayudarán a usar menos antibióticos con sus efectos indeseados (O'Neill, 2016).

1.1.2 Desarrollo de patógenos bacterianos hipervirulentos

La amenaza de los patógenos multirresistentes no es la única, sino que también está acompañada de cepas que han adquirido nuevos mecanismos de virulencia que provocan enfermedades más severas. Hay evidencia de que los antibióticos tienen otro efecto negativo, además de los habituales como resistencia a antibióticos y depleción de microbiota, el cual es potenciar la selección de microorganismos con alta virulencia. Incluso se ha reportado que, con valores menores a la concentración mínima inhibitoria, se modula el 5-10% del genoma bacteriano (Goneau et al., 2020). Un ejemplo clásico de la activación de factores de virulencia es la Shiga-toxina en *Escherichia coli* (EHEC) o la activación de fagos en lisogenia, pero la selección de cepas con mayor virulencia se puede deber a un aumento en la frecuencia de la mutación *de novo* o a un aumento de la adquisición de elementos foráneos (Blázquez et al., 2012). Hay que destacar que el estrés antibiótico provoca la escisión de genes de virulencia codificados en diferentes MGEs permitiendo que esta información sea movilizada por transferencia genética horizontal, en ello se ha reportado el involucramiento de fagos, secuencias de inserción, islas genómicas y elementos integrativos conjugativos (Beaber et al., 2004; Fothergill et al., 2011; Schreiber et al., 2013). La movilización se logra por la activación del sistema de respuesta a estrés SOS relacionado con quiebres de la doble hélice de DNA, la transcripción de integrasas y la liberación del MGE que potencialmente podría generar cepas hipervirulentas (Goneau et al., 2020).

Los mecanismos de hipervirulencia no están únicamente relacionados a antibióticos, también hay algunos involucrados en la inhibición del crecimiento de otras bacterias o en el mecanismo de secuestro de hierro. Por un lado, algunos patógenos son capaces

de producir bacteriocinas como estrategia para contrarrestar la resistencia a la colonización, puesto que estos péptidos antimicrobianos inhiben el crecimiento de especies relacionadas cercanamente, mientras las bacterias productoras se protegen a través de mecanismos de inmunidad (Lee, 2020). Por otro lado, los sideróforos son pequeñas moléculas microbianas que tienen una alta afinidad por hierro que son capaces de “robarlo” de moléculas del hospedero, afectando sus mecanismos defensivos. Debido a esto, los mamíferos han evolucionado otras moléculas que impiden el secuestro del hierro por parte de los sideróforos, por ejemplo, lipocalina-2, pero a su vez, los microorganismos desarrollaron una forma de evadir la acción de este último. Dentro de ellos están las enzimas y receptores codificados en el cluster génico *iroA* (que codifica la síntesis de salmoquelina) o el sideróforo yersiniabactina producido por especies de *Klebsiella* y *Yersinia*, entre muchos otros (Nairz et al., 2018).

1.2 Elementos genéticos móviles y su rol en el desarrollo de cepas multirresistentes e hipervirulentas

El genoma de las especies microbianas puede ir modificándose a través de varias maneras, ya sea reduciéndose o, por otro lado, sufriendo mutaciones, rearreglos o adquiriendo material genético mediante transferencia genética horizontal (Schmidt & Hensel, 2004). Se ha reconocido ampliamente que la transferencia horizontal ha tenido un rol trascendental en la evolución del genoma de distintos linajes bacterianos. Pues de esta manera las bacterias adquieren genes que pueden ser beneficiosos para ellas bajo ciertas condiciones ambientales. Existen varios tipos de MGE, algunos de los más estudiados en bacterias son los plásmidos que corresponden a moléculas de DNA que se replican autónomamente y pueden portar genes, asimismo, tienen la capacidad de diseminarse a otras células por transferencia horizontal donde nuevamente podrán

replicarse de forma autónoma (Dionisio et al., 2019). Por otro lado, están las secuencias de inserción y transposones, que son fragmentos de DNA capaces de automovilizarse que se escinden e integran casi aleatoriamente en nuevos lugares de la misma u otra molécula de DNA y, por otra parte, los integrones, que usan recombinación sitio-específica para mover genes entre dos sitios definidos. Estos MGE pueden estar en varias copias dentro de la misma célula. Hay otros MGE que utilizan la conjugación como los elementos ICE capaces de portar genes dentro de un esqueleto que se integra al genoma hospedero y se replica junto a él. Otros MGE usan la transducción como los fagos y la transformación como el DNA extracelular. Sin embargo, este trabajo se centrará en genes adquiridos por transferencia horizontal que forman bloques sinténicos conocidos como islas genómicas (Partridge et al., 2018; Schmidt & Hensel, 2004).

1.2.1 Islas genómicas (GIs)

Las islas genómicas han sido definidas como fragmentos de DNA cuya longitud es comúnmente entre 10 y 200 kb, los cuales están presentes de forma variable al comparar regiones equivalentes del cromosoma en distintas cepas. Algunas características principales de estos elementos son: 1) su contenido GC es menor que en el resto del genoma, 2) generalmente están flanqueados por repetidos directos de 16 a 20 pb, 3) frecuentemente están insertados en el extremo 3’ de genes tDNA (genes que codifican los tRNA), y 4) también pueden contener secuencias de inserción o transposones que, a su vez, pueden movilizar o causar deleciones de regiones aledañas de ADN (Juhas et al., 2009).

Las GIs además de integrarse en genes tDNAs, también pueden hacerlo en genes tmDNAs, los cuales codifican tmRNAs, que corresponden a RNAs con la doble función

de transferencia y mensajeros que participan en el control de calidad de la traducción. Estos, son RNAs que funcionan, por un lado, como tRNAs uniéndose a ribosomas que se traban durante la traducción y, por otro lado, como mRNAs al permitir la adición de una señal peptídica a la cadena truncada que actúa como codón stop que permite que la traducción finalice y se reciclen las subunidades del ribosoma. Esta señal que está codificada en el gen *ssrA*, indica que la traducción es incorrecta enviando a la cadena polipeptídica naciente a degradación proteolítica (Janssen & Hayes, 2012).

Además, se ha descrito que los genes portados en las GIs pueden ofrecer una ventaja adaptativa para la cepa portadora. De hecho, está aceptada la idea de que la transferencia genética horizontal ha tenido un rol preponderante en la evolución de las especies bacterianas y esto se ha debido en gran parte a la GIs. Pero la transferencia de este material genético no se limita exclusivamente a los genes que están dentro de la GI, sino también a parte del genoma núcleo de las cepas, que algunas veces, también se desprende junto a ellas siendo transferido a otras cepas filogenéticamente lejanas (Hochhut et al., 2000). Se piensa que las GIs han aparecido muchas veces de forma independiente durante la evolución, en vez de que todas estén filogenéticamente relacionadas, por esto, se les agrupa en una gran familia de elementos en base a su características estructurales y análogas. Desde hace algunos años se ha propuesto que el origen de algunas GIs está relacionado con profagos y algunos plásmidos conjugativos, de hecho, no es infrecuente ver profagos en varias etapas de funcionalidad integrados en el cromosoma (Juhas et al., 2009).

A pesar de lo que se pensaba hace un tiempo, las GIs no contienen únicamente genes relacionados con la patogenicidad, sino que en algunos casos están relacionados con la simbiosis, capacidades metabólicas (metabolismo de sacarosa, compuestos aromáticos, resistencia a mercurio o síntesis de sideróforos, entre otros),

fitness o resistencia. A diferencia de muchos plásmidos, las GIs no codifican todo el proceso de autotransferencia, sino que a veces aprovechan el empacamiento de fagos lisogénicos o son movilizados por plásmidos o el sistema conjugativo de ICEs (Juhas et al., 2009).

La virulencia y la resistencia antibiótica son una de las características más frecuentemente asociadas con GIs en bacterias patógenas. Es más, se ha visto que en las últimas décadas las poblaciones bacterianas han pasado de ser mayoritariamente sensibles hasta un aumento importante en la resistencia antibiótica, empeorando la crisis de resistencia antimicrobiana ya que en los últimos 30 o 40 años las GIs han contribuido a la diversificación y adaptación bacteriana (Juhas et al., 2009).

Dentro de las GIs hay otros grupos que son importantes, por ejemplo, los elementos integrativos y conjugativos (ICEs, por sus siglas en inglés) que son un grupo de GIs que tienen la capacidad de transferirse de forma horizontal, puesto que contienen módulos que codifican toda la maquinaria que les permite escindirse del cromosoma, y transferirse por conjugación a otra célula e integrarse en el cromosoma del nuevo hospedero. Se integran al cromosoma por recombinación sitio-específica entre repetidos directos y la acción de una integrasa. Los ICEs también incluyen a un grupo específico de transposones conjugativos que corresponden a elementos originados en bacterias grampositivas que pueden tener como blanco múltiples sitios de integración diferentes. La transferencia horizontal de cualquiera de estos elementos puede provocar cambios en la preferencia de nicho en una cepa (Botelho & Schulenburg, 2021; Burrus et al., 2002; Burrus & Waldor, 2004).

1.3 Género *Klebsiella* y *Klebsiella pneumoniae*

Hay ciertas especies de *Klebsiella* que son consideradas como patógenos oportunistas que colonizan superficies mucosas sin causar enfermedad. Sin embargo, desde la mucosa pueden diseminarse a otros tejidos causando enfermedades graves como neumonía, infecciones urinarias y sepsis (Paczosa & Mecsas, 2016). *Klebsiella pneumoniae* es una bacteria patogénica gramnegativa descrita por primera vez por Carl Friedlaender en 1882 cuando la aisló desde pulmones de pacientes muertos por neumonía (Friedlaender, 1882). Es parte de la familia *Enterobacteriaceae*, la cual también incluye *Escherichia coli* y diferentes especies de *Yersinia*, *Salmonella* y *Shigella*.

Las especies de *Klebsiella* se pueden encontrar habitando en plantas, animales y humanos, en estos últimos, son el agente causal de varios tipos de enfermedades como infecciones del tracto respiratorio, del tracto urinario y del torrente sanguíneo (Bengoechea & Sa Pessoa, 2019; Martin & Bachman, 2018). En 2017, la OMS publicó una lista de "patógenos prioritarios" resistentes a los antibióticos, en los cuales se debía enfocar la investigación y el desarrollo de nuevos fármacos. La lista estaba conformada por 12 familias de bacterias y dentro de la prioridad más alta, denominada prioridad crítica, consideraba a bacterias resistentes a carbapenémicos, una de ellas fue *K. pneumoniae* (Organización Mundial de la Salud, 2017). También esta especie fue incluida en el término "ESKAPE" que agrupa seis patógenos con aumentos en su virulencia y resistencia a antibióticos (Mulani et al., 2019).

A pesar de su ubicuidad en la naturaleza, la gran mayoría de los aislados de *K. pneumoniae* provienen de muestras clínicas humanas. Los muestreos en ambientes no humanos son muy escasos o con poca profundidad, aun así, se ha visto que hay

algún solapamiento entre los linajes aislados clínicos y los ambientales; y también entre humanos y de animales de compañía (Wyres et al., 2020).

1.3.1 Virulencia en *K. pneumoniae*

Según el tipo de infecciones que provoca, hay un primer grupo de *K. pneumoniae* que actúa como patógeno oportunista infectando pacientes severamente enfermos o inmunocomprometidos provocando neumonía, infecciones del tracto urinario y de la sangre. Un segundo grupo de *K. pneumoniae* son hipervirulentas, las cuales infectan a personas sanas en ambientes comunitarios causando infecciones severas como absceso hepático piogénico, endoftalmitis y meningitis. Un tercer grupo de *K. pneumoniae* son resistentes a antibióticos debido a que codifican, entre otras enzimas, carbapenemasas; estas cepas son oportunistas, pero difíciles de tratar. Todas las cepas dependen de su genoma accesorio para determinar si la infección es asintomática o si se vuelven patogénicas (Martin & Bachman, 2018).

Como ya se mencionó, las diferentes cepas de *K. pneumoniae* pueden actuar como comensales, oportunistas o patogénicas dependiendo de los genes que portan. Al actuar como comensales, es común que colonicen el sistema intestinal o el tracto respiratorio, variando su prevalencia según la edad, ubicación geográfica y reciente contacto con recintos hospitalarios. En segundo lugar, la mayoría de las infecciones son oportunistas y son denominadas infecciones clásicas, ellas se deben a infecciones asociadas a recintos hospitalarios (también llamadas infecciones nosocomiales) y la mayor preocupación se da cuando ellas provienen de cepas multirresistentes. En tercer lugar, fuera de los hospitales, esta especie puede actuar como una patógena verdadera e infecta a pacientes que no poseen los factores de riesgo a las infecciones clásicas. Una infección por cepa hipervirulenta se caracteriza por darse en múltiples

sitios, por provocar bacteremia y alta morbimortalidad (Bengoechea & Sa Pessoa, 2019; Wyres et al., 2020). Es importante mencionar que las infecciones generan complicaciones cuando ocurren en neonatos, ancianos o inmunocomprometidos (Lam et al., 2021).

Durante las décadas de 1980 y 1990, en algunos países asiáticos de la Cuenca del Pacífico, comenzaron a aparecer reportes de infecciones severas de *K. pneumoniae* que se alejaban del comportamiento común de las infecciones clásicas que ocurrían en los hospitales (Cheng et al., 1991; Liu et al., 1986; Wang et al., 1998). Estas infecciones comunitarias se debieron a cepas hipervirulentas de *K. pneumoniae* (hvKp) provocando las características ya mencionadas como absceso hepático piogénico, endoftalmitis, meningitis e infecciones del torrente sanguíneo (Fang et al., 2007; Martin & Bachman, 2018).

1.3.2 Convergencia de resistencia y virulencia

K. pneumoniae es una causa frecuente de infecciones oportunistas resistentes a antimicrobianos en pacientes hospitalizados. Esta especie es naturalmente resistente a penicilinas y algunas cepas pueden tener resistencias a múltiples antimicrobianos, conocidas como multirresistentes. Sin embargo, durante la última década se ha presenciado la aparición de cepas multirresistentes productoras de β-lactamasas de espectro extendido o de carbapenemasas asociadas a brotes de infecciones nosocomiales a nivel mundial y, en paralelo, se ha visto la aparición de cepas hipervirulentas asociadas a infecciones comunitarias donde expresan factores de virulencia adquiridos. Los datos hasta ahora muestran que se han agrupado en diferentes subpoblaciones, sin embargo, hay señales de que este aislamiento mutuo

puede estar perdiéndose y que podrían emerger cepas altamente peligrosas, puesto que originarían enfermedades severas debido a cepas altamente patogénicas y resistentes a todos los antibióticos conocidos hasta hoy (Wyres et al., 2020; Wyres & Holt, 2016)

1.4 Estudios bioinformáticos del genoma de *K. pneumoniae*

El genoma de una cepa típica de *K. pneumoniae* tiene una longitud de entre 5 y 6 Mb, conteniendo entre 5.000 y 6.000 genes, de ellos entre 1.700 y 2.000 genes se conservan en casi todos los miembros de la especie (Wyres et al., 2020), esta colección se conoce como el genoma núcleo (*core genome*), puesto que están presentes en >95% de las cepas. El resto de los genes se consideran parte del genoma accesorio y son compartidos con diferentes especies bacterianas de *Klebsiella* y otros Enterobacteriales.

Entre las cepas de *K. pneumoniae* publicadas, se ha visto que tienen una identidad nucleotídica solo entre el 95% y 96% en promedio, por esta razón, en la literatura se ha decidido utilizar una clasificación especial, puesto que no hay una nomenclatura específica. El conjunto de todas las cepas se conoce como el complejo de especies de *K. pneumoniae* (KpSC), donde el grupo principal es conocido como *K. pneumoniae sensu stricto* que corresponde a aproximadamente el 85% del total de cepas en las bases de datos. Los estudios genómicos también muestran que esta es una población diversa, pero estructurada, lo cual provee un marco útil para trabajar con ellas (Martin & Bachman, 2018; Wyres et al., 2020).

A pesar de que las técnicas bioinformáticas son capaces de entregar, en poco tiempo, cantidades masivas de datos, el tiempo necesario para realizar su análisis de forma manual es demasiado extenso. Para esto, se han diseñado una amplia variedad

de programas computacionales que ayudan en cada una de las etapas del análisis de los datos: desde la secuenciación hasta la anotación de cada uno de los genes contenidos en los genomas estudiados.

1.4.1 Herramientas especializadas para el estudio de bacterias en general y *Klebsiella* en particular

Con el paso del tiempo y con la mayor cantidad de investigaciones, han surgido diferentes bases de datos que contienen información, a nivel general, sobre una gran variedad de especies bacterianas relacionadas con factores de virulencia (VFDB), resistencia antibiótica (CARD), toxinas-antitoxinas (TADB) y varias otras. Por otro lado, existen bases de datos especializadas en ciertas especies, lo cual confiere a los investigadores simultáneamente un conocimiento amplio (a lo largo de varios taxones) y profundo (dentro de una especie en particular) del contenido codificado en los genomas, aunque se debe mencionar que aún se desconoce una cierta fracción de los genes portados en GIs y en los genomas (Seoane & Bou, 2021).

Una de las herramientas más importantes y especializada en analizar *K. pneumoniae* y su complejo de especies asociado es Kleborate (Lam et al., 2021) que fue diseñada para detectar y genotipificar loci claves en la virulencia y resistencia a antibióticos, además, de identificar directamente desde los ensamblajes genómicos las especies, los linajes y predecir sus serotipos según los antígenos K y O.

1.4.2 Análisis sistemático de múltiples genomas y MGEs simultáneamente

Como se mencionó previamente, los elementos genéticos móviles tienen una gran relevancia en la aparición de hipervirulencia y multirresistencia, por esto es tan importante conocerlos para que las estrategias de vigilancia y monitoreo permitan

identificar eficazmente cuando estos elementos estén insertados en microorganismos ambientales o que colonicen personas enfermas. Los MGEs que se adquieren a través de transferencia genética horizontal son de variados tipos como plásmidos, bacteriófagos, transposones e GIs (Marcoleta et al., 2016).

Específicamente, se ha encontrado evidencia que apoya la idea de que las islas genómicas tienen un papel importante en la evolución y patogénesis de *K. pneumoniae* (Berríos-Pastén et al., 2020), por esto es primordial estudiar las GIs y conocer su gran diversidad. Hasta ahora en nuestro grupo de trabajo la caracterización de MGEs se ha realizado a través del análisis individual de cada uno de los elementos identificados y surge la necesidad de automatizar este proceso para abarcar el análisis masivo de cromosomas de cepas de *K. pneumoniae* y sus MGEs. En este sentido, en este trabajo se presenta el desarrollo y resultado de una herramienta de automatización de este proceso, aunque previamente se debe generar una sistematización en su estudio para permitir análisis más rigurosos y comparables entre distintas colecciones de cepas. La sistematización debiera incluir una forma de identificar el lugar donde se inserta la isla genómica, cuántos son los genes presentes y cuáles son las funciones codificadas en aquellos genes. Como las funciones codificadas son tan diversas, se debería poder integrar esta información a partir de diferentes bases de datos para lograr finalmente una anotación integral que dé cuenta de la mayor cantidad de información disponible hasta el momento.

2. OBJETIVOS

2.1 Objetivo General

Desarrollar una herramienta bioinformática para comparar, clasificar y anotar islas genómicas asociadas a genes de tRNAs y tmRNAs de cepas del género *Klebsiella*.

2.2 Objetivos Específicos

2.2.1 Programar una herramienta de clasificación, comparación y anotación de islas genómicas. Evaluar su desempeño en cuanto a uso de recursos computacionales y tiempo de cómputo.

2.2.2 Evaluar los resultados de la herramienta desarrollada con los resultados previamente publicados para GIs de la especie *Klebsiella pneumoniae*.

2.2.3 Caracterizar las islas previamente predichas en un grupo de 1.004 cepas del género *Klebsiella* usando la herramienta desarrollada.

3. MATERIAL Y MÉTODOS

3.1 Secuencias genómicas de cepas de *K. pneumoniae*

Se consideraron dos conjuntos de genomas de *K. pneumoniae* para el desarrollo de este trabajo. En primer lugar, se utilizaron secuencias genómicas de 66 cepas de la especie provenientes de la base de datos del National Center for Biotechnology Information (NCBI). Esta colección fue previamente analizada en busca de MGEs y la información fue curada manualmente, por ello se considera como referencia en el grupo de investigación (Berríos-Pastén et al., 2020). El segundo conjunto de genomas analizados en este trabajo se trata de un conjunto ampliado de cepas de *Klebsiella*, construido según se detalla a continuación. Los cromosomas se descargaron el 27 de marzo de 2021 desde la base de datos del NCBI utilizando como criterio de inclusión cepas identificadas como parte del género *Klebsiella* y que los proyectos tuvieran como estado de avance “completo” (cromosomas cerrados y plásmidos). Como control, todos los genomas descargados se analizaron con Kleborate (Lam et al., 2021), programa que ha sido diseñado para la identificación de genes marcadores del género que permiten la identificación de la especie, el secuenciotipo y factores de virulencia. De esta manera, se descartaron secuencias que no pertenecieran a alguna de las especies conocidas de *Klebsiella*, quedando finalmente una colección de 1.004 genomas.

3.2 Predicción de islas genómicas asociadas a t(m)DNAs

La predicción de islas genómicas se realizó con una herramienta desarrollada por nuestro grupo de investigación llamada Kintun-VLI (Acevedo, 2019)

(https://github.com/GMI-Lab/Kintun-VLI). En resumen, la herramienta compara cada uno de los t(m)DNAs codificados en los cromosomas de las cepas analizadas con una colección de secuencias de referencia. Con esto puede detectar variaciones en la continuidad de la arquitectura genómica común en torno a los t(m)DNAs, estas secuencias variantes, luego, son analizadas en busca de marcadores de islas genómicas como los repetidos directos, presencia de integrasas y contenido GC distinto al del cromosoma hospedero. Para su posterior análisis, cada una de las islas genómicas identificadas fue exportada en archivos formato FastA.

3.3 Clusterización de secuencias nucleotídicas de islas genómicas

Para analizar la diversidad de las islas genómicas obtenidas y para disminuir el volumen de los análisis que le siguen a la predicción, las secuencias de las islas identificadas fueron clusterizadas para generar una colección no redundante de estos elementos. Este análisis fue realizado con el programa CD-HIT (Fu et al., 2012). En cuanto a los parámetros considerados, se utilizó el programa con la opción `-est`, para el análisis de secuencias nucleotídicas. En relación con los valores determinados para considerar dos islas como repetidas, se utilizaron las opciones `-c 0.80 -s 0.85`, así dejando secuencias que tuvieran al menos un 80% de identidad y un 85% de cobertura, como iguales.

3.4 Desarrollo y programación de Wapi, una herramienta de comparación y anotación de islas genómicas

La herramienta utiliza programas previamente publicados para cada una de sus funciones. Al inicio, se utilizó el programa PROKKA 1.14.6 (Seemann, 2014), con sus opciones por defecto, para la predicción de genes con PRODIGAL (Hyatt et al., 2010) y

la anotación de ellos en base a bases de datos especializadas en diferentes ámbitos. Las cinco bases de datos utilizadas fueron: (i) integrasas obtenidas el 27 de julio de 2021 de la base de datos UniProt con el buscador de NCBI, incorporando solamente las identificadas en genomas de bacterias y eliminando manualmente los genes identificados como transposasas; (ii) secuencias aminoacídicas y nucleotídicas de la base de datos de toxinas y antitoxinas TADB v2.0 de junio de 2017 (Xie et al., 2018); (iii) proteínas de factores de virulencia de la base de datos VFDB (B. Liu et al., 2022) obtenidas el 12 de agosto de 2021; (iv) la base de datos de resistencia antibiótica CARD obtenida el 13 de agosto de 2021 (Alcock et al., 2020) y; (v) la base de datos de genes de resistencia BacMet2 v2.0, predichos como biocidas antibacterianos y genes de resistencia a metales, publicada el 11 de marzo de 2018.

Para la predicción de fagos integrados en las islas genómicas, se utilizó el programa PhageBoost 0.1.7 (Sirén et al., 2021), con sus opciones por defecto, que entrega regiones rotuladas como profagos, que en cuyo interior presenten varios genes relacionados con profagos o fagos integrados.

Luego, se predice la presencia de clústeres biosintéticos con el programa antiSMASH 6.0.0 (Blin et al., 2021) con opción *knownclusters*. De forma similar al programa anterior, también predice regiones que abarcan varios genes, pero en este caso, relacionadas con clusters génicos biosintéticos (BGCs) para la producción de metabolitos secundarios tales como poliquétidos, péptidos no ribosomales, péptidos lasso (que incluyen las microcinas), terpenos, sideróforos, β-lactámicos y un largo etcétera.

Después, se realizó una anotación funcional de solo una parte de los genes predichos en las secuencias, para esto se tomaron todos los genes que no lograron ser anotados con las bases de datos con PROKKA, es decir, clasificados como

"*hypothetical proteins*" y se analizaron con el programa eggNOG-mapper 2.1.6 (Cantalapiedra et al., 2021; Huerta-Cepas et al., 2019), con sus opciones por defecto. Esto entregó los genes agrupados por categorías COG.

Finalmente, para complementar la identificación de los genes codificados en las islas genómicas, se utilizó el programa nhmmer 3.3.2 (Wheeler & Eddy, 2013) para identificar la presencia de ciertas secuencias nucleotídicas. El primer grupo contenía secuencias RNA de antitoxinas de la base de datos TADB que no podían ser analizadas por PROKKA puesto que sólo reconoce secuencias aminoacídicas y el segundo grupo contuvo secuencias de oriT de la base de datos oriTDB con evidencia experimental y otras predichas obtenidas el 13 de enero de 2022 (Li et al., 2018) para identificar orígenes de transferencia.

3.5 Islas genómicas previamente anotadas y disponibles en bases de datos

Con la herramienta ya programada se probó en otro conjunto de islas, en este caso, se utilizaron 4 islas genómicas publicadas previamente en la literatura, las cuales tienen diferencias en las especies que las portaban y del tejido donde fueron muestreadas. Ellas corresponden a la isla genómica GIE492 (Marcoleta et al., 2016), al ICE IHE3034-1 con número de acceso AM229678 proveniente de una cepa de *E. coli* (Putze et al., 2009), al elemento móvil ICEKp14 con número de acceso KY454638 proveniente de la cepa 16852116 de *K. pneumoniae* obtenida de una muestra de garganta (Lam et al., 2018) y el elemento móvil ICEKp1 con número de acceso KY454627 proveniente de *K. pneumoniae* de una muestra intestinal de mono (Lam et al., 2018).

3.6 Árbol filogenético y análisis de islotipos

Para los análisis filogenéticos se determinó el genoma núcleo (*core genome*) entre los cromosomas de las 66 cepas analizadas. Para esto, el pangenoma de las cepas fue calculado con Roary (Page et al., 2015) y los genes concatenados del genoma núcleo (presente en al menos un 95% de las cepas) fueron alineados con el software MUSCLE (Madeira et al., 2022) y la inferencia del árbol a partir del alineamiento se realizó con algoritmo de máxima verosimilitud con RaxML (Stamatakis, 2014).

Para estudiar la distancia entre las cepas respecto de las islas que portan, se calculó la distancia de Jaccard entre los islotipos de las cepas, esto es, la diferencia entre las clases de islas genómicas integradas en ambas cepas con respecto al conjunto completo entre las islas. De esta forma, valores de distancia con valor igual a cero implican cepas con las mismas islas genómicas mientras que valores mayores a cero implican diferencias entre los conjuntos. Para visualizar estos datos, se confeccionó un *heatmap* con la herramienta de anotación de iTOL (Letunic & Bork, 2021).

4. RESULTADOS

4.1 Creación de la herramienta

El primer resultado de este trabajo fue la creación de Wapi, un programa bioinformático escrito en lenguaje Python cuyo objetivo es comparar, clasificar y anotar islas genómicas asociadas a genes de t(m)RNAs, de forma automatizada en la línea de comandos, incorporando programas bioinformáticos disponibles de forma gratuita junto a bases de datos para generar una anotación especializada y funcional de las GIs. Wapi que significa “isla” en mapudungún, se basa en la herramienta Kintun (Acevedo, 2019) desarrollada dentro de nuestro grupo de trabajo, que significa “buscar”, en el mismo idioma. En la Figura 1 se muestra un esquema de la ejecución de cada programa con las bases de datos utilizadas, según corresponda. La ejecución de la herramienta se divide en 3 secciones: pasos previos, herramienta Wapi y el output o salida.

En primer lugar, **(1)** a la herramienta se le debe entregar un archivo de texto que contenga las secuencias nucleotídicas de las GIs a analizar. Como se mencionó en la sección de materiales y métodos, esas GIs fueron identificadas gracias a un método desarrollado por nuestro grupo de investigación y fueron nombradas según el tDNA en el que estaban insertadas, todas se almacenan en un único archivo en formato fastA. **(2)** Las secuencias de este archivo se procesaron por CD-HIT para generar un conjunto de GIs no redundantes con el fin de permitir una ejecución más eficiente del programa. Este conjunto no redundante también se almacenó en un único archivo de texto en formato multi fastA.

En la segunda sección, se realiza la ejecución propiamente tal de la herramienta Wapi donde se ejecutan 4 programas diferentes. **(3)** El primero de ellos es PROKKA

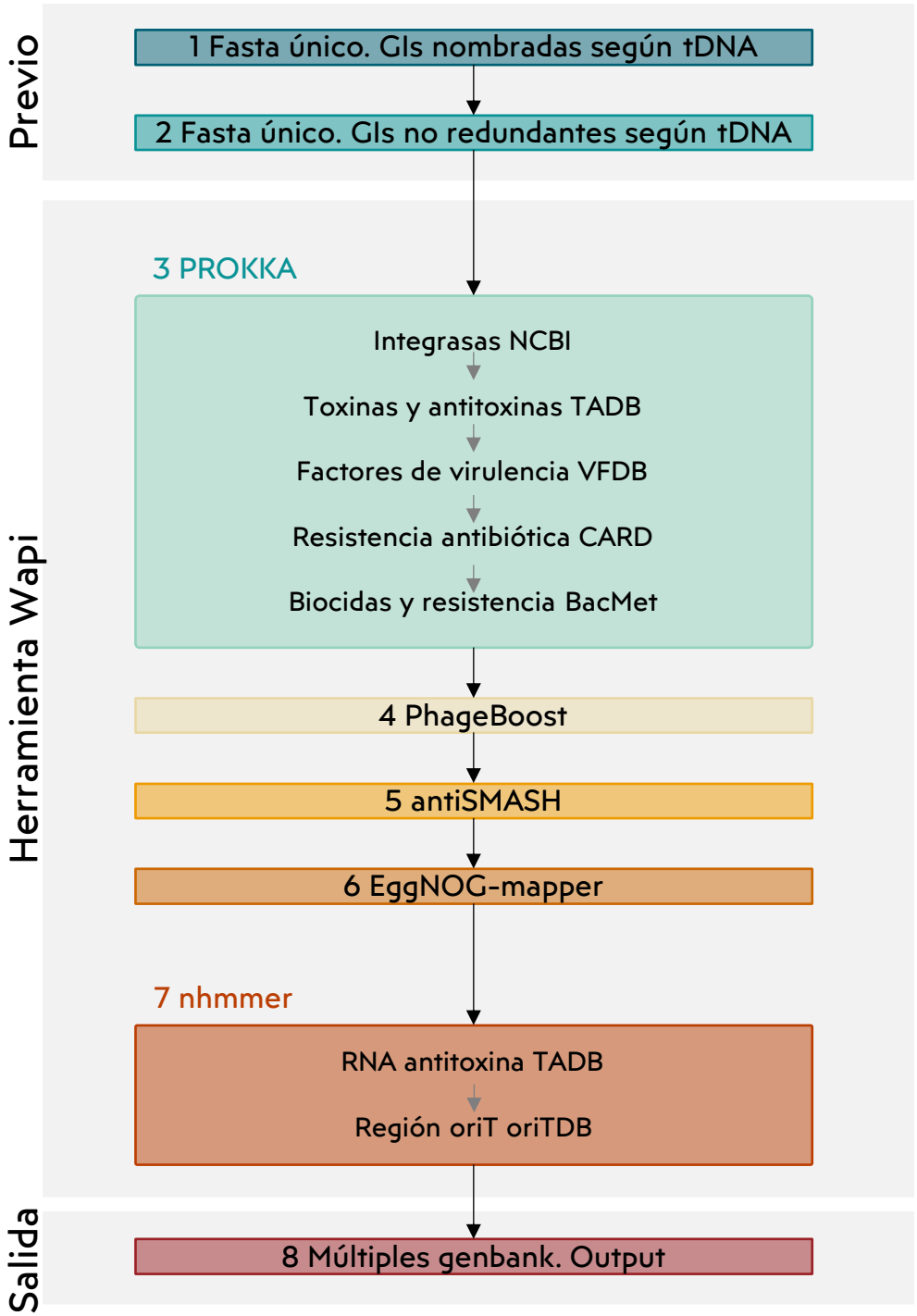

Figura 1. Esquema secuencial del funcionamiento de la herramienta Wapi. El programa utilizado en este trabajo tiene un funcionamiento secuencial dividido en tres grandes secciones: pasos previos, la herramienta Wapi y los archivos de salida.

que toma el archivo de las GIs no redundantes y lo procesa secuencialmente con 5 bases de datos distintas, según el orden ejecutado son: integrasas de NCBI, toxinas y antitoxinas de TADB, factores de virulencia VFDB, resistencia antibiótica CARD y biocidas BacMet2. PROKKA utiliza estas bases de datos para anotar las GIs, yendo de bases de datos más generales a más especializadas. Cuando hay genes que son anotados por más de una base de datos, se mantiene la anotación posterior por sobre la anotación previa. Al final de este procesamiento con PROKKA, se crean archivos de texto en formato genbank que contienen todos los genes predichos, a ellos se le agrega la información de su anotación o, si no pudieron ser anotados, se les identifica con "*hypothetical protein*". A pesar de que solo se mantiene la última anotación en cada gen, igualmente se registra la información si fue anotado por otra base de datos, por si fuera necesario conocer esa información. Estos archivos genbank irán incorporando la información de cada uno de los siguientes análisis hasta el último programa.

Luego, se realiza la predicción de regiones que corresponden a largas secuencias nucleotídicas que contienen varios genes a la vez. Estas regiones son reconocidas si codifican la presencia de profagos o de clusters biosintéticos. **(4)** El programa PhageBoost toma como insumo el mismo archivo de las GIs no redundantes (2) y lo analiza para encontrar regiones compatibles con fagos, similares a las que posee en su base de datos propia. Los archivos genbank incorporarán la información de cada región y de cada uno de los genes que están dentro de ellas. **(5)** Por otro lado, antiSMASH reconoce regiones compatibles con clusters biosintéticos, también tiene su propia base de datos y su información también se almacena en los genbanks correspondientes.

A continuación, **(6)** se analizan los todos genbanks obtenidos hasta el momento y se separan los genes que no hayan podido ser anotados hasta el momento, es decir, que sean identificados como "*hypothetical protein*" para ser procesados por EggNOG-

mapper, un programa especializado en realizar anotaciones funcionales. Este tipo de anotación se diferencia de la que realiza PROKKA porque asocia el gen a una función contenida en su base de datos propia, lo ventajoso de este programa es que es capaz de reconocer genes que no están presentes en ninguna otra base de datos, disminuyendo la cantidad de genes que quedan sin anotar.

(7) El último programa utilizado es nhmmer que reconoce ciertas secuencias nucleotídicas presentes en cualquier parte de los archivos. En primer lugar, se buscan secuencias de RNA presentes en la base de datos de TADB, si no se hiciera así podría perderse información valiosa puesto que PROKKA solo reconoce las secuencias aminoacídicas provenientes de DNA y, en segundo lugar, se buscan secuencias compatibles con la región oriT almacenadas en la base de datos oriTDB.

(8) Como *output* del programa se generarán varios tipos de archivos. Por un lado, se genera un archivo de texto donde se resumen el número de anotaciones realizadas y otros datos básicos sobre la información procesada. Por otro lado, todas las anotaciones realizadas por Wapi son almacenadas en un único archivo genbank que posee la información de todas las GIs no redundantes analizadas. Por comodidad, también se genera una carpeta que tendrá la misma información, pero separados en archivos genbank separados por cada GI.

4.2 Evaluación de la herramienta: comparación con islas genómicas individuales

Para comprobar que los resultados automatizados entregados por esta herramienta son concordantes a los resultados provenientes de análisis publicados anteriormente, se compararon en primer lugar, con la anotación de cuatro islas genómicas previamente reportadas en la literatura y, en segundo lugar, con un grupo de 66 cepas de *K. pneumoniae*.

Estas cuatro islas genómicas tienen su anotación incorporada en una base de datos de libre acceso, así que esos resultados se compararon con la anotación automatizada para probar la herramienta (datos en la Tabla 1 y esquema de los resultados en Figura 3).

En el primer caso, las anotaciones de la isla GIE492 entregadas por la herramienta Wapi fueron las menos similares a las anotaciones de referencia. Aquí se encontraron 4 CDS más que los originales, aunque cada uno con una longitud de unos pocos cientos de nucleótidos. La cantidad de genes a los que se le atribuyó función fue la misma (14), pero no correspondieron a los mismos ORFs. Se tuvieron pocas coincidencias (solo cinco) debido a varias razones: en primer lugar, es que ambas anotaciones tuvieron varios genes sin asignación de función, etiquetados como "*hypothetical protein*" (10 genes en la anotación Wapi y 7 en la referencia), pero entre sí no coincidieron, por lo tanto, no pudieron compararse sus anotaciones, siendo clasificados como "anotaciones diferentes", esto no ocurrió para las islas restantes. Entre los genes que pudieron ser anotados equivalentemente en ambos análisis estuvieron el gen de integrasa, el gen de producción de la microcina, el gen que codifica la inmunidad y el gen identificado en la publicación como u1 que codifica una metiltransferasa. Dentro de las diferencias, Wapi reconoció una región como profago, que no fue etiquetada como tal en la publicación y, además, contiene a los genes que no pudieron asignársele función en la publicación original (identificados con la letra "u" y números correlativos).

Para conocer qué tan importante fue esta diferencia en las anotaciones, en la Figura 2 se muestra un alineamiento de ambas anotaciones: la realizada por Wapi y la publicada como referencia. Aquí se ve que la discrepancia entre ellas no se debió a diferencias entre sus secuencias nucleotídicas (porque son prácticamente idénticas) o a una falla en la predicción y anotación de genes, sino que las bases de datos

utilizadas en cada caso tuvieron grandes diferencias, reflejando que la falta de coincidencia no indica un problema en el funcionamiento de Wapi. Además, se puede ver que la anotación de Wapi, fue más sensible en la identificación de genes presentes en la isla. Esta figura es un ejemplo de lo que ocurrió con las tres islas restantes durante la comparación de las anotaciones realizadas por Wapi y las anotaciones consideradas como referencia, pero como se mencionó anteriormente, en las tres restantes islas, la similitud entre la anotación automatizada de Wapi y la anotación de referencia fue mucho más alta.

Tabla 1. Comparación entre los resultados de la herramienta Wapi y el análisis previo correspondiente. Se muestran los datos para cada uno de los cuatro elementos genéticos móviles analizados: isla genómica GIE492, el elemento ICE IHE3034-1 de *E. coli*, el elemento ICEKp14 de la cepa 16852116 y el elemento ICEKp1. Entre paréntesis, el valor entregado por el análisis de referencia correspondiente.

	GIE492	**IHE3034-1**	**ICEKp14**	**ICEKp1**
N° de genes predichos	25 (21)	24 (24)	51 (49)	58 (52)
N° de genes anotados	14 (14)	22 (22)	44 (40)	45 (39)
N° de genes coincidentes	5	22	36	43
Presencia de integrasa	Sí (Sí)	Sí (Sí)	Sí (Sí)	Sí (Sí)
N° de clusters	1 (0)	1 (1)	2 (2)	3 (3)
N° de transposasas	0 (0)	2 (3)	0 (0)	1 (3)

La anotación del elemento genético móvil IHE3034-1 generada por la herramienta fue casi idéntica a la publicada previamente. Dentro de las similitudes están haber predicho 24 CDS (cada análisis encontró un CDS adicional que no tuvo el otro) donde 22 de las anotaciones fueron exactamente las mismas y una distinta, cuya diferencia es atribuible al algoritmo de predicción de ORFs y no a la base de datos utilizada o a la programación de la herramienta. Se observó que 18 de las anotaciones corresponden a una serie de genes pertenecientes al cluster de síntesis de colibactina (clbA-clbR), los restantes genes se atribuyeron a transposasas, integrasas y un tDNA-asn.

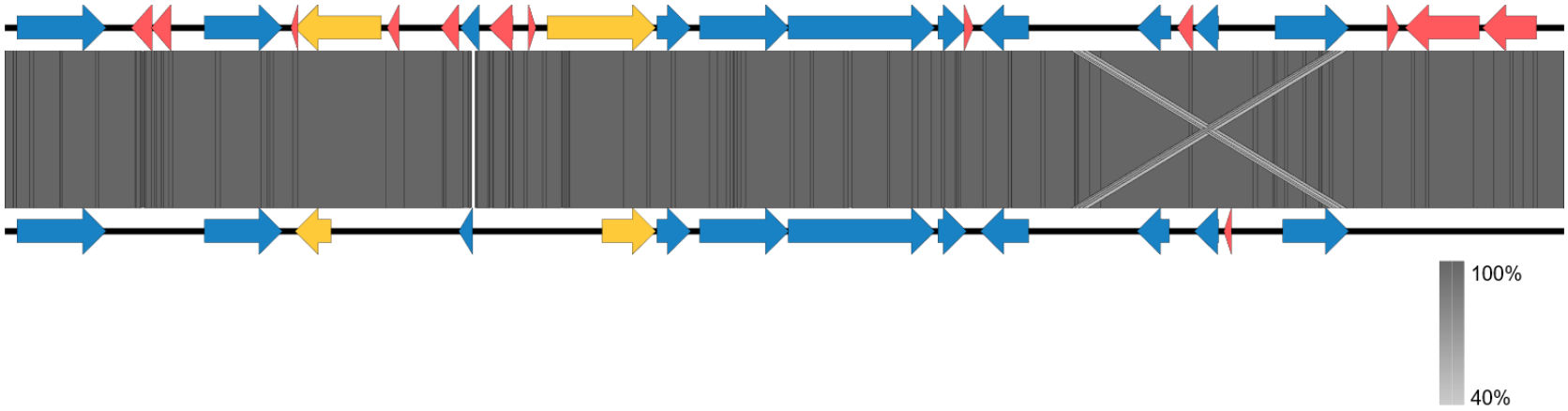

Figura 2. Alineamiento de las dos anotaciones de la isla genómica GIE492. Alineamiento de la anotación realizada por Wapi (arriba) y la anotación de referencia (abajo), las flechas indican los genes predichos en ambos casos y sus colores indican si fueron reconocidos en ambas anotaciones (azul), si fueron reconocidos en ambas anotaciones, pero con longitudes diferentes (amarillo) y si solo fueron identificados en una anotación (rojo). Hay que mencionar que el archivo `.gb` publicado por los autores de la referencia solo indicaba 14 genes y no los 21 genes descritos en la publicación, en esta figura solo se muestran los 14 genes que tuvieron sus coordenadas publicadas.

En el caso del ICEKp14, se reconocieron 51 CDS (dos más que en la publicación original), con la asignación de 39 funciones idénticas o similares; 3 asignaciones nuevas (que estaban como proteínas hipotéticas en la original); y el resto fueron asignaciones no concordantes que se explican por la diferencia de bases de datos utilizadas. En esta GI se detectaron clusters relacionados con la biosíntesis y el receptor de sideróforo (yersiniabactina) y con componentes del sistema de secreción tipo IV.

En el elemento ICEKp1, se identificaron 58 CDS (6 más que en la referencia), lográndose 45 asignaciones de función idénticas o similares y 3 asignaciones nuevas, mientras que las anotaciones restantes fueron diferentes a las originales. Al igual que en la isla anterior se identificaron algunos clusters como el de biosíntesis de yersiniabactina y el del sistema de secreción tipo IV, pero en este caso se encontró, además, el cluster de los genes *iro* correspondiente a la producción de salmoquelinas.

A manera de resumen, la herramienta Wapi anotó, de la misma manera que análisis publicados previamente, un alto porcentaje de los genes (67%), a excepción de la isla GIE492 (sin considerar esta isla, sería el 76% de coincidencias), por lo tanto, con esta pequeña muestra de secuencias nucleotídicas, se evidenció que tiene potencialidad de cumplir con su objetivo: anotar de forma automatizada, entregando resultados similares a anotaciones ya publicadas. Esto significó que la programación de la herramienta logró predecir, prácticamente, los mismos ORFs, e incluso algunos más, y a ellos, logró ir incorporando la información de las bases de datos.

De entre los genes anotados, se destacan integrasas en cada uno de los elementos genéticos móviles, clusters de síntesis de sideróforos y factores de virulencia como el sistema de secreción tipo IV. Es importante mencionar que las islas tuvieron diferentes genes codificados en su interior y diversos orígenes, puesto que provenían no solo de *K. pneumoniae*, sino también de *E. coli* y de muestras de diferentes tejidos y hospederos

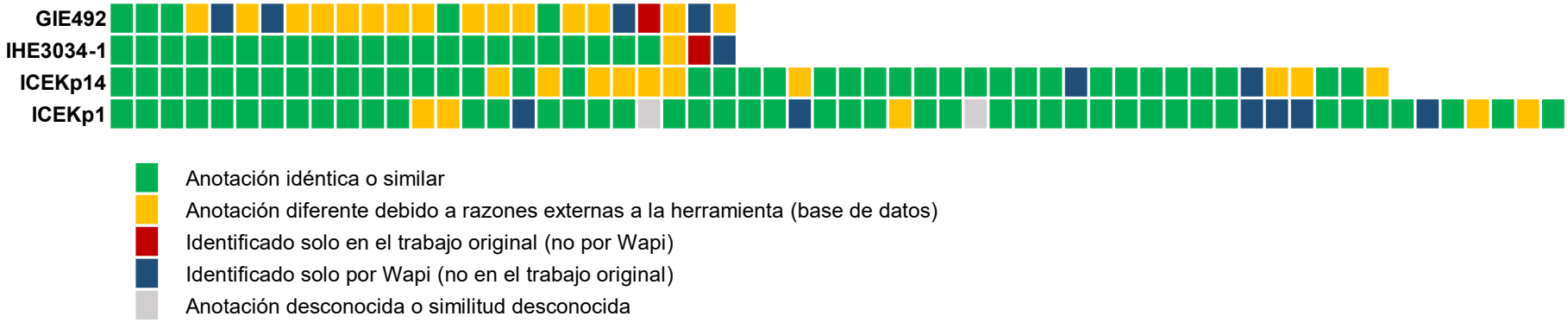

Figura 3. Anotación de genes codificados en los 4 elementos genéticos móviles en comparación a las referencias. En el esquema, cada bloque corresponde a un CDS identificado en cada MGE (número de CDS respectivamente: 25, 25, 51 y 58) y cómo fue la anotación de Wapi en comparación a la anotación de referencia. En verde están los genes que tuvieron una anotación idéntica o similar en Wapi con respecto a la referencia; en amarillo, cuando la anotación de Wapi fue diferente a la de la referencia debido a diferentes bases de datos utilizadas; en rojo, CDS que solo fueron reconocidos por la referencia, pero no por Wapi; en azul, CDS que solo fueron reconocidos por Wapi pero no por la referencia y; en gris, cuando la anotación de la referencia fue desconocida o no se pudo determinar si las anotaciones fueron similares.

demostrando que la programación de la herramienta y las bases de datos utilizadas no están limitadas a una única especie, como es *K. pneumoniae*.

4.3 Evaluación de la herramienta: conjunto de 66 cepas de *K. pneumoniae*

Una vez comprobado el éxito del análisis de algunas islas genómicas publicadas anteriormente, se prosiguió con el análisis de un grupo mayor de islas con el objetivo de, no solo comparar las anotaciones, sino también de generar agrupamiento de islas, para disminuir la redundancia de secuencias y así, hacer más eficiente el procesamiento computacional de conjuntos masivos de datos, lo cual se logró.

4.3.1 Clusterización de las islas

Se utilizó el análisis completo de un conjunto de 66 cepas (Berríos-Pastén et al., 2020) como referencia para comparar la anotación automatizada de Wapi, puesto que previamente se había reportado su anotación manual y cuidadosa. Solo hubo una excepción a esto, la correspondiente a los criterios de la clusterización, porque al inspeccionarla se observó que existían razones para programar el proceso de clusterización de las secuencias debido a que en la curación manual de los datos se consideraron parámetros demasiado laxos, es decir, que permitió dejar a dos secuencias dentro de un mismo cluster aunque tuvieran diferencias sustantivas. En algunos casos observados, se clasificaron dos secuencias como idénticas, a pesar de que existía un fragmento importante que estaba presente en una isla, mientras en la otra no. Esto es particularmente importante dado que cuando se busca obtener una colección no redundante de secuencias, si los criterios son muy laxos, al momento de seleccionar una representante de un subconjunto podrían enmascararse diferencias entre que podrían dar cuenta de la evolución de estos elementos. Por lo tanto, este

punto constituye una mejora importante introducida por la herramienta desarrollada, con respeto al análisis manual realizado previamente. De todos modos, Wapi permite al usuario seleccionar dichos parámetros de modo que se ajusten de mejor manera a los requerimientos particulares que pueda tener su análisis.

En la Figura 4A se muestran, a modo de ejemplo, los clusters asn1D-I y asn1D-IV de las islas genómicas insertadas en tRNA-asn extraídas de la clusterización elaborada por el análisis de referencia. En primer lugar, en el cluster asn1D-I, se ven alineadas las secuencias nucleotídicas correspondientes a cinco islas genómicas, las cuales presentan regiones que están presentes solo en algunas de ellas. Se ve claramente que la primera secuencia es mucho más larga que las cuatro restantes que, a su vez, se alinean de forma muy similar entre sí, pero no con la primera secuencia. Observando las cuatro regiones faltantes en las cuatro últimas secuencias, cuyo largo es aproximadamente 1 kb para cada una, se detectó que corresponden a genes de transposasa, que solo estaban presentes en la primera isla.

Algo similar se observó también en el cluster asn1D-IV. En el análisis anterior se determinó como un único grupo, pero los resultados de este trabajo apuntan a que debería considerarse como dos grupos, un cluster que contiene dos islas y un segundo en el que se agruparon otras cuatro. En este caso, hacer esta distinción resultó ser más importante puesto que los dos segmentos faltantes eran mucho más extensos que en el caso anterior y, de hecho, se predijo la presencia de 16 y 9 CDS, respectivamente (Figura 4A). Asimismo, los genes codificados en esos segmentos estaban relacionados con funciones muy relevantes como factores de virulencia (gen *vagC* y de producción de sideróforos) o resistencia a antibióticos (gen *oleI*). Es importante notar entonces que, si ellos fueran dejados dentro del mismo cluster, estas diferencias podrían ser ignoradas.

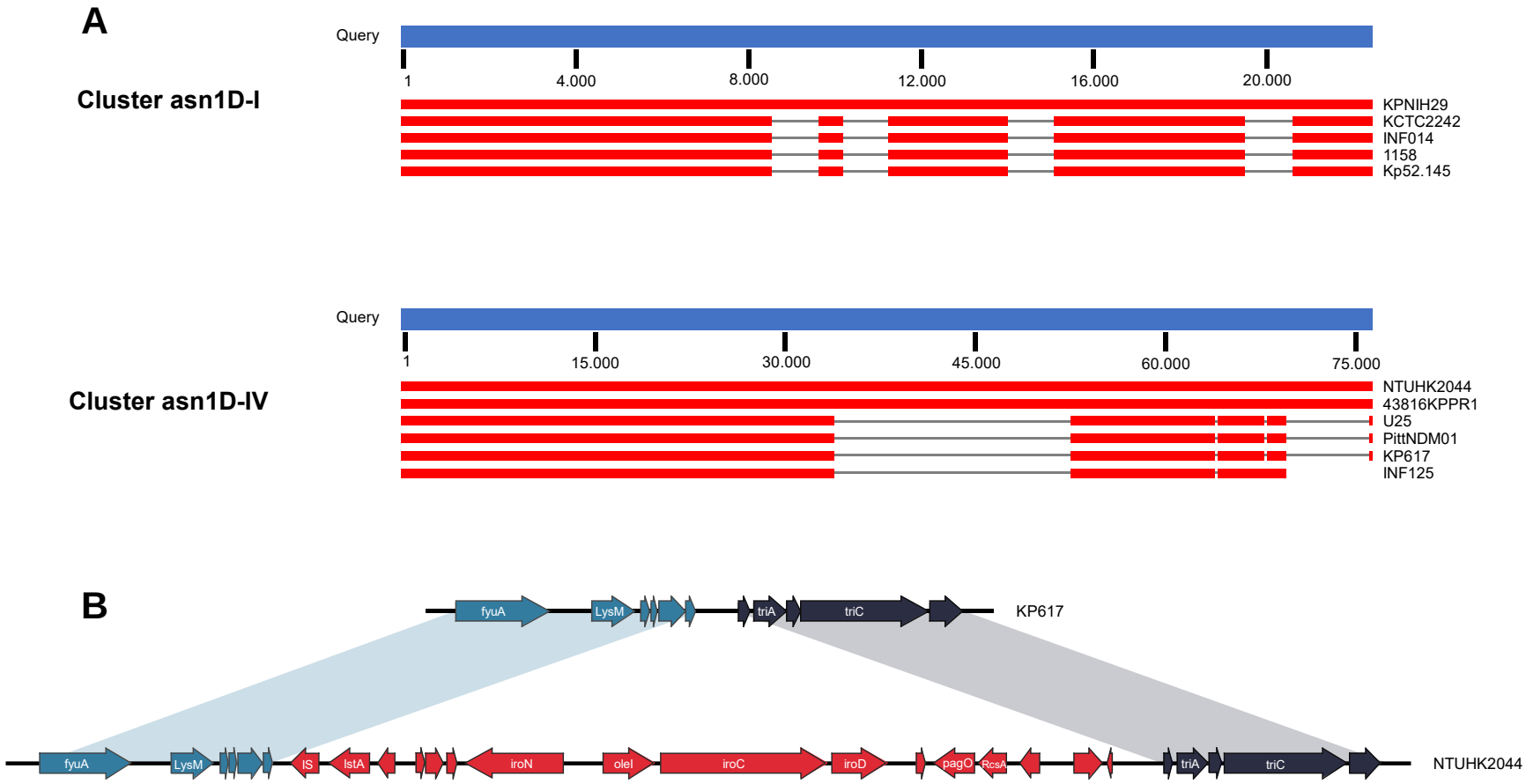

Figura 4. Diferencias entre islas similares que fueron clasificadas por el análisis de referencia como pertenecientes del mismo cluster. **(A)** Alineamiento por Blastn de secuencias de islas genómicas contenidas en los clusters I y IV insertadas en tRNA-asn1D. En cada cluster, la secuencia *query* fue la isla de mayor extensión, para evitar que se ignorara información al realizar el alineamiento. **(B)** Vista en detalle de donde se evidencia la cantidad de genes que son diferentes (en rojo) para islas del cluster asn1D-IV en el análisis de referencia (Berríos-Pastén et al., 2020), arriba la isla KP617 y abajo la isla NTUHK2044.

Para visualizar esto, se muestra en la Figura 4B un fragmento de dos islas del cluster asn1D-IV en las coordenadas donde está presente el segmento con 16 CDS antes mencionado (presente en cepas KP617 y NTUHK2044). En el análisis, estas islas fueron reconocidas como lo suficientemente similares como para dejarlas en un mismo cluster, pero se muestra en la figura en color rojo, un gran segmento que contiene 16 CDS que está presente en una isla, pero en la otra, no. Este segmento de casi 20 kb contiene genes que podrían entregarle una capacidad nueva a la cepa portadora. De hecho, si se observa la anotación de aquellos ORF realizada por la herramienta Wapi, se encontrará que hay genes relacionados con una secuencia de inserción (IS), proteínas asociadas a virulencia, proteína de mantención de plásmido (vapC), receptor de salmoquelina (iroN), transportador ABC (iroC), sideróforo esterasa (iroD) y algunas proteínas hipotéticas. Estas diferencias entre islas similares son valiosas para entender algunos fenómenos asociados a estos elementos genéticos móviles y, por lo tanto, es fundamental poder tener este tipo de islas en diferentes grupos porque así se evita perder información relevante.

Teniendo en cuenta esto, se decidió utilizar parámetros más estrictos para considerar dos secuencias como iguales en la clusterización, con el fin de disminuir al mínimo la ocurrencia de este fenómeno para evitar omisiones o confusiones importantes en los resultados. Hay que recordar que este paso inicial de clusterización se hace con la finalidad de hacer más eficiente el procesamiento de secuencias debido a que el objetivo último de este trabajo es analizar más de mil cepas, lo cual requeriría simultáneamente alto poder de procesamiento computacional y tiempo. Así, evitando repetir el análisis en islas iguales, se puede lograr tener un resultado en un tiempo adecuado sin requerir la utilización de un computador con un altísimo poder de procesamiento.

Determinar los parámetros exactos para lograr este objetivo, siempre es difícil, puesto que no hay un valor único que se pueda obtener de la literatura y que sea generalizable para todos los análisis bioinformáticos, sin importar cuál especie bacteriana se esté analizando. Debido a esto, se tuvo que elegir un valor que estuviera basado en este grupo de datos analizados y, por ello, podría no ser generalizable para otros conjuntos de datos para especies filogenéticamente alejadas. Basado en estas secuencias y buscando mayor estrictez, se eligieron los parámetros para considerar dos secuencias como iguales cuando tuvieran una identidad mayor a 80% y una cobertura de 85% (es decir, una diferencia entre las longitudes de 15%) en el procesamiento por el programa CD-HIT.

Como se esperaba, estos parámetros generaron más clusters en la colección de 66 cepas en comparación al análisis de referencia, es decir, se encontró una mayor diversidad de secuencias, lo que redundó en un mayor número de islas a analizar por la herramienta. Del total de 412 islas de la colección de 66 cepas, se obtuvo un conjunto de 212 islas no redundantes (a diferencia de las 161 del análisis considerado como referencia). Como se ejemplifica en los paneles de la Figura 5, en general, se detectaron más clusters debido a la fragmentación de ciertos clusters del análisis previo. No se observó una mezcla de islas provenientes de clusters diferentes. Esto último se debió principalmente al parámetro de alta cobertura, así dos islas con secuencias muy similares, pero que una tuviera algún fragmento extra (como en el ejemplo de la Figura 4) provocó que el procesamiento las reconociera como diferentes.

Para determinar si esta clusterización fue mejor o peor que la del análisis anterior, se utilizó un alineamiento de las secuencias con BLASTN (Johnson et al., 2008). Tomando nuevamente como ejemplo las islas insertadas en el tRNA-asn1D, se vio que las secuencias, que tienen mayor identidad entre ellas, están juntas en el mismo cluster

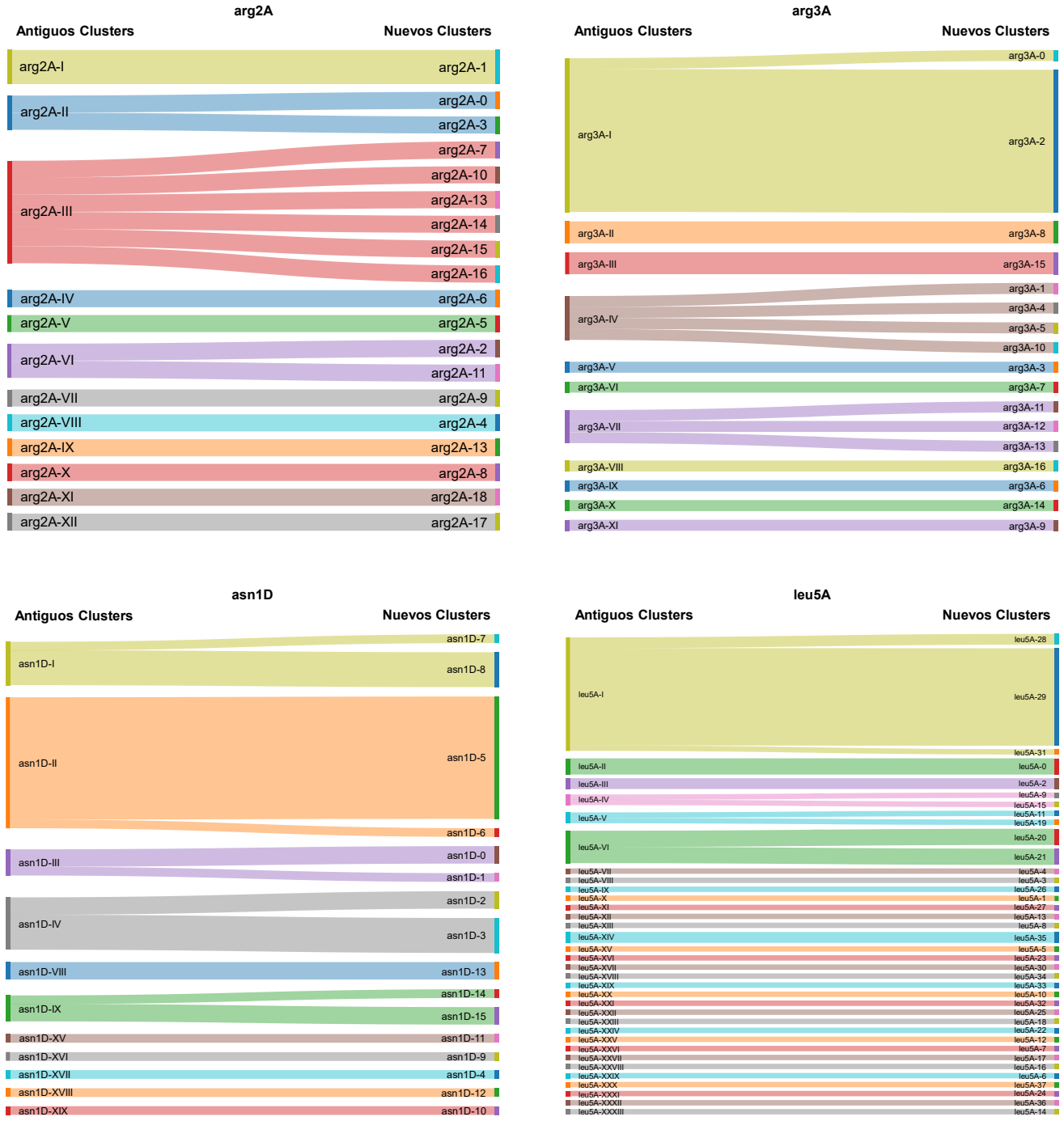

Figura 5. Gráficos de bandas que muestran cómo fue el cambio en la agrupación de las secuencias de islas genómicas entre el análisis de Berríos-Pastén et al. (izquierda) y el presente análisis (derecha). Ejemplo de solo 4 clusters encontrados en la referencia, el grosor de las bandas corresponde a cuántas islas están dentro de cada cluster. Partiendo por arriba a la izquierda: clusters insertados en los tDNA de arg2A, arg3A, asn1D y leu5A. Nótese que la clasificación aplicada en la referencia utilizó números romanos y que la clasificación aplicada en este trabajo utiliza números indoarábigos.

(Tabla 2). También se puede notar que las islas agrupadas por este análisis tienen un Max. Score similar, puesto que esta variable considera en su puntaje, una bonificación por cada base idéntica y una penalización por cada gap o base no coincidente. Por lo tanto, se desprende de estos resultados, que la clusterización realizada en este trabajo cumple mejor su objetivo, puesto que las GIs, dentro de un mismo cluster, son más similares entre sí que en los análisis previos realizados por el grupo, mejorando la identificación de sus genes, aunque esta mejora tenga una pequeña repercusión sobre el tiempo de cómputo necesario para la ejecución de la herramienta bioinformática.

Tabla 2. Datos entregados por la versión web de Blastn para los 4 clusters resultantes del análisis de Berríos-Pastén et al. Obsérvese la similitud en la identidad y en el Max Score de las islas que finalmente quedaron dentro del mismo cluster en el análisis de este trabajo.

Clusterización previa	Clusterización de este trabajo	Cepa hospedera de la GI	Identidad porcentual	Max. Score
asn1D-I	asn1D-7	KPNIH29	100,00%	49.440
	asn1D-8	KCTC2242	99,95%	32.704
		INF014	99,94%	15.597
		1158	99,93%	15.592
		Kp52.145	99,92%	15.586
asn1D-III	asn1D-0	1084	100,00%	2,658e+05
		RJF999	99,93%	2,634e+05
	asn1D-1	PMK1	99,92%	1,500e+05
asn1D-IV	asn1D-2	NTUHK2044	100,00%	1,418e+05
		43816KPPR1	99,90%	1,413e+05
	asn1D-3	U25	99,34%	86.319
		PittNDM01	99,34%	86.327
		KP617	99,34%	86.327
		INF125	99,27%	86.132
asn1D-IX	asn1D-15	960186733	100,00%	9.338
		CG43	99,40%	8.687
	asn1D-14	QMPB2170	98,96%	9.065

4.3.2 Anotación de las islas

El análisis del conjunto de 66 cepas realizado previamente (Berríos-Pastén et al., 2020) dio como resultado que ellas corresponden a cuatro especies diferentes dentro del complejo de especies de *K. pneumoniae* (KpSC). Gran parte de los análisis fueron realizados de forma manual y revisados con gran detalle, por esto, es un buen parámetro para comparar el análisis hecho por la herramienta Wapi desarrollada en este trabajo. A partir de los cromosomas completos de las 66 cepas, se extrajeron las islas genómicas gracias a un método desarrollado en nuestro grupo de investigación, donde se siguió la definición de segmentos de DNA variable adyacente a un t(m)DNA, con un mínimo de longitud de 3,5 kb.

Dado los resultados de la sección anterior, con las secuencias nucleotídicas de las GIs obtenidas se realizó la clusterización, es decir, se formaron grupos de islas similares entre sí para ir descartando secuencias que estuvieran repetidas. Para esto se utilizó el programa CD-HIT con los parámetros de 80% de identidad y 85% de cobertura, obteniéndose 212 islas no redundantes a partir de 412 islas totales después de 27 segundos.

El siguiente paso fue ejecutar la herramienta de anotación solamente sobre el conjunto no redundante de islas para generar la anotación de los genes predichos. Así, se lograron identificar 7.032 CDS entre las 212 islas genómicas, cuyos resultados están resumidos en la Tabla 3. La anotación realizada con PROKKA logró reconocer 75 CDS con la base de datos de integrasas, 62 CDS con la base de datos de toxinas y antitoxinas TADB, 695 CDS anotados con la base de datos de factores de virulencia VFDB, 64 CDS con la base de datos de resistencia antibiótica CARD y 185 CDS con la base de datos de biocidas antibacterianos y resistencia a metales predichos BacMet2.

Tabla 3. Tabla comparativa de los resultados entre los resultados del análisis previo y los resultados obtenidos con la herramienta de este trabajo. Se indican el conjunto no redundante de islas para cada análisis, las anotaciones de los genes e identificación de algunas regiones. Además, se indica el porcentaje de genes que quedaron sin anotar al final del análisis.

	Preprint	Este trabajo
GI no redundantes	161	212
% cobertura	60%	85%
% identidad	80%	80%
GI con integrasas	135 (84%)	133 (63%)
GI con oriT	-	19 (9%)
Profagos	Phaster	PhageBoost
Predicciones	32 intactas	34 profagos
	8 cuestionables	-
	10 incompletas	-
Clusters biosintéticos	-	>25 regiones
Genes (CDS) totales	3.540	7.032
Genes anotados totales	60,4%	77,5%
Primera anotación	eggNOG	PROKKA
N° CDS (%)	679 (36,9%)	1.080 (15,4%)
Segunda anotación	PROKKA	eggNOG
N° CDS (%)	- (23,5%)	2.902 (41,3%)
Genes sin anotar	39,6%	22,6%

Tabla 4. Resumen de los principales resultados entregados por la herramienta Wapi para los análisis de las 66 y 1.004 cepas. Realizado con una clusterización de 80% de identidad y 85% de cobertura. **(*)** Algunos CDS pudieron ser anotados por más de una base de datos.

	66 cepas	1.004 cepas
Islas genómicas no redundantes/totales	212 / 412	2.411 / 7.648
Genes predichos totales	7.032	79.293
CDS anotados como (*):		
Integrasas	75	666
Toxinas y antitoxinas (RNA)	62 (0)	822 (57)
Factores de virulencia	695	8.221
Resistencia a antibióticos	64	1.295
Biocidas o resistencia a metales	185	2.954
Regiones identificadas como:		
Profagos (CDS en su interior)	34 (600 CDS)	358 (6.006 CDS)
Clusters biosintéticos	25	107
Anotación funcional	2.902	29.898
CDS sin anotar (% del total)	1.584 (22,5%)	16.464 (20,8%)
Secuencias de oriT	32	123
Tiempo total de ejecución de herramienta	3 h 35 m	29 h 17 m 10 s

Luego en la identificación de regiones, se reconocieron 34 regiones compatibles con profagos por el programa PhageBoost, que en total contenían 600 CDS. Las regiones biosintéticas identificadas por antiSMASH fueron 25.

Después se tomaron todos los CDS que no pudieron ser anotados, los que en total fueron 5.340 CDS. De este grupo, 2.902 pudieron ser clasificados en alguna categoría COG gracias a la anotación funcional de eggNOG-mapper, mientras, los restantes 1.584 CDS no pudieron ser anotados por ningún programa, es decir, el 22,5% del total de genes predichos.

Como último paso de la herramienta, con el uso del programa nhmmer se pudieron identificar en total 32 secuencias de oriT entre todas las islas y ninguna antitoxina RNA según la base de datos TADB. La herramienta demoró en total 3 horas y 35 minutos.

Se observaron resultados similares a los conseguidos previamente, aunque hubo ciertas discrepancias relevantes. Las diferencias se deben principalmente a las bases de datos utilizadas o al momento en que se descargaron. Sobre la información nueva que fue incorporada en este trabajo, se puede mencionar las secuencias oriT debido a que en el análisis anterior no se ocupó una base de datos similar a esta para encontrar estas secuencias. Por otro lado, se encontraron nuevas o diferentes anotaciones en genes relacionados con factores de virulencia o resistencia a antibióticos, principalmente porque las bases de datos consultadas tenían más información y más reciente por la diferencia de tiempo que pasó entre el análisis original y este trabajo. Indicando que, como en todo análisis bioinformático, hay una parte importante de los resultados de los análisis que depende de elementos externos, como la actualización periódica de las bases de datos existentes.

Es importante también indicar el orden de utilización de las bases de datos, ya que, a diferencia de los análisis previos del grupo, inicialmente se utilizaron varias bases de

datos especializadas para anotar los genes, debido a que ellas son confiables porque tienen una dedicación exclusiva en su ámbito de registro, actualizando regularmente la información disponible. En este trabajo se dio prioridad a esta información por sobre otras más generales. A diferencia del análisis anterior que se le dio prioridad a la anotación funcional realizada por eggNOG-mapper, dejando estas bases de datos solamente para los genes que no pudieran ser anotados, en este trabajo, se dejó este programa para los genes que no tuvieran ninguna anotación, por dos razones, primero porque así quedarían menos genes a analizar por este programa, el cual requiere una mayor cantidad de tiempo para analizar. Y, en segundo lugar, porque como ya se mencionó, estos datos son más generales.

Esto no significa que el programa no tenga una buena calidad sino, todo lo contrario, tiene información que otras bases de datos no tienen. De hecho, hay una categoría especial, llamada categoría S, donde agrupa la información de genes cuando existe información escasa, es decir, se conoce la proteína que se genera, pero no su función exacta. Por esta última razón, eggNOG-mapper se utilizó en un paso posterior a las anotaciones iniciales porque la información de la categoría S, es mejor que se utilice en CDS que no hayan podido ser identificados por ninguna otra base de datos.

4.3.3 Análisis del genoma núcleo y del islotipo

Este conjunto de 66 cepas también permite deducir transferencias de islas por vía vertical y horizontal. Con este fin, primero se generó un árbol filogenético a partir del genoma núcleo (*core genome*, en inglés) de las 66 cepas, como se ve a la izquierda en la Figura 6. En este árbol se distinguen tres ramas principales, las que están coloreadas en azul, verde y rojo, lo que indica tres grandes grupos de cepas de *K. pneumoniae* que, a su vez, se pueden asociar con la clasificación de Kp. En la rama azul hay cepas de

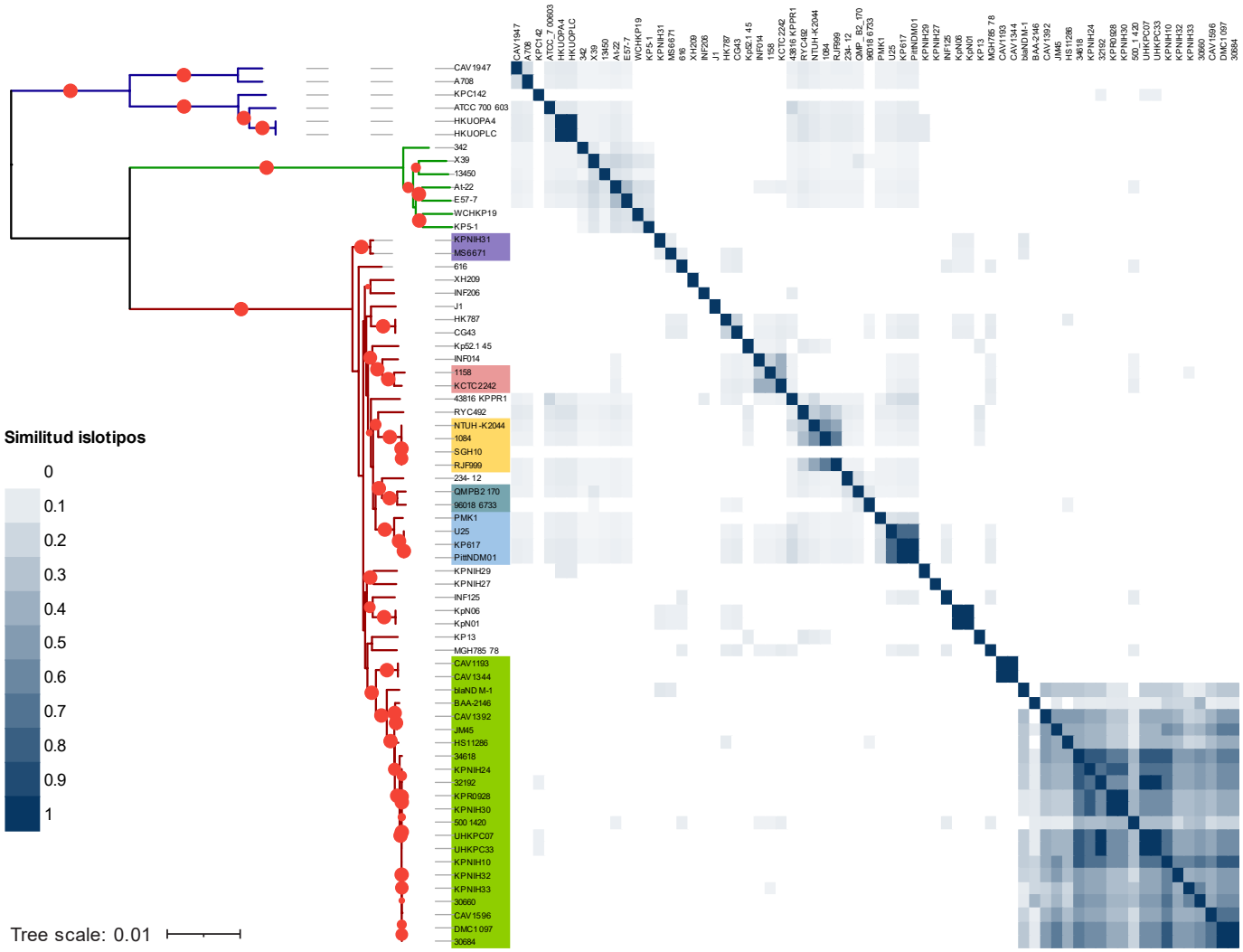

Figura 6. Árbol filogenético construido en base al genoma núcleo de las 66 cepas de *K. pneumoniae* analizadas y su relación con las islas insertadas en ellas. A la izquierda, árbol filogenético confeccionado con el genoma núcleo de las 66 cepas analizadas, se colorearon las tres ramas principales (azul, verde y rojo). Sobre el árbol se indica el bootstrap mayor a 75%. A la derecha, heatmap con la similitud de los islotipos entre cada par de islas, el color es más intenso cuando la similitud es mayor y viceversa. Se indican algunos grupos clonales (CG) en cuadros coloreados sobre el nombre de las cepas correspondientes, de arriba hacia abajo son: CG147 en morado, CG65 en rosado, CG23 en amarillo, CG43 en verde azulado, CG14/15 en celeste y CG258 en verde.

Kp2 (cepa CAV19547) y Kp4 (A708), en la rama verde Kp3 y en rojo, Kp1. En las ramas azul y verde hay una cierta diversidad que permite diferenciar las cepas dentro de cada grupo, mientras en la rama roja hay un grupo donde ocurre lo mismo, pero además, se ve un grupo de cepas muy cercanas, lo que indica que a pesar de que provienen de diferentes muestras, son cepas altamente similares entre sí.

A la derecha de la Figura 6, se muestra cuán similares son los islotipos (conjunto particular de islas) insertados en cada cepa. Tal como se mencionó en los materiales y métodos, el valor que representa la similitud se calculó como la distancia de Jaccard para cada par, es decir, la razón entre el número de islas que comparten y el total de islas que tienen entre ambas cepas. Para mostrar esto se utiliza un *heatmap* que indica con la intensidad de color cuán similar es el islotipo de una cepa, en comparación con todo el resto de las cepas. Así se ve que en torno a la diagonal de la figura hay colores más intensos, lo que muestra que cuando las cepas son cercanas, tienen islas similares entre sí. Esto se aprecia principalmente en las últimas cepas del árbol (abajo, destacadas en color verde, indicativo de que son parte del grupo clonal 258) donde se observa que cada cepa tiene una alta similitud con un gran grupo de cepas cercanas, lo que permite deducir que este fenómeno se debería en gran parte a la transferencia vertical de las islas, puesto que todas ellas tendrían un ancestro común cercano que portaba estas islas.

Además, se puede observar que hay algunas cepas que tienen similitud en su islotipo con otras cepas alejadas del árbol, como es el caso de las cepas CAV1947 y A708, que tienen cercanía con otras cepas de la rama azul, con cepas de la rama verde (cepas 342, X39, 13450, At-22 y E57-7) e incluso con otras cepas más alejadas de la rama roja (43816KPPR1, RYC492, NTUH-K2044, 1084, entre otras) lo que hace poco probable la transferencia vertical. Así, permite suponer que islas similares se

transfirieron horizontalmente entre estas cepas o que cada cepa recibió islas similares de forma independiente a partir de otras cepas donantes.

Varias cepas están destacadas en cuadros con colores que indican a cuál grupo clonal (CG, por sus siglas en inglés) pertenecen. Yendo de arriba hacia abajo, las cepas en morado pertenecen al CG147, en rosado a CG65, en amarillo a CG23, en verde azulado a CG43, en celeste a CG14/15 y en verde a CG258. Se observa que las cepas integrantes de cada grupo se mantienen juntas dentro del árbol, lo cual es lo esperado porque son cepas muy similares.

4.4 Análisis de una colección ampliada de cepas de *K. pneumoniae*

Una vez comprobado que la herramienta entrega anotaciones informativas para grupos de cientos de islas genómicas, se procedió a analizar un conjunto ampliado de cepas de *K. pneumoniae* y algunas otras especies dentro del género. Primero, se predijeron las islas genómicas entre los genomas obtenidos a partir del filtro realizado a la base de datos de NCBI, usando nuestra herramienta Kintun-VLI. Como resultado se obtuvieron 7.648 GIs, observando que las cepas tuvieron entre 25 y 2 islas genómicas cada una, con un promedio de 8 islas por cepa, con una desviación estándar de 3 islas, aproximadamente. El largo de los cromosomas se observó entre 4,57 Mpb y 6,35 Mpb con un promedio de 5,36 Mpb. Las islas correspondieron a entre un 0,09% y un 11,62% de la extensión del genoma, con un promedio de 3,25%. Se observó una correlación positiva entre la longitud de los cromosomas y el número de GIs por cromosoma y la longitud total de ellas por cromosoma, también una mayor cantidad de CDS en función de la longitud de las GIs. Todo esto apoya que los cromosomas más largos contienen más GIs.

Se realizó una clusterización por medio de CD-HIT con 80% de identidad y 85% de cobertura, obteniendo un conjunto no redundante de 2.411 islas, del total de 7.648 GIs. Dentro del conjunto no redundante se identificaron 79.293 CDS, cuyos resultados están resumidos en la Tabla 4. La anotación de ellos realizada con PROKKA entregó 666 CDS identificados como integrasas, 822 por la base de datos de toxinas y antitoxinas TADB, 8.221 como factores de virulencia por VFDB, 1.295 CDS relacionados con resistencia a antibióticos por CARD y 2.954 por la base de datos de genes predichos de biocidas antibacterianos y resistencias a metales BacMet2.

En cuanto a la identificación de regiones, en primer lugar, se reconocieron 358 probables profagos que contenían en total 6.006 CDS. En segundo lugar, también se reconoció la presencia de 107 regiones compatibles con clústeres de genes biosintéticos dentro de 106 islas genómicas diferentes. Considerando el total de las 107 regiones, mayoritariamente los CDS fueron catalogados como NRPS (Non-ribosomal peptide synthetase cluster) y T1PKS (Type I Polyketide synthase). En resumen, las regiones biosintéticas identificadas fueron asociadas a 56 regiones de producción de yersiniabactina (N° de acceso MIBiG BGC0001055), 19 asociadas a la biosíntesis de N-miristoil-D-asparragina (BGC0000972), 10 asociadas a turnerbactina (BGC0000451), 1 región a gobiquelina A y B (BGC0000366), 1 región a amonabactina (BGC0001502), 1 región a rifamicina (BGC0000137), 1 región a tilivallina (BGC0000446), y 18 regiones a otros clusters no especificados.

A partir de los 54.855 CDS no anotados y el procesamiento de eggNOG-mapper, se pudo anotar funcionalmente 29.898, mientras 16.464 quedaron como proteínas hipotéticas (20,8%).

Finalmente, nhmmer identificó 123 secuencias de oriT y 57 antitoxinas RNA entre todas las islas no redundantes. La herramienta demoró 29 horas, 17 minutos y 10 segundos en realizar todo el análisis.

El tiempo de procesamiento estuvo dentro de los valores esperados para un análisis de esta magnitud. El servidor utilizado para realizar este análisis es más poderoso que un computador personal, pero no es inalcanzable para cualquier laboratorio microbiológico que esté estudiando cepas similares, así que, en ese sentido, es una herramienta adecuada para realizar una anotación general de un conjunto masivo de islas genómicas.

La cantidad de genes que no pudieron ser anotados por Wapi, fue un poco menor a la anotación automatizada de las islas encontradas en las 66 cepas (20,8% vs 22,5%), pero fue bastante menor a la anotación manual realizada de esas mismas 66 cepas (39,6%). Por lo tanto, al comparar resultados, a la herramienta no solo le tomó poco tiempo, sino que también, pudo lograr una alta proporción de genes anotados, principalmente gracias a que combina diferentes bases de datos y programas de anotación. En este punto es necesario recordar que, dentro de la comunidad científica, la anotación de genes bacterianos sigue siendo una materia activa de investigación. Dado que aún no ha sido descrita la función de una gran cantidad de genes, entonces, que en este trabajo queden sin anotar un poco más del 20% de los genes analizados, es una cantidad razonable.

Con respecto a los resultados de la anotación, se encontró una cantidad apreciable de genes asociados a virulencia (según VFDB) porque fueron identificados como tales un poco más de un 10% de los CDS y en relación con los genes de resistencia a antibióticos, a metales o biocidas antibacterianos fue casi el 5% de los CDS (CARD + BacMet2). En el ámbito de las regiones identificadas, casi el 15% fueron

predichas como profagos y un poco más del 4% relacionadas con clusters biosintéticos.

Analizando los resultados generales obtenidos y comparándolos con la información disponible, podría concluirse que las islas genómicas portan solo algunos de los genes de resistencia (hay mayor cantidad en plásmidos), en cambio, son importantes pues serían los principales reservoreos de genes de virulencia en la especie. Esto se confirmaría, en parte, con los estudios de estos elementos en plásmidos. Sin embargo, dos aspectos claves de estos resultados deben ser revisados antes de concluir lo anterior. Por un lado, los resultados están sesgados por la información disponible en las bases de datos, por lo que la función específica de proteínas u otros elementos codificados en los elementos genéticos móviles no puede ser asociada directamente con los fenotipos. De hecho, entre las islas genómicas es posible encontrar sistemas de secreción de proteínas incompletos, principalmente del tipo IV y tipo VI, ¿Son, en efecto, sistemas de secreción truncos o se tratan de otras funciones mal anotadas? ¿Preservan al menos parte de su funcionalidad?

Por otro lado, los fenotipos de relevancia clínica podrían no estar necesariamente determinados por la presencia o ausencia de algunos genes, sino que también por la regulación de su expresión, que los genes codificados en islas participen en este tipo de procesos podría explicar la presencia frecuente de proteínas con funciones asociadas a la regulación transcripcional y a la modificación de la conformación del nucleoide bacteriano. Finalmente, conviene considerar estos resultados como preliminares o como punto de partida para el análisis en profundidad de estos elementos en la evolución de las cepas de *K. pneumoniae*.

5. DISCUSIÓN

5.1 Clusterización

La disminución de los costos de la secuenciación está permitiendo que cada año se publique una cantidad cada vez mayor de datos. Para poder analizar de forma más rápida y precisa las secuencias acumuladas, ha sido necesario utilizar herramientas que agrupen las secuencias y puedan identificar las que estén repetidas para no redundar innecesariamente en los análisis. Esta clusterización es un paso muy importante previo a la anotación de las secuencias. En este trabajo se utilizó el programa CD-HIT (Fu et al., 2012) que ha sido ampliamente utilizado en estudios similares, eso sí, conociendo las limitaciones que tienen este tipo de programas porque debido a su naturaleza de algoritmo incremental voraz (*greedy incremental algorithm*) el orden en que procesan los datos impacta en el resultado final (Zou et al., 2018). En este trabajo, además, se utilizaron valores *ad hoc* de identidad y cobertura debido a que fueron los que se adaptaron de mejor manera a las secuencias analizadas. En otros conjuntos de datos, estos valores podrían servir para secuencias similares –es decir, provenientes de especies cercanas– y para taxones más lejanos habría que evaluar caso a caso si se ajustan o no. Para estos casos, en esta herramienta da la opción de modificar los valores de cobertura e identidad, para ajustarse a las necesidades del usuario.

Aún así, se pudo mostrar que es preferible utilizar valores más estrictos para que secuencias, lo suficientemente diferenciadas, no queden dentro de los mismos clusters, perdiéndose información valiosa. En el conjunto analizado en este trabajo (Berríos-Pastén et al., 2020) se logró verificar que esta mayor estrictez no altera la

estructura general de los clusters, mezclando sus secuencias, sino que solamente separa algunos grupos, aumentando su número final. Es más, esto mejoró la identidad de las secuencias dentro de cada grupo, lo que produce una mejor clusterización. Un ejemplo de esto se ve en la Figura 4, donde las islas dentro de un mismo cluster, alineaban perfectamente sus secuencias, aunque tenían ciertos fragmentos que estaban presentes en unas, pero no en otras. Analizando los genes anotados en esos segmentos se vio la presencia frecuente de transposasas, evidenciando la presencia de secuencias de inserción y transposones. En el caso de haber utilizado un criterio menos estricto, podría haberse pasado por alto estas pequeñas diferencias, perdiendo información eventualmente útil para análisis posteriores.

5.2 Anotación

La anotación automatizada de secuencias depende principalmente de dos variables: el programa y la base de datos utilizada. Existen varios programas para realizar la anotación y entre ellos las principales diferencias son el algoritmo de identificación de los marcos abiertos de lectura (ORFs), el tiempo que toma en procesar todas las secuencias y si logra asociar las secuencias traducidas de los ORFs con las proteínas de las bases de datos, es decir, dependen en gran medida de estas últimas. Asimismo, las bases de datos son fundamentales porque si ellas son limitadas en cantidad de información o están desactualizadas, la anotación será deficiente. Por esta razón, se usó en este trabajo diferentes bases de datos que estuvieran especializadas en diferentes ámbitos, como factores de virulencia, resistencia a antibióticos, toxinas-antitoxinas, integrasas, entre otras, para lograr una anotación lo más completa posible que abarcara una amplia variedad de funciones. Se sabe que las bases de datos están limitadas a la información publicada hasta el momento y que muchos genes de

K. pneumoniae son desconocidos porque algunos de sus ambientes han sido poco estudiados, por esto con el paso del tiempo podrán obtenerse resultados más completos y actualizados utilizando la misma herramienta Wapi. Dada esta situación, se realizó una anotación funcional que complementara las bases de datos utilizadas, disminuyendo todo lo posible la cantidad de proteínas sin anotar.

Al comparar la cantidad de genes que pudieron asociarse a alguna función en este trabajo con el estudio de referencia de las 66 cepas (Berríos-Pastén et al., 2020), se pudo ver que disminuyó significativamente el número de genes sin anotación (de un 39,6% a un 22,6%), además, de obtener información nueva como las secuencias oriT. En el conjunto de las 1.004 cepas la cantidad de genes sin lograr anotar por Wapi estuvo cercano al anterior (20,8%).

Al ejecutar la herramienta Wapi con las cuatro GIs individuales, quedó en evidencia la gran similitud entre sus resultados con los resultados previos considerados. Si se analiza el número de CDS predichos, fueron bastante similares (como se aprecia en la Tabla 1), las pequeñas diferencias pueden deberse a los criterios aplicados por los investigadores cuando se analizaron detenidamente los resultados de forma manual. Un ejemplo de esto es la isla GIE492 donde Wapi identificó 2 CDS cuya longitud variaba entre 100 y 300 pb que no fueron declarados como tal en la publicación correspondiente (Marcoleta et al., 2016) o en ICEKp1 con 3 CDS que no superaban los 300 pb que tampoco fueron identificados como tal (Lam et al., 2018). A pesar de que no está mencionado en las publicaciones respectivas, es posible que estos CDS hayan sido identificados, pero no fueran declarados debido a su escasa longitud porque no serían capaces de codificar proteínas funcionales. También es importante mencionar que en los cuatro MGEs se pudo reconocer la presencia de integrasas, es decir, estas islas son potencialmente móviles, al escindirse y luego integrarse en el cromosoma de

otra célula gracias a la actividad de dichas enzimas. Por otra parte, las pocas discrepancias en la anotación de las secuencias se debieron a las diferentes bases de datos utilizadas entre la publicación de referencia y este trabajo. La utilización de versiones actualizadas de PROKKA, VFDB, CARD, eggNOG y otras, produjeron algunas discordancias entre las anotaciones debido a que cada base de datos, utilizada en este trabajo, tenía más información que las utilizadas en los estudios de referencia. Por esto cuando las diferencias entre la publicación y este trabajo se debieron a lo actualizadas que estaban las bases de datos, es preferible utilizar la información más actual, como en estos últimos casos. Pero cuando las diferencias entre ellas se deben a diferentes herramientas, se debiera considerar a los resultados como complementarios.

5.3 Comentarios finales

El poder de procesamiento computacional necesario para analizar los datos estudiados en este trabajo no es excesivamente alto. Este procesamiento podría hacerlo un computador accesible para la mayoría de los laboratorios que se dedican a estudios bioinformáticos en la actualidad. Para ejemplificar, se utilizó un computador con CPU Intel Xeon Silver 4114 2.20 GHz (10 núcleos) y 96 Gb RAM, al cual el análisis de las 4 islas genómicas iniciales le tomó un tiempo de 19 minutos y 14 segundos. Para el conjunto de 66 cepas que contenía 212 GIs no redundantes demoró 3 horas y 35 minutos, mientras que para el conjunto de 1.004 cepas que poseía 2.411 GIs no redundantes demoró 29 horas y 17 minutos. Por lo tanto, estos tiempos son razonables dado a la gran magnitud de información que se procesa y si se compara con el tiempo que se requeriría para realizar el mismo análisis de forma manual, hay una enorme diferencia, permitiendo aprovechar el tiempo ahorrado en análisis de los resultados y

no en el simple procesamiento de los datos. El tiempo de computo escala de manera lineal con el número de islas usado como *input*.

Herramientas bioinformáticas, como la desarrollada en este trabajo, serán las que permitirán en el futuro profundizar el conocimiento que se tiene sobre las diferentes cepas de *K. pneumoniae*, así como también, ampliar el conocimiento que se tiene sobre los elementos genéticos móviles (islas genómicas, en particular) y la relación entre ellas en la aparición de cepas hipervirulentas y/o multirresistentes. Dentro de lo que queda por descubrir está la forma en que se producen muchos fenómenos relacionados con la transferencia genética horizontal, solo por mencionar algunos está: explicar el porqué de la inserción preferente de las GIs en los t(m)DNA, la razón de porqué algunos t(m)DNA son *hotspots* de inserción de GIs, mientras que otros t(m)DNA similares no lo son. Además, aún queda por resolver la relación de las KpSC con el resto de las enterobacteriales y otras gramnegativas que se puede plantear en la pregunta ¿cuáles son los límites entre ellos, si es que los hay, al momento de intercambiar GIs y los elementos que se codifican en ellas?

5.4 Proyecciones

Como se vio en la Figura 4, hay algunas islas genómicas muy similares entre sí, pero que difieren en ciertas regiones. En estos casos, estas regiones correspondieron a secuencias de inserción que podrían o no tener un impacto en el fenotipo de la cepa que porta la GI. Un análisis bioinformático como este, es insuficiente por sí solo para asegurar si esto provocará un cambio en el comportamiento de la bacteria, pero es un buen punto de partida para estudios de laboratorio cuyo fin sea el de averiguar la verdadera función de aquellos fragmentos. Esto es solo un caso particular, pero podrían darse casos similares en otros conjuntos de cepas.

El procesamiento de clusterización y de anotación automatizada de MGE, no es el fin del estudio de estas cepas, sino es el inicio de estudios más profundos que podrán realizarse de forma más eficiente debido al procesamiento automatizado, evitando el laborioso estudio manual de las islas que se debía hacer hace algunos años. Un ejemplo de los análisis posteriores que se pueden hacer con estos datos está el dendrograma de la Figura 6 que indica la relación filogenética entre el genoma de las cepas y la comparación con la similitud de las GI insertadas en ellas, lo que permite analizar diferentes tipos de transferencia genética (vertical y horizontal). El análisis bioinformático podrá ser profundizado con el uso de otras herramientas bioinformáticas al evaluar los genes codificados en las islas, lo cual podría ser de gran ayuda para la vigilancia permanente que debiera existir sobre las cepas de *K. pneumoniae* y varias otras especies en nuestro país para advertir de la llegada de cepas hipervirulentas y/o multirresistentes que pondrán un gran desafío al sistema de salud.

6. CONCLUSIONES

1. Se creó una herramienta de clasificación, comparación y anotación de islas genómicas que es capaz de integrar resultados de múltiples programas y bases de datos en una única ejecución.

2. La herramienta se utilizó satisfactoriamente para analizar tres sets de islas genómicas. Para los dos primeros sets (de 4 islas individuales y de 66 cepas), los resultados de la anotación fueron altamente similares al compararlos con los análisis publicados. Y además, permitió analizar, por primera vez, un gran número de islas genómicas provenientes de un conjunto ampliado de cepas del género *Klebsiella* correspondiente a 1.004 cepas.

3. El análisis de las 7.648 islas genómicas, provenientes de las 1.004 cepas estudiadas, permitió identificar 2.411 tipos de islas diferentes (denominadas como conjunto no redundante de islas). Entre sus 79.293 CDS se encontraron diferentes genes relacionados con la codificación de integrasas, toxinas, antitoxinas, factores de virulencia, resistencia a antibióticos, biocidas y resistencia a metales. Además, se identificaron regiones que abarcaban varios genes a la vez relacionadas con profagos y clústeres biosintéticos, incluyendo secuencias de oriT.

4. El resultado de la anotación de genes codificados dentro de islas genómicas depende, en parte, de la base de datos utilizada y cuán actualizada está. Así se pudieron identificar más genes que análisis publicados en el pasado, debido simplemente a que las bases de datos incorporan en la actualidad más información.

Por lo tanto, en el futuro esta misma herramienta podrá tener mejores resultados si se utilizan bases de datos actualizadas.

7. BIBLIOGRAFÍA

Acevedo, R. (2019). *Desarrollo de una herramienta informática para la identificación de Islas Genómicas asociadas a genes que codifican tRNAs en el patógeno bacteriano Klebsiella pneumoniae* [Seminario de título, Ingeniería en Biotecnología Molecular, Facultad de Ciencias, Universidad de Chile].

Alcock, B. P., Raphenya, A. R., Lau, T. T. Y., Tsang, K. K., Bouchard, M., Edalatmand, A., Huynh, W., Nguyen, A. L. v., Cheng, A. A., Liu, S., Min, S. Y., Miroshnichenko, A., Tran, H. K., Werfalli, R. E., Nasir, J. A., Oloni, M., Speicher, D. J., Florescu, A., Singh, B., … McArthur, A. G. (2020). CARD 2020: Antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Research*, *48*(D1), D517–D525.

Beaber, J. W., Hochhut, B., & Waldor, M. K. (2004). SOS response promotes horizontal dissemination of antibiotic resistance genes. *Nature*, *427*(6969), 72–74.

Bengoechea, J. A., & Sa Pessoa, J. (2019). Klebsiella pneumoniae infection biology: Living to counteract host defences. *FEMS Microbiology Reviews*, *43*(2), 123–144.

Berríos-Pastén, C., Acevedo, R., Arros, P., Varas, M. A., Wyres, K. L., Lam, M. M. C., Holt, K. E., Lagos, R., & Marcoleta, A. E. (2020). Properties of genes encoding transfer RNAs as integration sites for genomic islands and prophages in Klebsiella pneumoniae. *BioRxiv*. https://doi.org/10.1101/2020.11.02.365908

Blázquez, J., Couce, A., Rodríguez-Beltrán, J., & Rodríguez-Rojas, A. (2012). Antimicrobials as promoters of genetic variation. En *Current Opinion in Microbiology* (Vol. 15, Issue 5, pp. 561–569).

Blin, K., Shaw, S., Kloosterman, A. M., Charlop-Powers, Z., van Wezel, G. P., Medema, M. H., & Weber, T. (2021). AntiSMASH 6.0: Improving cluster detection and comparison capabilities. *Nucleic Acids Research*, *49*(W1), W29–W35.

Botelho, J., & Schulenburg, H. (2021). The Role of Integrative and Conjugative Elements in Antibiotic Resistance Evolution. En *Trends in Microbiology* (Vol. 29, Issue 1, pp. 8–18). Elsevier Ltd.

Burrus, V., Pavlovic, G., Decaris, B., & Guédon, G. (2002). Conjugative transposons: the tip of the iceberg. En *Molecular Microbiology* (Vol. 46, Issue 3).

Burrus, V., & Waldor, M. K. (2004). Shaping bacterial genomes with integrative and conjugative elements. En *Research in Microbiology* (Vol. 155, Issue 5, pp. 376–386). Elsevier Masson SAS.

Bush, K., Courvalin, P., Dantas, G., Davies, J., Eisenstein, B., Huovinen, P., Jacoby, G. A., Kishony, R., Kreiswirth, B. N., Kutter, E., Lerner, S. A., Levy, S., Lewis, K., Lomovskaya, O., Miller, J. H., Mobashery, S., Piddock, L. J. V., Projan, S., Thomas, C. M., … Zgurskaya, H. I. (2011). Tackling antibiotic resistance. *Nature Reviews Microbiology*, *9*(12), 894–896.

Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P., & Huerta-Cepas, J. (2021). eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Molecular Biology and Evolution*, *38*(12), 5825–5829.

Cheng, D.-L., Liu, Y.-C., Yen, M.-Y., Liu, C.-Y., & Wang, R.-S. (1991). Septic Metastatic Lesions of Pyogenic Liver Abscess. *Archives of Internal Medicine*, *151*(8), 1557.

Dionisio, F., Zilhão, R., & Gama, J. A. (2019). Interactions between plasmids and other mobile genetic elements affect their transmission and persistence. En *Plasmid* (Vol. 102, pp. 29–36).

Duval, R. E., Grare, M., & Demoré, B. (2019). Fight against antimicrobial resistance: We always need new antibacterials but for right bacteria. En *Molecules* (Vol. 24, Issue 17). MDPI AG.

Fang, C. T., Lai, S. Y., Yi, W. C., Hsueh, P. R., Liu, K. L., & Chang, S. C. (2007). Klebsiella pneumoniae genotype K1: An emerging pathogen that causes septic ocular or central nervous system complications from pyogenic liver abscess. *Clinical Infectious Diseases*, *45*(3), 284–293.

Fothergill, J. L., Mowat, E., Walshaw, M. J., Ledson, M. J., James, C. E., & Winstanley, C. (2011). Effect of antibiotic treatment on bacteriophage production by a cystic fibrosis epidemic strain of Pseudomonas aeruginosa. *Antimicrobial Agents and Chemotherapy*, *55*(1), 426–428.

Friedlaender, C. (1882). Ueber die Schizomyceten bei der acuten fibrösen Pneumonie. *Archiv Für Pathologische Anatomie Und Physiologie Und Für Klinische Medicin*, *87*(2), 319–324.

Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics*, *28*(23), 3150–3152.

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).

Hochhut, B., Marrero, J., & Waldor, M. K. (2000). Mobilization of Plasmids and Chromosomal DNA Mediated by the SXT Element, a Constin Found in Vibrio cholerae O139. *Journal of Bacteriology*, *182*(7), 2043–2047.

Hoskisson, P. A., & Seipke, R. F. (2020). Cryptic or silent? The known unknowns, unknown knowns, and unknown unknowns of secondary metabolism. *MBio*, *11*(5), 1–5.

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., Mende, D. R., Letunic, I., Rattei, T., Jensen, L. J., von Mering, C., & Bork, P. (2019). EggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research*, *47*(D1), D309–D314.

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.

Janssen, B. D., & Hayes, C. S. (2012). The tmRNA ribosome-rescue system. En *Advances in Protein Chemistry and Structural Biology* (Vol. 86, pp. 151–191). Academic Press Inc.

Juhas, M., van der Meer, J. R., Gaillard, M., Harding, R. M., Hood, D. W., & Crook, D. W. (2009). Genomic islands: Tools of bacterial horizontal gene transfer and evolution. *FEMS Microbiology Reviews*, *33*(2), 376–393.

Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., & Madden, T. L. (2008). NCBI BLAST: a better web interface. *Nucleic Acids Research*, *36*(Web Server issue).

Lam, M. M. C., 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).

Lam, M. M. C., Wick, R. R., Wyres, K. L., Gorrie, C. L., Judd, L. M., Jenney, A. W. J., Brisse, S., & Holt, K. E. (2018). Genetic diversity, mobilisation and spread of the yersiniabactin-encoding mobile element ICEKp in klebsiella pneumoniae populations. *Microbial Genomics*, *4*(9).

Letunic, I., & Bork, P. (2021). Interactive tree of life (iTOL) v5: An online tool for phylogenetic tree display and annotation. *Nucleic Acids Research*, *49*(W1), W293–W296.

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.

Liu, B., Zheng, D., Zhou, S., Chen, L., & Yang, J. (2022). VFDB 2022: a general classification scheme for bacterial virulence factors. *Nucleic Acids Research*, *50*(D1), D912–D917.

Liu, Y.-C., Cheng, D.-L., & Lin, C.-L. (1986). Klebsiella pneumoniae Liver Abscess Associated With Septic Endophthalmitis. *Archives of Internal Medicine*, *146*(10), 1913.

Madeira, F., Pearce, M., Tivey, A. R. N., Basutkar, P., Lee, J., Edbali, O., Madhusoodanan, N., Kolesnikov, A., & Lopez, R. (2022). Search and sequence analysis tools services from EMBL-EBI in 2022. *Nucleic Acids Research*.

Marcoleta, A. E., Berríos-Pastén, C., Nuñez, G., Monasterio, O., & Lagos, R. (2016). Klebsiella pneumoniae asparagine tDNAs are integration hotspots for different genomic islands encoding microcin E492 production determinants and other putative virulence factors present in hypervirulent strains. *Frontiers in Microbiology*, 7, 849.

Martin, R. M., & Bachman, M. A. (2018). Colonization, infection, and the accessory genome of Klebsiella pneumoniae. *Frontiers in cellular and infection microbiology*, 8, 4..

Mulani, M. S., Kamble, E. E., Kumkar, S. N., Tawre, M. S., & Pardesi, K. R. (2019). Emerging strategies to combat ESKAPE pathogens in the era of antimicrobial resistance: a review. *Frontiers in microbiology*, 10, 539.

Murray, C. J., Ikuta, K. S., Sharara, F., Swetschinski, L., Robles Aguilar, G., Gray, A., Han, C., Bisignano, C., Rao, P., Wool, E., Johnson, S. C., Browne, A. J., Chipeta, M. G., Fell, F., Hackett, S., Haines-Woodhouse, G., Kashef Hamadani, B. H., Kumaran, E. A. P., McManigal, B., … Naghavi, M. (2022). Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *The Lancet*, *399*(10325), 629–655.

Nairz, M., Dichtl, S., Schroll, A., Haschka, D., Tymoszuk, P., Theurl, I., & Weiss, G. (2018). Iron and innate antimicrobial immunity—Depriving the pathogen, defending the host. En *Journal of Trace Elements in Medicine and Biology* (Vol. 48, pp. 118–133). Elsevier GmbH.

O'Neill, J. (2016). *Tackling drug-resistant infections globally: Final report and recommendations*. https://amr-review.org/sites/default/files/160518_Final%20paper_with%20cover.pdf

Organización Mundial de la Salud. (2017). *La OMS publica la lista de las bacterias para las que se necesitan urgentemente nuevos antibióticos*. https://www.who.int/es/news/item/27-02-2017-who-publishes-list-of-bacteria-for-which-new-antibiotics-are-urgently-needed

Paczosa, M. K., & Mecsas, J. (2016). Klebsiella pneumoniae: Going on the Offense with a Strong Defense. *Microbiology and Molecular Biology Reviews*, *80*(3), 629–661.

Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T. G., Fookes, M., Falush, D., Keane, J. A., & Parkhill, J. (2015). Roary: Rapid large-scale prokaryote pan genome analysis. *Bioinformatics*, 31(22), 3691–3693.

Partridge, S. R., Kwong, S. M., Firth, N., & Jensen, S. O. (2018). *Mobile Genetic Elements Associated with Antimicrobial Resistance*.

Putze, J., Hennequin, C., Nougayrède, J. P., Zhang, W., Homburg, S., Karch, H., Bringer, M. A., Fayolle, C., Carniel, E., Rabsch, W., Oelschlaeger, T. A., Oswald, E., Forestier, C., Hacker, J., & Dobrindt, U. (2009). Genetic structure and distribution of the colibactin genomic island among members of the family Enterobacteriaceae. *Infection and Immunity*, *77*(11), 4696–4703.

Schmidt, H., & Hensel, M. (2004). Pathogenicity Islands in Bacterial Pathogenesis. *Clinical Microbiology Reviews*, *17*(1), 14–56.

Schreiber, F., Szekat, C., Josten, M., Sahl, H. G., & Bierbaum, G. (2013). Antibiotic-induced autoactivation of IS256 in Staphylococcus aureus. *Antimicrobial Agents and Chemotherapy*, *57*(12), 6381–6384.

Seemann, T. (2014). Prokka: Rapid prokaryotic genome annotation. *Bioinformatics*, *30*(14), 2068–2069.

Seoane, A., & Bou, G. (2021). Bioinformatics approaches to the study of antimicrobial resistance. *Revista Española de Quimioterapia*, *34*, 15–17.

Sirén, K., Millard, A., Petersen, B., Gilbert, M. T. P., Clokie, M. R. J., & Sicheritz-Pontén, T. (2021). Rapid discovery of novel prophages using biological feature engineering and machine learning. *NAR Genomics and Bioinformatics*, *3*(1), 1–10.

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9), 1312–1313.

Wang, J., Liu, Y., Lee, S. S., Yen, M., Wang, Y. C., Wann, S., & Lin, H. (1998). Primary Liver Abscess Due to Klebsiella pneumoniae in Taiwan. *Clinical Infectious Diseases*, *26*(6), 1434–1438.

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

World Bank Group. (2017). *Drug-resistant infections: a threat to our economic future*. http://documents.worldbank.org/curated/en/323311493396993758/final-report

World Health Organization (WHO), Food and Agriculture Organization (FAO), World Organisation for Animal Health (OIE), & UN Environment Programme (UNEP). (2021). *Antimicrobial resistance and the United Nations sustainable development cooperation framework*. https://www.who.int/publications/i/item/9789240036024

Wyres, K. L., & Holt, K. E. (2016). Klebsiella pneumoniae Population Genomics and Antimicrobial-Resistant Clones. En *Trends in Microbiology* (Vol. 24, Issue 12, pp. 944–956). Elsevier Ltd.

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

Xie, Y., Wei, Y., Shen, Y., Li, X., Zhou, H., Tai, C., Deng, Z., & Ou, H. Y. (2018). TADB 2.0: An updated database of bacterial type II toxin-antitoxin loci. *Nucleic Acids Research*, *46*(D1), D749–D753.

Zou, Q., Lin, G., Jiang, X., Liu, X., & Zeng, X. (2020). Sequence clustering in bioinformatics: an empirical study. Briefings in bioinformatics, 21(1), 1-10.