Flujo de trabajo para el ensamblaje, anotación y comparación de genomas de bacteriófagos

Estudiantes:

Luis Guillermo Gómez Orozco

Johnnatan Ruiz Escobar

Asesor:

Juan Esteban Pérez Jaramillo, Ph, D.

Co-asesor:

Cristian David Grisales, Biólogo

Instituto de Biología

Facultad de Ciencias Exactas y Naturales

Universidad de Antioquia

2021

Resumen

El complejo de especies *Ralstonia solanacearum* (RSSC) es el causante de una grave enfermedad de plantas denominada marchitez bacteriana, la cual afecta una gran variedad de cultivos de importancia económica. Las estrategias actuales para el control de este patógeno son limitadas y poco eficientes, haciendo necesario la búsqueda de nuevas alternativas. Los bacteriófagos, virus que parasitan bacterias, tienen un enorme potencial para el control de este patógeno, y en los últimos años se han aislado y caracterizado “fagos” como una posible herramienta de biocontrol de *R. solanacearum*. La caracterización genómica de los bacteriófagos es un prerrequisito fundamental para su posterior implementación en procesos de fagoterapia, y por ende es importante desarrollar protocolos de procesamiento y análisis de datos que permitan procesar las secuencias de una manera ágil y rigurosa. En este trabajo se presenta un flujo de trabajo bioinformático basado en el uso de plataformas web y programas de libre acceso para el procesamiento y análisis de secuencias producto del secuenciamiento de genomas, con la ventaja de que puede ser utilizado por personas sin experiencia en programación. Este flujo de trabajo permite realizar el pre-procesamiento, ensamblaje y anotación de genomas, así como la ejecución de análisis de genómica comparativa, mediante el uso de las mejores herramientas disponibles para cada proceso. Para probar su funcionalidad, se ejecutaron todos los pasos con el genoma del fago de *R. solanacearum* NJ-P3, obteniendo un adecuado ensamblaje, así como una anotación exitosa. Para el análisis de sintenia se incluyeron también los fagos de *R. solanacearum* DU_RP_II y RpY1, los cuales resultaron ser genómicamente similares a NJ-P3, con algunas variaciones en la organización del genoma. Este flujo de trabajo es una alternativa rápida y rigurosa para el análisis de datos de genomas bacteriófagos, y es un aporte desde la bioinformática para potenciar el uso de la fagoterapia para el control de patógenos bacterianos en cultivos de importancia económica.

Palabras clave: *Ralstonia solanacearum*, bacteriófagos, ensamblaje, anotación, flujo de trabajo, comparación genómica.

Tabla de contenido

1. Marco teórico

1.1 Los bacteriófagos

Características e importancia ecológica

Los bacteriófagos, o comúnmente también llamados “fagos”, son virus que infectan bacterias y fueron descritos inicialmente por Frederick Twort en 1915 (Twort, 1915; Keen, 2015) y posteriormente, en 1917, por Félix d’Herelle, quien también les dio el nombre de bacteriófagos por su capacidad de lisar bacterias (etimológicamente del griego phagein: “comer”) (D’Herelle, 1917; Keen, 2015). Los fagos poseen un alto nivel de especificidad, capacidad de persistir a largo plazo en la mayoría de los ambientes, y tienen la facultad de reproducirse rápidamente en huéspedes apropiados. Estas características contribuyen al equilibrio y evolución de las bacterias en los ecosistemas, puesto que actúan como organismo de control en el crecimiento poblacional bacteriano, y algunos pueden alterar genéticamente la composición nucleotídica de las bacterias mediante el ciclo lisogénico (profagos) (Kutter & Sulakvelidze, 2005). Se estima que los bacteriófagos, con un número alrededor de 10^{30} a 10^{32}, son los organismos biológicos más abundantes de la tierra (Ackermann, 2001). Se han aislado fagos de todos los entornos en los que existen bacterias, donde cada especie bacteriana es infectada al menos por un tipo de fago, y muy probablemente muchas bacterias son infectadas por más de un fago (Comeau *et al.,* 2008).

Los fagos varían ampliamente en tamaño, forma y complejidad, y los genomas que contienen son aún más diversos (Ackermann, 2001). Pueden presentar genomas compuestos por ADN o ARN mono o bicatenario, con tamaños que oscilan de las 3.400 pb hasta 735.000 pb y, a diferencia de las bacterias, no hay un solo gen (por ejemplo, ARNr 16S) presente en todos los genomas de los fagos (Ackermann, 2001). La conformación proteica y su organización en el genoma es igualmente variable, y es probable que estos genomas representen el mayor depósito de genes y proteínas que existen (Keen, 2015). Actualmente existen 13 familias de fagos que se clasifican de acuerdo a la naturaleza de su genoma (de cadena sencilla (ss), de doble cadena (ds), ARN o ADN) y a sus características

morfológicas. Según Ackermann (2001), de los aproximadamente 5100 bacteriófagos analizados por microscopía electrónica desde el año 1959, los fagos con cola representan el 96% del total y dentro de este grupo la distribución de los aislados experimentales es de 60,8% a la Siphoviridae, 25,1% perteneciente a la familia Myoviridae, y 14,1% a la Podoviridae. Myoviridae tiene una cola contráctil larga, mientras que Sipho- y Podoviridae tienen colas no contráctiles largas y cortas, respectivamente. Si bien esta clasificación es bastante popular, su significado evolutivo está lejos de ser claro (Brüssow & Hendrix, 2002).

Mecanismos de replicación de los bacteriófagos

Ciclo lítico

Para nuestro caso describiremos el mecanismo de replicación de virus de ADN, específicamente de bacteriófagos. En la síntesis de los fagos líticos, el virus se replica en el interior de la bacteria y el proceso finaliza con la lisis (destrucción) del huésped inducido por el fago y la liberación de las partículas virales. El ciclo lítico para lo bacteriófagos está dividido en cinco etapas bien definidas:

I. *Adsorción*: El ciclo lítico se inicia a través de la adsorción de un fago a la superficie de una bacteria sensible, denominada “hospedador homólogo”. Este es un evento altamente específico, que depende de la presencia de receptores adecuados ubicados en la pared celular bacteriana. Los fagos se adsorben a la célula por medio de la cola, la cual generalmente posee una placa basal, y estructuras adicionales como fibras que otorgan estabilidad a esta unión (Neve, 1996).

II. *Inyección del ADN viral*: Luego de la adsorción del fago a una célula sensible, el ADN del fago es inyectado desde la cabeza a través de la cola hacia el interior de la célula bacteriana, mientras la partícula del fago vacía o “fantasma” queda en la superficie externa de la bacteria (Neve, 1996).

III. Biosíntesis de componentes virales: Una vez que el ADN del fago ha sido inyectado en la célula bacteriana, el metabolismo normal de la célula se detiene y toda la maquinaria biosintética bacteriana estará dedicada a sintetizar las nuevas

moléculas de ADN del fago (replicación del ADN) y por consiguiente la síntesis de las proteínas del fago (Neve, 1996). Durante este período, las moléculas precursoras de la progenie viral se encuentran en etapa de síntesis o dispersas en el interior de la célula, dicha fase está contenida dentro del período de latencia (Klaenhammer & Fitzgerald, 1994).

IV. Ensamblaje *de novo*: Durante esta etapa, las proteínas estructurales se ensamblan y las nuevas moléculas de ADN del fago son empaquetadas en el interior de las cabezas, y finalmente se obtienen las partículas virales completas (Neve, 1996). El proceso de ensamblaje es de tipo secuencial, y se sintetizan en forma aislada las colas y las proteínas de la cápside.

V. Lisis de la célula bacteriana huésped: El ciclo lítico se completa cuando se produce la ruptura de la pared o lisis celular, a través de enzimas llamadas lisinas, que están codificadas en el genoma del fago. Las nuevas partículas virales ya ensambladas son liberadas al medio. El período durante el cual se produce la liberación de la progenie viral se conoce como “período exponencial”; mientras que el número de partículas fágicas liberadas a partir de una única célula infectada se denomina burst-size o número de explosión (Garvey *et al.,* 1995).

Ciclo lisogénico

En este mecanismo de replicación el ADN del fago se replica en el interior de la célula, sin embargo, la bacteria huésped se mantiene viable y no hay liberación de viriones. Los virus que presentan ciclos lisogénicos son muy diferentes a aquellos bacteriófagos virulentos que provocan lisis instantánea en la célula huésped con posterior liberación de partículas de la progenie del fago. En este caso, una vez el genoma del fago ingresa a la célula procariota puede persistir en forma de plásmido o incorporarse al ADN cromosómico de la bacteria y permanecer latente en forma de profago, el cual posteriormente puede replicarse en sincronía con el ADN del huésped a medida que este se divide (Buttimer *et al*., 2017). El profago puede permanecer por un tiempo indefinido en este estadio y replicarse sin ser patógeno para la célula huésped. Una vez se presenta un cambio en el medio celular, este puede provocar la liberación y activación del profago, promoviendo el inicio del ciclo lítico (Buttimer *et al.,* 2017). Estos desencadenantes pueden ser de

naturaleza química o física, como cambios en la temperatura o exposición a luz ultravioleta (Brunner & Pootjes, 1969; Guglielmotti *et al.,* 2012). En algunos casos la incorporación del genoma del virus en el material genético procariota puede provocar que se codifiquen genes que optimicen algunas características del huésped bacteriano como pueden ser la resistencia a antibióticos y/o el aumento en la virulencia (Lin *et al.,* 2017). Lo anteriormente mencionado puede ser un problema a la hora de utilizar bacteriófagos para el control biológico de bacterias, y por tal motivo uno de los principales factores que se tiene en cuenta para decidir si un fago es utilizable para biocontrol es que sean específicamente líticos.

Fagos del complejo de especies de Ralstonia solanacearum

La bacteria *Ralstonia solanacearum,* considerada un complejo de especies debido a su alta plasticidad fenotípica y genotípica, es una bacteria Gram negativa aeróbica de la familia *Burkholderiaceae*, agente causal de la marchitez bacteriana, enfermedad muy limitante y de difícil control en cultivos de importancia económica (Barrios *et al.,* 2008; Martínez, 2010). Una potencial alternativa para el manejo de la enfermedad causada por *R. solanacearum* es el uso de bacteriófagos líticos. Una amplia variedad de bacteriófagos que infectan específicamente cepas del complejo de especies de *R. solanacearum* han sido aisladas e identificadas*,* pertenecientes a las familias Inoviridae, Myoviridae, Podoviridae o Siphoviridae (Addy *et al.,* 2018). Inicialmente, fagos infectando a *R. solanacearum* fueron aislados de suelos rizosféricos de plantas de tabaco (Tanaka *et al.,* 1990). Posteriormente, fagos infectando a la bacteria en diversos tipos de suelos y cultivos fueron reportados y se comenzó a validar experimentalmente su capacidad bactericida. Por ejemplo, tres fagos líticos (RSA1, RSB1 y RSL1), aplicados solos o en mezcla, fueron usados para tratar células de *R. solanacearum* en condiciones in vitro, resultando en una drástica y rápida reducción de las poblaciones de la bacteria. Posteriormente, la aplicación individual del fago RSL1 en plántulas de tomate infectadas con la bacteria mostró un control del 100% de la marchitez bacteriana durante el tiempo de la evaluación (Fujiwara *et al.,* 2011). Aplicaciones sucesivas de otro bacteriófago lítico,

denominado PE204, en la rizosfera de plántulas de tomate inoculadas con cepas virulentas de *R. solanacearum*, bajo condiciones de invernadero, inhibieron la aparición de síntomas de marchitez bacteriana (Bae *et al.,* 2012).

A la par del aislamiento de bacteriófagos y de la caracterización de su capacidad lítica para el control de cepas del complejo de especies de *R. solanacearum*, la caracterización genómica es fundamental para determinar su potencial como posibles agentes de biocontrol, y actualmente se pueden encontrar en las bases de datos públicas varios genomas de fagos específicos de *R. solanacearum*. Por ejemplo, recientemente se reportaron los genomas de los fagos RsoM1USA (Addy *et al.,* 2019), considerado el primer fago aislado en suelos de Estados Unidos de América, y el bacteriófago RsoP1EGY (Ahmad *et al.,* 2018) aislado en África (Egipto). También se cuenta con datos recientes de genomas de 23 fagos aislados en islas del océano indico (Mauricio y Reunión) que infectan al complejo de especies de *Ralstonia solanacearum* (Trotereau *et al.,* 2021). Estos genomas van desde tamaños pequeños como el genoma del inovirus PE226, fago de ARNss con 5.475 pb que codifican únicamente 9 ORF, o genomas más grandes como el del fago DU_RP_ll (Park, 2018) perteneciente a la familia Podoviridae, con una longitud de secuencia de 42,091pb (38 ORF), o el fago RpY1 (Lee *et al.,* 2021) con un tamaño de genoma de 43.284 pb (53 ORF). Por último, podemos encontrar bacteriófagos infectantes de *R. solanacearum* que son considerados fagos jumbo (fagos con un tamaño de genoma de mínimo 200 kpb) como es el caso de los fagos RSF1 y RSL2 (Bhunchoth *et al.,* 2016), cada uno con 237 ORF, ambos de la familia Myoviridae. Este amplio rango de tamaño en los genomas de los fagos que infectan al complejo de *Ralstonia solanacearum* (RSSC) da evidencia de la gran diversidad que se puede encontrar dentro de este grupo en términos de organización del genoma, estructura, propagación y evolución (Yuan & Gao, 2017).

La base de datos de GenBank incluye la taxonomía actualizada de 35 fagos que infectan al complejo de especies de *Ralstonia solanacearum* (RSSC), los cuales están clasificados de acuerdo con las guías del Comité Internacional de

Taxonomía de Virus (ICTV). Este número es bajo en comparación, por ejemplo, con los 224 genomas de fagos de *Pseudomonas* o los fagos que infectan bacterias asociadas con humanos como *Escherichia*, con 270 genomas reportados (Trotereau *et al.,* 2021). Sin embargo, es de esperar que la cantidad de genomas de fagos reportados contra *R. solanacearum* aumente, debido a su potencial para ser usados como principal alternativa para el biocontrol de la marchitez bacteriana causado por la bacteria *(*Álvarez & Biosca, 2017), así como al constante desarrollo de las tecnologías de secuenciación de nueva generación (NGS) que han facilitado la secuenciación masiva de múltiples genomas, en poco tiempo y a un menor costo por base (Rihtman *et al.,* 2016). Este último factor ha permitido la generación de una mayor cantidad de datos genómicos y metagenómicos que son indispensables para extender nuestro conocimiento sobre la ecología y evolución de los fagos en diferentes ecosistemas (Dion *et al*, 2020).

1.2 Manejo de datos producto de secuenciamiento de próxima generación (NGS)

Análisis y calidad de datos NGS

Un factor fundamental es la evaluación de la calidad de datos de secuenciamiento sin procesar. Para llevar a cabo esta evaluación se efectúa el control de calidad de las secuencias o Quality Check (QC), donde el objetivo principal es proporcionar un chequeo general sobre el estado de las lecturas generadas en la secuenciación. Estas lecturas son definidas como R1 o *forward* para las que codifican en dirección 5'-3' y R2 o *reverse* para las lecturas con direccionalidad 3´-5', para lecturas pareadas. Actualmente existen diversas herramientas que permiten realizar el control de calidad. Entre las de uso común se encuentra FastQC (Andrews, 2010), herramienta que proporciona una forma sencilla de evaluar las secuencias en bruto, y donde se presenta un amplio rango de características asociadas a los datos ingresados y a su calidad. El software muestra gráficamente métricas como la longitud de las lecturas, calidad de

secuencia por base, puntuaciones de calidad por secuencia, contenido de Guanina-Citosina (%GC), niveles de duplicación de secuencia, contenido de *k-mer*, entre otros parámetros. Una gran ventaja que presenta FastQC, además de sus componentes gráficos, es que permite la entrada de secuencias Fastq en formato SAM/BAM. Adicionalmente, FastQC permite el análisis de lecturas de diferentes plataformas de secuenciación como 454, PacBio e Illumina (Loganantharaj *et al.,* 2016).

Gracias al control de calidad o "QC" se puede tener una visión general de los datos obtenidos y evaluar si son adecuados para los pasos posteriores del procesamiento de secuencias. En caso de que la calidad de las lecturas no sea óptima, se deben filtrar aquellos elementos que puedan estar perjudicando la calidad de los datos, como por ejemplo, presencia de adaptadores, secuencias de baja calidad, extremos de baja calidad y lecturas contaminantes que pueden generar errores en posteriores análisis (Zhou *et al.,* 2014). El filtrado consiste inicialmente en recortar fragmentos de baja calidad de las secuencias o eliminar secuencias completas. Dentro de las herramientas disponibles para perfilar la calidad de las secuencias generadas mediante secuenciación de alto rendimiento, *filter by quality*, basado en el kit de herramientas FASTX (Gordon, 2010) filtra las lecturas según los puntajes de calidad basándose en el valor *phred*. En este punto del procesamiento es fundamental la elección de un límite de calidad balanceado, ya que un umbral demasiado alto puede descartar una gran cantidad de datos, mientras que un umbral demasiado bajo puede introducir errores en el ensamblaje (Liao *et al.,* 2017). La distribución de los puntajes de calidad para cada posición nucleotídica se calcula para cada lectura y si la calidad de una lectura en particular es menor que el valor de corte de calidad elegido, la lectura se descarta. Es importante aclarar que no existe un valor universalmente aceptado para realizar el recorte o el descarte de la secuencia, pues esto dependerá en gran parte de la naturaleza de los datos y de los objetivos de la investigación.

Posteriormente y con el fin de garantizar la buena calidad de las secuencias se utiliza Trimmomatic (Bolger *et al.,* 2014), que es una herramienta flexible y

eficiente de pre-procesamiento. Este software rápido y multiproceso, además de incluir gran variedad de pasos para el recorte y filtrado de lecturas como *Headcrop* (elimina el número deseado de bases al inicio de todas las lecturas), *Trailing* (corta los fragmentos finales de las lecturas que presentan un valor phred inferior al deseado) entre otras, también posee algoritmos que permiten localizar y eliminar secuencias de adaptadores utilizados en la secuenciación (Bolger *et al.,* 2014). Adicionalmente, y como complemento de Trimmomatic se utiliza un software especializado en la eliminación de adaptadores denominado Cutadapt (Martin, 2011), el cual busca y eliminar las secuencias de adaptadores, colas poli-A, cebadores y otro tipo de secuencias no deseadas.

Un último paso en la etapa del pre-procesamiento consiste en identificar y eliminar contaminantes genéticos que provienen de fuentes diferentes al organismo de interés. Estos pueden ser secuencias del huésped o secuencias que se introdujeron en el laboratorio en la etapa de extracción del ADN (Domínguez *et al.,* 2018). Este tipo de secuencias pueden ser detectadas generando alineamientos de las lecturas contra bases de datos de referencia. Bowtie2 (Langmead & Salzberg, 2012) es una herramienta de rápida ejecución para alinear las lecturas obtenidas luego de los pasos previos de procesamiento con secuencias de referencia, normalmente del hospedero o posibles contaminantes. Las secuencias que no se alinean con los datos de referencia corresponden a la secuencia de interés, las cuales son utilizadas para los procesos de ensamblaje y la anotación.

Ensamblaje

Una vez se depuran las lecturas, el siguiente paso es descifrar la secuencia genómica a partir de pequeños fragmentos de ADN, los cuales se fusionan en secuencias más largas o contiguas (contigs) para reconstruir el genoma de interés, proceso conocido como ensamblaje. Gracias al ensamblaje del genoma se puede contar con el catálogo de genes que un organismo puede expresar durante su ciclo de vida. Existen principalmente dos estrategias bajo las cuales se genera el ensamblaje de las lecturas. La primera estrategia es el ensamblaje basado en

referencia, donde las lecturas se ensamblan teniendo un genoma previamente ensamblado como referencia, los cuales se pueden obtener en bases de datos de ADN de libre acceso, como por ejemplo en el GenBank. Sin embargo, muchas veces no se cuenta con los genomas de referencia o estos se encuentran incompletos, lo cual presenta una limitante a la hora de utilizar este método (Lischer & Shimizu, 2017). El ensamblaje *de novo* es la segunda estrategia para reconstruir genomas, en este caso no se utiliza un genoma referencia, sino que el ensamblaje final se lleva a cabo únicamente con las secuencias obtenidas, sin tener un conocimiento previo de la organización del mismo (Aguilar & Falquet, 2015).

Para el ensamblaje de genomas a partir de datos de secuenciación de alto rendimiento se han desarrollado diversos programas. No obstante, la elección del programa a utilizar para ensamblar un genoma depende mucho del modelo con el que se está trabajando, la cantidad de datos e inclusive los recursos computacionales con que se cuenta (Domínguez *et al.,* 2018). Aunque el ensamblaje puede ser un proceso difícil para genomas muy grandes, en el caso de los fagos puede ser relativamente más sencillo ya que los genomas son más pequeños y menos complejos (Domínguez *et al.,* 2018). Una herramienta de código abierto ampliamente utilizada para el ensamblaje *de novo* a partir de lecturas generadas por tecnologías NGS es SPAdes (Bankevich *et al.,* 2012). Esta herramienta genera muy buenos resultados de ensamblaje trabajando con genomas de fagos, superando a otros algoritmos como Velvet y Ray que se utilizan para el mismo fin (Rihtman *et al.,* 2016), debido a que genera mayor integridad del genoma ensamblado (% del genoma representado en 1 solo contig). Inicialmente SPAdes fue diseñado para trabajar con lecturas de Illumina o IonTorrent (Bankevich *et al.,* 2012), sin embargo, actualmente es capaz de proporcionar ensamblajes híbridos utilizando lecturas generadas a partir de diferentes tecnologías de secuenciación como PacBio, Oxford Nanopore y Sanger. Es importante mencionar que SPAdes admite lecturas pareadas y no pareadas en diferentes formatos, por ejemplo, FastQ. Esta herramienta se desarrolló

principalmente para ser ejecutada desde líneas de comando, sin embargo, algunas plataformas web de libre acceso han incorporado versiones de SPAdes.

Una vez ensamblado el genoma el siguiente paso es verificar la calidad de este proceso, ya que el éxito de la posterior anotación depende en gran medida de la calidad del ensamblaje (Domínguez *et al.,* 2018). Diversas estadísticas son utilizadas para describir la integridad y la contigüidad de un ensamblaje. Estas permiten tener un criterio objetivo a la hora de elegir entre ensamblajes cuando se generan diferentes posibilidades a partir de un mismo conjunto de datos y pueden ayudar a identificar y eliminar potenciales problemas como gaps, fragmentación del genoma, contaminantes, sobrerrepresentación de regiones codificantes entre otras (Treangen & Salzberg, 2011; Yandell & Ence, 2012). Entre las diferentes métricas para evaluar la calidad del ensamblaje se puede encontrar el número de contigs, longitud total del ensamblaje, longitud de contig más grande, % GC, entre otras. Las regiones ricas en Guanina y Citosina GC son las que presentan mayor densidad de genes y dan estabilidad al genoma y por el contrario, regiones bajas en estas bases nitrogenadas pueden componer islas de patogenicidad (fracción del ADN genómico de un microorganismo patógeno que le faculta como virulento) (Piña-Iturbe *et al.,* 2020). Los porcentajes de GC también son importantes para soportar y establecer análisis comparativos entre bacteriófagos (Jung *et al.,* 2020). Si bien todas estas métricas son relevantes, el N50 es la medida que se utiliza a menudo para describir la “integridad” de un ensamblaje de genoma (Yandell & Ence, 2012). N50 es definido como la longitud de los contigs tal que usando contigs de igual o mayor tamaño produce la mitad de las bases del ensamblaje, y aunque este valor constituye un indicador acerca de la contigüidad del genoma, no es una señal de precisión y calidad del genoma ensamblado (Aguilar & Falquet, 2015). Se han desarrollado múltiples métodos para comparar la calidad de los ensamblajes, sin embargo, la herramienta QUAST es la más utilizada (Gurevich *et al.,* 2013). Esta herramienta de rápida ejecución presenta una interfaz de usuario amigable la cual facilita la interpretación de las métricas resultantes. Adicionalmente QUAST permite evaluar la calidad de los ensamblajes sin el uso de un genoma de referencia, permitiendo el análisis de ensamblajes *de novo*.

Para finalizar el proceso de ensamblaje es necesario detectar errores en el ensamblaje, como la presencia de regiones sin información (gaps), inserciones o deleción (indels), regiones repetidas, polimorfismos, etc. Para este fin, la herramienta automatizada Pilon (Walker *et al.,* 2014) permite realizar una mejora del borrador del genoma, debido a que integra algoritmos que permiten corregir paralelamente múltiples errores que reducen la calidad del ensamblaje. Este programa mejora significativamente los borradores, detectando y corrigiendo ensamblajes incorrectos, errores de bases, polimorfismos y llenando gaps (Walker *et al.,* 2014). La aplicación de esos cambios generados por Pilon sobre el borrador del ensamblaje produce genomas más contiguos y con un menor número de errores, factor clave para posteriormente extraer información biológicamente relevante de dichos genomas. Esta herramienta, que se encuentra disponible gratuitamente como software de libre acceso, funciona para datos generados a partir de diferentes tecnologías de secuenciación.

Anotación del genoma

Una vez que se ha obtenido el conjunto ordenado de contigs, el siguiente paso es anotar el borrador del genoma. La anotación es el proceso de búsqueda de "genes" y también puede incluir la identificación de ARN ribosómico y de transferencia codificados en el genoma. La anotación del genoma se puede lograr de manera relativamente sencilla cargando un ensamblaje del genoma en la herramienta web PATRIC (https://patricbrc.org/). Esta plataforma utiliza el algoritmo de RASTtk (Rapid Annotation using Subsystem Technology tool kit) como motor de anotación, ofreciendo varias ventajas que incluyen velocidad y facilidad al usuario (Brettin *et al.,* 2015). RASTtk permite a los usuarios optimizar y personalizar los pasos de anotación para un genoma determinado. La modularidad de RASTtk también hace que sea mucho más fácil desarrollar e incorporar software para mejorar las anotaciones del genoma (Brettin *et al.,* 2015). Esta herramienta ha integrado en su algoritmo un llamador de genes adicional denominado Prodigal (Hyatt *et al.,* 2010), el cual posee una alta precisión con

genes cortos y posiciones de inicio, y porque es más robusto a las diferencias en el contenido de Guanina y Citosina. Así mismo, incluye un algoritmo de anotación basado en *k –mers* y scripts que encuentran regiones de repetición, CRISPR y secuencias de inserción (Lagesen *et al.,* 2007).

Es importante mencionar que la anotación de un genoma tiene dos objetivos a realizar que son la anotación estructural (¿dónde están los genes y cómo se ordenan en el genoma?), y la anotación funcional (¿para qué es cada gen?). A continuación se define cada tipo de anotación:

Anotación estructural: El proceso de encontrar particularidades en el ADN como exones, intrones, promotores, transposones, etc., es conocido como anotación estructural. Un gen puede definirse como "una región de secuencia necesaria para generar productos funcionales". Los productos funcionales de los genes son ARN y proteínas y los genes que conducen a la producción de proteínas se denominan genes codificantes de proteínas. Otros genes que no codifican proteínas, sino moléculas de ARN funcionales, se denominan genes no codificantes (Spieth & Lawson, 2006). Las anotaciones estructurales también identifican pseudogenes, los cuales inicialmente se consideraron como carentes de función y que evolutivamente eran irrelevantes. Sin embargo, actualmente se conoce su participación en mecanismos de regulación genética, y su predicción puede mejorar el entendimiento de los genomas (Xiao *et al.,* 2016). La identificación de genes que codifican proteínas y otros elementos reguladores es un proceso complejo que ocupa un lugar central en la anotación de genes (Yandell & Ence, 2012).

Anotación funcional: La asociación de información biológica con secuencias de genes o proteínas identificadas por anotación estructural se denomina anotación funcional. Básicamente la anotación funcional implica la asociación de una descripción funcional con un gen, después de identificar una secuencia similar utilizando bases de datos, como BLAST (Ejigu & Jung, 2020).

Las anotaciones requieren datos de apoyo que se puedan utilizar o presentar como evidencia de las asignaciones previstas. Actualmente, los métodos basados en homología juegan un papel central en la anotación del genoma debido a la gran cantidad de secuencias de ADN disponibles (Mathé *et al.,* 2002; Domínguez *et al.,* 2018). La secuencia o estructura de nucleótidos y proteínas se puede encontrar fácilmente en bases de datos de dominio público, por ejemplo, GenBank (Sayers *et al.,* 2019), European Nucleotide Archive (ENA) (Brooksbank *et al.,* 2014), DNA Databank of Japan (DDBJ) (Kodama *et al.,* 2018), entre otras. En estas bases de datos también se pueden descargar los genomas necesarios para realizar análisis comparativos. Estos se realizan una vez se ensambla y se anota un genoma, con el fin de identificar las relaciones evolutivas del organismo secuenciado y ensamblado con otros organismos.

Análisis comparativo de sintenia

Una vez se ha ensamblado y anotado el genoma se pueden realizar alineamientos *in silico* contra genomas de otros organismos, con el fin de comparar y establecer relaciones filogenéticas. La esencia de la genómica comparativa radica en cómo comparamos los genomas para revelar las relaciones evolutivas de las especies. Desafortunadamente, en la mayoría de los casos, alinear correctamente incluso solo dos genomas a una resolución de unos pocos pares de bases puede ser un desafío. Un genoma generalmente contiene miles, millones o miles de millones de nucleótidos y es diferente del genoma de una especie estrechamente relacionada como resultado de diferentes procesos evolutivos como mutaciones en las secuencias, reordenamientos cromosómicos y expansión o pérdida de familias de genes (Alkan, Coe, & Eichler, 2011).

Los bloques de sintenia se definen formalmente como regiones de cromosomas entre genomas que comparten un orden común de genes homólogos derivados de un ancestro común (Vergara & Chen, 2010). Las comparaciones de la sintenia del genoma entre y dentro de las especies han brindado la oportunidad de estudiar los procesos evolutivos que conducen a la diversidad del número y la estructura de

los cromosomas en muchos linajes en el árbol de la vida (Howe *et al.,* 2016). El análisis de sintenia en especies estrechamente relacionadas es actualmente la norma para cada nuevo genoma publicado (Trotereau *et al.,* 2021). En general, la identificación de sintenia es un proceso de filtrado y organización de todas las similitudes locales entre las secuencias del genoma en una imagen global relativamente fácil de interpretar (Batzoglous, 2005). El análisis de sintenia depende en gran medida de la calidad del ensamblaje. Por ejemplo, las secuencias que faltan en un ensamblaje pueden llevar a que falten anotaciones de genes y, posteriormente, a que falten relaciones ortólogas (Bhutkar *et al*., 2006). En este trabajo se evaluaron las diferentes opciones de análisis de datos, ensamblaje, anotación y de genómica comparativa para el análisis de genomas de bacteriófagos mediante el uso de herramientas disponibles en la web.

2. Objetivos

2.1 Objetivo general

Diseñar un flujo de trabajo para el ensamblaje, anotación y análisis de sintenia de genomas bacteriófagos, mediante el uso de herramientas web gratuitas.

2.2 Objetivos específicos

2.2.1. Efectuar el análisis de calidad de secuencias obtenidas mediante NGS.

2.2.2. Realizar el ensamblaje *de novo* de las secuencias previamente depuradas.

2.2.3. Identificar los marcos abiertos de lectura (ORF), las regiones de codificación (CDs), y los ARN de transferencia (tARN) en el genoma ensamblado.

2.2.4. Realizar análisis de sintenia del bacteriófago NJ-P3 contra fagos que presentaron similitud de secuencias.

3. Métodos

3.1 Datos y Pre-procesamiento

Con el fin de probar las diferentes herramientas y elegir las más adecuadas para construir el flujo de trabajo se utilizaron lecturas pareadas (paired-end) provenientes de la plataforma Illumina HiSeq del bacteriófago NJ-P3 de *Ralstonia solanacearum* (código de acceso ENA SRR8402465). El preprocesamiento se realizó por medio de la herramienta web de libre acceso Galaxy Versión 0.8.1 (https://usegalaxy.org/). Inicialmente se utilizó FastQC (Andrews, 2010) para examinar la calidad de los datos provenientes de secuenciación de alto rendimiento. Una vez se determinó la calidad de los datos se procedió a ejecutar la herramienta Cutadapt Version 1.16.6 (Martin, 2011) en Galaxy, para eliminar diferentes contaminantes provenientes de la secuenciación. Se realizó un filtrado o *trimming* de los datos con baja calidad o que presentaban valores *phred* inferiores a 30. Para este filtrado se ejecutó Filter by quality Version 1.0.2 (Gordon, 2010) en Galaxy, con los parámetros *Quality cut-off value* de 30 (valor phred) en el 90% de las bases. Se utilizó Trimmomatic Version 0.38.0 (Bolger *et al.,* 2014) en Galaxy para eliminar fragmentos de lecturas con valores phred inferiores a 30, mediante el uso de las siguientes herramientas: *Slidingwindow* y un valor medio de calidad de 30; *Headcrop* con un valor de 10 para eliminar los primeros 10 nucleótidos de todas las lecturas; y *Trailing* con una puntuación *phred* mínima de 30 para eliminar los segmentos finales cuyo valor *phred* es inferior a 30. Por último, se determinó la presencia de contaminantes y se efectuó la posterior eliminación de secuencias contaminantes procedentes de otros organismos diferentes al de interés. Esto se llevó a cabo con el software Bowtie2 (Langmead & Salzberg, 2012) Galaxy Version 2.4.2, donde los datos fueron alineados contra los genomas del aislado bacteriano *Ralstonia solanacearum* UY031 (código de acceso en GenBank CP012687.1), y contra el genoma preestablecido en Galaxy para el *Homo sapiens*.

3.2 Ensamblaje

Para ensamblar las lecturas resultantes del pre-procesamiento se usó la herramienta web de libre acceso PATRIC (patricbrc.org). El ensamblaje se llevó a

cabo mediante una estrategia *de novo* utilizando SPAdes (Bankevich *et al.,* 2012). Esta plataforma realiza un trimaje previo por defecto. Para evaluar cuantitativamente la calidad del ensamblaje se utilizó QUAST Version 5.0.2 (Gurevich *et al.,* 2013) en Galaxy, y para evaluarla cualitativamente se utilizó Bandage Version 0.8.1 (Wick *et al.,* 2015) en Galaxy. Ambas herramientas se ejecutaron con los parámetros preestablecidos. Posteriormente se utilizó Pilon Version 1.20.1 (Walker *et al.,* 2014) en Galaxy para refinar y corregir errores sobre el borrador del genoma, buscando generar el mejor ensamblaje posible.

3.3 Anotación

La anotación del genoma se llevó a cabo mediante el uso de la plataforma PATRIC (Pathosystems Resource Integration Center), dentro de la cual se utilizó el motor de anotación de genomas RASTtk (Brettin *et al.,* 2015) para realizar la auto anotación del ensamblaje. Con el fin de asignar posibles funciones a los CDS (coding sequences) se evaluó la similitud de cada una de las secuencias resultantes de la auto anotación, con secuencias ya conocidas que se encuentran depositadas en la base de datos del NCBI (National Center for Biotechnology Information). Se realizaron alineamientos a través del servidor BLAST (Basic Local Alignment Search Tool) en búsqueda de secuencias homólogas, usando principalmente las secuencias de aminoácidos (aa), seleccionando el algoritmo de BLASTp (protein-protein), y utilizando la base de datos de proteínas no redundantes (nr). Posteriormente y con el mismo fin de asignar posibles funciones, se utilizó el servidor interactivo HHpred (Zimmermann *et al.,* 2018), para la detección de homología de proteínas y predicciones estructurales. Allí se proporcionó la secuencia de aminoácidos de cada CDS generado en la auto-anotación por PATRIC y se realizó una búsqueda en múltiples bases de datos tales como Protein Data Bank (PDB), Conserved Domains Database (CDD) y Pfam. Los genes relacionados a tRNA se identificaron usando tRNAscan-SE 2.0 (opción: -B para tRNA bacterianos) (Chan & Lowe, 2019). Las secuencias relacionadas con proteínas de membrana fueron predichas utilizando la herramienta TMHMM v 2.0 (Krogh *et al.,* 2001) y SOSUI (Hirokawa *et al.,* 1998).

3.4 Comparación genómica

Con el fin de realizar una breve aproximación a los análisis de comparación genómica, se optó por realizar un acercamiento a los análisis de sintenia mediante el uso del programa de libre accesos Artemis y su herramienta para la comparación genómica: Artemis Comparison Tool o ACT (Carver *et al.,* 2005). Para tal fin se descargaron los genomas anotados de los bacteriófagos de *R. solanacearum* DU_RP_II (MF150911) y el RpY1 (MN996301). Posteriormente se generaron los archivos de comparación necesarios mediante la herramienta Blast (Ver el archivo complementario del tutorial, apartado 7. **Comparación genómica ACT**), y se cargaron a la plataforma ACT. El genoma del bacteriófago NJ-P3 se tomó como referente y fue ubicado en el medio por ser el genoma de interés. Se establecieron los porcentajes de identidad (similitud entre genomas) de corte en valores mínimos y máximos entre 0 y 100 respectivamente, y se asignó el valor de corte como predeterminado entre mínimos 291 y máximo 20,000. Se activó la opción grafica de picos y valles para los tres genomas lo cual permite visualizar los porcentajes GC a lo largo del genoma. Se realizó una comparación similar mediante el programa Easyfig (Sullivan *et al.,* 2011) con un mínimo valor de identidad de 85%.

4. Resultados

Se descargó las lecturas en crudo desde el Archivo Europeo de Nucleótidos (ENA por sus siglas en inglés) del Bacteriófago NJ-P3 de *R. solanacearum* QL-RS1115, clasificado como *Podoviridae* (Fagos de cola corta), vinculado al Bioproyecto "*Podoviridae* strain: NJ-P3" (código de acceso NCBI SAMN10698423, ID: 10698423). Este bacteriófago fue aislado en China en 2015, en la provincia de Nanjing, Jiangsu, y secuenciado por la Universidad Agrícola de Nanjing (Fecha de subida al NCBI 2019-01-06; subido al ENA en 2019-01-08). La plataforma utilizada para el secuenciamiento fue Illumina HiSeq 4000, con un total de lecturas de 6.002.639, una cobertura aproximada de 150X y un recuento de bases de 1.800.791.700. A la fecha no se conocen los genomas completos, ni la publicación sobre este Bioproyecto.

4.1 Evaluación de calidad en las secuencias crudas.

Una vez descargado el genoma, se cargó a la plataforma Galaxy para realizar el pre-procesamiento y análisis de calidad de las lecturas (Ver el archivo complementario del tutorial, apartados **1, 2 y 3**). Para comenzar se realizó un análisis de calidad de los datos mediante la herramienta FastQC (Fig. 1).

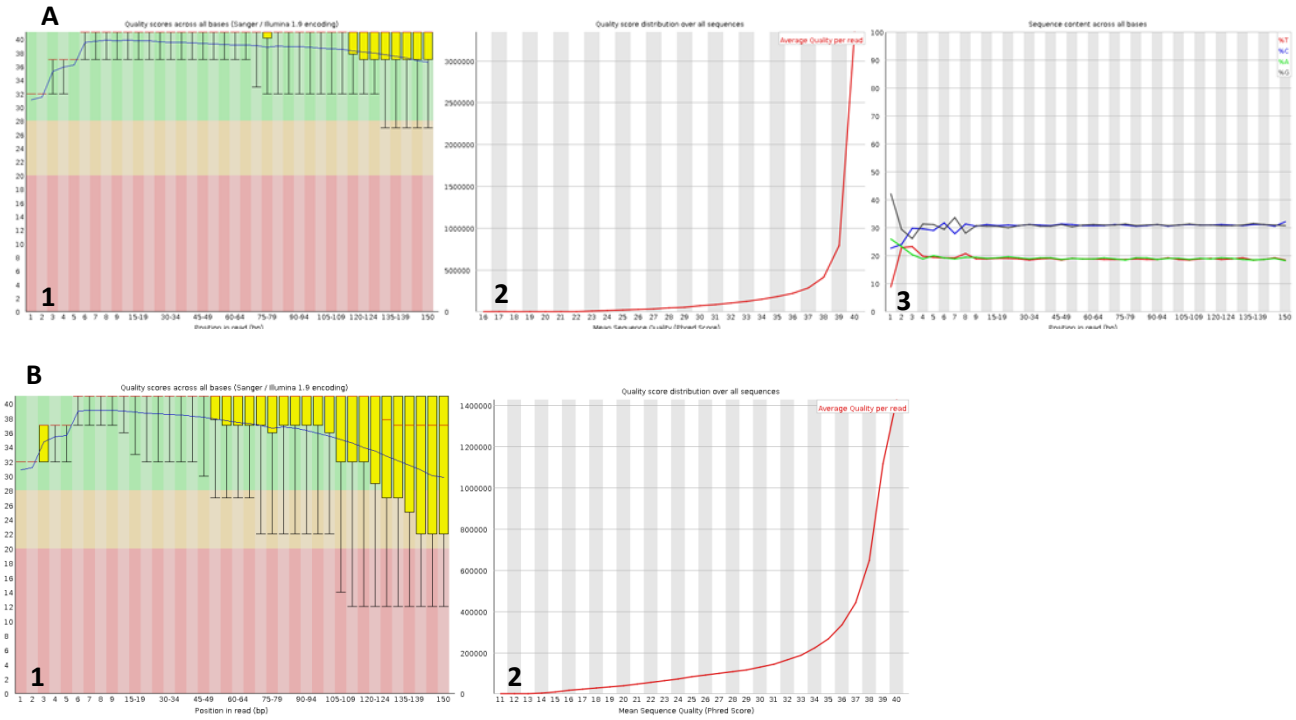


Fig. 1. Análisis de calidad de lecturas mediante FastQC. (A) Representación gráfica de los diferentes parámetros para la lectura correspondiente al R1; **1**) Calidad por secuencias de base; **2**) Puntuaciones de calidad por secuencia; **3**) Contenido por secuencia base. **(B)** Representación gráfica de los datos R2; **1**) Calidad por secuencias de base; **2**) Puntuaciones de calidad por secuencia. Las demás graficas pertenecientes a los datos R2 son significativamente semejantes a las gráficas de R1, por lo cual se omitieron.

Los datos R1 presentaron un porcentaje de GC correspondiente a 61% y los R2 a 62%, con fluctuaciones entre 40% y 87% de GC en las lecturas. La calidad por secuencia de base en R1 inicia con un valor *phred* de 31 que aumenta al inicio y disminuye hacia el final de las lecturas hasta un valor *phred* de 30. Para R2 se evidencia un comportamiento similar pero con una caída más abrupta en los valores *phred* hacia el final de las lecturas. Se encontró que en las puntuaciones de calidad por secuencia los datos del R2 presentan un mayor número de

secuencias con calidades bajas con respecto a los datos de R1. Según el análisis de contenido por secuencia base, la calidad de las primeras 10 pares de bases no es buena, por lo que se realizó un recorte en esta región de la secuencia. La distribución de la longitud de la secuencia está normalizada a 150pb en todas las lecturas (longitud por defecto en la secuenciación para el genoma del fago NJ-P3). Con el fin de eliminar cualquier adaptador utilizado durante el proceso de secuenciación, se recurrió a la herramienta Cutadapt, ejecutada con valores predeterminados para secuenciación con tecnología Illumina. Se procesaron un total de 6.002.639 lecturas, con un recuento de 900.395.850 pb, de manera independiente para R1 y R2. Así mismo se ejecutó de forma pareada examinando un total de 1.800.791.700 pb, hallando cero adaptadores en ambos casos.

Los datos que presentaron una calidad en valor phred inferior a 30 en el 90% de la lectura se eliminaron con ayuda de la herramienta Filter by quality. Una vez procesados los datos se obtuvieron los archivos de salida con 4.801.611 y 3.434.513 lecturas para R1 y R2 respectivamente. Se eliminaron en total 1.201.028 y 2.568.126 de lecturas, equivalentes al 20% y 42% de las lecturas en R1 y R2, respectivamente.

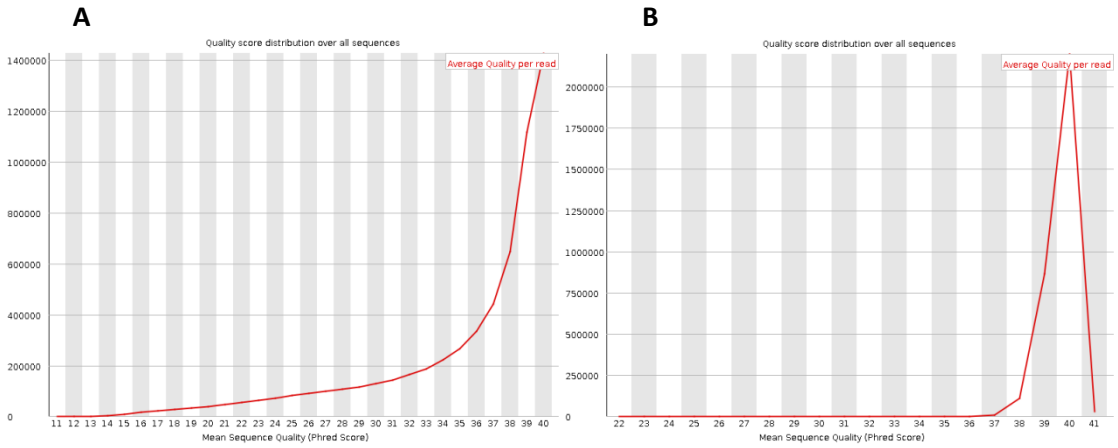


Fig. 2. Comparación en la Puntuación de calidad por secuencia, datos crudos vs filtrados por calidad. Se presentan en el eje X el valor phred de los datos, en el eje Y el número de lecturas que representan dicho valor. **(A)** FastQC datos en crudo. **(B)** FastQC datos procesados por Filter by quality

Una vez realizado el trimado de los datos, se emparejaron todas las secuencias, y fueron eliminadas aquellas lecturas únicas sin emparejar. El archivo de salida presentó un total de 3.230.006 lecturas, de las cuales el 100% son datos emparejados, presentando una longitud promedio de 138 pb. Se suprimieron los primeros diez pares de bases en todas las lecturas, debido a que esta región presentaba irregularidad en los porcentajes de GC mediante *Headcrop.* Igualmente se eliminaron aquellos fragmentos dentro de las lecturas cuyos últimos pares de bases presentaban valores phred inferiores a 30 con la herramienta *Trailing*.

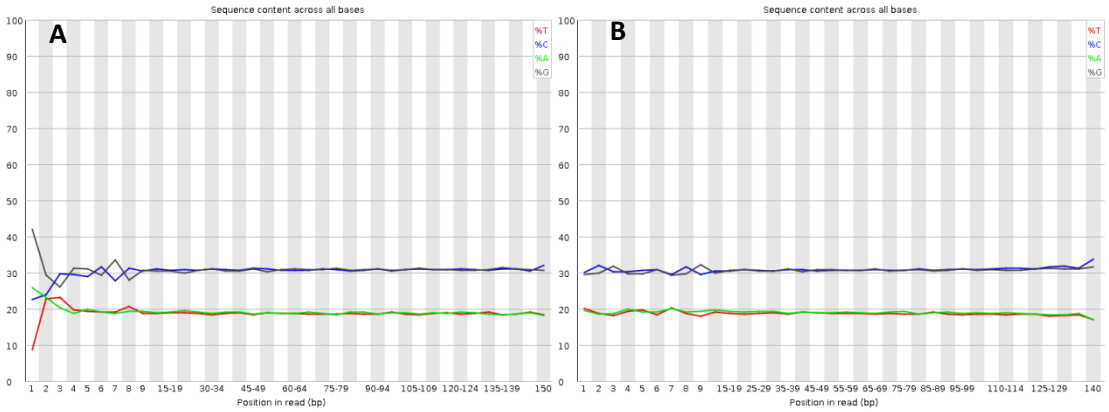


Fig. 3. Comparación en el contenido por secuencia base, datos crudos vs datos trimados. En el eje X se representan los pb en las lecturas, en el eje Y el procentaje. La linea roja representa el porcentaje de Timina a lo largo de la lectura, en verde la Adenina, en azul la Citocina y en gris la Guanina. **(A)** FastQC datos crudos. **(B)** FastQC datos Trimmados.

Como último paso del preprocesamiento se utilizó Bowtie2 para mapear el genoma del bacteriófago NJ-P3 al genoma de referencia del aislado bacteriano *R. solanacearum* UY031. Se alinearon concordantemente 101 lecturas al menos una vez, 34 lecturas se alinearon de manera discordante, y un total de 48.108 *mates* (lecturas con insertos largos) fueron mapeados al genoma de referencia una o más veces. Un total de 0.75% de las lecturas se mapearon y alinearon con el genoma de referencia, las cuales fueron eliminadas mediante el programa Bowtie2. En este paso se obtuvieron los archivos de salida con un total de 3.229.905 lecturas pareadas procesadas, con contenidos de GC de 61% y 62% para R1 y R2 respectivamente. Finalmente, tras mapear el genoma del

bacteriófago NJ-P3 contra el genoma *Homo sapiens*, se encontraron 0 concordancias.

4.2 Ensamblaje *de novo* y análisis de calidad del ensamblaje.

Se ensambló el genoma mediante el software SPAdes versión 3.12.0 disponible en PATRIC. Se obtuvieron 22 contigs con una longitud total de 65.922 pb. De éstos, 9 presentaron una longitud igual o superior a 1.000pb; dos mostraron una longitud mayor o igual a 5.000pb; y un único contig resultó poseer más de 10.000pb. Este último, de 38.385 pares de bases según el análisis de Bandage exhibió una topología circular (Fig. 4). Los resultados del análisis Quast muestran un porcentaje de GC de 61,78%, un valor N50 de 38.385, valor equivalente a la longitud del ensamblaje del fago. Tras realizar el refinamiento con el software Pilon desde la plataforma Galaxy, la herramienta halló una única región problemática entre las bases 651-674 con secuencia nucleotídica "CGACGATTTTCCCGACGATTTTCC" y fue eliminada. Cabe aclarar que esta eliminación se realizó en el décimo (10) contig, equivalente al contig_433 del archivo contig.fasta. Este contig que puede ser omitido dadas sus características de longitud y probablemente es un contaminante, por esta razón no se generó cambios en el contig de interés a ser anotado.

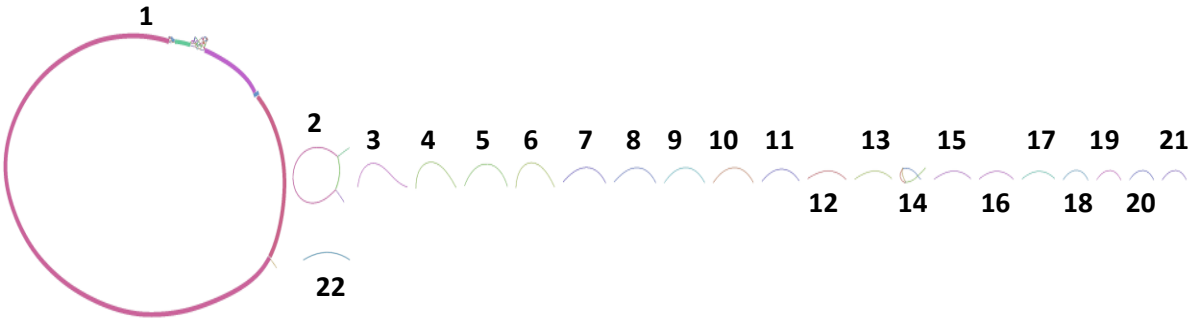


Fig. 4. Grafico del Bandage, ensamblaje. (A). Topología de los 22 contigs ensamblados mediante SPAdes v3.12.0 (PATRIC). La topología circular de mayor tamaño (contig 1) es el contig de interés a ser anotado (por sus características y parentesco genómico con los bacteriófagos DU_RP_II y RpY1). Los contigs de menor tamaño, circulares y no circulares, representan fragmentos contaminantes provenientes del hospedero *Ralstonia solanacearum,* según análisis en BLAST.

4.3 Anotación y verificación de tARNs.

En total la auto-anotación en PATRIC identificó 49 CDS con longitudes que iban desde 102 a 2639 pares de bases, las cuales codificaban polipéptidos de 34 hasta 879 aminoácidos. Entre los 49 CDS predichos más del 80% de las secuencias representaban proteínas hipotéticas y solo a una pequeña porción de éstas (4 de 49), representando aproximadamente el 8%, se les asignaron posibles funciones relacionadas a fagos. Gracias a los alineamientos realizados en BLASTp y en HHpred en las diferentes bases datos, se pudo aumentar el número de las posibles funciones para los CDS predichos, pasando de 4 a 25, equivalente al 51% de secuencias codificantes con posibles funciones putativas. Dentro de éstas, se encontraron genes con funciones importantes para los fagos con ciclo lítico como lo son la holina y las lisozimas. También, se encontraron genes relacionados con proteínas estructurales de cápside y cola. Además, se identificaron genes implicados en la replicación del fago como por ejemplo genes que codifican para integrasas (Tabla 1). La mayoría de las secuencias encontradas en BLASTp presentaban alta similitud con los fagos de *R. solanacearum* DU_RP_II (MF150911) y el fago RpY1 (MN996301).

tRNAscan-SE

La auto-anotación realizada con PATRIC logró detectar una única secuencia correspondiente a un tRNA en la posición 2870-2945 en pares de bases, dentro del genoma. Una vez realizada la búsqueda de este tipo de secuencias en todo el genoma utilizando tRNAscan-SE se corroboró la presencia de esta secuencia en la misma posición mencionada anteriormente, y se determinó como tRNA (tRNA - Asp, GTC). Sin embargo, nótese que el isotipo inferido por el anticodón (Fig. 3A) difiere del modelo específico de isotipo que arrojó la puntuación más alta (Fig. 3B), por tal razón se debe ratificar que este no corresponda a un pseudogen u otra estructura genómica.

A

Sequence Name	tRNA #	tRNA Begin	tRNA End	tRNA Type	Anticodon	Infernal Score	Isotype Model	Isotype Score
Ralstonia_phagetutorial_contig_1	1	2870	2945	Asp	GTC	70.2	Thr	79.4

Isotype-Specific Model Scores:

B

tRNAscan ID	Anticodon predicted isotype	Met	Phe	Pro	SeC	Ser	Thr	Trp	Tyr	Val	fMet
Ralstonia_phagetutorial_contig_1.trna1	Asp	76.4	71.3	59.1	No Hit	27.5	79.4	53.7	25.4	64.7	20.5

Top score | 2nd highest score | 3rd highest score

C

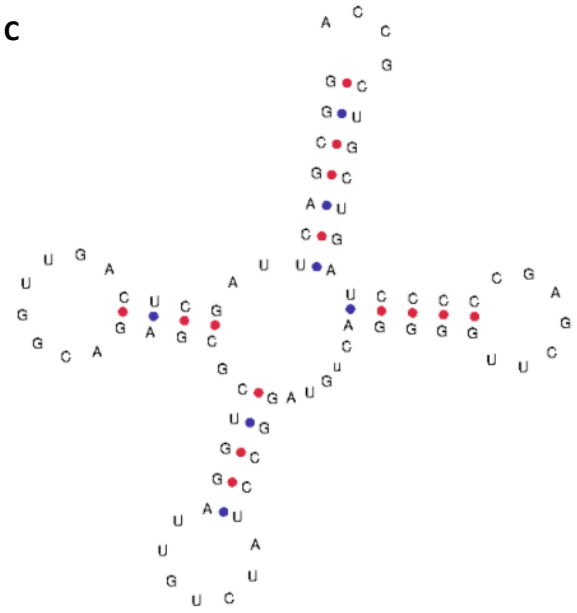


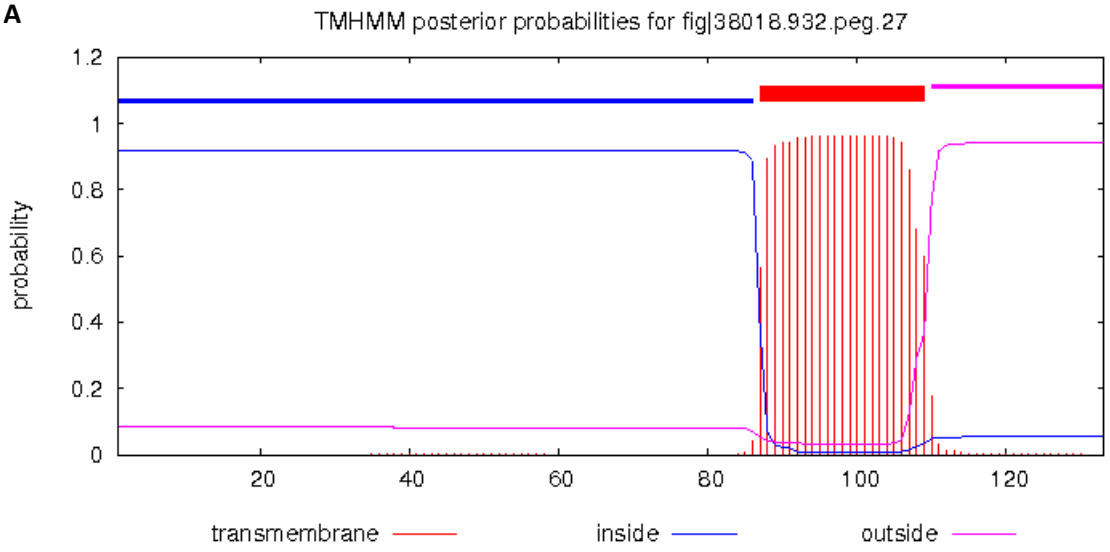
Fig. 5. Resultados tRNAscan-SE. **A)** Tabla con información relacionada del tRNA predicho en el genoma del bacteriófago NJ-P3. **B)** Tabla con puntuaciones de isotipos más probables. **C)** Estructura secundarias del tRNA predicho.

Identificación de proteínas de membrana con TMHMM y SOSUI

TMHMM logró identificar 6 secuencias las cuales podrían estar relacionadas con proteínas de membrana, de longitudes desde 47 hasta 174 aminoácidos y con 1 a 3 segmentos trans-membranales **(Fig. 6A)**. Todas las secuencias identificadas por TMHMM excepto 1 (CDS 28 - Holina) se reportaban como proteínas hipotéticas después de los alineamientos realizados en BLASTp. Cada una de estas secuencias se analizó mediante el programa web SOSUI para corroborar la

presencia de fragmentos trans-membranales dentro de la secuencia de consulta. Se encontró que al interior del CDS 28, una secuencia continua con 23 amino ácidos correspondían a una hélice trans-membranal **(Fig. 6B)**.

A



B

This amino acid sequence is of a MEMBRANE PROTEIN which have 1 transmembrane helix.

No.	N terminal	transmembrane region	C terminal	type	length
1	88	WLWKLVGTGLLACLGVIGWASTQ	110	PRIMARY	23

Fig. 6. Identificación de proteínas de membrana. **A)** Segmentos transmembranales predichos por la herramienta TMHMM. Las barras de color rojo representan las regiones transmembranales dentro de la secuencia consulta. **B)** Breve descripción de las propiedades de hélice transmembranal encontrada, información proporcionadas por el programa SOSUI.

Se predijo ocho CDS implicados en la morfogénesis del fago NJ-P3, tres de ellos relacionados con proteínas de la cápside y 5 relacionados con proteínas de cola **(Tabla 1).** Los CDS 1, 5 y 15 se incluyeron dentro de los CDS que codifican proteínas relacionadas a la cápside. Los CDS que se predijeron como codificantes de proteínas de cola: 8, 19, 23, 25 y 31. De igual manera se anotaron 4 CDS implicados en la replicación del fago NJ-P3, comprendidos entre los CDS 34 y 37 correspondientes a integrasas, y CDS 35 y 45 correspondientes a polimerasa y

helicasa de ADN, respectivamente. Se presentan 3 CDS consecutivos (CDS 28 a 30) que hacen parte de proteínas implicadas en el ciclo lítico de bacteriófagos, representados como holinas, lisozimas y proteínas de lisis. Además, se predijeron 4 CDS (9, 16, 27 y 40) que corresponden a proteínas transmembrana. Otras posibles funciones putativas para CDS anotados son endonucleasa (CDS 3 y 11), proteasa (CDS12), deacetilasa (CDS 32), nucleasa (CDS 42) y proteína RNAseH-like (CDS 48), las cuales están implicadas en el metabolismo (transcripción y regulación) del fago. Los 24 CDS restantes del fago NJ-P3 se anotaron como proteínas hipotéticas, puesto que con los programas utilizados no fue posible definir su función putativa.

4.4 Análisis comparativo entre genomas (sintenia).

Se encontró una alta similitud en la identidad genética entre los tres bacteriófagos comparados, donde se presentan regiones de sintenia continua entre los fagos DU_RP_II y NJ-P3 compartiendo la conformación proteica y distribución a lo largo del genoma (**Fig. 8**). Gran parte de las proteínas involucradas en procesos relacionados con la generación de la cápside (CDS 1, 5 y 15 proteínas putativas codificantes para cápside y terminasas estructurales), y proteínas de cola (8, 19, 23, 25 y 31), se ven representadas y compartidas, con porcentajes de identidad superiores al 90% y sin presentar grandes variaciones en la estructura general del genoma. Se presenta una región ampliamente compartida, la cual presenta una identidad superior al 95%, comprendida entre las bases 2.974-19.340 para NJ-P3 y 2.557-17.979 para DU_RP_II. Esta región se interrumpe por un gap de 750 pb en la proteína R2B_p015 del fago DU_RP_II y dos inserciones de proteínas en el fago NJ-P3 correspondientes a los CDS 7 y 11. Las proteínas que codifican en dirección 3’-5’ (reverse), se encuentran hacia la parte final de ambos genomas y codifican para el metabolismo de los fagos (transcripción y regulación). Esta región presenta continuos de colinealidad interrumpidos por aparentes inserciones de proteínas en el fago DU_RP_II entre las bases 32.464-32.904 correspondiente a la proteína de locus tag R2B_p028, y la inserción en las bases 36.370-37.830 equivalente a la proteína de locus tag R2B_p033. En comparación a los fagos DU_RP_II y NJ-P3, el fago RpY1 presenta la región codificante 3’-5’ al inicio del genoma. Esta región muestra bloques de colineali-

Tabla 1. Posibles funciones putativas encontradas para algunos de los cds predichos en la anotación. El % de identidad y E-value corresponden a aquellas secuencias que presentaron una mayor homología con las secuencias de consulta (query cover), estas búsquedas se realizaron con BLASTn desde el servidor del NCBI. Aquellos CDS que muestran valores de probabilidad y E-value pertenecen a posibles funciones encontradas mediante la herramienta web HHPRED. Los CDS que aparecen en color rojo indican las secuencias en las cuales TMHMM detectó secuencias de hélices transmembrana. La tabla completa se encuentra en el material complementario.

CDS	POSIBLE FUNCIÓN	% ID	PROBABILIDAD	E-VALUE
CDS 1	Putative bacteriophage terminase small subunit	-	95.35%	0.13
CDS 3	Putative HNH homing endonuclease	-	100%	1.6e-31
CDS 5	Putative terminase large subunit	99.60%	-	0.0
CDS 8	Head-tail connector protein	98.76%	-	0.0
CDS 9	Putative transmembrane protein	-	-	-
CDS 11	Putative HNH endonuclease	44.44%	-	1,00E-27
CDS 12	Putative protease	97.63%	-	0.0
CDS 15	Putative major capsid protein	98.67%	-	0.0
CDS 16	Putative transmembrane protein	-	-	-
CDS 19	Tail tubular protein A	-	100%	1.5e-36
CDS 23	Minor structural protein	99.32%	-	0.0
CDS 25	Structural protein	96.07%	-	0.0
CDS 27	Putative transmembrane protein	-	-	-
CDS 28	Holin	91.67%	-	1,00E-43
CDS 29	Lysozyme	98.82%	-	1,00E-119
CDS 30	Bacteriophage lysis protein	-	98.85	1.7e-7
CDS 31	Minor tail protein	97.95%	-	0.0
CDS 32	Putative polysaccharide deacetylase	98.48%	-	1,00E-84
CDS 34	Integrase	97.00%	-	0.0
CDS 37	Integrase	98.60%	-	0.0
CDS 39	Putative DNA polymerase	83.38%	-	0.0
CDS 40	Putative transmembrane protein	-	-	-
CDS 42	PD-(D/E)XK nuclease superfamily protein	88.70%	-	0.0
CDS 45	Putative DNA helicase	97.83%	-	0.0
CDS 48	Putative RNaseH-like domain protein	81.76%	-	5,00E-80

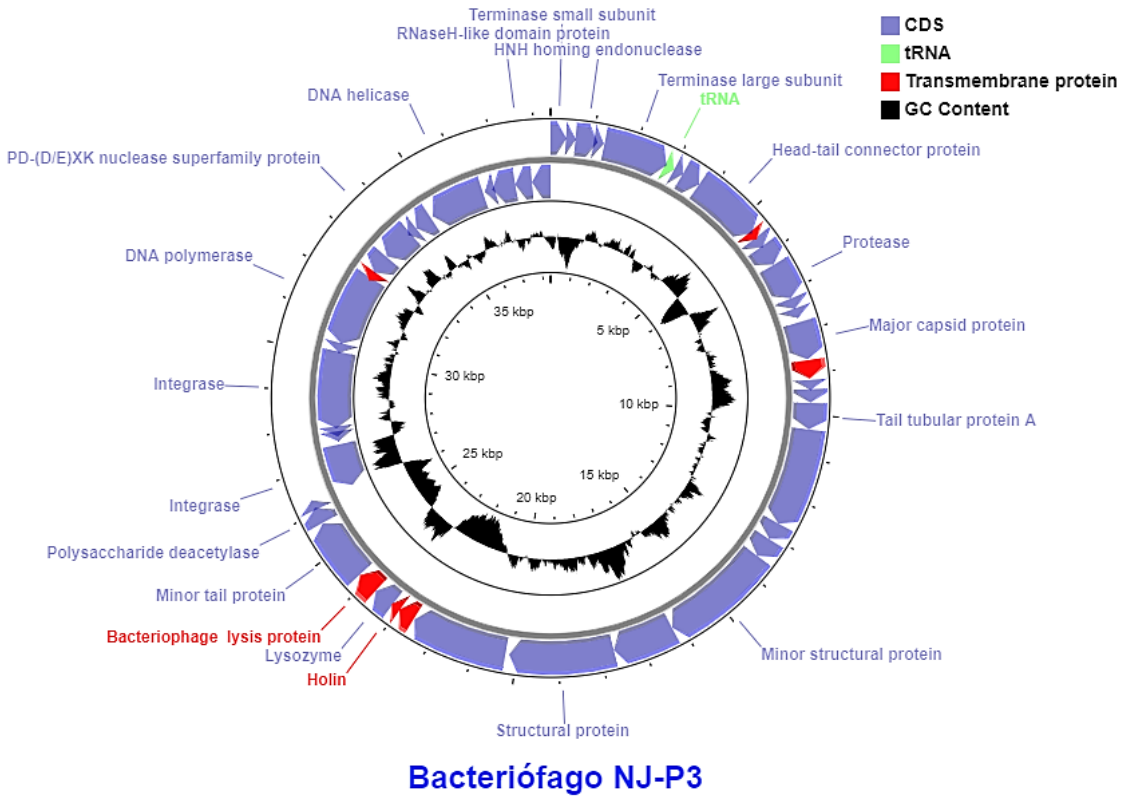


Fig. 7. Representación gráfica del bacteriófago NJ-P3 con CGView Server. Esta gráfica se encuentra divida en 4 círculos: el círculo más externo muestra el marco de lectura que se lee en dirección 5´-3´ (forward); en el segundo círculo la hebra que codifica en sentido 3´-5´ (reverse); el tercer círculo representa mediante un gráfico de picos y valles el %GC (62.2%); el círculo interno refleja la longitud total del genoma. Las flechas de color violeta corresponden a los CDS predichos para el genoma, algunos de estos con sus funciones putativas correspondientes (etiquetas fuera del círculo). Aquellos CDS que no poseen etiquetas se anotaron como proteínas hipotéticas. Las barras de color rojo indican las secuencias correspondientes a proteínas de transmembrana predichas por TMHMM. La región de color verde que aparece en la gráfica representa el único tRNA (2870pb-2945pb) encontrado para el fago NJ-P3.

4.4 Análisis comparativo entre genomas (sintenia).

Se encontró una alta similitud en la identidad genética entre los tres bacteriófagos comparados, donde se presentan regiones de sintenia continua entre los fagos DU_RP_II y NJ-P3 compartiendo la conformación proteica y distribución a lo largo del genoma (**Fig. 8**). Gran parte de las proteínas involucradas en procesos relacionados con la generación de la cápside (CDS 1, 5 y 15 proteínas putativas

codificantes para cápside y terminasas estructurales), y proteínas de cola (8, 19, 23, 25 y 31), se ven representadas y compartidas, con porcentajes de identidad superiores al 90% y sin presentar grandes variaciones en la estructura general del genoma. Se presenta una región ampliamente compartida, la cual presenta una identidad dad con identidad superior al 95% y están involucradas proteínas de transcripción y regulación, que comprende los CDS 42, 45 y 48 y algunas proteínas hipotéticas. El fago RpY1 presenta un gap entre las bases 11.700-16.200, el cual no exhibe homología alguna con el fago NJ-P3, curiosamente este vacío de colinealidad se inserta en la región contigua a la codificante 3'-5'. A partir del CDS 27 del fago RpY1, se retoma una sintenia continua con respecto al fago NJ-P3, donde se presentan grandes bloques de colinealidad. Esta región muestra dos aparentes proteínas insertadas en el fago NJ-P3 entre las bases 3.254-3.709 codificante para una proteína hipotética y en las bases 5.790-6.302 codificante para una endonucleasa, interrumpiendo así el continuo de colinealidad entre los CDS 8 y 9 del fago NJ-p3. Al final de este continuo, entre las bases 32.803-36.012, se presenta una alteración en los aminoácidos del ORF 46 del fago RpY1, el cual guarda homología con el ORF 26 del fago NJ-P3, esta alteración es una inserción que se da entre las bases 34.450-35.220 e interrumpe la homología entre los organismos.

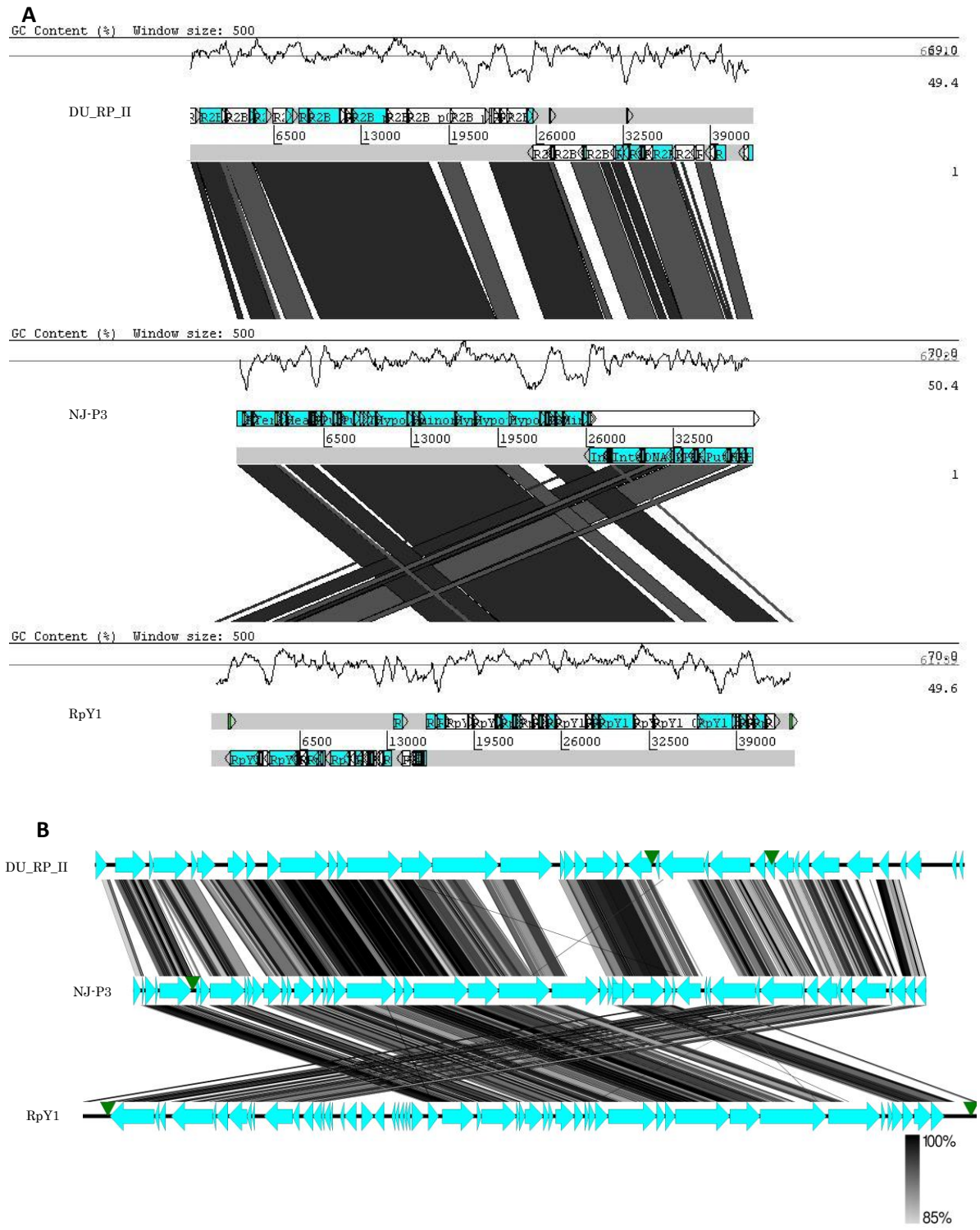


Fig. 8. Comparación genómica entre los bacteriófagos de *R. solanacearum* DU_RP_II, NJ-P3 y RpY1. (**A**) Gráfico generado con ACT donde se muestra la distribución de GC a lo largo del genoma mediante un diagrama de picos y valles para cada uno de los tres genomas. **(B)** Gráfico generado con EasyGraph donde se visualiza la comparación de sintenia entre los tres genomas. Para (**A** y **B**) se muestran las proteínas anotadas en color azul claro, la flecha indica la direccionalidad en la trascripción de los genes. Los ARNt se marcan de color verde. Se establece

una relación por homología en las regiones conservadas mediante bloques colineales entre los genomas, con un valor mínimo de identidad del 85%, donde la intensidad del color en escala de grises denota mayor similitud entre las secuencias.

5. Discusión

En esta guía se busca proporcionar una alternativa simple y basada en herramientas en línea para el procesamiento, ensamblaje, anotación y análisis de genómica comparativa de genomas de bacteriófagos. La guía no está dirigida a quienes deseen realizar un procesamiento automatizado de cientos de genomas a la vez. Si bien para el uso de la guía, los lectores deben estar familiarizados con conceptos de genética y con la naturaleza básica de los datos de secuencia, no se requieren habilidades de programación y todos los ejemplos que usamos se pueden realizar en una computadora de escritorio (Mac, Windows o Linux). La guía no pretende ser exhaustiva, sino presentar un conjunto de herramientas simples, flexibles y gratuitas para el análisis de genomas. Gran parte de este trabajo está basado en las pautas brindadas por Jung y colaboradores (2020), donde se enmarca todo el proceso de análisis de datos, ensamblaje y anotación en 12 sencillos pasos. No hacemos uso de las herramientas allí propuestas, pues nos enfocamos en aplicaciones gratuitas y alojadas en la web.

Se desarrolló un flujo de trabajo con diversas herramientas, con el fin de preparar las lecturas para un óptimo ensamblaje del genoma. Se utilizaron un conjunto de métricas basadas en la evaluación cualitativa de los datos producidos por fastQC, que incluye puntuaciones de calidad *Phred*, contenido de GC, distribuciones de *K-mer* e información de sobrerrepresentación de secuencias. Con el fin de eliminar todas las lecturas que generan ruido para optimizar el ensamblaje, se eliminaron los datos con valores phred inferiores a 30 en el 90% de la secuencia, descartando así 20% y 42% de las lecturas en R1 y R2 respectivamente. Esto promovió un reajuste en la cobertura total de nuestro genoma, pasando del 150X a 100X aproximadamente, el cual es el valor sugerido para una óptima secuenciación de bacteriófagos según Sims y colaboradores (2014). Por otra parte, la eliminación de adaptadores de secuenciación se realizó con Cutadapt, encontrando 0 adaptadores como resultado. Esto puede deberse a que muchas

empresas tras realizar la secuenciación eliminan los adaptadores utilizados (MacConaill *et al.,* 2018), o a que los adaptadores que utiliza el programa Cutadapt por defecto no son los mismos que se manipularon para la secuenciación del genoma NJ-P3 y posiblemente estos adaptadores no eliminados son las secuencias sobre-representadas. En el análisis de datos es importante eliminar fragmentos de lecturas con baja calidad (valor *phred* bajo 30), así como cortar las primeras 10-15 bases de las lecturas como sugiere Russell (2018), ya que al inicio de la secuenciación el llamado de las bases es impreciso y a menudo erróneo. Esta irregularidad también se presenta en el tramo final de las lecturas, donde se pierde fidelidad del llamado de bases y cae la calidad de las mismas (Russell, 2018). Por otra parte, en ocasiones se omiten contaminantes externos, razón por la cual es necesario mapear las lecturas a genomas de referencia, principalmente contra el hospedero en cuestión o posibles fuentes de contaminación (Domínguez *et al.,* 2018). El análisis mediante Bowtie2 halló un total de 0,75% de las lecturas alineadas al genoma de *R. solanacearum* aislado UY031, las cuales fueron eliminadas con el fin de evitar contigs de carácter contaminante durante el ensamblaje (Russell, 2018). Con el fin de eliminar fragmentos de lecturas o lecturas que pudiesen haber sido contaminadas por ADN de humano mediante la manipulación de las muestras, se mapearon las lecturas al genoma de referencia *Homo sapiens*, hallando 0 concordancias. Este resultado sugiere que las lecturas no presentaban contaminantes de origen humano.

Normalmente, la evaluación de la calidad de los borradores de ensamblajes se lleva a cabo mediante mediciones estadísticas y alineación con un genoma de referencia (si está disponible). En este trabajo no contamos con un genoma de referencia, por lo tanto, es importante tomar en consideración las métricas estadísticas como N50, el número de contigs, longitud de contig, y longitud media de contig. Se realizaron diversos ensamblajes con métricas de cobertura y longitud mínima de contig, y en el mejor de los resultados se obtuvieron un total de 22 contigs con una longitud de 65.922 pb y un valor N50 de 38.385, equivalente a la longitud del contig de mayor tamaño. Alhakami y colaboradores (2017), encontraron que los ensamblajes presentan una mayor contigüidad y un menor

número de ensamblajes incorrectos cuando el valor N50 es semejante a la longitud del genoma, lo cual indica claramente que el ensamblaje presentó una métrica N50 adecuada. Aun así, el genoma ensamblado resulto estar fragmentado, lo cual pudo deberse a posibles contaminantes del hospedero, pues se encontró un contig con identidad 99,67%, cobertura 100% y E-value de 0 perteneciente un cromosoma de *R. solanacearum* según BLAST. Es posible que por esta razón se hayan generado diversos contigs, complicando el ensamblaje, pues lo ideal es obtener un número de contigs igual al número de cromosomas que posea el organismo en cuestión (Aguilar & Falquet, 2015). Según las notas del trabajo realizados por Russell (2018) los pequeños contigs tienden a ser contaminantes y pueden ser ignorados, razón por la cual se omitieron estos pequeños contigs en los posteriores análisis. Por otra parte, el análisis realizado por Bandage arrojó un gráfico circular con una interrupción, gap que probablemente se deba a que en la región se presentó una baja cobertura en las lecturas o exceso de repeticiones (Treangen & Salzberg, 2011). Esta topología circular es propia de los bacteriófagos que infectan *R. solanacearum,* según los aportes del trabajo realizado por Trotereau y colaboradores (2021), en el que se describen 23 nuevos fagos de *R. solanacearum,* todos ellos de ADN y doble cadena circular, con una longitud media de 41.536 pb y contenido medio de GC del 63%. Estas métricas resultaron ser semejantes al producto obtenido en el ensamblaje del fago NJ-P3, por lo que se omitió el inconveniente con el genoma fragmentado y continuamos trabajando con el contig de 38.385 pb, asumiendo este como el ensamblaje final del genoma del fago NJ-P3.

La caracterización genómica de los fagos presenta cierta limitante debido a que cuando se realizan las anotaciones se encuentra que más del 50% de los CDS predichos corresponden a proteínas hipotéticas. Tal es el caso del fago NJ-P3, donde (25 de 49) CDS que representan el 51% de los CDS predichos se anotaron como proteínas hipotéticas. Los bacteriófagos con los cuales NJ-P3 presentó una mayor homología de secuencias también cumplen con estas proporciones. Para el bacteriófago DU_RP_ll, el 58% de los CDS (22 de 38), se determinaron como proteínas hipotéticas (Park, 2018). Asimismo, en la anotación del genoma

realizada para el fago RpY1, de los 53 CDS predichos 18 se anotaron con posibles funciones y los restantes 35 se determinaron como proteínas hipotéticas, representando entonces el 66% del genoma anotado (Lee *et al.,* 2021). Considerando la proporción de genes con función desconocida es pertinente resaltar que los fagos además de ser el mayor reservorio genético en la tierra también son los representantes de los genomas con el mayor número de funciones no caracterizadas (Hatfull, 2008).

Se ha descrito que una característica distintiva de muchos fagos es que las distribuciones de algunos genes en los genomas de los fagos se agrupan en especies de módulos, y dentro de cada módulo se encuentran genes que codifican para cierta característica en específico, como síntesis de la cápside, generación de la cola, replicación de ADN, etc., (Casjens, 2005; Cazares *et al.,* 2014). Se ha demostrado que este tipo de organización también se cumple para fagos que infectan a cepas de *Ralstonia solanacearum*, un ejemplo de esto son los fagos RsoM1USA (Addy *et al.,* 2019) y RPSC1 (Liao, 2018) que muestra una distribución bien definida para sus genes en relación a su función. Sin embargo, el fago NJ-P3 no muestra una distribución tan organizada de sus genes con respecto a las funciones generadas en la anotación. La única región que demuestra este tipo de orden es la comprendida entre los CDS 28 a 30 que corresponde a CDS que codifican proteínas relacionadas con la lisis del huésped. La organización de las demás regiones en algunos casos se ve interrumpida por CDS anotados como hipotéticos o por inserciones. Este patrón se puede deber a errores en la anotación que corresponden a la poca resolución por la falta de información dentro de las bases de datos que se utilizaron para anotar o inclusive puede ser debido a un error cometido en la organización de los contigs durante el ensamblaje (Addy *et al.,* 2019).

Los alineamientos realizados en Blast para la búsqueda de posibles funciones a los CDS predichos, en muchas ocasiones arrojaron resultados que mostraban que dichos alineamientos presentaban homología de secuencia con la especie bacteriana *Ensifer adhaerens.* Cabe aclarar que independiente de los resultados

de búsqueda y para fines de este trabajo, se eligieron aquellos que estaban relacionados con bacteriófagos y sus proteínas. No obstante, creemos que realizar un trabajo teniendo en cuenta los datos relacionados con *Ensifer adhaerens* es fundamental ya que en algunos alineamientos se presentaron porcentajes de cobertura e identidad de secuencias del 100% con E-value de 0.0 representando una alta homología entre las secuencias. Sería sumamente importante abordar este punto para evaluar si estas secuencias hacen parte de algún evento de transferencia horizontal entre *Ensifer adhaerens* y el fago NJ-P3. En este punto es importante mencionar que en la anotación generada para el fago NJ-P3 no se encontraron genes vinculados con un ciclo de vida lisogénico o la conversión a profago, sin embargo, no se debe descartar el hecho de que alguno de los CDS anotados como proteínas hipotéticas puedan estar involucrados en este ciclo de vida.

En este trabajo, con el propósito de corroborar los análisis de sintenia realizados y la veracidad en las relaciones de los bloques de colinealidad generados, se optó por ejecutar dicho análisis en los programas ACT (**Fig. 8A**) y Easyfig (**Fig. 8B**). Como resultado se obtuvo una identidad compartida superior al 80%, al igual que la presencia de continuos bloques de colinealidad entre los genomas de los bacteriófagos de *R. solanacearum* NJ-P3, DU_RP_II y RpY1. Nuestro análisis final sugiere un estrecho parentesco en la historia evolutiva de los tres virus. Esta posición podemos asumirla gracias al estudio realizado por Brüssow y colaboradores (2002), quienes señalan que bacteriófagos que infectan a la misma especie bacteriana suelen mostrar identidad de secuencia nucleotídica limitada a ciertas regiones en el genoma o en ocasiones nula. En este trabajo los genomas comparados presentaron una alta homología entre ellos. Por otra parte, nuestros resultados concuerdan parcialmente con los resultados obtenidos por Trotereau y colaboradores (2021), quienes encontraron que grupos de bacteriófagos de *R. solanacearum* no presentaron ninguna homología, en otros se evidenció una homología significativamente débil, y sólo se encontró similitud entre los fagos Cimandef y Gervaise. Esto sugiere que bacteriófagos que infectan un mismo hospedero pueden o no compartir homología en las secuencias. Por otra parte, se

presentaron inserciones y deleciones de proteínas, alguna codificantes, a lo largo de los genomas de los fagos NJ-P3, DU_RP_II y RpY1, lo cual es una característica común en los bacteriófagos, y puede darse, al menos en parte, por la aparente estructura de mosaico propia de los fagos, pues cada genoma representa una combinación única (Hatfull, 2008). Es así que puede explicarse la singularidad en el genoma del fago RpY1, cuya disposición de proteínas codificantes en dirección 3´-5’ difiere de la distribución estructural proteica de los bacteriófagos NJ-P3 y DU_RP_II. Aunque esta reestructuración del genoma también puede verse impulsada por errores en el ensamblaje o la anotación del genoma como se discutió previamente.

Los estudios de genómica cuentan con una limitante, pues existen diversas herramientas con las que realizar el análisis y el tratamiento de datos, es así que se torna complejo discernir entre cual podemos utilizar y bajo qué parámetros hacerlo. Algunas de estas herramientas son mejores y más complejas que otras, y van desde gratuitas hasta plataformas de pago. Para sacar el máximo provecho de éstas, es necesario en ocasiones, tener un nivel alto en lenguajes de programación, pues la mayoría de los flujos de trabajo se encuentran alojados en sistemas de códigos. Este es un proyecto enfocado hacia personas, que como nosotros, se está adentrando en el mundo de la biología computacional. Pretendemos incentivar a los nuevos investigadores en el área, para abrir los horizontes hacia nuevas herramientas de mayor complejidad que permitan realizar un análisis de datos más riguroso, un ensamblaje más acertado y una anotación más óptima, lo cual puede contribuir con el mejoramiento y actualización de los datos alojados en las bases de datos.

6. Conclusión

Existen diversas herramientas y algoritmos enfocados al análisis de datos, ensamblaje, anotación y comparación de genomas, todas ellas con características específicas y orientadas hacia grupos en particular de organismos. Depende de las necesidades del investigador, de los objetivos de la investigación y de la

naturaleza de los datos elegir cuales herramientas y bajo qué parámetros deben ser ejecutadas para obtener los resultados deseados. Este trabajo es un “abrebocas” en materia de genómica de fagos, en el que se exploran algunas herramientas de libre acceso y de fácil uso que permiten contar con un punto de partida para ensamblar y anotar genomas de bacteriófagos. Particularmente este grupo de organismos presenta una alta diversidad genética, razón por la cual es importante continuar con esta línea de investigación. Adicionalmente, los bacteriófagos han tomado protagonismo en diferentes áreas de la investigación como las ciencias médicas, la biología evolutiva y la fagoterapia. Por último, las tecnologías y las herramientas analíticas emergentes podrían mejorar drásticamente los ensamblajes y anotaciones de los genomas, ayudando así a fortalecer las bases de datos, y por consiguiente la comprensión de las entidades biológicas.

7. Agradecimientos

En primer lugar nos gustaría agradecer a la Universidad de Antioquia por la formación y el desarrollo nuestras capacidades como profesionales en el área de las Ciencias Biológicas. Así mismo, es de agradecer a la Unidad de Bio-prospección y Estudio de Microbiomas (BIOEM) del PECET, por su acogida y acompañamiento durante este proceso. Agradecemos de una manera muy especial el apoyo brindado por nuestras familias y compañeros. Nos gustaría brindar nuestra gratitud a los docentes David Andrés Borrego Muñoz, Juan Esteban Pérez Jaramillo y Cristian David Grisales Vargas por el apoyo, las ideas, el ánimo y la disposición de enseñar, seguramente sin su acompañamiento este proyecto no hubiese sido posible. Y como último, pero no menos importante, agradecer a Ricardo Callejas por enseñarnos que la Biología es una pasión que se vive en el día a día. Mil y mil gracias a todos y todas.

8. Referencias

Ackermann, H. W. (2001). Frequency of morphological phage descriptions in the year 2000. Archives of Virology, 146(5), 843–857. https://doi.org/10.1007/s007050170120

Addy, H. S., Farid, M. M., Ahmad, A. A., & Huang, Q. (2018). Host range and molecular characterization of a lytic Pradovirus-like Ralstonia phage RsoP1IDN isolated from Indonesia. Archives of virology, 163(12), 3409–3414.

Addy, H. S., Ahmad, A. A., & Huang, Q. (2019). Molecular and Biological Characterization of Ralstonia Phage RsoM1USA, a New Species of P2virus, Isolated in the United States. Frontiers in microbiology, 10, 267. https://doi.org/10.3389/fmicb.2019.00267

Aguilar-Bultet, L., & Falquet, L. (2015). Secuenciación y ensamblaje de novo de genomas bacterianos: una alternativa para el estudio de nuevos patógenos. Revista de salud animal, 37(2), 125-132.

Álvarez, B., & Biosca, E. G. (2017). Bacteriophage-Based Bacterial Wilt Biocontrol for an Environmentally Sustainable Agriculture. Frontiers in plant science, 8, 1218. https://doi.org/10.3389/fpls.2017.01218.

Ahmad, A. A., Elhalag, K. M., Addy, H. S., Nasr-Eldin, M. A., Hussien, A. S., & Huang, Q. (2018). Sequencing, genome analysis and host range of a novel Ralstonia phage, RsoP1EGY, isolated in Egypt. Archives of virology, 163(8), 2271–2274.

Alhakami, H., Mirebrahim, H., & Lonardi, S. (2017). A comparative evaluation of genome assembly reconciliation tools. Genome biology, 18(1), 93. https://doi.org/10.1186/s13059-017-1213-3

Alkan, C., Coe, B. P., & Eichler, E. E. (2011). Genome structural variation discovery and genotyping. Nature reviews. Genetics, 12(5), 363–376. https://doi.org/10.1038/nrg2958

Andrews S.2010. FastQC: a quality control tool for high throughput sequence data. [Internet]. Fecha de acceso: 20 de marzo de 2021. Disponible en: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Bae, J. Y., Wu, J., Lee, H. J., Jo, E. J., Murugaiyan, S., Chung, E., & Lee, S. W. (2012). Biocontrol potential of a lytic bacteriophage PE204 against bacterial wilt of tomato. Journal of microbiology and biotechnology, 22(12), 1613–1620. https://doi.org/10.4014/jmb.1208.08072

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., & Pevzner, P. A. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. Journal of computational biology: a journal of computational molecular cell biology, 19(5), 455–477. https://doi.org/10.1089/cmb.2012.0021

Barrios, M. O., Andrea Rodríguez Gaviria, P., Gonzalo, J., Osorio, M., & Yepes, M. S. (2008). Hospedantes de Ralstonia solanacearum en plantaciones de banano y platano en Colombia. Rev.Fac.Nal.Agr.Medellín, 61(2), 4518–4526.

Batzoglou S. (2005). The many faces of sequence alignment. Briefings in bioinformatics, 6(1), 6–22. https://doi.org/10.1093/bib/6.1.6

Bhunchoth, A., Blanc-Mathieu, R., Mihara, T., Nishimura, Y., Askora, A., Phironrit, N., Leksomboon, C., Chatchawankanphanich, O., Kawasaki, T., Nakano, M., Fujie, M., Ogata, H., & Yamada, T. (2016). Two asian jumbo phages, ϕRSL2 and ϕRSF1, infect Ralstonia solanacearum and show common features of ϕKZ-related phages. Virology, 494, 56–66. https://doi.org/10.1016/j.virol.2016.03.028

Bhutkar, A., Russo, S., Smith, T. F., & Gelbart, W. M. (2006). Techniques for multi-genome synteny analysis to overcome assembly limitations. Genome informatics. International Conference on Genome Informatics, 17(2), 152–161. https://doi.org/10.11234/gi1990.17.2_152

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, 30(15), 2114-2120.

Brettin, T., Davis, J. J., Disz, T., Edwards, R. A., Gerdes, S., Olsen, G. J., Olson, R., Overbeek, R., Parrello, B., Pusch, G. D., Shukla, M., Thomason, J. A., 3rd, Stevens, R., Vonstein, V., Wattam, A. R., & Xia, F. (2015). RASTtk: a modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. Scientific reports, 5, 8365. https://doi.org/10.1038/srep08365

Brooksbank, C., Bergman, M. T., Apweiler, R., Birney, E., & Thornton, J. (2014). The European Bioinformatics Institute's data resources 2014. Nucleic acids research, 42(Database issue), D18–D25. https://doi.org/10.1093/nar/gkt1206

Brunner, M., & Pootjes, C. F. (1969). Bacteriophage release in a lysogenic strain of Agrobacterium tumefaciens. Journal of virology, 3(2), 181-186.

Brüssow, H., & Hendrix, R. W. (2002). Phage genomics: small is beautiful. Cell, 108(1), 13–16. https://doi.org/10.1016/s0092-8674(01)00637-7

Buttimer, C., McAuliffe, O., Ross, R. P., Hill, C., O'Mahony, J., & Coffey, A. (2017). Bacteriophages and bacterial plant diseases. Frontiers in microbiology, 8, 34.

Carver, T. J., Rutherford, K. M., Berriman, M., Rajandream, M. A., Barrell, B. G., & Parkhill, J. (2005). ACT: the Artemis Comparison Tool. Bioinformatics (Oxford, England), 21(16), 3422–3423. https://doi.org/10.1093/bioinformatics/bti553

Casjens S. R. (2005). Comparative genomics and evolution of the tailed-bacteriophages. *Current opinion in microbiology*, *8*(4), 451–458. https://doi.org/10.1016/j.mib.2005.06.014

Cazares, A., Mendoza-Hernández, G., & Guarneros, G. (2014). Core and accessory genome architecture in a group of Pseudomonas aeruginosa Mu-like phages. *BMC genomics*, *15*(1), 1146. https://doi.org/10.1186/1471-2164-15-1146

Chan, P. P., & Lowe, T. M. (2019). tRNAscan-SE: Searching for tRNA Genes in Genomic Sequences. Methods in molecular biology (Clifton, N.J.), 1962, 1–14.

Comeau, A. M., Hatfull, G. F., Krisch, H. M., Lindell, D., Mann, N. H., & Prangishvili, D. (2008). Exploring the prokaryotic virosphere. Research in Microbiology, 159(5), 306–313. https://doi.org/10.1016/j.resmic.2008.05.001

D'Herelle F. (1917). Sur un microbe invisible antagoniste des bacillus dysentérique. Acad Sci Paris. 1917;165:373–5.

Dion, M. B., Oechslin, F., & Moineau, S. (2020). Phage diversity, genomics and phylogeny. Nature reviews. Microbiology, 18(3), 125–138. https://doi.org/10.1038/s41579-019-0311-5

Domínguez Del Angel, V., Hjerde, E., Sterck, L., Capella-Gutierrez, S., Notredame, C., Vinnere Pettersson, O., Amselem, J., Bouri, L., Bocs, S., Klopp, C., Gibrat, J. F., Vlasova, A., Leskosek, B. L., Soler, L., Binzer-Panchal, M., & Lantz, H. (2018). Ten steps to get started in Genome Assembly and

Annotation. *F1000Research*, *7*, ELIXIR-148. https://doi.org/10.12688/f1000research.13598.1

Ejigu, G. F., & Jung, J. (2020). Review on the Computational Genome Annotation of Sequences Obtained by Next-Generation Sequencing. Biology, 9(9), 295. https://doi.org/10.3390/biology9090295

Fujiwara, A., Fujisawa, M., Hamasaki, R., Kawasaki, T., Fujie, M., & Yamada, T. (2011). Biocontrol of Ralstonia solanacearum by treatment with lytic bacteriophages. Applied and environmental microbiology, 77(12), 4155–4162. https://doi.org/10.1128/AEM.02847-10

Garvey, P., Van Sinderen, D., Twomey, D. P., Hill, C., & Fitzgerald, G. F. (1995). Molecular genetics of bacteriophage and natural phage defence systems in the genus Lactococcus. International Dairy Journal, 5(8), 905-947.

Gordon, A. (2010). FASTQ/A short-reads pre-processing tools. [Internet]. Fecha de acceso: 20 de marzo de 2021. Disponible en: http://hannonlab.cshl.edu/fastx_toolkit/

Guglielmotti, D. M., Mercanti, D. J., Reinheimer, J. A., & Quiberoni, A. (2012). Review: efficiency of physical and chemical treatments on the inactivation of dairy bacteriophages. Frontiers in microbiology, 2, 282. https://doi.org/10.3389/fmicb.2011.00282

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. Bioinformatics, 29(8), 1072-1075.

Hatfull G. F. (2008). Bacteriophage genomics. Current opinion in microbiology, 11(5), 447–453. https://doi.org/10.1016/j.mib.2008.09.004

Hirokawa, T.; Boon-Chieng, S.; Mitaku, S. (1998). SOSUI: classification and secondary structure prediction system for membrane proteins. Bioinformatics, 14(4), 378–379.

Howe, K. L., Bolt, B. J., Cain, S., Chan, J., Chen, W. J., Davis, P., Done, J., Down, T., Gao, S., Grove, C., Harris, T. W., Kishore, R., Lee, R., Lomax, J., Li, Y., Muller, H. M., Nakamura, C., Nuin, P., Paulini, M., Raciti, D., … Sternberg, P. W. (2016). WormBase 2016: expanding to enable helminth genomic research. Nucleic acids research, 44(D1), D774–D780. https://doi.org/10.1093/nar/gkv1217

Hyatt, D., Chen, G. L., Locascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site

identification. BMC bioinformatics, 11, 119. https://doi.org/10.1186/1471-2105-11-119

Jung, H., Ventura, T., Chung, J. S., Kim, W. J., Nam, B. H., Kong, H. J., Kim, Y. O., Jeon, M. S., & Eyun, S. I. (2020). Twelve quick steps for genome assembly and annotation in the classroom. PLoS computational biology, 16(11), e1008325. https://doi.org/10.1371/journal.pcbi.1008325

Keen, E. C. (2015). A century of phage research: Bacteriophages and the shaping of modern biology. BioEssays, 37(1), 6–9. https://doi.org/10.1002/bies.201400152.

Klaenhammer, T. R., & Fitzgerald, G. F. (1994). Bacteriophages and bacteriophage resistance. In Genetics and biotechnology of lactic acid bacteria (pp. 106-168). Springer, Dordrecht.

Krogh, A., Larsson, B., von Heijne, G., & Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. Journal of molecular biology, 305(3), 567–580.

Kodama, Y., Mashima, J., Kosuge, T., Kaminuma, E., Ogasawara, O., Okubo, K., Nakamura, Y., & Takagi, T. (2018). DNA Data Bank of Japan: 30th anniversary. Nucleic acids research, 46(D1), D30–D35. https://doi.org/10.1093/nar/gkx926

Kutter, E. & Sulakvelidze, A. (2005). Bacteriophages: biology and applications (págs. 510). CRC Press, Boca Raton, FL.

Lagesen, K., Hallin, P., Rødland, E. A., Staerfeldt, H. H., Rognes, T., & Ussery, D. W. (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic acids research, 35(9), 3100–3108. https://doi.org/10.1093/nar/gkm160

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. Nature methods, 9(4), 357.

Lee, S. Y., Thapa Magar, R., Kim, H. J., Choi, K., & Lee, S. W. (2021). Complete Genome Sequence of a Novel Bacteriophage RpY1 Infecting Ralstonia solanacearum Strains. Current microbiology, 78(5), 2044–2050

Liao, P., Satten, G. A., & Hu, Y. J. (2017). PhredEM: a phred-score-informed genotype-calling approach for next-generation sequencing studies. Genetic epidemiology, 41(5), 375–387. https://doi.org/10.1002/gepi.22048

Liao M. (2018). Genomic characterization of the novel Ralstonia phage RPSC1. *Archives of virology*, *163*(7), 1969–1971. https://doi.org/10.1007/s00705-018-3713-1

Lin, D. M., Koskella, B., & Lin, H. C. (2017). Phage therapy: An alternative to antibiotics in the age of multi-drug resistance. World journal of gastrointestinal pharmacology and therapeutics, 8(3), 162–173. https://doi.org/10.4292/wjgpt.v8.i3.162

Lischer, H. & Shimizu, KK (2017). El enfoque de ensamblaje de novo guiado por referencias mejora la reconstrucción del genoma de especies relacionadas. Bioinformática de BMC, 18 (1), 474. https://doi.org/10.1186/s12859-017-1911-6

Loganantharaj, R., & Randall, T. A. (2016). An Overview And Comparison of Tools for RNA-Seq Assembly. Computational Methods for Next Generation Sequencing Data Analysis, 269–286.

MacConaill, LE, Burns, RT, Nag, A., Coleman, HA, Slevin, MK, Giorda, K., Light, M., Lai, K., Jarosz, M., McNeill, MS, Ducar, MD, Meyerson , M. y Thorner, AR (2018). Los adaptadores de secuenciación de índice dual únicos con UMI eliminan eficazmente la interferencia de índices y mejoran significativamente la sensibilidad de la secuenciación masivamente paralela. Genómica de BMC, 19 (1), 30. https://doi.org/10.1186/s12864-017-4428-5

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. journal, 17(1), 10-12.

Martínez-Bolaños L. 2010. Moko bacteriano del plátano. Ficha técnica SINAFEV. México.

Mathé, C., Sagot, M. F., Schiex, T., & Rouzé, P. (2002). Current methods of gene prediction, their strengths and weaknesses. Nucleic acids research, 30(19), 4103–4117. https://doi.org/10.1093/nar/gkf543

Neve, H. (1996). Bacteriophages. In: Dairy Starter Cultures. Cogan and T. M. Accolas, J.M. (Ed.) VCH Publisher Inc. Cap. 6. 157-190.

Park T. H. (2018). Complete genome sequence of DU_RP_II, a novel Ralstonia solanacearum phage of the family Podoviridae. *Archives of virology*, *163*(1), 269–271. https://doi.org/10.1007/s00705-017-3577-9

Piña-Iturbe, A., Suazo, I. D., Hoppe-Elsholz, G., Ulloa-Allendes, D., González, P. A., Kalergis, A. M., & Bueno, S. M. (2020). Horizontally Acquired Homologs

of Xenogeneic Silencers: Modulators of Gene Expression Encoded by Plasmids, Phages and Genomic Islands. Genes, 11(2), 142. https://doi.org/10.3390/genes11020142

Rihtman, B., Meaden, S., Clokie, M. R., Koskella, B., & Millard, A. D. (2016). Assessing Illumina technology for the high-throughput sequencing of bacteriophage genomes. PeerJ, 4, e2055. https://doi.org/10.7717/peerj.2055

Russell D. A. (2018). Sequencing, Assembling, and Finishing Complete Bacteriophage Genomes. Methods in molecular biology (Clifton, N.J.), 1681, 109–125. https://doi.org/10.1007/978-1-4939-7343-9_9

Sayers, E. W., Cavanaugh, M., Clark, K., Ostell, J., Pruitt, K. D., & Karsch-Mizrachi, I. (2019). GenBank. Nucleic acids research, 47(D1), D94–D99. https://doi.org/10.1093/nar/gky989

Sims, D., Sudbery, I., Ilott, N. E., Heger, A., & Ponting, C. P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. Nature reviews. Genetics, 15(2), 121–132. https://doi.org/10.1038/nrg3642

Spieth, J., & Lawson, D. (2006). Overview of gene structure. WormBook: the online review of C. elegans biology, 1–10. https://doi.org/10.1895/wormbook.1.65.1

Sullivan, M. J., Petty, N. K., & Beatson, S. A. (2011). Easyfig: a genome comparison visualizer. Bioinformatics (Oxford, England), 27(7), 1009–1010.

Tanaka H, Negishi H, Maeda H. 1990. Control of tobacco bacterial wilt by an avirulent strain of Pseudomonas solanacearum M4S and its bacteriophage. Annals Phytopathological society. 56:243-246.

Treangen, T. J., & Salzberg, S. L. (2011). Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nature reviews. Genetics, 13(1), 36–46. https://doi.org/10.1038/nrg3117

Trotereau, A., Boyer, C., Bornard, I., Pécheur, M. J. B., Schouler, C., & Torres-Barceló, C. (2021). High genomic diversity of novel phages infecting the plant pathogen Ralstonia solanacearum, isolated in Mauritius and Reunion islands. Scientific reports, 11(1), 5382. https://doi.org/10.1038/s41598-021-84305-7

Twort F. (1915). An investigation on the nature of ultra-microscopic viruses. The Lancet. 1915; 186: 4814.

Vergara, I. A., & Chen, N. (2010). Large synteny blocks revealed between Caenorhabditis elegans and Caenorhabditis briggsae genomes using

OrthoCluster. BMC genomics, 11, 516. https://doi.org/10.1186/1471-2164-11-516

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S. K., & Earl, A. M. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PloS one, 9(11), e112963. https://doi.org/10.1371/journal.pone.0112963

Wick, R. R., Schultz, M. B., Zobel, J., & Holt, K. E. (2015). Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics (Oxford, England)*, *31*(20), 3350–3352. https://doi.org/10.1093/bioinformatics/btv383

Xiao, J., Sekhwal, M. K., Li, P., Ragupathy, R., Cloutier, S., Wang, X., & You, F. M. (2016). Pseudogenes and Their Genome-Wide Prediction in Plants. International journal of molecular sciences, 17(12), 1991. https://doi.org/10.3390/ijms17121991

Yandell, M., & Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. Nature Reviews Genetics, 13(5), 329-342.

Yuan, Y., & Gao, M. (2017). Jumbo Bacteriophages: An Overview. Frontiers in microbiology, 8, 403. https://doi.org/10.3389/fmicb.2017.00403

Zimmermann, L., Stephens, A., Nam, S. Z., Rau, D., Kübler, J., Lozajic, M., Gabler, F., Söding, J., Lupas, A. N., & Alva, V. (2018). A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. Journal of molecular biology, 430(15), 2237–2243. https://doi.org/10.1016/j.jmb.2017.12.007

Zhou, Q., Su, X. & Ning, K. Assessment of quality control approaches for metagenomic data analysis. Sci Rep 4, 6957 (2014).