UNIVERSIDAD DE CHILE - FACULTAD DE CIENCIAS - ESCUELA DE PREGRADO

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

Rodolfo Esteban Acevedo Carvajal

Director del Seminario de Título: Dr. Andrés Marcoleta Caldera

Co-Director: Dra. Rosalba Lagos Mónaco

Abril 2019

Santiago – Chile

BIOGRAFÍA

Nací el 7 de diciembre de 1994, desde pequeño me interesé por el área de la biología, donde también apunté siempre a estudiar en la Universidad de Chile, a la edad de 19 años, luego de dar por segunda vez la PSU, logré mi objetivo de estudiar en la universidad que quería una carrera que se dedicara a la investigación en el área de la biología. En mi paso por esta universidad, he conocido a gente realmente inteligente y talentosa que tiene mucho que aportar, aquí aprendí y me formé como científico, con el futuro objetivo de usar todo lo aprendido aquí, para contribuir a mejorar la sociedad chilena.

AGRADECIMIENTOS

Quiero empezar agradeciendo a la Dra. Rosalba Lagos y al Dr. Octavio Monasterio por haberme dado la oportunidad de formar parte del BEM. También al Dr. Andrés Marcoleta por la oportunidad de trabajar en el área bioinformática, aportando a su investigación.

Agradezco en gran parte a Camilo Berríos, quien fue mi maestro, mi compañero y mi amigo en todo este proceso, sin todas sus enseñanzas y ayuda, este trabajo jamás hubiese sido realizado. Además, agradezco a Nicolás, Álvaro, Valentina, José Ignacio, Johanna y Juan, mis compañeros de laboratorio, por darme ese ambiente de compañerismo y apoyo que fue de gran ayuda en este proceso. También debo incluir al resto de integrantes del BEM quienes aportaron con un ambiente grato de trajo.

Quiero agradecer a mis padres por su apoyo incondicional en este largo proceso, sin ellos jamás hubiese llegado hasta aquí. A mi hermana, quien siempre apoyo con su alegría, cosa que yo no pude. A mis tíos y primos, en especial a mi primo Adolfo quien a pesar de su escaso tiempo se dio la molestia de ayudarme con problemas informáticos que escapaban de mi conocimiento. También agradezco a Claudia Álvarez, quien me ha apoyado y ha estado dándome ánimos y devolviéndome la esperanza en gran parte de este proceso, y se ha vuelto indispensable en mi vida.

Finalmente, agradezco a todos los académicos que me hicieron clases durante la carrera, quienes me entregaron las herramientas necesarias para llegar hasta aquí y seguir adelante, ellos me hicieron comprender realmente lo que es ser un científico.

Muchas gracias a todos.

A mis amados padres

ÍNDICE DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

RESUMEN

En las últimas dos décadas se ha registrado un incremento significativo en la detección de cepas hipervirulentas y multi-resistentes del patógeno bacteriano *Klebsiella pneumoniae,* lo cual estaría ligado con el alto dinamismo y la rápida evolución de su genoma. Dentro de los elementos genéticos móviles (EGMs) que determinan dicho dinamismo se encuentran las islas genómicas (IGs), que corresponden a fragmentos de DNA de hasta 200 kb que se integran preferentemente en genes que codifican tRNAs (tDNAs). Se ha demostrado que las IGs de *K. pneumoniae* y otras enterobacterias portan una gran variedad de genes relacionados con virulencia y resistencia a antibióticos, por lo que su identificación y caracterización a partir del creciente número de genomas secuenciados de esta especie resulta altamente relevante. En esta dirección, el presente trabajo describe el desarrollo de una herramienta para identificar IGs asociadas a tDNAs que se basa en el análisis del contexto genómico de dichos genes, es decir, la organización e identidad de los genes que se encuentran adyacentes en el cromosoma. En una primera etapa, el programa identifica los tDNAs y su contexto para luego clasificarlos y nombrarlos. En una segunda etapa, el programa analiza la continuidad del contexto genómico conservado de cada tDNA y determina la presencia de IGs cuando una disrupción es identificada. El programa creado, se utilizó para analizar el genoma de 50 cepas de *K. pneumoniae*, donde los resultados obtenidos fueron comparados con la identificación de IGs realizada manualmente y curada en un estudio previo realizado por nuestro grupo. Como resultado, se logró identificar correctamente un 99,3% de todos los tDNAs presentes en dichas cepas, además de 303 IGs de un total de 308 IGs detectadas manualmente. Adicionalmente, el programa mostró un desempeño satisfactorio, aunque mejorable en la detección de integrasas y repetidos directos, detectando ambos

elementos en 148 de las IGs identificadas. Finalmente, se comparó el desempeño del programa desarrollado y otros programas para la identificación de IGs en la detección y delimitación de la isla GIE492 caracterizada experimentalmente en un trabajo previo de nuestro grupo, determinando que la detección hecha por el programa desarrollado por nuestro grupo es más sensible y precisa que cualquiera de los otros programas utilizados.

ABSTRACT

In the last two decades, there has been a significant increase in the detection of hypervirulent and multi-resistant strains of the bacterial pathogen *Klebsiella pneumoniae*, which would be related to the high dynamism and rapid evolution of its genome. Within the mobile genetic elements (EGMs) that determine this dynamism are Genomic Islands (IGs), which correspond to DNA fragments up to 200 kb that are preferentially integrated into genes encoding tRNAs (tDNAs). It has been shown that IGs from *K. pneumoniae* and other species from *Enterobacteriaceae* carry a wide variety of genes related to virulence and antimicrobial drug resistance. Thus, their identification and characterization among the increasing number of sequenced genomes from this species is highly relevant. In this direction, this work describes the development of a bioinformatics tool intended to identify IGs associated to tDNAs based on the analysis of the genomic context of these genes, that is, the organization and identity of the genes that are adjacent to tDNAs in the chromosome. In the first step, the program identifies all the tDNAs and their contexts to classify and name them. In a second step, the program analyzes the continuity of the conserved genomic context for each tDNA, determining the presence of a putative IG when a disruption of such continuity is detected. Our program was used to analyze the genome of a set of 50 *K. pneumoniae* strains, where the results obtained were compared with the identification of IGs performed and cured manually in a previous study conducted by our group. As a result, 99.3% of all the tDNAs present in the strains were correctly identified. Also, the software predicted 303 IGs from a total of 308 IGs detected manually. Additionally, the program showed a good performance in the identification of integrases and direct repeats, detecting both elements in 148 of the IGs, although there is still room for improvement. Finally, we compared the performance of our software and other programs for the identification of IGs, in the detection and precise delimitation of the

GIE492 island, which was characterized experimentally in a previous work of our group. We determined that the detection made by our program was more sensitive and precise than any of the other programs tested.

ABREVIATURAS

DNA	**Ácido desoxirribonucleico**
RNA	**Ácido ribonucleico**
tRNA	**RNA de transferencia**
mRNA	**RNA mensajero**
tmRNA	**RNA mensajero y de transferencia tmRNA**
tDNA	**Gen que codifica un tRNA**
tmDNA	**Gen que codifica un tmRNA**
EGM	**Elemento genético móvil**
IG	**Isla Genómica**
CDS	**secuencias codificantes de proteínas**
FDL	**Fragmento disruptivo del *locus virgen***

CÓDIGO DE AMINOÁCIDOS

Ala	Alanina	**Lys**	Lisina
Arg	Arginina	**Met**	Metionina
Asn	Asparagina	**Phe**	Fenilalanina
Asp	Ácido Aspártico	**Pro**	Prolina
Cys	Cisteína	**Sec**	Selenocisteína
Gln	Glutamina	**Ser**	Serina
Glu	Ácido Glutámico	**Thr**	Treonina
Gly	Glicina	**Trp**	Triptófano
His	Histidina	**Tyr**	Tirosina
Ile	Isoleucina	**Val**	Valina
Leu	Leucina		

1. INTRODUCCIÓN

1.1 *Klebsiella pneumoniae*: una amenaza urgente para la salud humana

Klebsiella pneumoniae es una bacteria Gram-negativo perteneciente a la familia de las enterobacterias clásicamente considerada como un patógeno oportunista, la cual puede ser portada de manera asintomática en el tracto intestinal, la piel, la nariz y la garganta en individuos sanos, pero que puede causar una variedad de infecciones en pacientes hospitalizados, principalmente neumonía e infecciones del tracto urinario (Podschun & Ullmann, 1998).

En las últimas décadas la rápida aparición de cepas de *K. pneumoniae* resistentes a múltiples antibióticos ha sido catalogada como una amenaza urgente para la salud global, de acuerdo con informes recientes sobre resistencia a antimicrobianos emitidos por el Centro para el Control y la Prevención de Enfermedades de EEUU (CDC) y el Departamento de Salud del Reino Unido. Ambas entidades destacan una alta prevalencia de resistencia a antibióticos carbapenémicos y β-lactámicos de amplio espectro (Nordmann, et al., 2009; Sirot, et al., 1988). Además, a mediados de los años ochenta se describió un nuevo síndrome clínico asociado a infecciones por cepas de *K. pneumoniae* adquiridas en la comunidad, consistente en el desarrollo de abscesos hepáticos que derivan en infecciones metastáticas del torrente sanguíneo y otros órganos. Aunque estas observaciones se hicieron inicialmente en la Cuenca del Pacífico Asiático (principalmente Taiwán, Corea, Vietnam y Japón), actualmente casos similares están siendo reportados en distintas regiones del planeta. Las causantes de estos fenómenos son una nueva variante de cepas de *K. pneumoniae* denominadas hipervirulentas, las que han sido definidas por una combinación de características fenotípicas y genotípicas (Shon, et al., 2013). Algunos genes esenciales para el fenotipo

hipervirulento de *K. pneumoniae*, se agrupan en elementos genéticos móviles, facilitando así transferencia de factores de virulencia a otras cepas (Lam et al., 2018). A pesar de que el conjunto completo de productos genéticos responsables de este fenotipo todavía se encuentra bajo investigación, se ha observado que en las cepas hipervirulentas hay una alta prevalencia de factores asociados a una mayor producción de cápsula, distintos sistemas de sideróforos que mediarían la captación de hierro durante el proceso infectivo (Marr & Russo, 2018). Además de los abscesos hepáticos, las infecciones derivadas de las cepas hipervirulentas también pueden derivar en abscesos no hepáticos, neumonía y, con menos frecuencia, fascitis necrosante, endoftalmitis y meningitis (Shon et al., 2013).

1.2 Las islas Genómicas y su impacto en la virulencia de *K. pneumoniae*

La rápida aparición de cepas hipervirulentas e hiperresistentes de *K. pneumoniae* ha sido asociada al alto dinamismo de su genoma, donde la adquisición y pérdida de elementos genéticos móviles y la transferencia horizontal de los mismos han permitido la evolución de nuevas capacidades biológicas, incluyendo mecanismos de resistencia a antibióticos y de patogénesis (Holt et al., 2015). Los elementos genéticos móviles (EGMs) son segmentos de DNA que pueden ser transferidos desde una ubicación dentro del genoma de una célula a otro lugar del mismo genoma o bien a una nueva célula. Normalmente, codifican enzimas y otras proteínas involucradas en la transferencia intracelular o intercelular del elemento. Se ha descrito que los EGMs de distintos patógenos bacterianos albergan un amplio repertorio de factores de virulencia y de resistencia a antibióticos, y que debido a su alta dinámica, estos elementos promueven una variación significativa en los genes portados por distintas cepas (Frost, et al., 2005).

Entre los EGMs que tendrían una alta relevancia en la evolución de *K. pneumoniae* se encuentran las islas genómicas (IGs). Las IGs corresponden a fragmentos de DNA cromosómico cuya presencia varía entre cepas bacterianas estrechamente relacionadas. Estos elementos poseen un tamaño de alrededor de 5 a 200 kb, presentan un contenido GC y uso de codones que difiere del promedio cromosómico, están flanqueados por repetidos directos, y suelen codificar una integrasa que cataliza la integración y escisión de la isla (Boyd, et al., 2009; Dobrindt, et al., 2004; Juhas et al., 2009). Además, algunas de ellas portan genes relacionados con su movilidad, por ejemplo factores que permiten su transferencia a través de conjugación (Marcoleta, et al., 2016). Existen diferentes tipos de IGs, donde la diferencia radica principalmente en las funciones que están codificadas en la isla (set de genes asociados a determinadas capacidades metabólicas), por ejemplo la capacidad de utilizar nuevas fuentes de carbono y nitrógeno, la capacidad de metabolizar nuevos compuestos, resistencia a antibióticos y metales pesados, o la capacidad de causar enfermedades (Dobrindt et al., 2004; Hacker & Carniel, 2001).

En el caso de *K. pneumoniae*, se ha propuesto que las IGs constituyen uno de los principales reservorios de factores de virulencia en esta especie (Marcoleta et al., 2016). Entre los factores de virulencia codificados por las IGs se encuentran los factores de adherencia, que permiten que las bacterias se puedan adherir a las superficie del hospedero; sideróforos, que se encargan de la captación de iones Fe^{3+}; reguladores que median la sobreproducción de polisacáridos capsulares, los cuales otorgan resistencia a la fagocitosis y a otros mecanismos de defensa del sistema inmunitario del hospedero; y el sistema de secreción de tipo IV (T4SS), que se requiere para la administración dirigida de toxinas o modulinas a las células eucariotas y que modulan el contacto del

huésped, interfieren con la señal de transducción y las vías apoptóticas del huésped y promueven la entrada en células no fagocíticas, además el T4SS presenta una estructura compleja de al menos 10 subunidades y es similar a los sistemas de conjugación para la transferencia de DNA (Schmidt & Hensel, 2004).

1.3 tDNAs como sitios de integración de IGs

Las IGs frecuentemente se integran en genes que codifican tRNAs (tDNAs), y en algunos casos en genes que codifican tmRNAs (tmDNAs). Por un lado, el RNA de transferencia es una molécula adaptadora esencial que se encuentra presente en los tres dominios de la vida. El papel principal del tRNA es participar en la traducción de los codones del RNA mensajero durante la síntesis de proteínas en los ribosomas, donde la adición de un aminoácido específico al polipéptido naciente depende del apareamiento entre el anticodón presente en tRNA y el codón del mRNA que está siendo decodificado (Shepherd & Ibba, 2015). Por otro lado, el RNA de transferencia y mensajero (tmRNA) actúa primero como un tRNA para unirse a los ribosomas estancados, y luego como un RNA mensajero para agregar una señal peptídica a la región C-terminal de la cadena polipeptídica naciente, donde el polipéptido con la señal será dirigido la proteólisis, lo que garantiza una rápida degradación de polipéptidos truncados potencialmente nocivos (Shepherd & Ibba, 2015).

Principalmente en enterobacterias, existe evidencia de que los tDNAs y tmDNAs son usados como sitios de integración no sólo de IGs, sino también de otros EGMs tales como profagos (Reiter, et al., 1989). Si bien aún no se ha establecido claramente cuáles son los factores que determinan la preferencia por los tDNAs como sitios de integración de IGs, en la literatura se ha sugerido que ciertas características de estos elementos podrían favorecer su uso: (i) los tDNAs se encuentran normalmente en múltiples copias

por cromosoma, (ii) son genes altamente conservados y (iii) son genes constantemente transcritos. Las características mencionadas, se atribuyen al hecho de que los tRNAs participan en un proceso indispensable para el funcionamiento de los organismos vivos. Estudios previos sobre las propiedades de los tDNAs y tmDNAs como sitios de integración de IGs, han establecido que las integrasas codificadas en estos elementos reconocen una secuencia diana de entre 3 posibles sublocaciones al interior de estos genes, lo que ha dado lugar a clasificar las integrasas filogenéticamente dependiendo de la sublocación que reconozcan (Williams, 2002). A su vez, en un estudio sobre IGs usando la cianobacteria *Prochlorococcus marinus* como modelo propuso ciertas características presentes en tDNAs que son usados como sitios de integración de estos elementos: (i) son solo unos cuantos tDNAs específicos; (ii) el uso de codones de los genes presentes en las IGs no estaría relacionado con la identidad del anticodón codificado por el tDNA donde se integra la IG; (iii) la mayoría de los extremos 3' de los tRNAs codificados por los tDNAs que son usados como sitio de integración carecen del extremo CCA; (iv) las regiones de los extremos 3' de los tDNAs usados como sitios de integración, mostraron tanto un alto contenido GC como una estructura palindromica en su secuencia (Liu & Zhu, 2010). A pesar de que se ha demostrado la preferencia del uso de tDNAs como sitios de integración, no serían los únicos sitios posibles, ya que estudios en *Escherichia coli* han demostrado que existen otros sectores en el genoma que son usados como sitio de integración de EGMs (Touchon et al., 2009).

1.4 La influencia del contexto genómico en la inserción de IGs asociadas a tDNAs

Previamente en nuestro laboratorio, se estudiaron los tDNAs de asparagina (asn-tDNAs) como sitios de integración de islas genómicas en *K. pneumoniae*. En total, se

analizaron los cromosomas de 50 cepas de esta especie, estudiando la presencia de IGs en cada una de las cuatro copias del asn-tDNA típicamente presentes, así como también el contexto genómico de estos genes, entiéndase por contexto genómico a las secuencias de nucleótidos situadas río arriba como río abajo de un gen en particular, en este caso al asn-tDNA. En este sentido, fue creado un nuevo método de identificación de estos elementos, que considera si el contexto genómico de los tDNAs estudiados ha sufrido una alteración en su secuencia, lo que se conoce como la interrupción del contexto genómico (Figura 1A). Las observaciones realizadas permitieron determinar que cuando una IG se integra en un tDNA, se produce un corrimiento de la región río abajo del mismo, la cual permanece inalterada. Por lo tanto, es posible definir un límite aproximado de las IGs, comenzando justo río abajo del tDNA y extendiéndose hasta el inicio de la región conservada río abajo. El análisis comparativo de los 50 cromosomas de *K. pneumoniae* realizado por nuestro grupo de investigación permitió identificar el contexto genómico de cada asn-tDNA en ausencia de IGs integradas, lo que denominamos la configuración de *locus virgen* de un tDNA. A partir de esto, se ideó una estrategia de identificación de IGs basados en la comparación de cada asn-tDNA presente en un genoma dado con su respectivo *locus virgen*, donde la interrupción de su continuidad (desplazamiento del contexto río abajo del tDNA) es evidencia de la presencia de un EGM integrado en dicho tDNA (Marcoleta, et al., 2016).

Como parte del análisis anteriormente descrito, se observó que a pesar de tener una secuencia 100% idéntica, las cuatro copias los asn-tDNAs (ubicadas en distintos contextos genómicos) son utilizadas como sitios de integración con una frecuencia muy dispar (Figura 1B). Lo anterior cuestiona el entendimiento actual del uso de tDNAs como sitio de integración, el cual se pensaba sólo depende de la secuencia del sitio de

recombinación (generalmente el extremo 3' del tDNA) y su reconocimiento por parte de una proteína integrasa específica. Nuestros resultados previos sugieren que la frecuencia de integración de IGs podría estar determinada también por el contexto genómico en el que se encuentran los distintos tDNAs.

Para conseguir una comprensión más profunda y transversal de este fenómeno, se evaluó la presencia de IGs integradas en la totalidad de tDNAs presentes en la muestra de 50 cepas de *K. pneumoniae* (Berríos, 2018). En total fueron analizados 4.281 *loci* de tDNAs, consiguiendo identificar 119 tipos de tDNAs según su contexto virgen. Adicionalmente, se confirmó que el contexto genómico en el que se encuentra cada tDNA (principalmente el contexto río arriba) está altamente conservado entre las distintas cepas, observando que en algunos casos existen en la región río abajo pequeños elementos variables que podrían estar relacionados con el alto dinamismo de estas regiones (Figura 1C). Respecto de las frecuencias de integración de IGs, se observó que del total de tipos de tDNAs identificados (119), sólo en 18 se identificó la integración de IGs (Figura 1D). Por otra parte, se constató que la disparidad en el uso como sitio de integración de tDNAs idénticos ubicados en diferentes contextos genomicos observada para el caso de los asn-tDNAs, ocurre también con otros tipos de tDNAs. Por ejemplo, de las dos copias idénticas del phe-tDNA presentes en el cromosoma de *K. pneumoniae*, sólo una de ellas es utilizada como sitio de integración de IGs (Figura 1B). Todo lo anterior corrobora la hipótesis que de alguna manera el contexto genómico de los tDNAs podría determinar o condicionar la integración de IGs en los tDNAs. Sin embargo, se desconoce la naturaleza molecular de este efecto.

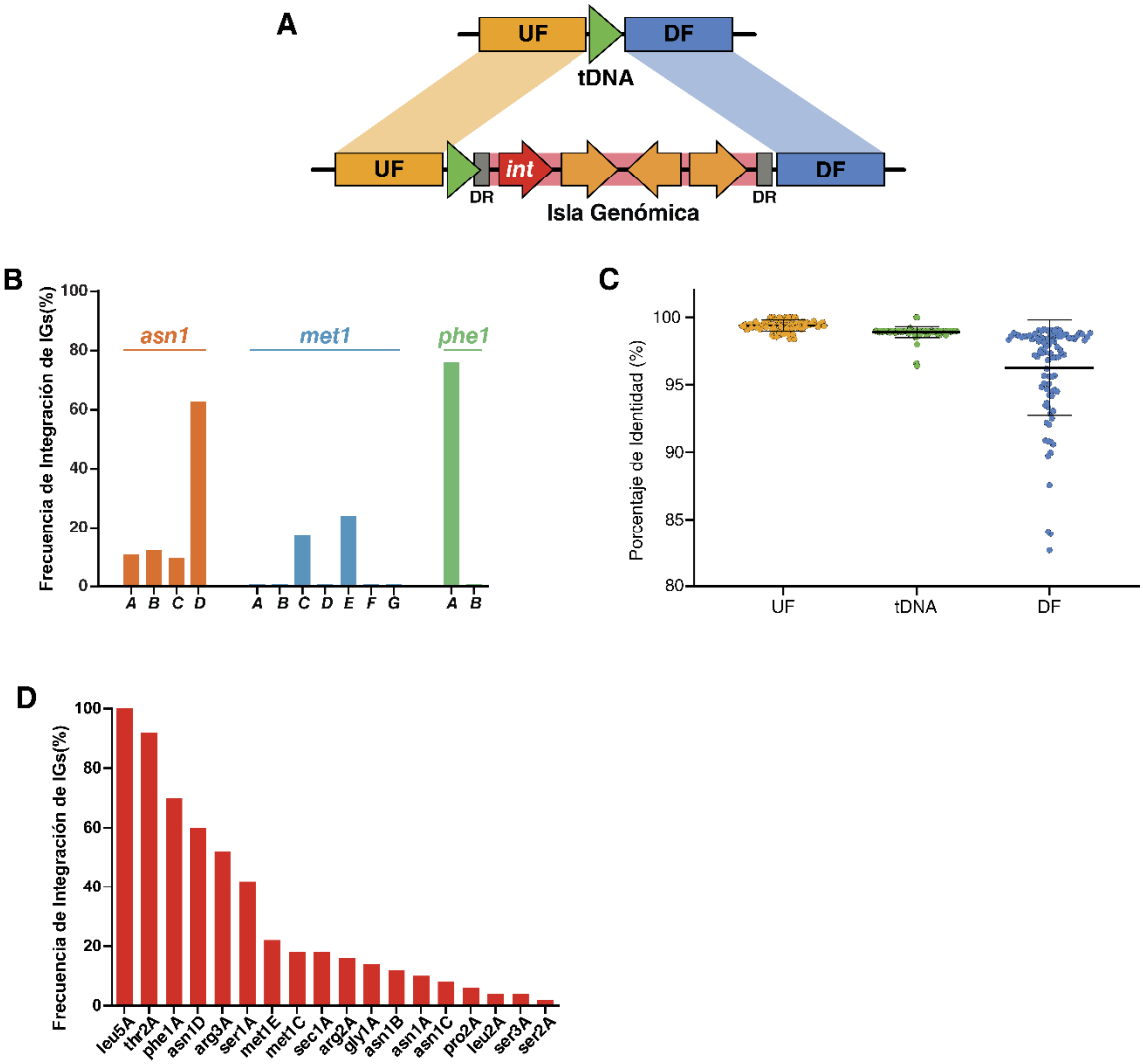


Figura 1: Resultados del estudio previo sobre el uso de tDNAs como sitios de integración de IGs en *K. pneumoniae* realizado por nuestro laboratorio.
(A) Esquema de la inserción de una IG en un cromosoma hospedero. Mientras que la región río arriba (UF) del tDNA queda intacta, la región río abajo (DF) es desplazada producto de la integración. Las IGs frecuentemente portan un gen que codifica una integrasa (int) ubicado en el extremo 5' de la isla (próximo al tDNA), la cual cataliza la integración y escisión del elemento. **(B)** Comparación de las frecuencias de integración de IGs en *loci* idénticos ubicados en distintas regiones del cromosoma, realizado para los tDNAs *asn1*, *phe1* y *met1*. **(C)** Conservación promedio de los elementos que constituyen los *loci vírgenes* de los tDNAs analizados: el contexto río arriba, el tDNA, y el contexto río abajo. **(D)** Frecuencia de uso de distintos tDNAs como sitio de integración de IGs (Berrios, 2018).

1.5 Herramientas de identificación de IGs

Actualmente, existen herramientas que en general permiten realizar la identificación de IGs para distintas especies bacterianas. Una de las herramientas es Islander, una base de datos de islas genómicas con su propio programa predictor que encuentra IGs basadas en las consecuencias, a nivel de secuencia, de su integración específica en tDNAs. Esta herramienta se centra en analizar fragmentos del cromosoma bacteriano en búsqueda de integrasas y repetidos directos (Hudson, et al., 2015). tRNAcc es otra herramienta de predicción de IGs, la que utiliza las regiones río arriba y río abajo de los tDNAs, buscando secuencias conservadas de estas regiones y de los tDNAs que flanquean. Esta comparación debe hacerse especificando los *loci* de los tDNAs y una secuencia de referencia, ambas definidas por el usuario (Ou et al., 2006). Otra herramienta, es AlienHunter, la cual se encarga de analizar la composición de nucleótidos de regiones específicas de una secuencia. Para ello, utiliza distribuciones de motivos de orden variable para analizar la composición local de una secuencia, y luego determinar si esa región local corresponde a una secuencia proveniente de un evento de transferencia horizontal (Vernikos & Parkhill, 2006). Así mismo, el programa MSGIP también se encarga de analizar la composición nucleotídica de una determinada secuencia utilizando la clasificación a través del método *mean-shift* para detectar la presencia de IGs. El método mean-shift es un método matemático iterativo que permite la agrupación de datos, sin la previa definición del números de grupos ni de las características de estos (De Brito et al., 2016). La herramienta más reciente es xenoGI, la cual identifica las IGs comparando secuencias de genes presentes en el genoma de diferentes bacterias estrechamente relacionadas, determinando si los genes analizados fueron adquiridos por transferencia horizontal o fueron heredados, lo que también permite situar el origen de estos genes en un mapa filogenético (Bush et al., 2018). Por

otra parte, existe una aplicación web que utiliza diferentes algoritmos de búsqueda para identificar IGs llamada IslandViewer, específicamente, utiliza las herramientas IslandPath-DIMOB (Bertelli & Brinkman, 2018), IslandPick (Langille, Hsiao, & Brinkman, 2008) y SIGI-HMM (Waack et al., 2006) las cuales se encargan de buscar genes de movilidad, realizar alineamientos múltiples con secuencias de otros genomas y analizar la preferencia de uso de codones, respectivamente, de esta manera se tiene un análisis mas completo para identificar IGs (Langille & Brinkman, 2009). El principal problema de las herramientas mencionadas es que dependen de “elementos supuestamente típicos” de IGs, tales como repetidos directos o integrasas, entre otros, pero ninguna de ellas toma ventaja de la gran conservación del contexto genómico en el que se encuentran normalmente los tDNAs en esta especie bacteriana, lo cual no ha sido descrito en la literatura y forma parte de un artículo actualmente en redacción por parte de nuestro grupo. Considerando que los contextos podrían resultar determinantes en los patrones de integración de IGs, además de la relevancia de este tipo de elementos para la evolución de patógenos bacterianos de importancia clinica global, se hace necesaria una herramienta que permita la identificación y clasificación automática de los tDNAs en base a su contexto genómico conservado, y que saque partido de dicha clasificación para la detección altamente específica de IGs.

1.6 Un nuevo sistema de nomenclatura para los tDNAs de *K. pneumoniae*

Previamente, nuestro grupo de investigación desarrolló una nueva nomenclatura para poder nombrar y clasificar los tDNAs de distintas cepas de *K. pneumoniae*, y de esta manera poder diferenciarlos (Marcoleta et al., 2016). Gracias a esta nomenclatura, fue posible realizar un estudio para definir familias de tDNAs y compararlos entre genomas de distintas cepas de esta especie, evaluando sus propiedades como sitios de

integración de IGs (Berríos, 2018). Esta nomenclatura considera no solo la secuencia del tDNA y el anticodón que codifica, sino que además toma en cuenta el contexto genómico en el que se encuentra, permitiendo individualizar tDNAs que están presentes en múltiples copias idénticas localizadas en distintas regiones del cromosoma. La nomenclatura propuesta consta de 3 identificadores (Figura 2), según se describe a continuación. El primero consiste en el código de tres letras para identificar el aminoácido que es portado por el tRNA codificado en el tDNA en análisis (por ejemplo, asparagina = asn). El segundo es un número arábigo que especifica uno de todos los posibles anticodones que permiten decodificar dicho aminoácido (por ejemplo, para asparagina hay dos posibles anticodones: GUU, representado por el número “1” y AUU representado por el número “2”). Finalmente, el tercer identificador es una letra mayúscula que da cuenta del contexto genómico típico en el que se encuentra el tDNA. De acuerdo a lo anterior y a modo de ejemplo, en el cromosoma de *K. pneumoniae* existen sólo 4 genes que codifican tRNAs de asparagina, que corresponden a 4 copias idénticas del tDNA que codifica el anticodón GUU, ubicadas en distintos contextos genómicos. En consecuencia, dichos genes fueron nombrados *asn1A*, *asn1B*, *asn1C* y *asn1D*. Por otra parte, para el caso del tDNAs que codifican para tRNAs de treonina (para el cual existen 4 posibles anticodones) existen cepas que poseen 2 copias de este tDNA, sin embargo, uno de ellos codifica para el anticodón GGT representado por el número “1” y el otro el anticodón TGT representado por el número “3”. Además, ambos tDNAs se ubican en distintos contextos genómicos, por lo que estos genes fueron nombrados *thr1A* y *thr3A*.

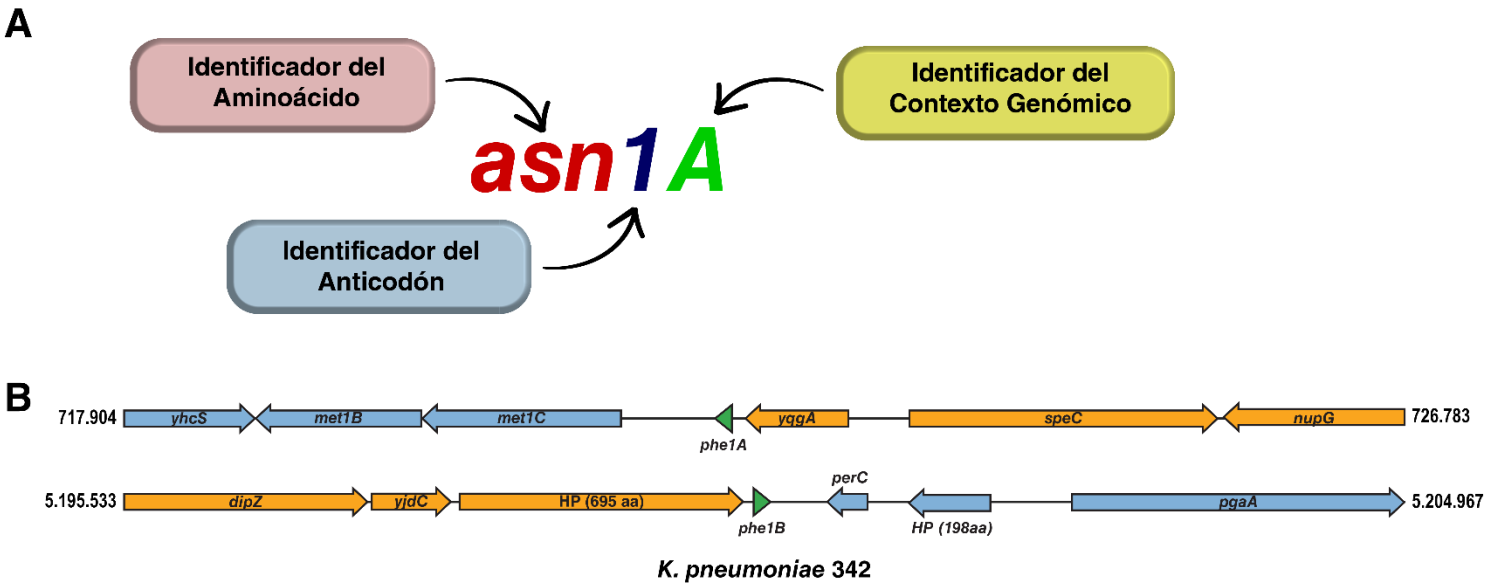


Figura 2: Nuevo método de anotación para los tDNAs de *K. pneumoniae* basado en el contexto genómico de los mismos. (A) Esquema que muestra los componentes del nuevo método de anotación. En este caso, las tres primeras letras indican que el aminoácido que será portado por el tRNA codificado es asparagina (asn), el número indica el anticodón basado en una asignación previamente realizada (para los tDNAs-asn, 1 indica que el anticodón es GTT) y el indicador del contexto en forma de una letra ("A"). **(B)** Ejemplo del uso de la nomenclatura para diferenciar a los dos tDNAs-phe comúnmente encontrados en los cromosomas de *K. pneumoniae* En este caso, ambos tDNAs son idénticos a nivel de secuencia (codifican el mismo anticodón), pero pueden ser diferenciados por su contexto genómico.

1.7 Un nuevo algoritmo para la identificación de IGs

Los estudios sobre la influencia del contexto genómico de los tDNAs en la inserción de IGs permitieron el desarrollo de un algoritmo para la identificación de estos elementos en cepas de *K. pneumoniae*. Este algoritmo, consiste en el reconocimiento de todos los tDNAs del cromosoma de una cepa problema, donde los tDNAs identificados son clasificados según la nueva nomenclatura mediante el reconocimiento del contexto rio arriba de estos. Luego de la clasificación, el siguiente paso consiste en analizar el contexto rio abajo de cada tDNA encontrado y compararlo con el contexto correspondiente al *locus virgen*, de tal manera de pesquisar si existe un corrimiento de dicho contexto rio abajo, indicativo de la presencia de una IG integrada en el tDNA en cuestión (Berríos, 2018). A pesar de que este algoritmo cumplía satisfactoriamente con el objetivo de identificar IGs, el tiempo de análisis por cada cromosoma resultó ser demasiado alto. Considerando que el objetivo a futuro será utilizar el algoritmo sobre una gran cantidad de genomas, se hace necesario contar con un programa que sea capaz de ejecutar el algoritmo de manera automática y en un corto periodo de tiempo. En vista de lo anterior, este proyecto tiene por objeto el desarrollo de una nueva herramienta de identificación de IGs con un algoritmo distinto a cualquiera anteriormente desarrollado, ya que aprovecha toda la información que se ha generado en el laboratorio respecto de la organización de los tDNAs, sus contextos, la nueva nomenclatura propuestas y los patrones de inserción de IGs, haciendo la detección de estos elementos altamente sensible y especializada para el cromosoma de *K. pneumoniae*.

2 OBJETIVOS

2.1 Objetivo General

Desarrollar una herramienta bioinformática para la identificación de Islas Genómicas asociadas a tDNAs en genomas de *Klebsiella pneumoniae*, en base al análisis del contexto genómico de dichos genes.

2.2 Objetivos específicos

2.2.1. Organizar la información obtenida previamente por nuestro grupo sobre los distintos tDNAs presentes en el cromosoma de *Klebsiella pneumoniae* y su contexto genómico, construyendo una base de datos accesible por la herramienta de identificación de Islas Genómicas.

2.2.2. Desarrollar una herramienta bioinformática que permita la identificación y anotación de tDNAs en genomas de *K. pneumoniae*, aplicando una nomenclatura basada en la información del contexto genómico.

2.2.3. Desarrollar una herramienta que permita la identificación de Islas Genómicas basada en detectar la interrupción de los contextos vírgenes previamente definidos y organizados en una base de datos.

2.2.4. Someter a prueba el desempeño de la herramienta bioinformática sobre secuencias genómicas de *K. pneumoniae* previamente curadas, evaluando su precisión en la anotación de tDNAs y la identificación de Islas Genómicas.

3 MATERIALES Y MÉTODOS

3.1 Origen de las secuencias analizadas

Las secuencias consideradas en este trabajo provienen de una colección de 50 cepas de *K. pneumoniae* previamente secuenciadas, ensambladas y anotadas (Tabla Anexo I). Las secuencias utilizadas en este trabajo fueron obtenidas desde la base de datos de secuencias de la NCBI (Coordinators, 2016). Previo a este trabajo, las islas genómicas presentes en estas cepas fueron identificadas y curadas manualmente (Berríos, 2018; Marcoleta et al., 2016).

3.2 Visualización de secuencias

El análisis manual de las secuencias se realizó utilizando el programa SnapGene Viewer™ (GSL Biotech LLC.). A pesar de que esta herramienta está diseñada para la visualización y diseño de plásmidos, es una herramienta versátil que también permite la visualización de genomas completos, con una interfaz gráfica que permite una cómoda inspección, en este caso, de los *loci* de tDNAs y su contexto.

3.3 Construcción de bases de datos de distintos elementos de secuencia

La secuencia del *locus* virgen de cada tDNA, incluyendo los genes anotados que forman parte de su contexto genómico, se almacenaron en formato Genbank y FastA, permitiendo su uso posterior por las distintas herramientas bioinformáticas creadas o utilizadas. Para definir los distintos *loci vírgenes*, se consideraron las secuencias codificantes de proteínas (CDS), los genes de rRNAs y otros tDNAs presentes en las cercanías del tDNA en análisis. En este sentido, fueron considerados tantos de estos elementos como fueran necesarios para diferenciar los distintos *loci*. Para su descripción y análisis fueron evaluadas las distancias nucleotídicas y el sentido de los elementos codificados en los contextos genómicos río arriba y río abajo de los tDNAs. Todos los

elementos identificados se almacenaron en archivos de secuencias para su uso posterior. En el caso de los rRNAs, para que las herramientas construidas pudieran identificarlos, se buscaron las secuencias consenso de los tres tipos de rRNAs identificados utilizando la herramienta de alineamiento múltiple de secuencias MUSCLE (Edgar, 2004).

Además, fue construida una base de datos de integrasas para la identificación de estos genes en los cromosomas sobre los que se usa la herramienta. Las secuencias nucleotídicas que componen dicha base de datos se obtuvieron a partir de la búsqueda de las integrasas anotadas en todas las enterobacterias presente en la base de datos de la NCBI.

3.4 Desarrollo de herramientas bioinformáticas para la aplicación de una nueva nomenclatura e identificación de Islas Genómicas asociadas a tDNAs

Las herramientas fueron escritas en lenguaje de programación Python, utilizando los paquetes BioPython (Cock et al., 2009) y Pandas (McKinney, 2011). Además, se utilizó parte del código del script run_MLST, perteneciente al repositorio *bioinformatics* de Leighton Pritchard (disponible en https://github.com/widdowquinn) para la identificación de los distintos tipos de contextos genómicos. Para la identificación de tDNAs se utilizó la herramienta ARAGORN (Laslett & Canback, 2004), la que hace uso de algoritmos heurísticos y proporciona información respecto del anticodón del tRNA codificado, las coordenadas de los tDNAs y el sentido del tDNA en el cromosoma. Para identificar las secuencias flanqueantes a los tDNAs se utilizó la herramienta BLAST (Altschul, et al., 1990), la cual permitió la comparación de las secuencias de las cepas analizadas y las bases de datos de contextos genómicos creada a partir del análisis exhaustivo del genoma de *K. pneumoniae* realizado previamente (Berríos, 2018).

Para analizar la presencia de un corrimiento del contexto rio abajo y de esta manera detectar IGs integradas, se utilizó la herramienta progressiveMauve (Darling, et al., 2010). Esta herramienta permite comparar genomas bacterianos mediante la búsqueda de bloques de secuencias conservadas mediante alineamientos locales múltiples, lo que a su vez permite detectar regiones que interrumpen la continuidad del *locus virgen* (región denominada Fragmento disruptivo del *locus virgen* o FDL, considerada como una potencial isla genómica), entregando las coordenadas de las regiones conservadas y no conservadas. Por otra parte, tanto la detección de genes que codifican integrasas como la de repetidos directos, fueron realizadas utilizando la herramienta BLASTn. La detección de integrasas se consiguió comparado las secuencias nucleotídicas de los FDLs identificados con las secuencias presentes en la base de datos de integrasas construida como parte de este trabajo. Para la identificación de repetidos directos, fue comparada la secuencia del FDL y la secuencia del tDNA donde se encuentra, incluyendo un segmento de 1 kb de la región río abajo del mismo. Los criterios utilizados para decidir si se encontró o no un repetido directo fueron: 1) un porcentaje de identidad superior al 75% entre ambas copias del repetido, 2) que posea el mismo sentido que el tDNA, 3) que el tamaño del repetido sea superior a 15 pb e inferior a 200 pb, y 4) que el repetido se encuentre en el último 20% de la extensión total del FDL predicho.

3.5 Comparación de tamaños para IGs y repetidos directos predichos

Para comparar los tamaños de las IGs y los repetidos directos predichos tanto por el programa como de forma manual se calculó la Diferencia absoluta porcentual, mediante Ecuación (1). Siendo T_p el tamaño predicho por el programa y T_m el tamaño predicho de forma manual.

$$\text{Diferencia absoluta porcentual} = \frac{|T_p - T_m|}{T_p} \times 100 \quad (1)$$

3.6 Evaluación de distintos programas para la identificación de IGs

El desempeño del programa desarrollado para la identificación de islas genómicas fué evaluado y comparado con otros programas disponibles para fines similares. Para ello, se determinó la precisión y la sensibilidad en la identificación de la isla genómica GIE492 en un genoma problema. Dicha isla fue previamente delimitada y caracterizada experimentalmente por nuestro grupo de investigación (Berríos, 2018; Marcoleta et al., 2016). La precisión, la sensibilidad y el valor F1 que relaciona las dos primeras, se calcularon de la siguiente manera (Lu & Leong, 2016):

$$\text{Presición} = \frac{TP}{TP+FP}; \ \ \text{Sensibilidad} = \frac{TP}{TP+FN}; \ \ F1 = \frac{2 \times \text{Presición} \times \text{Sensibilidad}}{\text{Presición} + \text{Sensibilidad}} \quad (2)$$

La evaluación se basó en analizar las CDS que forman parte de la región detectada como “isla genómica” por cada programa (en las cercanías de la región donde se ubica efectivamente la GIE492 en el genoma problema) y compararlas con las CDS que efectivamente forman parte de la GIE492 (IG de referencia). La detección de una CDS se consideró como “verdaderamente positiva” cuando al menos el 50% de su secuencia está presente en la IG de referencia. En la ecuación (2), “TP” corresponde al número total de detecciones positivas. Por otra parte, “FP” corresponde al número de falsos positivos, que corresponde al total de CDS identificadas que no se encuentran presentes en la IG de referencia. Por último, “FN” es el número de falsos negativos, representado por el total de CDS que se encuentran presentes en la IG de referencia, pero que no fueron identificadas por los programas (para más detalles ver Figura Anexo 1).

4 RESULTADOS

4.1 Construcción de una base de datos que comprende los diferentes *loci vírgenes* de tDNAs en *K. pneumoniae*

Con objetivo de organizar en una base de datos la secuencia de los diferentes *loci vírgenes* de tDNAs previamente identificados, se extrajeron las secuencias de los *loci vírgenes* completos, es decir, incluyendo las regiones flanqueantes y los tDNAs. Las secuencias fueron obtenidas a partir de archivos de secuencias cromosómicas ensambladas de diferentes cepas de *K. pneumoniae*, las cuales fueron estudiadas con anterioridad (Berríos, 2018). En total, fueron creados 70 archivos correspondientes a los *loci vírgenes* de los tDNAs. Vale destacar que entre estos archivos hay un total de 119 tipos de tDNAs, los que fueron clasificados en unidades monocistrónicas o policístrónicas de acuerdo con trabajos anteriores (Berríos et al., 2018). Para el uso ulterior de esta información, las secuencias fueron exportadas en formato FastA y Genbank usando programas para este propósito escritos en Python. Además, fueron separadas y almacenadas las secuencias de las regiones río arriba y río abajo de los tDNAs.

A partir de la información almacenada, se construyó una base de datos con las secuencias codificantes de proteínas (CDS) presentes en las regiones flanqueantes de los tDNAs. Para esto, se consideraron sólo los tres CDS más próximos a cada tDNA, tanto para la región río arriba como para la río abajo. Para simplificar esta base de datos, todas las CDS fueron almacenadas en el sentido positivo de la hebra de DNA, por lo que el reverso complementario se determinó para aquellas CDS que estuviesen codificadas en la hebra negativa. Las CDS fueron clasificadas según la posición que ocupan con respecto de los tDNAs. De esta forma, se determinaron tres posiciones posibles (I, II y

III) dependiendo de si la CDS se encuentra justo después del tDNA (I), alejado por una CDS (II) o por dos CDS (III) (Figura 3). Las secuencias fueron almacenadas en formato multiFastA para cada posición.

Para analizar y almacenar las secuencias que codifican rRNAs presentes en las regiones río arriba y río abajo en los *loci* de los tDNAs, se realizaron alineamientos múltiples para cada tipo de estas secuencias (5S, 16S y 23S). Se observó una alta conservación de estas secuencias (sobre un 95% de identidad) entre las cepas analizadas, por lo que se determinó la secuencia consenso de estos genes y se almacenaron en formato FastA.

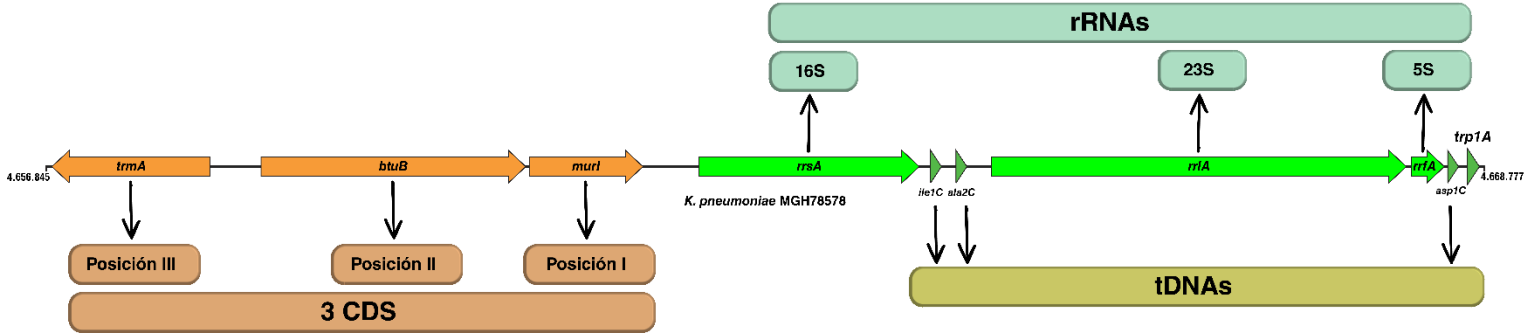


Figura 3: Los distintos elementos que pueden encontrarse en las regiones río arriba de los tDNAs en los *loci vírgenes*. En la figura se presenta como ejemplo el locus del tDNA *trp1A* en el cromosoma de la cepa *K. pneumoniae* MGH78578. En la mayoría de los casos, cada *locus virgen* presente en la base de datos creada posee 3 CDS, organizadas según su proximidad con el tDNA, en este caso los genes *murI, btuB* y *trmA* ocupan las tres posiciones en el locus de *trp1A* (extremo derecho en la figura). Además, se registró la presencia o ausencia de genes que codifican rRNAs y también de otros tDNAs.

4.2 Desarrollo de la herramienta bioinformática que permite la identificación y anotación de los tDNAs en cepas de *K. pneumoniae*, aplicando la nueva nomenclatura propuesta

Haciendo uso del lenguaje de programación Python y las herramientas bioinformáticas Biopython, BLAST y ARAGORN, se desarrolló un programa capaz de identificar y clasificar los diferentes tDNAs presentes en el cromosoma de una cepa problema de *K. pneumoniae* y anotarlos de acuerdo a la nueva nomenclatura basada en el contexto genómico propuesta por nuestro grupo de investigación (Berríos, 2018; Marcoleta et al., 2016).

El programa desarrollado requiere como archivo de entrada la secuencia nucleotídica del cromosoma ensamblado de una cepa *K. pneumoniae* en formato FastA. En esta primera versión, el programa ha sido optimizado para trabajar sobre cromosomas completamente ensamblados (en un solo fragmento). Antes de comenzar los análisis posteriores, el programa incluye una función que permite identificar que se cumplan las condiciones mencionadas. El siguiente paso en el programa es la identificación de los tDNAs codificados en el cromosoma, para esto utiliza la herramienta ARAGORN (Laslett & Canback, 2004). Luego de esta identificación, cada tDNA es almacenado en listas dentro del programa. Como resultado de esta etapa, para cada tDNA identificado se registra y almacena la secuencia del anticodón codificado, las coordenadas que definen la posición del tDNA en el cromosoma y si este se ubica en la hebra sentido o antisentido (“+” o “-“).

Luego, el programa analiza el contexto genómico de cada tDNA identificado en la primera etapa, de modo de nombrarlo aplicando la nomenclatura para tDNAs previamente desarrollada por nuestro grupo de investigación (Marcoleta et al., 2016). Cabe señalar que para la asignación del número que representa el anticodón, se

construyó un diccionario que considera un número para todas las posibilidades que permite el código genético, el cual fue incluido como parte del código del programa. Luego, fue programado el módulo para la identificación del contexto genómico que flanquea a cada tDNA. En primera instancia, el programa intenta resolver el contexto genómico sólo considerando la región conservada río arriba del tDNA. Para ello, se probaron distintas estrategias que permitieran comparar la región río arriba de todos los tDNAs presentes en un cromosoma problema con las secuencias presentes en la base de datos de *loci vírgenes* construida.

En primera instancia se utilizó la herramienta BLASTn (Altschul et al., 1990) para realizar comparaciones por alineamientos entre una región de 10 kpb río arriba de los tDNAs y las secuencias nucleotídicas de los *loci vírgenes*. Sin embargo, esta estrategia mostró un desempeño insuficiente, principalmente en la identificación y clasificación de aquellos tDNAs que se encuentran formando clusters u ocupando posiciones muy cercanas en el cromosoma. Para superar este obstáculo, como estrategia alternativa se desarrolló un programa que también analiza una región de 10 kbp río arriba del tDNA, pero ahora a través de la identificación y anotación de los CDS, rRNAs y otros tDNAs presentes en dicha secuencia. De esta manera, el programa conseguiría identificar un contexto determinado dirimiendo la presencia/ausencia de estos elementos conservados. Esta estrategia resultó ser más precisa que la anterior y permitió la identificación de la mayoría de los tDNAs en secuencias problemas (aproximadamente un 99,9%, tal como se describe en la sección 4.4).

Para superar conflictos relacionados con la presencia de otros tDNAs cercanos, se creó una tabla que organizara de manera numérica los patrones de presencia de tDNAs en los *loci vírgenes*. Para ello se analizaron cada uno de estos *loci*, registrando

la presencia de otros tDNAs en los contextos río arriba. En total se encontraron 73 diferentes patrones de presencia de tDNAs (P. tDNAs). Cada P. tDNAs fue identificado con un código numérico, donde la ausencia de otros tDNAs en el contexto virgen fue identificada con el valor 1 (Tabla I).

Para el funcionamiento de esta parte del programa, los elementos presentes en la región río arriba de los tDNAs en los *loci vírgenes* fue empleada para construir los denominados *perfiles de referencia.* Los perfiles contienen la presencia o ausencia, identidad y sentido de los elementos codificados de forma alfanumérica dentro de una matriz, de manera que esta información pueda ser procesada e interpretada por el programa. A cada tDNA le corresponde un *perfil de referencia* y cada *perfil de referencia* posee en total 8 elementos, Los primeros elementos son “PosI”, “PosII” y “PosIII” y corresponden a las CDS que se encuentran en la posición I, posición II y posición III, respectivamente, siendo la posición I la más cercana al tDNA. Estos elementos “Pos” son definidos como un valor alfanumérico, que comienza con el número asignado en la base de datos a las secuencias codificantes identificadas, seguido de una letra la cual puede ser una letra “p” o “n” que denotan el sentido de la secuencia con respecto al tDNA en el *loci vírgenes*, donde “p” indica que la secuencia posee el mismo sentido que el tDNA y “n” indica que la secuencia posee sentido contrario al tDNA. Los siguientes elementos son el “5S”, “16S” y “23S”, los cuales indican la presencia o ausencia de los rRNAs 5S, 16S y 23S respectivamente, estos elementos son de valor numérico donde los valores son 0 y 1, donde el 1 indica la presencia del rRNA y el 0 la ausencia de este. El siguiente elemento es el “P. tDNA”, el cual es el valor numérico asignado al conjunto de tDNAs presente en la secuencia analizada (Tabla I). El último elemento es “ST” (“Sequence Type”), el cual corresponde a un valor numérico que identifica a cada *perfil*

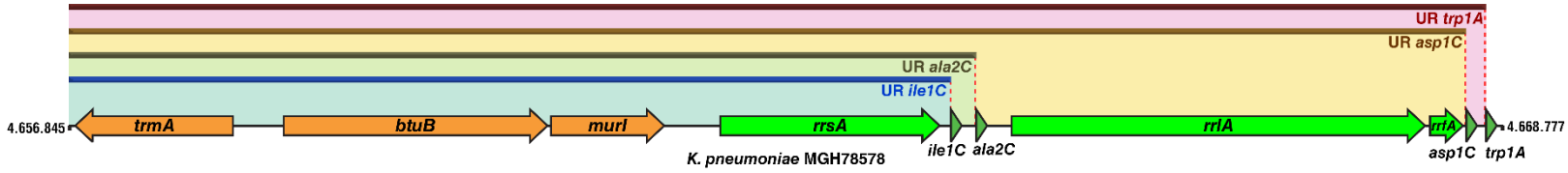
de referencia asociado a cada uno de los tDNAs registrados en la base de datos. Por ejemplo, en el caso de *trp1A* (Figura 4A), la región río arriba definida como parte del *locus virgen* comprende en las Posiciones I, II y III los CDS que tienen asignado el número 3 en la base de datos, lo que se combina con el sentido de cada elemento para así obtener tres identificadores representativos de estos CDS, también presenta los rRNAs 5S, 16S y 23S, y finalmente presenta un P. tDNAs de *ala2-ile1-asp1*. De esta manera, el *perfil de referencia* correspondiente al contexto río arriba del *locus virgen* de *trp1A* (Figura 4B), se construye a partir de los siguientes elementos: las CDS de las Posiciones I, II y III (*trmA*, *btubB* y *murI*) tienen asignado el número "3" en la base de datos, lo que sumado al sentido de cada CDS con respecto al tDNA, hace que las Posiciones I y II, sean representadas por un "3p", mientras que la Posición III con un "3n". La presencia de rRNAs 5S, 16S y 23S es representada con un número "1" para cada uno de ellos, además el P. tDNAs correspondiente a *ala2-ile1-asp1* tiene asignado el número "70". Finalmente, en base a la combinación de todos los elementos anteriormente descritos, el número asignado al *perfil de referencia* (ST) de *trp1A* es "115". En las cercanías de *trp1A* se encuentran los tDNAs *asp1C*, *ala2C* e *ile1C*, donde estos tDNAs comparten exactamente las mismas CDS en su región río arriba, sin embargo, poseen diferencias con respecto a la presencia de rRNAs y al P. tDNAs, por lo que sus "ST" son diferentes (Figura 4B).

Tabla I: Patrones de presencia de tDNAs detectados en las regiones río arriba de cada tDNA presente en el cromosoma de *K. pneumoniae*.

A cada patrón de presencia de tDNAs (P. tDNA) presente en las regiones río arriba de cada *locus virgen* presentes en la base de datos le fue asignado un número, con el objetivo de simplificar la información.

P. tDNAs	Código	P. tDNAs	Código
Sin tDNAs	1	*met1*	46
lys1-val1-val1-val1	2	*leu4*	47
ala1-lys1-val1-val1-val1	3	*leu4-leu4*	48
ala1-ala1-lys1-val1-val1-val1	4	*arg4-his1*	49
ile1	5	*lys1-val1*	50
ala1	6	*lys1-lys1-val1-val1*	51
ala2-ala2-ile1-ile1-ile1	8	*lys1-lys1-lys1-val1-val1*	52
ala2-ile1-ile1	9	*lys1-lys1-lys1-lys1-val1-val1*	53
ala1-val1	10	*ala1-ala1-ala1-val1-val1-val1*	54
ala1-val1-val1	11	*ala1-ala1-val1-val1-val1*	55
ala1-val1-val1-val1	12	*lys1-lys1-lys1-val1*	56
ser4	16	*met1-met1*	57
arg1-ser4	17	*leu1*	58
arg1-arg1-arg1-ser4	18	*gln1-gln1-leu3-met1*	59
ser3	19	*arg4-his1-leu4*	60
asn1-ser3	20	*ser2*	61
asn1	21	*asn1-asn1*	62
ala2-asp1-ile1-trp1	22	*glu1-gly3-thr3-tyr1*	64
ala2-asp1-ile1-ile1-trp1	23	*gly3-thr3-tyr1*	65
ala2-ile1	24	*ala2-ala2-asp1-ile1-ile1*	69
glu1	25	*ala2-asp1-ile1*	70
asp1	26	*asp1-glu1*	71
ala2-ile1-thr3	27	*asp1-asp1*	72
ala2-ala2-ile1-ile1	28	*tyr1*	73
ala2-ile1-thr3-tyr1	29	*glu1-thr3*	74
ala2-gly3-ile1-thr3-tyr1	30	*thr3*	75
arg1-arg1-ser4	31	*lys1*	76
gly2	32	*lys1-lys1-val1*	77
leu3-met1	33	*ala1-ala1*	78
gln1-leu3-met1	34	*ala1-ala1-ala1*	79
gln1-gln1-leu3-met1-met1	35	*ala1-ala1-ala1-val1*	80
gln1-gln1-gln2-leu3-met1-met1	36	*ala1-ala1-val1*	81
gly2-gly2	37	*ala1-ala1-ala1-val1-val1*	82
glu1-thr3-tyr1	39	*ala1-ala1-val1-val1*	83
thr3-tyr1	40	*val2*	84
arg4	41	*val2-val2*	85
cys1-gly2	45		

A



B

tDNA	Pos. I	Pos. II	Pos. III	5S	16S	23S	P. tDNA	ST
trp1A	3p	3p	3n	1	1	1	70	115
asp1C	3p	3p	3n	1	1	1	24	32
ala2C	3p	3p	3n	0	1	0	5	9
ile1C	3p	3p	3n	0	1	0	1	47

Figura 4: Ejemplo del uso de los *perfiles de referencia* para describir los elementos de secuencia presentes en los contextos río arriba de los tDNAs. (A) Ejemplo de la delimitación de los contextos río arriba de tDNAs codificados en una región del cromosoma de la cepa *K. pneumoniae* MGH78578. El límite de esta región está dada por la presencia de 3 CDS. **(B)** Ejemplo de la forma en que los elementos codificados en las regiones río arriba son convertidos en los *perfiles de referencia*. En este caso, los genes presentes en este *locus* son identificados con el número 3 y la letra que le sigue indica la orientación positiva (p) o negativa (n) de las CDS. Dentro de la base de datos, cada *perfil de referencia* posee un valor de referencia (ST).

De acuerdo con todo lo señalado anteriormente, se describirá la estrategia con la que opera el programa para nombrar los tDNAs considerando su contexto genómico (Figura 5). Para cada tDNA identificado en el cromosoma problema, el programa utiliza BLASTn de modo de identificar elementos conservados en su región río arriba, comparándola con las bases de datos que albergan los CDS y rRNAs que forman parte de los *loci vírgenes*. Como resultado, se declara la presencia de un elemento cuando se registra una identidad superior a un 80%, al alinear la secuencia nucleotídica del elemento presente en la secuencia problema y el correspondiente elemento presente en la base de datos. Así, al analizar la secuencia del tDNA a anotar, el programa genera un perfil de los elementos identificados en la región río arriba, el cual es luego comparado con los *perfiles de referencia* de la base de datos de *loci vírgenes*, buscando el *perfil de referencia* que posea mayor cantidad de elementos comunes con el perfil problema. La comparación entre perfiles es realizada de manera tal que el programa analiza que los elementos del perfil problema sean los mismos que el *perfil de referencia*. Por cada elemento que coincida entre los perfiles, el programa le asignará un punto a la comparación, donde el puntaje máximo para una comparación será de 8 puntos. Una vez asignado el puntaje, el programa calcula un puntaje porcentual de similitud entre el perfil problema y los *perfiles de referencia* (8 = 100%). Para resolver qué contexto flanquea al tDNA, el programa considera dos criterios: (i) el perfil problema y el *perfil de referencia* deben tener un porcentaje de similitud de al menos un 80% entre sus componentes, y (ii) entre los que cumplan con la condición anterior, el programa resolverá que el perfil problema es igual al *perfil de referencia* con el que comparte un mayor porcentaje de similitud. Luego, el programa le asignará a este último el número que identifica a ese *perfil de referencia* dentro de la base de datos (ST), para luego anexar al nombre del tDNA en análisis la letra que identifica el respectivo contexto

genómico. En caso de que el programa no encuentre un *perfil de referencia* con una similitud superior a un 80%, se le asigna a dicho tDNA la palabra ‘NEW’ como identificador del contexto genómico, dando cuenta de que se trata de un nuevo contexto que aún no ha sido registrado en la base de datos. Una vez que el programa ha clasificado todos los tDNAs encontrados en la secuencia problema de acuerdo con la nueva nomenclatura, éste guardará los resultados en un archivo CSV para su posterior análisis.

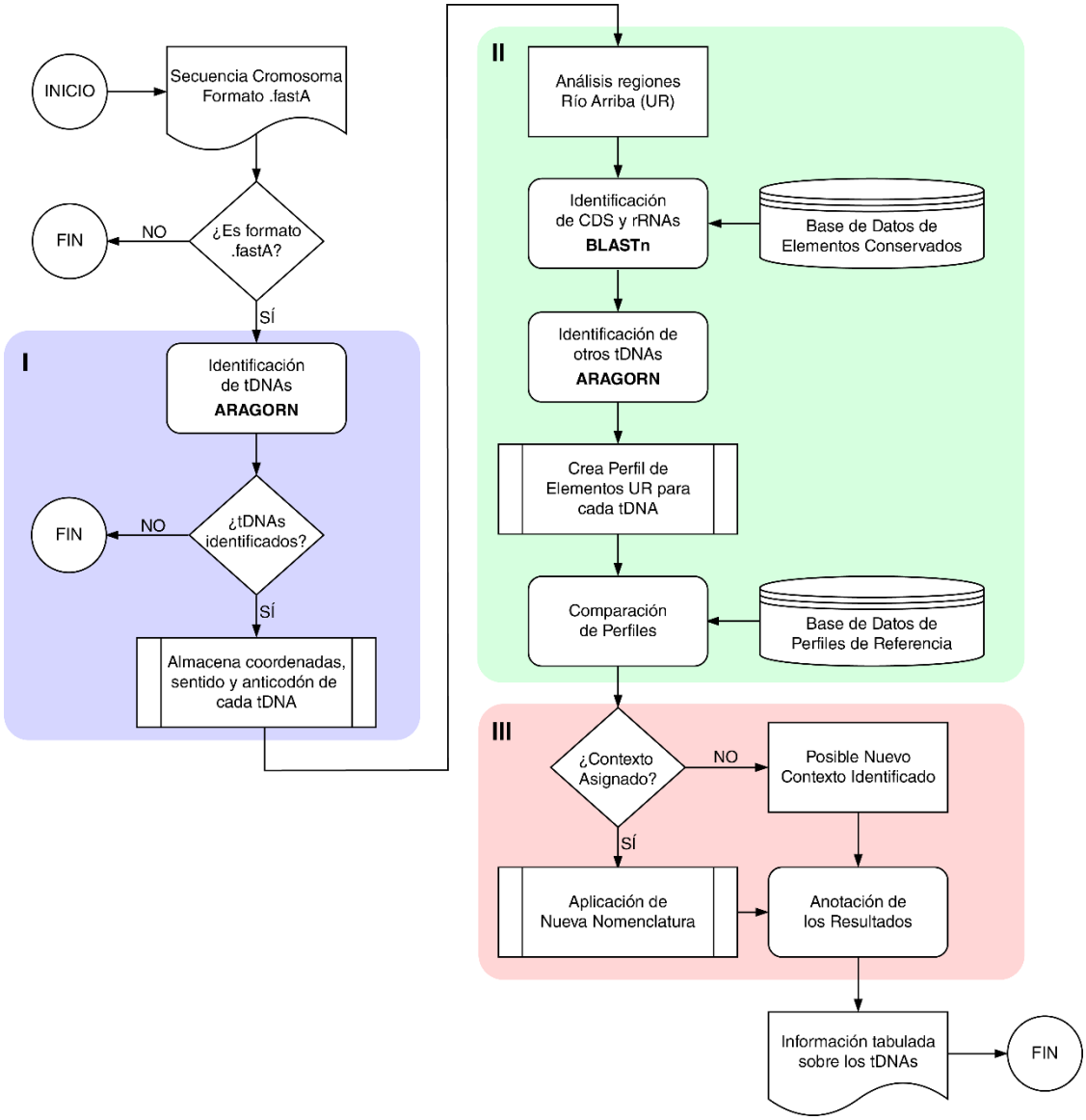


Figura 5: Diagrama de flujo que muestra las distintas etapas para la identificación de tDNAs y su contexto. En el resumen pueden observarse las distintas etapas del proceso programado. **(I)** Identificación de los tDNAs utilizando la herramienta ARAGORN; **(II)** Análisis de las regiones río arriba (UR) en búsqueda de elementos conservados; y (**III**) Combinación de toda la información recopilada para asignar un número de perfil de referencia y la posterior aplicación de la nueva nomenclatura propuesta.

Con el objetivo de evaluar el correcto funcionamiento del programa y su comunicación con la base de datos de *loci vírgenes*, se generaron un total de 70 archivos en formato FASTA, los cuales contienen los *loci vírgenes* de los 119 tDNAs identificados a partir de los 50 genomas de *K.pneumoniae* previamente utilizados en la elaboración de dicha base de datos. Estos nuevos archivos contenían la secuencia de los tDNAs registrados en la base de datos más una secuencia de un tamaño de 10 kpb de la región río arriba de estos tDNAs. Los 70 archivos generados fueron analizados por el programa, luego de lo cual se evaluó que la nomenclatura asignada a los tDNAs, así como sus coordenadas y sentido, correspondieran a los asignados en el análisis manual de estas secuencias realizadas previamente (Berríos, 2018). Se observó que para la totalidad los archivos analizados, la identificación y anotación de tDNAs realizada por el programa fue correcta en la nomenclatura, coordenadas y sentido. Estos resultados indican que el programa construido permite realizar satisfactoriamente la identificación y anotación según la nueva nomenclatura de todos los tDNAs presentes en un determinado genoma de *K. pneumoniae*.

4.3 Desarrollo de una herramienta que permita la identificación de Islas Genómicas integradas en tDNAs, basada en detectar la interrupción de los contextos vírgenes de dichos genes

Nuevamente, haciendo uso del lenguaje de programación Python y las herramientas bioinformáticas Biopython y BLAST, además del algoritmo progressiveMauve que forma parte del programa Mauve (Darling et al., 2010), se desarrolló un programa que permite la identificación de Islas Genómicas integradas en tDNAs basada en detectar la interrupción del contexto genómico virgen definido para dichos genes.

La estrategia general implementada en esta herramienta consiste en analizar la región río abajo de un tDNA problema, en búsqueda del contexto río abajo conservado definido para dicho tDNA, mediante su comparación con la base de datos de los *loci vírgenes*. Para lograr dicho propósito, en una primera instancia se probó utilizar alineamientos globales (con Biopython) para detectar la presencia del contexto río abajo del *loci vírgenes* inmediatamente junto al tDNA problema. En caso de que el contexto conservado no se encontrase, el programa debía avanzar por la secuencia (alejándose del tDNA) hasta conseguir que el alineamiento superara cierto umbral de identidad, lo cual sería indicativo de haber encontrado la región conservada. Luego el programa analizaría la región comprendida entre el tDNA y el contexto río abajo conservado, para identificar la presencia en dicha región de elementos comúnmente encontrados en Islas Genómicas. Sin embargo, la estrategia fue descartada debido a que el análisis para un solo tDNA tardaba alrededor de una hora en completarse, lo que llevaría a que el programa demorase más de cuatro días en analizar una sola cepa. Frente a la imposibilidad de optimizar el programa anterior, se optó por otro tipo de análisis.

Para superar el obstáculo del excesivo tiempo de proceso que requiere el análisis, se desarrolló un programa que se basa en el uso de ProgressiveMauve, una herramienta que permite identificar los fragmentos conservados y no conservados (ausentes) entre dos o más secuencias problema. Considerando dichos atributos, se propuso adaptarla para la identificación de IGs, específicamente para delimitar posibles bloques de secuencia que interrumpen el *locus virgen* definido para un tDNA dado, causando el desplazamiento de la correspondiente región río abajo conservada, a cada uno de estos bloques se les denominará Fragmento disruptivo del *locus virgen* (FDL). Al implementar esta estrategia se observó que el análisis realizado usando progressiveMauve tarda menos de 10 segundos por tDNA, por lo que se decidió incorporar al programa el análisis realizado por esta herramienta.

En base a lo anteriormente descrito, se explicará en mayor detalle el funcionamiento del programa. Como sustrato, el programa utiliza los resultados obtenidos a partir de la primera sección del análisis, es decir, de la identificación y anotación de los tDNAs presentes en el cromosoma. Para identificar la continuidad o interrupción de cada *locus* de tDNA (como producto de la integración de una posible IG), el programa realiza los siguientes pasos: (i) comprueba si se cumple como requisito que el tDNA tenga asignado un contexto para comparar, (ii) determina la presencia o ausencia de genes que codifican rRNAs en la región río abajo, (iii) extrae la secuencia del tDNA en análisis y de su contexto, específicamente 5 kpb de la región río arriba, 250 kpb de la región río abajo, siempre que dicha región no comprenda genes de rRNAs. En el caso de que sí posea este tipo de elementos, sólo son consideradas 40 kpb. Lo anterior es realizado dadas las múltiples copias de *loci* con rRNAs en el cromosoma, lo que imposibilita la identificación del contexto río abajo en el caso de que la región

analizada comprendiera más de 40 kpb. Respecto de este último punto, se decidió utilizar esta distancia luego de ensayar distintos tamaños, siendo 40 kpb el tamaño con el cual se obtuvo una menor tasa de error en la identificación del contexto río abajo.

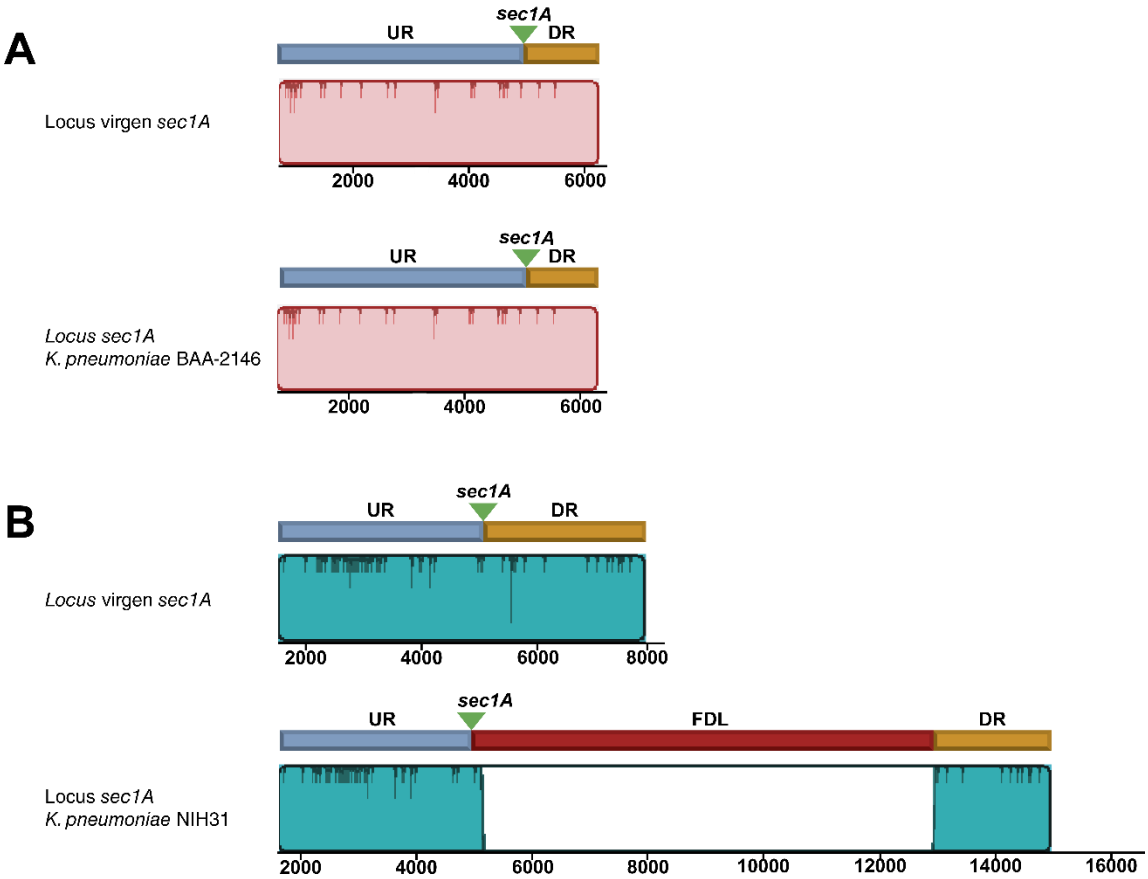


Figura 6: Análisis con progressiveMauve para la detección de FDLs situados río abajo del tDNA *sec1A* en el cromosoma de dos cepas de *K. pneumoniae.* La herramienta progressiveMauve permite la identificación de bloques de secuencia conservada. En este sentido, este programa es utilizado para la identificación de interrupciones del contexto genómico vírgen definido para cada tDNA, como señal de la presencia de algúna IG integrada en dicho locus. **(A)** Alineamiento entre la región que comprende el tDNA *sec1A* en el cromosoma de la cepa *K. pneumoniae* BAA-2146, y el *locus virgen* definido para dicho tDNA. La identidad entre ambas secuencias está representada por el alto de la superficie coloreada dentro de los bloques graficados. **(B)** Alineamiento entre el *locus sec1A* de la cepa *K. pneumoniae* NIH31 y el respectivo contexto virgen. Se puede notar que hay un alto nivel de identidad en los extremos de las secuencias, a la vez que se evidencia una interrupción de la continuidad del *locus virgen*, que es interpretada como un FDL de ~8 kpb. Las escalas indican el tamaño de los bloques en pb.

Mediante el algoritmo ProgressiveMauve, el programa compara el contexto genómico del tDNA analizado y su contexto equivalente en la base de datos de secuencias de *loci vírgenes*. De esta manera, se logran identificar fragmentos conservados entre las secuencias, además de interrupciones de la continuidad del *locus virgen* (FDLs), los que corresponden a potenciales IGs integradas en dichos sitios (Figura 6). Luego, de entre los FDLs identificados, el programa selecciona aquellos que: (i) poseen una extensión superior a 1 kpb y, (ii) su inicio se sitúe entre las bases -50 y +500 con respecto a la primera base del tDNA. Como resultado de estos análisis se genera una lista de los tDNAs analizados, y la presencia o no de FDLs y las coordenadas de estos en el cromosoma. Posterior al registro de las coordenadas de los FDLs predichos, se implementó en el programa un módulo para la identificación de características comúnmente asociadas a IGs: (i) la presencia de genes que codifiquen integrasas, (ii) repetidos directos que las delimiten, (iii) una medida de la diferencia entre el contenido GC del FDL y el contenido GC promedio del cromosoma hospedero.

Para detectar la presencia de integrasas, el programa utiliza BLASTn y una base de datos de genes que codifican integrasas construida para este propósito. La base de datos fue construida a partir de descargar desde la NCBI todas las secuencias codificantes que cumplieran con estar anotadas como integrasas y haber sido identificadas en genomas de especies de la familia *Enterobacteriaceae*. Considerando lo anterior, se obtuvo un total de 1478 secuencias nucleotídicas, las que se organizaron en una base de datos empleando scripts de Python para esta tarea. Luego, para determinar la presencia de genes codificantes de integrasas en los FDLs predichos, la secuencia completa del FDL es comparada con la base de datos de integrasas y los

posibles genes de integrasas son anotados en el FDL, si cumplen con tener más de un 85% de identidad y más de un 80% de cobertura.

Por otra parte, para la identificación de repetidos directos el programa utiliza la herramienta BLASTn, la cual permite comparar la secuencia del FDL predicho y la secuencia del tDNA donde se encuentra integrado, junto con un segmento de 1 kb de la región río abajo del mismo, lo cual resultó aumentar la precisión en la detección de repetidos directos. Para que una región sea considerada como un repetido directo debe cumplir las siguientes condiciones: (i) un porcentaje de identidad superior al 75% entre ambas copias del repetido, (ii) que se encuentre en el cromosoma con el mismo sentido que el tDNA, (iii) y que el tamaño del repetido sea superior a 15 pb e inferior a 200 pb. En esta primera versión del programa, la búsqueda de repetidos directos tiene como objetivo la delimitación de las islas genómicas predichas, por ello además el programa solo considera como repetidos directos a aquellos que se encuentran más alejados de los tDNAs, en el último 20% de la extensión total del FDL predicho. Si no son encontrados repetidos directos el programa acusa su ausencia como parte de los resultados y cuando son encontrados, el programa corrige las coordenadas del FDL predicho para coincidir con la primera base del repetido directo cercano al tDNA analizado hasta el fin del repetido directo identificado.

Finalmente, para el cálculo del contenido GC, el programa emplea funciones de la librería Biopython. El programa calcula el %GC de los FDLs predichos y los compara con el %GC promedio calculado para el cromosoma. Como resultado, el programa indica los porcentajes calculados, y si el %GC de cada FDL es mayor o menor que el del cromosoma hospedero.

Considerando que resultados no publicados de nuestro laboratorio sugieren que en algunos casos (menos frecuentes) la integración de IGs puede ocurrir en la región río arriba de los tDNAs, el programa desarrollado incluye un módulo que permite la identificación de este tipo de integraciones. Este análisis es realizado para todos los tDNAs a los que no se les pudo asignar algún contexto en la sección de identificación, considerando que el contexto río arriba podría haberse alterado producto de la integración de una IG en dicha región. Para cada uno de estos tDNAs sin contexto asignado, el programa realiza las siguientes operaciones: (i) se define una región que inicia 250 kpb río arriba y 250 kpb río abajo del tDNAs bajo análisis, (ii) utilizando progressiveMauve, cada una de estas secuencias es comparada con las secuencias presentes en la base de datos de *loci vírgenes* cuyos tDNAs posean el mismo identificador de aminoácido y de anticodón que el tDNA en análisis; (iii) cuando se identifican regiones conservadas con una identidad mayor al 85% entre el contexto de dicho tDNA y una secuencia de la base de datos de *loci vírgenes*, el programa asigna al tDNA bajo análisis el identificador de contexto correspondiente a dicha secuencia. Es decir, el contexto genómico y con ello el nombre del tDNA es resuelto mediante este análisis. Finamente, el programa analiza las interrupciones presentes en ambas regiones flanqueantes del tDNA en busca de FDLs. En este sentido, los FDLs deben cumplir con las condiciones previamente descritas, y de ser así, es analizada la presencia de integrasas y repetidos directos, además del contenido GC. En el caso de que ningún contexto de los *loci vírgenes* pueda ser asignado para algún tDNA, el programa reporta que se trata de un *locus* nuevo.

Finalmente, el programa utiliza los resultados de estos análisis para escribir tablas en formato CSV que el usuario puede revisar, y también para escribir un archivo

del cromosoma en formato GenBank, que posee la secuencia con la anotación de los FDLs predichos. Todos los pasos que componen la estrategia de identificación de IGs implementada en el programa desarrollado se encuentran esquematizadas en la Figura 7. De esta manera, la herramienta desarrollada presenta al usuario los FDLs encontrados y si estos presentan o no características típicas de IGs, de manera que el usuario es quién debe decidir cuáles de los FDLs encontrados corresponden a IGs en base a las características detectadas por el programa.

Con el objetivo de comprobar el correcto funcionamiento del programa en la identificación de IGs, se analizaron un total de 20 *loci* de dos tDNAs previamente identificados como sitios frecuentes de integración de IGs: 10 *loci phe1A* y 10 *loci asn1D*, donde cada *locus* pertenecía a una cepa de *K. pneumoniae* distinta. Los *loci* fueron seleccionados de manera que 5 de cada grupo de *loci* no tuvieran una IG integrada y los otros sí la tuvieran, de acuerdo a la identificación de IGs y curación manual de estos cromosomas realizada previamente (Berríos, 2018). Luego, los *loci* fueron analizados con el programa, observando que en el 100% de los casos el programa identificó las IGs en los *loci* donde previamente habían sido identificadas, a la vez que reportó la ausencia de IGs en el resto de los cromosomas.

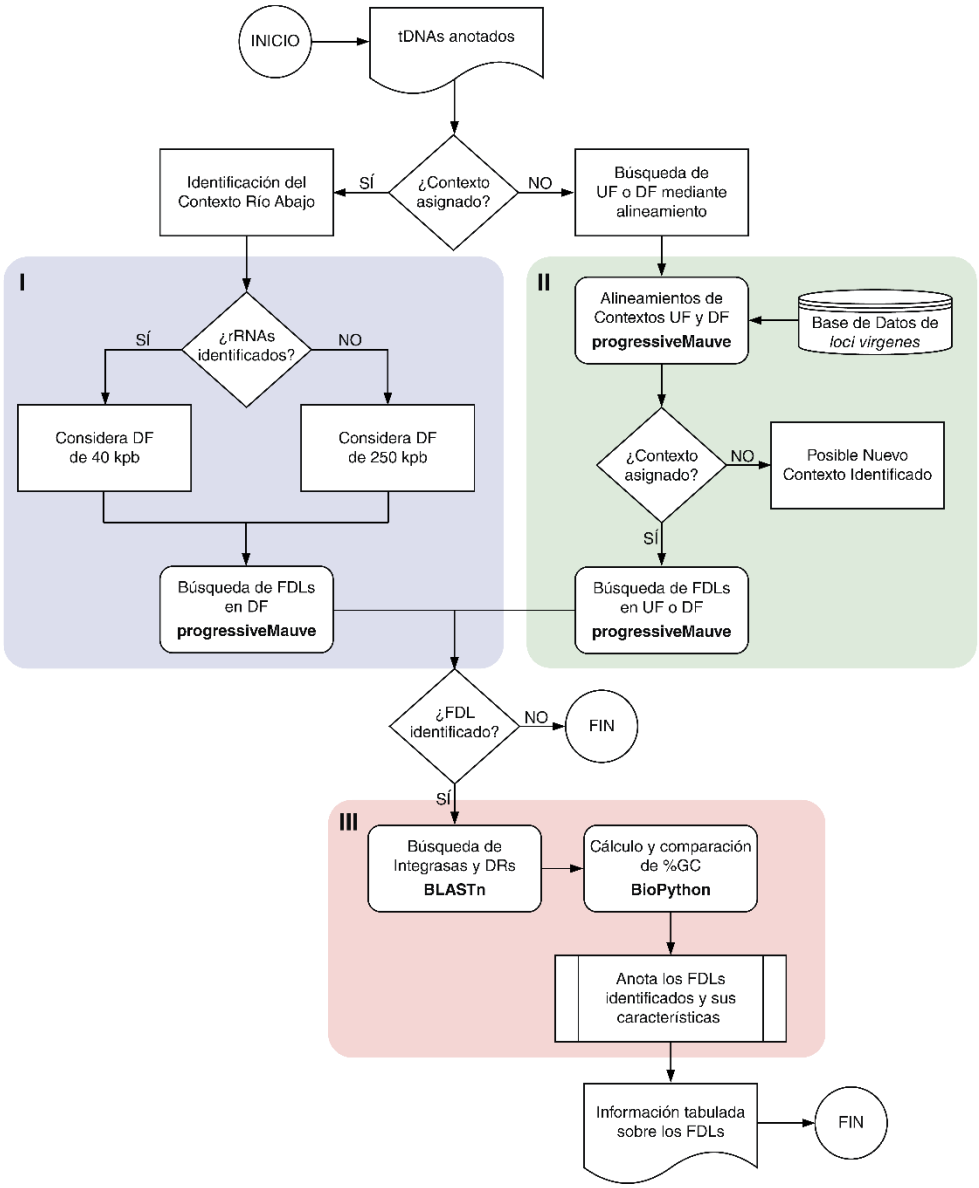


Figura 7: Diagrama de flujo que muestra las distintas etapas de la herramienta que identifica la presencia de IGs en los tDNAs presentes en el cromosoma de *K. pneumoniae*. En el resumen pueden observarse los pasos que sigue la herramienta desarrollada para la identificación de IGs. **(I)** En primera instancia el programa evalúa la presencia de rRNAs en las cercanías del tDNA en análisis, de modo de definir el fragmento de la región río abajo (DF) que se empleará para llevar a cabo los alineamientos progresivos. **(II)** En el caso en que los contextos genómicos no pudieron ser determinados, la herramienta analiza la presencia de FDLs integrados río arriba (UF) de los tDNAs. Finalmente, **(III)** los FDLs identificados son analizados en busca de integrasas y repetidos directos (DR) y los resultados son anotados.

4.4 Evaluación de los resultados de ambas herramientas sobre secuencias de cromosomas de *K. pneumoniae* previamente curadas

Las dos herramientas desarrolladas en este trabajo fueron combinadas para generar un único programa que hiciera ambos análisis de forma progresiva. El programa construido fue usado para analizar las 50 cepas de *K. pneumoniae* que se analizaron y curaron manualmente en un trabajo previo realizado por nuestro grupo de investigación (Berríos, 2018), de manera de poder comparar las predicciones de IGs realizadas manualmente con las que realiza el programa desarrollado. La comparación estuvo centrada en dos aspectos: la correcta identificación de los contextos y la predicción de IGs.

En total, el programa creado identificó y anotó un total de 4.285 tDNAs en el cromosoma de las 50 cepas analizadas. Respecto de la asignación de contextos, el programa asignó contextos a 4.282 tDNAs (99,9%). Luego, la asignación de contextos fue comparada con las realizadas manualmente, analizando si la asignación realizada por el programa coincide con las previamente hechas. En total, la asignación coincidió en 4.251 tDNAs (99,2%), por lo tanto, en solo 31 tDNAs (0,7%) el contexto fue asignado erróneamente (Tabla II).

Luego, fueron comparadas las predicciones de IGs del programa creado y la anotación manual. En total, el programa predijo 336 FDLs entre las 50 cepas de *K. pneumoniae*. Cuando se comparan estas predicciones con las realizadas manualmente se obtiene que la predicción coincide en 303 IGs predichas de forma manual. De forma particular, 33 de las predicciones del programa no fueron consideradas previamente por el análisis manual debido a su tamaño inferior a 5 kb, mientras que 5 predicciones realizadas manualmente no fueron detectadas por el programa.

Tabla II: Desempeño del programa en la anotación automática de tDNAs considerando el contexto genómico, en un set de 50 cromosomas previamente analizados de manera manual.

Resultado Asignación Contextos	Número tDNAs	% tDNAs
Asignados Correctamente[1]	4251	99,2
Asignados Incorrectamente	31	0,7
No Asignados	3	0,1
Total	4285	100

[1]Coincidencia entre el análisis manual de los tDNAs y la asignación realizada por el programa.

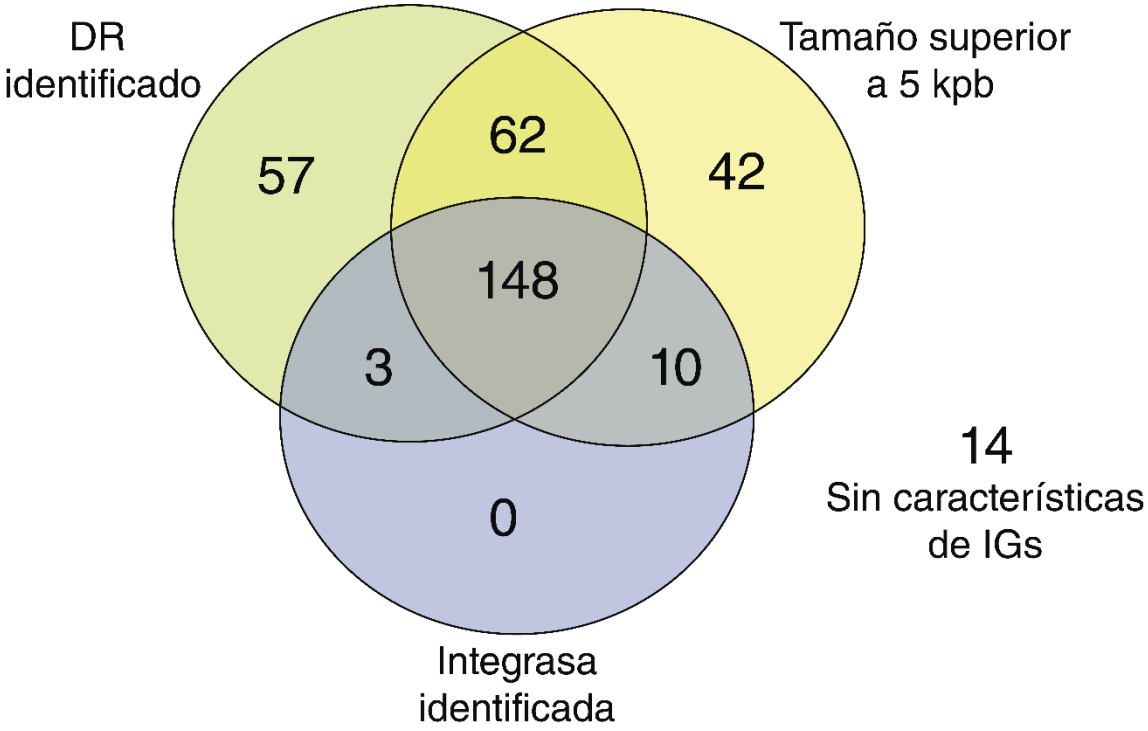


Figura 8: Agrupación de los FDLs encontrados según las características de IGs identificadas. Los 336 FDLs identificados por el programa fueron agrupados según la presencia de elementos característicos de IGs. A partir de esto se observa que del total de FDLs, 148 (44%) poseen todas las características comúnmente encontradas en IGs.

Luego, entre los 336 FDLs identificados por el programa, se determinó cuántos poseían características comunes de Islas Genómicas: (i) tamaño superior a 5 kpb, (ii) presencia de integrasa y (iii) presencia de repetidos directos. Se decidió considerar como IGs a todos los FDLs con un tamaño mayor a 5 kpb, en total, se encontraron 262 IGs y de estas, 148 que coincidieron en la presencia de las tres características, y sólo 14 de los FDLs predichos no poseían ninguna de las características (Figura 8).

La extensión de las predicciones realizadas, sobre *loci* equivalentes, por ambos métodos fue comparada. En este sentido, para cada *locus* se calculó la diferencia de los tamaños predichos por ambos métodos (Figura 9A-C), además de la Diferencia absoluta porcentual, que es el porcentaje del valor absoluto de la diferencia de tamaños obtenidos por ambos métodos con respecto al tamaño predicho por el programa, dicho porcentaje se calculó mediante el uso de la ecuación (1). De este análisis se obtuvo que, en promedio, la Diferencia absoluta porcentual fue de un 7,21% con una desviación estándar de 37,99%. Lo anterior confirma que la predicción de IGs entre ambos métodos es generalmente similar. Cabe destacar que para las IGs asociadas al tDNA *leu5A* es donde se registraron las mayores diferencias de tamaños predichos por ambos métodos (Figura 9C) y también las Diferencias absolutas porcentuales más alejadas del promedio.

De forma análoga, fue puesta a prueba la predicción de repetidos directos realizada por el programa. Se compararon los resultados de análisis del programa sobre los FDLs en las 50 cepas de *K. pneumoniae* y la identificación manual de repetidos directos realizada previamente sobre dicho set. Solo en 156 de los FDLs fueron detectados repetidos directos por ambos métodos y en 65 FDLs el programa detectó repetidos directos que la anotación manual de los datos no detectó. Adicionalmente, en el caso de los repetidos detectados por ambos métodos, se comparó el tamaño predicho

para dichos elementos por ambas aproximaciones (Figura 9D) y se calculó la Diferencia absoluta porcentual, utilizando nuevamente la ecuación (1). Se obtuvo que en promedio la Diferencia absoluta porcentual fue de un 47,44%, con una desviación estándar de 125,06%. Este resultado indica que hubo varios casos en que los tamaños de los repetidos directos predichos por el programa no fueron iguales o cercanos a sus contrapartes predichos por el método manual.

También fueron comparadas las predicciones de integrasas, comparando las realizadas por el programa con la anotación de los cromosomas realizada anteriormente. Fueron consideradas 10 de las cepas de *K. pneumoniae* y se analizó la anotación de las integrasas adyacentes a los tDNAs. Se determinó que en 56 de los tDNAs con IGs en las cepas, el programa acertó a la predicción de integrasas en 40 casos (71,4%) y en 16 casos la predicción de los genes de integrasas fue errónea o parcial. Además, se observó que en algunos casos de predicciones parciales, los genes identificados como integrasas realmente eran genes que codificaban transposasas.

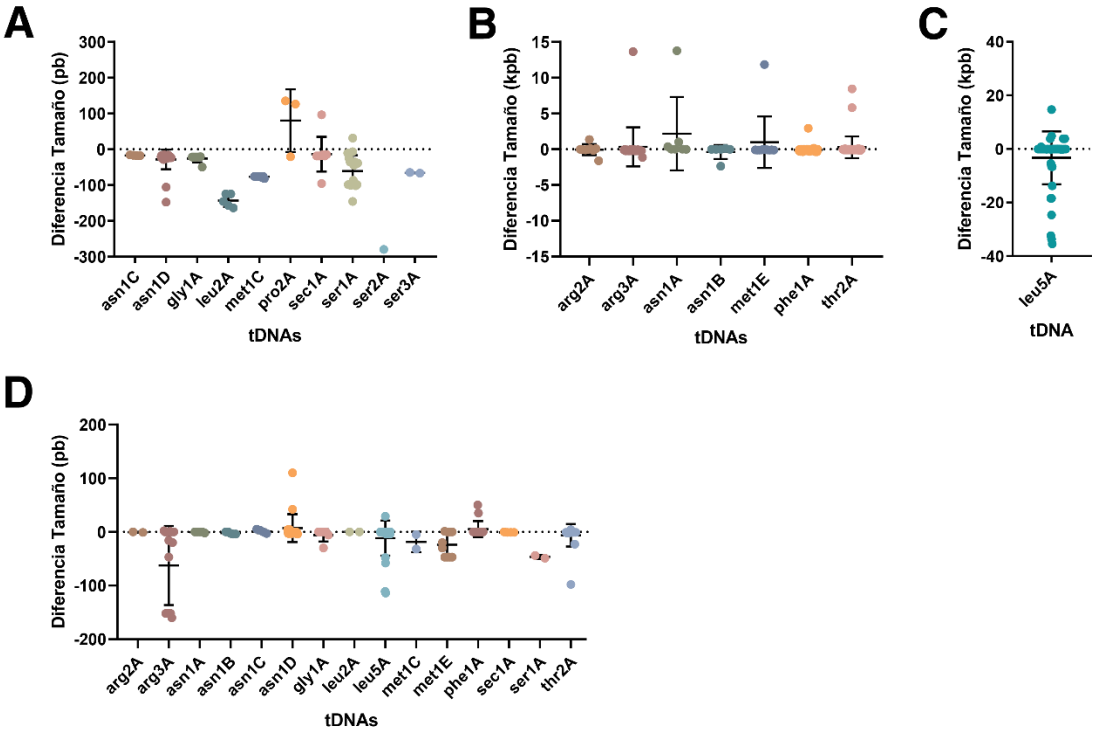


Figura 9: Comparación de los tamaños de las IGs y sus repetidos directos (DR) entre la identificación realizada por el programa y el análisis manual realizado previamente. (A-C) La diferencia entre los tamaños fue calculada como la sustracción del tamaño predicho por el programa y el tamaño predicho de forma manual de las islas integradas en los *loci* en cada uno de los tDNAs analizados. Los tDNAs fueron agrupados en tres grupos según la magnitud de la diferencia. **(D)** El tamaño de los DRs fueron comparados de forma análoga que el tamaño de las islas. En los gráficos, los tDNAs fueron ordenados de forma alfabética y la línea punteada muestra la posición del valor cero (sin diferencias). Junto a los puntos, la barra presenta el promedio de las diferencias y las barras la desviación estándar de los promedios.

Finalmente, se realizó una evaluación comparativa de la capacidad de nuestro programa para identificar IGs, con respecto al desempeño de otros programas que se han utilizado para la identificación de este tipo de elementos. Para ello, fue usada como modelo la isla GIE492, que se encuentra integrada en el *locus asn1C* del cromosoma de la cepa *K. pneumoniae* RYC492. Esta isla fue seleccionada porque previamente se determinó de manera experimental la extensión de la misma, se identificaron los repetidos directos que la limitan y se corroboró que ésta se escinde del cromosoma (Marcoleta et al., 2016). Por lo tanto, se evaluó la capacidad de cinco programas, además del programa desarrollado en este trabajo, para identificar y delimitar la isla GIE492 a partir del cromosoma no anotado de *K. pneumoniae* RYC492. Los programas MSGPI (De Brito et al., 2016) y AlienHunter (Vernikos & Parkhill, 2006) identifican diferencias en la composición nucleotídica del cromosoma. También, se utilizaron los programas IslandPath-DIMOB (Bertelli & Brinkman, 2018), IslandPick (Langille et al., 2008) y SIGI-HMM (Waack et al., 2006), los cuales son forman parte de las herramientas para la búsqueda de islas genómicas que ofrece la base de datos IslandViewer (Langille & Brinkman, 2009), uno de los repositorios de IGs más citados. Al comparar la identificación de la GIE492 realizada por los distintos programas (Figura 10), se observó que todos reportaron la presencia de una o dos IGs en el *locus* bajo análisis. Sin embargo, el nuevo programa desarrollado realizó la identificación más precisa, considerando las coordenadas de la isla determinadas experimentalmente. Para sustentar esta afirmación, fueron calculados los siguientes evaluadores: (i) la precisión, la cual permite conocer en proporción que tan confiable es la predicción realizada, (ii) la sensibilidad, la cual nos permite conocer en proporción que tan exacta es la predicción realizada (ver Figura Anexo 1) y (iii) el valor F1 que indica el equilibrio entre la precisión y la sensibilidad de la predicción, o de alguna manera, es una media de la exactitud de

la predicción. En el caso de cualquiera de los parámetros calculados, un valor de 1 indica la predicción exacta de la IG y un valor de 0 indica que la isla no fue predicha correctamente. Como puede observarse por los valores de los parámetros, el nuevo programa diseñado (Tabla III) consigue predecir la GIE492 exactamente y mejor que cualquiera de los programas utilizados en esta evaluación.

Tabla III: Evaluación de la identificación de la Isla Genómica GI-E492 de la cepa *K. pneumoniae* RYC492 realizada por distintos programas.

Programa	Parámetros de Evaluación		
	Precisión	Sensibilidad	F1
MSGPI	0,47	1	0,63
AlienHunter	0,88	1	0,93
IslandPath-DiMOB	0,95	1	0,97
SIGI-HMM	1	0,71	0,83
IslandPick	1	0,50	0,67
Nuevo Programa	1	1	1

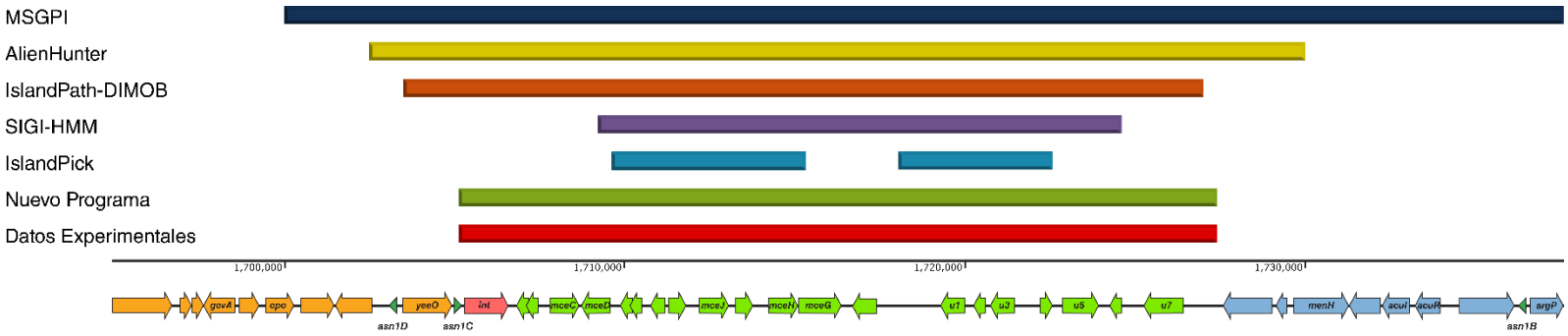


Figura 10: El nuevo programa predice exactamente las coordenadas de la isla GI-E492 de la cepa *K. pneumoniae* RYC492. En el esquema se presenta la posicicion y extension de las islas genómicas predichas por los distintos programas usados para esta comparación. La barra inferior roja muestra la extension de la GI-E492 determinada de forma experimental (Marcoleta, et al., 2016). En la parte inferior del esquema se muestra la posición de las CDS y tDNAs en el locus.

5 DISCUSIÓN Y PROYECCIONES

Actualmente los programas que buscan identificar IGs no consideran o no le dan la debida importancia a los contextos genómicos en el que se encuentran los tDNAs donde se integran dichas islas. En este trabajo, se buscó desarrollar un programa bioinformático capaz de detectar IGs que estuviesen asociadas a tDNAs en genomas de *K. pneumoniae* y, que para ello, el programa utilizase la información de los contextos genómicos de dichos tDNAs. El programa desarrollado se compone de dos herramientas, una se encarga de identificar los tDNAs, y sus respectivos contextos en el genoma, y la otra herramienta se encarga de buscar la presencia de IGs. Diferentes evaluaciones y pruebas indicaron que el programa tiene un muy buen desempeño para la anotación de tDNAs usando una nomenclatura que considera el contexto genómico, además de para identificar IGs integradas en los mismos, mostrando ser mucho más preciso y sensible que otros programas desarrollados previamente por otros investigadores para la búsqueda de IGs.

Previo al trabajo bioinformático propiamente tal, fue necesario registrar y organizar todos los datos de los *loci vírgenes* para que el programa pudiese considerarlos en los análisis que realiza. La curación de estos datos fue ardua, pero se consiguió no solo obtener una colección ordenada de secuencias y anotaciones, también se consiguió que las estrategias de organización de los datos sean lo suficientemente sencillas para que los programas construidos interactúen con las bases de datos y para posibilitar la inspección manual de los datos contenidos. Lo anterior se refleja en que se redujeron a cero los problemas asociados al acceso y consulta del programa a dichas bases de datos. Sin embargo, en este trabajo sólo se consideró un total de archivos correspondiente a secuencias cromosómicas de 50 cepas de *K. pneumoniae*, siendo

que existe una cantidad mucho mayor de secuencias cromosómicas ya ensambladas de otras cepas, y un número mucho mayor de no ensambladas, lo que sugiere que aún debe haber muchos contextos que no se hayan descrito y será necesario tener esta información en la base de datos para futuros análisis, por ello es evidente que en el futuro habrá que agregar más información a la base de datos construida.

Con respecto a la primera parte del programa creado, la herramienta para la identificación y anotación de tDNAs, el objetivo de esta parte del trabajo era construir y evaluar la predicción de los genes que codifican tRNAs en los cromosomas y de esta manera conseguir aplicar una nueva nomenclatura para tDNAs que permitiera identificarlos según su contexto. Para esto fue considerada la base de datos de *loci vírgenes* construida en este trabajo y mediante la comparación de los *loci* se desarrolló un programa que consigue la identificación de los tDNAs con gran precisión.

La comparación de los resultados obtenidos con el programa de identificación y anotación de tDNAs, con respecto a los obtenidos en estudios anteriores, entrego como resultado que existen 3 tDNAs a los cuales el programa no fue capaz de asignarles un identificador del contexto genómico, la razón de este resultado se debe a que los tDNAs predichos son falsos positivos del programa ARAGORN, ya que las coordenadas de estos supuestos tDNAs se encuentran en regiones de la secuencia que ya están ocupadas por CDS. Además se tiene conocimiento que ARAGORN no está exento de hacer predicciones erróneas, la que aumenta en zonas de alto contenido GC (Laslett & Canback, 2004). Para conseguir deshacerse de estos falsos positivos, se podría incorporar a la herramienta de predicción de tDNAs un módulo que permita hacer la anotación de todos las CDS del cromosoma y filtras y descartar aquellos tDNAs cuyas coordenadas se sobrepongan con coordenadas de una región codificante.

Con respecto a los 31 tDNAs a los cuales se les asignó incorrectamente alguno de los identificadores, sólo en 3 de estos casos hubo una incorrecta asignación del contexto genómico. No obstante, se observó que estos 3 tDNAs se encuentran al interior de profagos. La explicación más probable del error en la asignación es que los profagos encontrados en las cepas analizadas sean profagos cercanos filogenéticamente a los profagos de los cuales se obtuvieron los *loci vírgenes* de estos tDNAs, lo que explicaría por qué comparten sólo algunos de los CDS. Sin embargo, el identificador de contexto genómico en estos 3 casos corresponde de igual manera a tDNAs asociados a profagos, por lo que aunque afecta la predicción correcta del tDNA según la nueva nomenclatura, no impide reconocer que estos tDNAs son portados por profagos. En su versión actual, el programa no informa que estos tDNAs se encuentran asociados a profagos, lo que será implementado en futuras versiones.

Con respecto al resto de los casos donde se asignó un identificador erróneamente, en todos ellos los tDNAs se encontraban agrupados formando clusters, formando posibles unidades policistrónicas de tDNAs, en las cuales había ausencia o re-arreglos de uno o más tDNAs en el operón. La ausencia de esos tDNAs provocó que el programa registrara un nuevo P. tDNAs presente en la secuencia actual. Esto a su vez conlleva a que cuando es realizada la comparación de los perfiles le fuera asignado un identificador de contexto incorrecto. Sin embargo, a pesar de la errónea identificación del contexto, éste fue identificado como algún contexto vecino de la misma estructura de tipo operón. La resolución de este problema no es sencillo, dado a que entonces versiones futuras del programa tendrían que considerar la variabilidad de los *loci* de tDNAs. Actualmente se trabaja para conseguir superar este obstáculo.

Adicionalmente, fue desarrollada la herramienta que permite la identificación de Islas Genómicas detectando la disrupción de la continuidad de los *loci vírgenes* definidos previamente a partir de la comparación de un set de 50 cromosomas de *K. pneumoniae*. La comparación entre las predicciones del nuevo programa y la anterior identificación manual de las islas resultó en que sólo 5 de las IGs predichas con el método manual no fueron predichos por el programa. Esta diferencia se debe a que para 3 de estos casos, las IGs tienen su inicio sobre las 1,5 kpb en la región río abajo de sus respectivos tDNAs y el programa considera el inicio de los FDLs dentro de la primera 1kpb, por lo tanto el programa no puede identificar a estas IGs, es evaluada actualmente la solución para este problema. Para los 2 casos faltantes, donde ambos resultaron ser el tDNA *thr2A*, el programa no logro encontrar el contexto correspondiente de la región río abajo, por lo que no fue capaz de realizar la búsqueda de IGs en ambos casos. Al analizar la secuencia de los contextos río abajo de ambos casos, se observó que en ninguno de ellos se encontraba la secuencia correspondiente al contexto río abajo del *locus virgen* de *thr2A*, por lo que en estos dos casos se trata de variantes del *locus virgen* de *thr2A*, los cuales probablemente presentan una deleción en esa región del cromosoma.

El programa identificó 33 FDLs que no fueron considerados por la curación manual de los cromosomas. Todos estos elementos poseían un tamaño inferior a las 5 kpb, y es por esta razón que en la curación manual fueron excluidas. Lo anterior es muestra de que el programa es capaz de identificar FDLs de un tamaño inferior a la de una isla genómica, probablemente estos FDLs sean vestigios de alguna inserción o algún evento de recombinación, los cuales podrían ser analizados en estudios futuros.

En 303 casos, la predicción de Islas Genómicas realizada por el programa coincidió con la realizada de forma manual. Al comparar el tamaño de estas predicciones

y calcular las diferencias absolutas porcentuales, se comprobó que en general la diferencia entre ambas predicciones es baja. Sin embargo, se observó un grupo en que las diferencias entre el tamaño de las predicciones fueron las más altas registradas y con las Diferencias absolutas porcentuales más alejadas al promedio. Estos casos fueron analizados y se encontraron dos patrones entre estas diferencias: (i) en varios casos el programa consiguió detectar un repetido directo que no fue detectado de forma manual, lo que conlleva a que existan diferencias entre los límites de los IGs detectados; (ii) en el caso de tDNAs como *leu5A* y *thr2A*, existe río abajo de estos elementos una región variable. Según resultados previos de nuestro grupo, dichos tDNAs corresponden a dos de los sitios del cromosoma de *K. pneumoniae* con mayor tráfico de IGs, donde en la mayoría de las cepas analizadas se observaron este tipo de elementos integrados (Berríos, 2018). Por lo tanto, la presencia de regiones variables de hasta 15 kb río abajo de dichos tDNAs detectada podrían corresponder a vestigios de IGs que se integraron en estos *loci* en el pasado, y que fueron degenerando producto de posibles re-arreglos por recombinación homóloga entre secuencias de inserción u otro tipo de repetidos. En su versión actual, el programa no reporta la presencia de estas regiones variables, lo cual podría ser implementado en futuras versiones. Lo anterior significa un obstáculo desde el punto de vista del funcionamiento del programa, ya que la región río abajo del tDNA incluye una secuencia que no forma parte del respectivo *locus virgen* y, por lo tanto, afecta a la determinación de la extensión de las IGs. Una forma de poder superar este obstáculo sería analizar en detalle y clasificar las regiones variables encontradas y ver si estas regiones se repiten con cierto patrón en el cromosoma de otras cepas. Luego, esta información podría ser incorporada a la base de datos de *loci vírgenes* y así lograr que el programa reconozca y considere estas regiones variables al momento de realizar los análisis, tal como si fuesen otros posibles contextos para el tDNA en cuestión.

Por otro lado, el programa fue diseñado para la identificación de elementos que frecuentemente se encuentran en IGs: integrasas y repetidos directos. Con respecto a los repetidos directos, el promedio de las Diferencias absolutas porcentuales calculadas al comparar los tamaños predichos por ambos métodos no fue el esperado, en parte este resultado se obtuvo ya que los tamaños de los repetidos directos predichos fueron muy variados, lo que se comprueba al observar la desviación estándar del promedio de las Diferencias absolutas porcentuales. Por otro lado, hubo casos en donde los tamaños predichos por el programa fueron iguales a los tamaños predichos por el método manual, lo que demuestra que el programa puede realizar buenas predicciones de estos elementos. También no se debe descartar que haya errores en los tamaños predichos por el método manual, ya que este método implica un error humano asociado. Por otro lado, se obtuvo que la predicción de integrasas por parte del programa tiene un acierto sobre el 70%, esto implica que tiene buena confiabilidad, sin embargo, ésta puede seguir mejorándose. Además, se obtuvo que secuencias correspondientes a transposasas fueron reconocidas como integrasas, sin embargo, estas predicciones fueron parciales, ya que en ninguno de los casos se predijo la secuencia nucleotídica completa, de hecho nunca fue más de la mitad de la secuencia de la transposasa. Lo anterior no revierte mayor misterio, dado a que tanto integrasas como transposas median la integración de EGMs en el cromosoma, lo que explica que posean regiones de sus genes que son similares. La razón de los resultados obtenidos con respecto a la predicción de integrasas por parte del programa se debe probablemente a que existe un problema con el procedimiento utilizado por parte de la herramienta BLASTn, donde se deba ser más estricto con los parámetros de similitud y de cobertura, lo que evitaría que parte de la secuencia de genes de transposasas fuesen predichos como integrasas. Otra posible solución pudiese ser utilizar las secuencias proteicas de integrasas, y luego mediante la

herramienta tBLASTn buscar en las secuencias de los FDLs encontrados posibles secuencias de integrasas. Otra estrategia puede ser anotar los genes en las secuencias de los FDLs encontrados utilizando la herramienta PROKKA (Seemann, 2014), que fue diseñada para la anotación de genes, de esta manera se podrá conocer los lugares de la secuencia de los FDLs encontrados que posean integrasas, además de otros genes. En resumen, a pesar de que la predicción de repetidos directos e integrasas en esta versión del programa es correcta en la mayoría de los casos, aún deben mejorarse las estrategias de identificación de estos elementos para mejorar la cobertura y precisión de la identificación.

Respecto de las proyecciones de este trabajo, conviene señalar que la herramienta desarrollada se construyó específicamente para su uso con cepas de *K. pneumoniae*, puesto que se basa en la organización particular de su cromosoma y en particular en el contexto en el que se encuentran los tDNAs. Sin perjuicio de lo anterior, la única barrera para ampliar el uso de la herramienta desarrollada para el análisis de otras especies bacterianas es la ampliación de la base de datos de *loci vírgenes*, que en su primera versión fue creada sólo a partir de secuencias de *K. pneumoniae*. Para ampliar la base de datos, deben llevarse a cabo análisis masivos de otras especies para resolver la configuración de los *loci vírgenes* de tDNAs, y de esa manera permitir el análisis de los contextos en secuencias problema de otras especies. Por otro lado, la base de datos podría fragmentarse en secciones, de manera que a través del programa el usuario pueda definir si utilizar toda la base de datos o una subdivisión a nivel de familias o especies de bacterias.

Considerando todo lo expuesto en este trabajo, el objetivo de desarrollar un programa para la aplicación de la nueva nomenclatura y la predicción de Islas

Genómicas se cumplió exitosamente. Luego de utilizar esta primera versión del programa sobre los cromosomas de las cepas de *K. pneumoniae* consideradas en este trabajo se determinó que nuevas versiones de este programa deben considerar las mejoras que se mencionan a continuación. En primer lugar, la posibilidad de que los *loci vírgenes* nuevos sean registrados por el programa. De esta manera, al analizar cientos o miles de cromosomas, el programa pueda registrar en las bases de datos nuevos *loci vírgenes*, mejorando así la cobertura del programa. Además, es deseable una optimización adicional para trabajar con genomas ensamblados en múltiples segmentos o contigs, que corresponden al estado de completitud en el ensamblaje más frecuente entre los genomas disponibles en bases de datos públicas. Por otro lado, es deseable implementar en el programa la posibilidad de detectar la integración de IGs en tándem (dos IGs se integran secuencialmente en el mismo tDNA, quedando una tras de otra), lo que no es tan extraño entre las IGs que ocupan los tDNAs como sitios de integración, y que fue detectado en varios casos por la anotación manual de IGs. También, se podría implementar una función para la anotación de todos los posibles genes portados por los IGs o incluso de todo el cromosoma para análisis ulteriores de estos elementos y las funciones codificadas en ellas. También, se podría considerar la opción de que al inicio del programa el usuario defina qué clase de FDLs deba reportar el programa y cuáles no, de esta manera se evitaría que el usuario deba analizar todos los FDLs detectados por el programa y determinar cuáles de ellos corresponden a IGs.

Finalmente, se espera implementar prontamente esta herramienta de forma online en la plataforma de la *Klebsiella pneumoniae tDNA-associated Genomic Islands Database* (KleptGI-DB), la cual se espera pronto esté disponible para la comunidad científica como uno de los repositorios más grandes de Islas Genómicas de esta especie

bacteriana. Además, no se descarta que el análisis de secuencias de otras enterobacterias permita ampliar la base de datos y, por ende, la capacidad del programa de hacer predicciones de IGs en otras especies bacterianas.

6 CONCLUSIONES

1.- A partir de este trabajo, se desarrolló una herramienta altamente sensible y precisa para la anotación de tDNAs y la identificación de IGs en *K. pneumoniae*.

2.- El programa desarrollado, basado en el análisis del contexto genómico de cada tDNA presente en el cromosoma, no requiere de la presencia de elementos de secuencia comúnmente asociados a IGs para su identificación. Esto permite superar el obstáculo de la alta variabilidad en la presencia de este tipo de elementos.

3.- La primera versión de este programa mostró un desempeño muy satisfactorio en la anotación de tDNAs y la identificación de IGs en un set de 50 genomas de *K. pneumoniae* analizados y curados previamente en búsqueda de dichos elementos. Además, mostró una gran precisión y sensibilidad, al analizar la isla genómica GIE492 delimitada y caracterizada experimentalmente, teniendo un mejor desempeño que cualquiera de los otros programas utilizados.

4.- La evaluación del desempeño del programa permitió identificar aspectos relevantes que pueden ser mejorados, entre ellos: a) mejores estrategias para la identificación de repetidos directos e integrasas; b) optimizaciones para el análisis de cromosomas fragmentados en múltiples contigs; c) optimizaciones que le permitan lidiar de mejor manera con regiones río abajo variables, como aquellas que se observan en tDNAs con álto tráfico de IGs tales como *thr2A* y *leu5A*.

5.- El programa diseñado en este trabajo es de gran utilidad para el análisis masivo de cromosomas de *K. pneumoniae* secuenciados, de modo de aumentar nuestro conocimiento sobre IGs y dinámica del genoma en esta especie, así como también para

determinar qué funciones están codificadas en estos elementos y evaluar su importancia en el desarrollo y evolución de cepas multi-resistentes e hipervirulentas.

7 REFERENCIAS

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410.

Berríos, C. (2018). *Análisis de los genes que codifican RNAs de transferencia (tDNAs) como sitios de integración de islas genómicas en Klebsiella pneumoniae* (seminario de título). Facultad de ciencias, Universidad de Chile. Santiago, Chile.

Berríos, C., Rosalba, L., & Marcoleta, A. E. (Noviembre, 2018). Transfer RNA Genes Transcribed as Monocistronic Units are Preferred for Genomic Islands Integration in *Klebsiella pneumoniae*. Poster presentado en el XXIV Congreso Latinoamericano de Microbiología. Santiago, Chile.

Bertelli, C., & Brinkman, F. S. L. (2018). Improved genomic island predictions with IslandPath-DIMOB. *Bioinformatics*, *1*, 7.

Boyd, E. F., Almagro-Moreno, S., & Parent, M. A. (2009). Genomic islands are dynamic, ancient integrative elements in bacterial evolution. *Trends in Microbiology*, *17*(2), 47–53.

Bush, E. C., Clark, A. E., DeRanek, C. A., Eng, A., Forman, J., Heath, K., … Wu, H. (2018). xenoGI: Reconstructing the history of genomic island insertions in clades of closely related bacteria. *BMC Bioinformatics*, *19*(1), 32.

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., … De Hoon, M. J. L. (2009). Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*(11), 1422–1423.

Coordinators, N. R. (2016). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, *44*(Database issue), D7.

Darling, A. E., Mau, B., & Perna, N. T. (2010). Progressivemauve: Multiple genome alignment with gene gain, loss and rearrangement. *PloS One*, *5*(6), e11147.

De Brito, D. M., Maracaja-Coutinho, V., De Farias, S. T., Batista, L. V., & Do Rego, T. G. (2016). A Novel method to predict genomic islands based on mean shift clustering algorithm. *PLoS ONE*, *11*(1), e0146352.

Dobrindt, U., Hochhut, B., Hentschel, U., & Hacker, J. (2004). Genomic islands in pathogenic and environmental microorganisms. *Nature Reviews Microbiology*, *2*(5), 414–424.

Edgar, R. C. (2004). MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, *5*(1), 113.

Frost, L. S., Leplae, R., Summers, A. O., & Toussaint, A. (2005). Mobile genetic elements: The agents of open source evolution. *Nature Reviews Microbiology*, *3*(9), 722–732.

Hacker, J., & Carniel, E. (2001). Ecological fitness, genomic islands and bacterial pathogenicity: A Darwinian view of the evolution of microbes. *EMBO Reports*, *2*(5), 376–381.

Holt, K. E., Wertheim, H., Zadoks, R. N., Baker, S., Whitehouse, C. A., Dance, D., … Thomson, N. R. (2015). Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in *Klebsiella pneumoniae* , an urgent threat to public health. *Proceedings of the National Academy of Sciences*, *112*(27), E3574–E3581.

Hudson, C. M., Lau, B. Y., & Williams, K. P. (2015). Islander: A database of precisely mapped genomic islands in tRNA and tmRNA genes. *Nucleic Acids Research*, *43*(D1), D48–D53.

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.

Lam, M. M. C., Wyres, K. L., Duchêne, S., Wick, R. R., Judd, L. M., Gan, Y. H., … Holt, K. E. (2018). Population genomics of hypervirulent *Klebsiella pneumoniae* clonal-group 23 reveals early emergence and rapid global dissemination. *Nature Communications*, *9*(1), 2703.

Langille, M. G. I., & Brinkman, F. S. L. (2009). IslandViewer: An integrated interface for computational identification and visualization of genomic islands. *Bioinformatics*, *25*(5), 664–665.

Langille, M. G. I., Hsiao, W. W. L., & Brinkman, F. S. L. (2008). Evaluation of genomic island predictors using a comparative genomics approach. *BMC Bioinformatics*, *9*(1), 329.

Laslett, D., & Canback, B. (2004). ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Research*, *32*(1), 11–16.

Liu, H., & Zhu, J. (2010). Analysis of the 3′ ends of tRNA as the cause of insertion sites of foreign DNA in Prochlorococcus. *Journal of Zhejiang University SCIENCE B*, *11*(9), 708–718.

Lu, B., & Leong, H. W. (2016). GI-SVM: a sensitive method for predicting genomic islands based on unannotated sequence of a single genome. *Journal of bioinformatics and computational biology*, *14*(01), 1640003.

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.

Marr, C. M., & Russo, T. A. (2018). *Hypervirulent Klebsiella pneumoniae*: a new public health threat. *Expert Review of Anti-Infective Therapy*.

McKinney, W. (2011). pandas: a Foundational Python Library for Data Analysis and Statistics. *PyHPC*, 1–9.

Nordmann, P., Cuzon, G., & Naas, T. (2009). The real threat of *Klebsiella pneumoniae* carbapenemase-producing bacteria. *The Lancet Infectious Diseases*, *9*(4), 228–236.

Ou, H. Y., Chen, L. L., Lonnen, J., Chaudhuri, R. R., Thani, A. Bin, Smith, R., … Rajakumar, K. (2006). A novel strategy for the identification of genomic islands by comparative analysis of the contents and contexts of tRNA sites in closely related bacteria. *Nucleic Acids Research*, *34*(1), e3–e3.

Podschun, R., & Ullmann, U. (1998). *Klebsiella spp.* as Nosocomial Pathogens: Epidemiology, Taxonomy, Typing Methods, and Pathogenicity Factors. *Clinical Microbiology Reviews*, *11*(4), 589–603.

Reiter, W., Palm, P., & Yeats, S. (1989). Transfer RNA genes frequently serve as integration sites for prokaryotic genetic elements. *Nucleic Acids Research*, *17*(5), 1907–1914.

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

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

Shepherd, J., & Ibba, M. (2015). Bacterial transfer RNAs. *FEMS Microbiology Reviews*, *39*(3), 280–300.

Shon, A. S., Bajwa, R. P. S., & Russo, T. A. (2013). Hypervirulent (hypermucoviscous) *Klebsiella Pneumoniae*: A new and dangerous breed. *Virulence*, *4*(2), 107–118.

Sirot, J., Chanal, C., Petit, A., Sirot, D., & Labia, R. (1988). *Klebsiella pneumoniae* and Other Enterobacteriaceae Producing Novel Plasmid-Mediated β- Lactamases Markedly Active against Third-Generation Cephalosporins : Epidemiologic Studies Author (s): J . Sirot , C . Chanal , A . Petit , D . Sirot , R . Labia and. *REVIEWS OF INFECTIOUS DISEASE*, *10*(4), 850–859.

Struve, C., Roe, C. C., Stegger, M., Stahlhut, S. G., Hansen, D. S., Engelthaler, D. M., … Krogfelt, K. A. (2015). Mapping the evolution of hypervirulent *Klebsiella pneumoniae*. *MBio*, *6*(4), e00630-15.

Touchon, M., Hoede, C., Tenaillon, O., Barbe, V., Baeriswyl, S., Bidet, P., … Denamur, E. (2009). Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLoS Genetics*, *5*(1), e1000344.

Vernikos, G. S., & Parkhill, J. (2006). Interpolated variable order motifs for identification of horizontally acquired DNA: Revisiting the Salmonella pathogenicity islands. *Bioinformatics*, *22*(18), 2196–2203.

Waack, S., Keller, O., Asper, R., Brodag, T., Damm, C., Fricke, W. F., … Merkl, R. (2006). Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models. *BMC Bioinformatics*, *7*(1), 142.

Williams, K. P. (2002). Integration sites for genetic elements in prokaryotic tRNA and tmRNA genes: sublocation preference of integrase subfamilies. *Nucleic Acids Research*, *30*(4), 866–875.

8 ANEXOS

Tabla Anexo I: Cepas de *Klebsiella pneumoniae* analizadas en este trabajo.

KpI	KpN06	ST278	n/a	CP012992.1	2013	Canada	Humano	Sangre
KpI	KPNIH10	ST258	CG258	NZ_CP007727.1	2011	USA	Humano	Ingle
KpI	KPNIH24	ST258	CG258	NZ_CP008797.1	2012	USA	Humano	Garganta/Ingle
KpI	KPNIH27	ST34	CG34	NZ_CP007731.1	2012	USA	Humano	Ingle
KpI	KPNIH29	ST1518	n/a	NZ_CP009863.1	2013	USA	Humano	Hisopo perirrectal
KpI	KPNIH30	ST258	CG258	NZ_CP009872.1	2013	USA	Humano	Región perirrectal
KpI	KPNIH31	ST392	CG147	NZ_CP009876.1	2013	USA	Humano	Tracto urinario
KpI	KPNIH32	ST258	CG258	NZ_CP009775.1	2013	USA	Humano	Hisopo perirrectal
KpI	KPNIH33	ST258	CG258	NZ_CP009771.1	2013	USA	Humano	Hisopo perirrectal
KpI	KPR0928	ST258	CG258	NZ_CP008831.1	2012	USA	Humano	Esputo
KpI	MGH78578	ST38	n/a	CP000647.1	1994	USA	Humano	Sangre
KpI	MS6671	ST147	CG147	NZ_LN824133.1	2014	Emiratos	Humano	Orina
KpI	NTUH-K2044	ST23	CG23	AP006725.1	2008	Taiwan	Humano	Meningitis
KpI	PittNDM01	ST14	CG14/15	CP006798.1	2013	USA	Humano	Tracto urinario
KpI	PMK1	ST15	CG14/15	NZ_CP008929.1	2011	Nepal	Humano	Tracto urinario
KpI	RJF999	ST23	CG23	CP014010.1	2015	China	Humano	Tracto urinario
KpI	RYC492	ST35	CG35	APGM01000001.1	1984	España	Humano	Heces
KpI	U25	ST14	CG14/15	CP012043.1	2010	India	Humano	Tracto urinario
KpI	UHKPC07	ST258	CG258	CP011985.1	2015	USA	Humano	Orina
KpI	UHKPC33	ST258	CG258	CP011989.1	2014	USA	Humano	Orina
KpI	XH209	ST17	CG17/20	CP009461.1	2013	China	Humano	Sangre
KpII	HKUOPLC	ST480-2LV	n/a	CP012300.1	2013	Hong Kong	Panda gigante	Heces
KpIII	342	ST146	n/a	CP000964.1	2008	USA	Maíz	Tejido del tallo de *Zea mays*
KpIII	KP5-1	novel ST	n/a	CP008700.1	2005	USA	*Nezara viridula* (insecto)	Desconocida

KpI	KCTC2242	ST375	CG65	CP002910.1	2010	Republica de Corea	Ambiental	Productora de 2,3-butanodiol
KpI	KP13	ST442	n/a	NZ_CP003999.1	2009	Brasil	Humano	Sangre
KpI	Kp52.145	ST66	CG66	NZ_FO834906.1	1935	Indonesia	Humano	Tracto respiratorio
KpI	KP617	ST14	CG14/15	CP012753.1	2013	Corea del Sur	Humano	Paciente con quemaduras
KpI	KpN01	ST278	n/a	CP012987.1	2013	Canada	Humano	Tracto urinario
KpI	KpN06	ST278	n/a	CP012992.1	2013	Canada	Humano	Sangre
KpI	KPNIH10	ST258	CG258	NZ_CP007727.1	2011	USA	Humano	Ingle
KpI	KPNIH24	ST258	CG258	NZ_CP008797.1	2012	USA	Humano	Garganta/Ingle
KpI	KPNIH27	ST34	CG34	NZ_CP007731.1	2012	USA	Humano	Ingle
KpI	KPNIH29	ST1518	n/a	NZ_CP009863.1	2013	USA	Humano	Hisopo perirrectal
KpI	KPNIH30	ST258	CG258	NZ_CP009872.1	2013	USA	Humano	Región perirrectal
KpI	KPNIH31	ST392	CG147	NZ_CP009876.1	2013	USA	Humano	Tracto urinario
KpI	KPNIH32	ST258	CG258	NZ_CP009775.1	2013	USA	Humano	Hisopo perirrectal
KpI	KPNIH33	ST258	CG258	NZ_CP009771.1	2013	USA	Humano	Hisopo perirrectal
KpI	KPR0928	ST258	CG258	NZ_CP008831.1	2012	USA	Humano	Esputo
KpI	MGH78578	ST38	n/a	CP000647.1	1994	USA	Humano	Sangre
KpI	MS6671	ST147	CG147	NZ_LN824133.1	2014	Emiratos	Humano	Orina
KpI	NTUH-K2044	ST23	CG23	AP006725.1	2008	Taiwan	Humano	Meningitis
KpI	PittNDM01	ST14	CG14/15	CP006798.1	2013	USA	Humano	Tracto urinario
KpI	PMK1	ST15	CG14/15	NZ_CP008929.1	2011	Nepal	Humano	Tracto urinario
KpI	RJF999	ST23	CG23	CP014010.1	2015	China	Humano	Tracto urinario
KpI	RYC492	ST35	CG35	APGM01000001.1	1984	España	Humano	Heces
KpI	U25	ST14	CG14/15	CP012043.1	2010	India	Humano	Tracto urinario
KpI	UHKPC07	ST258	CG258	CP011985.1	2015	USA	Humano	Orina
KpI	UHKPC33	ST258	CG258	CP011989.1	2014	USA	Humano	Orina
KpI	XH209	ST17	CG17/20	CP009461.1	2013	China	Humano	Sangre

n/a: No disponible

Figura Anexo 1: Precisión y Sensibilidad

La precisión es calculada como la proporción entre los elementos correctamente predichos (verdaderos positivos) con respecto a todos los elementos predichos (verdaderos positivos y falsos positivos). La sensibilidad es calculada como la proporción entre los elementos correctamente predichos con respecto al total de todos los elementos realmente positivos (verdaderos positivos y falsos negativos).

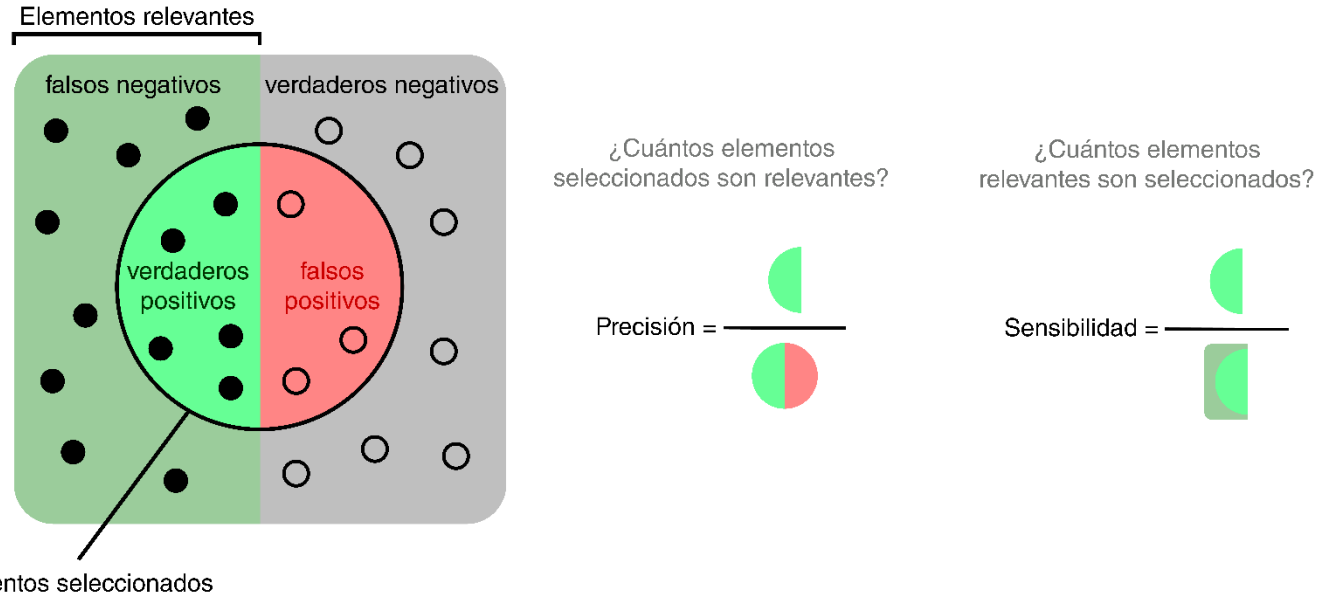


Figura adaptada de https://commons.wikimedia.org/wiki/File:Precisionrecall.svg (Abril, 2019)